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

Tape Challenge 2022 @CC-IN2P3

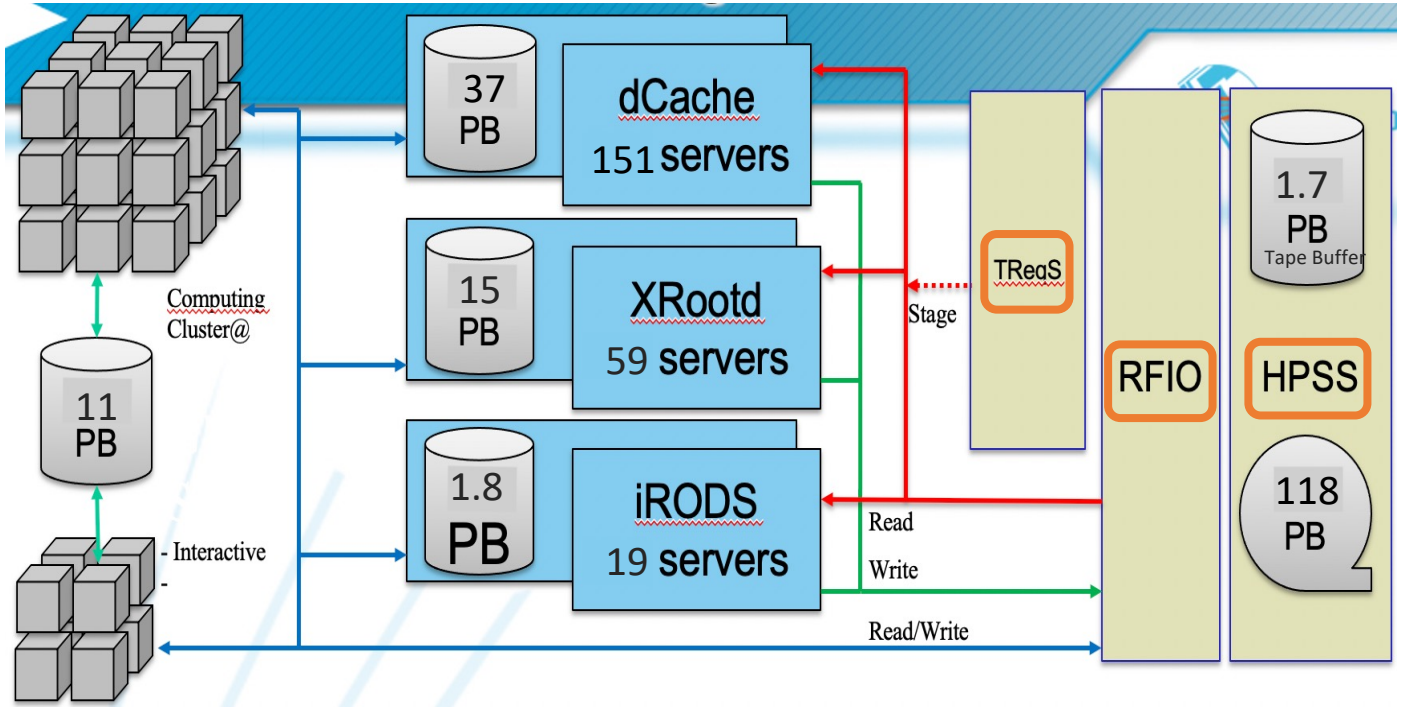
Aresh VEDAEE (21.04.2022)

TAPE INFRASTRUCTURE (1/2)

- Storage System : HPSS 8.3.0u15
  - 2 core server DELL R640
  - 8 Tape Mover DELL R640
    - + 9 Tape Mover R720/R730 (only for T10K-D)
  - 12 Disk Mover DELL R720

- Tape Libraries:
  - T10KD: ~all LHC data migrated to Jaguar E
  - Jaguar E:
    - 48 TS1160 drives (nominal speed 400 MB/s)
    - Spectralogic TFinity 9 Frames
    - A dual-arm robot
    - 755 terapack chambers (9 tapes per terapack)
    - 6795 tape cartridges (20TB per cartridge)
    - Capacity : ~136 PB (Used: 118 PB / 92%)
    - 1.7PB Tape Buffer

- Additional services:
  - RFIO Service (for staging and migrations)
  - TReqS: **T**ape **Req**uest **S**cheduler (for staging requests)



TAPE INFRASTRUCTURE (2/2)

- Acquisition of new Jaguar E tape library with similar features to the current one
  - 9 frames
  - 755 terapack chambers
  - 6795 cartridges
  - 48 drives
  - Dual-arm robot

- <u>Expected deployment Q2 2022</u>
  - Initial config:
    - 12 TS1160 drives
    - 1500 cartridges

TAPE CONFIGURATION

Migration Policy

- Min 10 drives out of 46 R/W drives in production only for migration
- Drive distribution not based on VO-specific file families but on COS file families (wrt combination of 2 factors: file size & VO)

	COS12 (file size < 2GB)	COS14 (file size > 2GB)
ALICE&LHCB	MAX 2 drives	**MAX 4 drives**
ATLAS	**MAX 4 drives**	**MAX 6 drives**
CMS	MAX 2 drives	**MAX 2 drives**
GENERAL (all others)	MAX 4 drives	MAX 4 drives

- LHCB&CMS: ~100% in COS14
- ATLAS data: 11% COS12 89% COS14 => ATLAS could get up to 10 drives

Now: MAX 4 drives

Staging Policy

- Staging scheduler requests max 36 drives (out of 46 R/W drives in production) at each staging pass (4min)
- Pending time for staging requests set up to minimum 4min (but there is no max and the staging file can be served after hours)

RFIO service LCG/EGEE VOs

	Read	Write
During TC '22	**1 server with 400 simultaneous connections**	
Now	**1 server with 1024 simultaneous connections.**	**1 server with 400 simultaneous connections**

TAPE CHALLENGE '22 @ CC-IN2P3: OVERVIEW

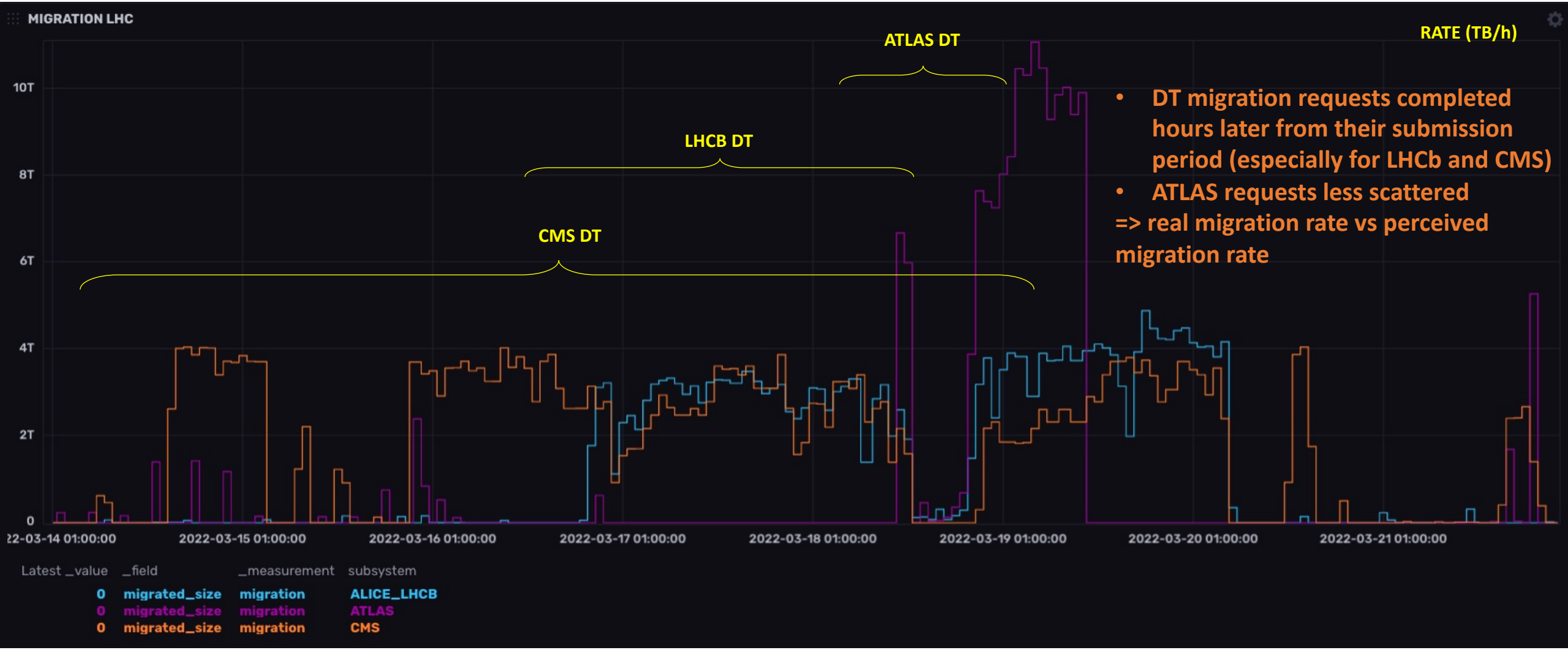
	DT (migrations)	A-DT (stagings)
ATLAS	from 18/03/22 8h00 to 18/03/22 24h00 [CET]	from 21/03/22 11h00 to 24/03/22 12h00 [CET]
CMS	from 14/03/22 10h00 to 19/03/22 14h00 [CET]	from 29/03/22 15h00 to 30/03/22 05h00 [CET]
LHCB	from 16/03/22 14h00 to 18/03/22 09h00 [CET]	from 22/03/22 11h00 to 23/03/22 18h00 [CET]

	DT (migrations)			OK?	A-DT (stagings)			OK?
	MAX	AVG	TARGET		MAX	AVG	TARGET	
ATLAS	4GB/s	1.8GB/s	3.5GB/s	✓	6.9GB/s	2.6GB/s	1.2GB/s	✓
CMS	7GB/s	0.9GB/s	0.29GB/s	✓	8.3GB/s	3.4GB/s	?	?
LHCB	4GB/s	1.4GB/s	1.2GB/s	✓	4.5GB/s	2GB/s	0.98GB/s	✓

Source:

Data Challenges / FTS Status Board (March '22)
Tape Challenge '22 - VO post-mortem report
Tape Challenge '22 - Outline (googledoc)

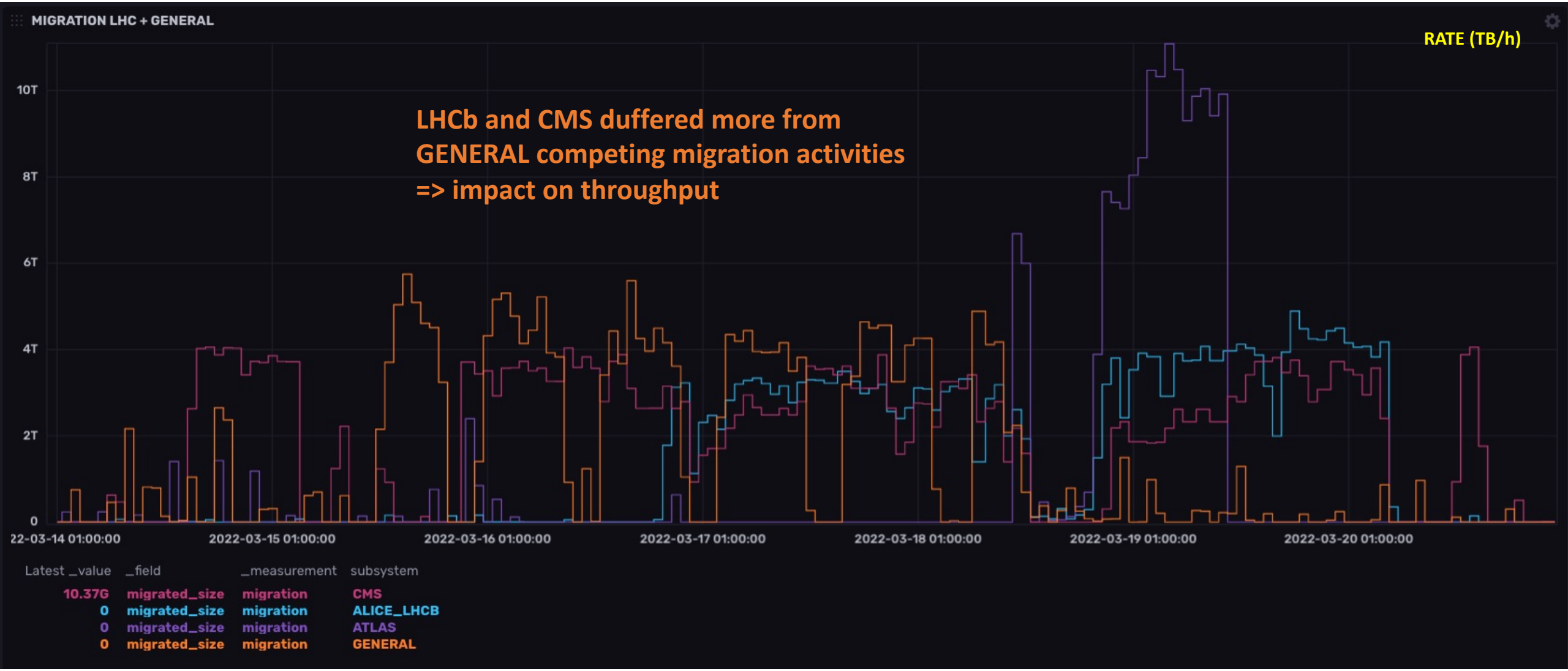
DT: LHC MIGRATIONS



- DT migration requests completed hours later from their submission period (especially for LHCb and CMS)
- ATLAS requests less scattered

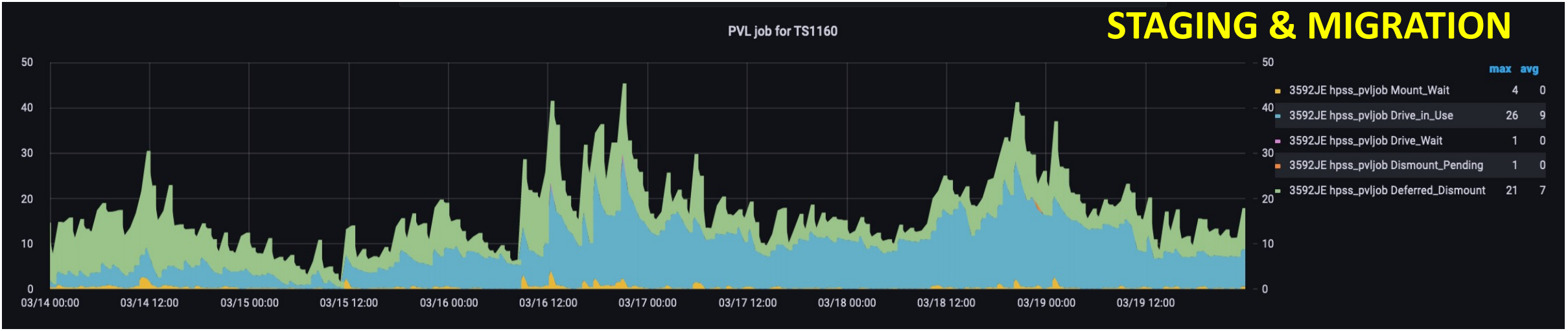
=> real migration rate vs perceived migration rate

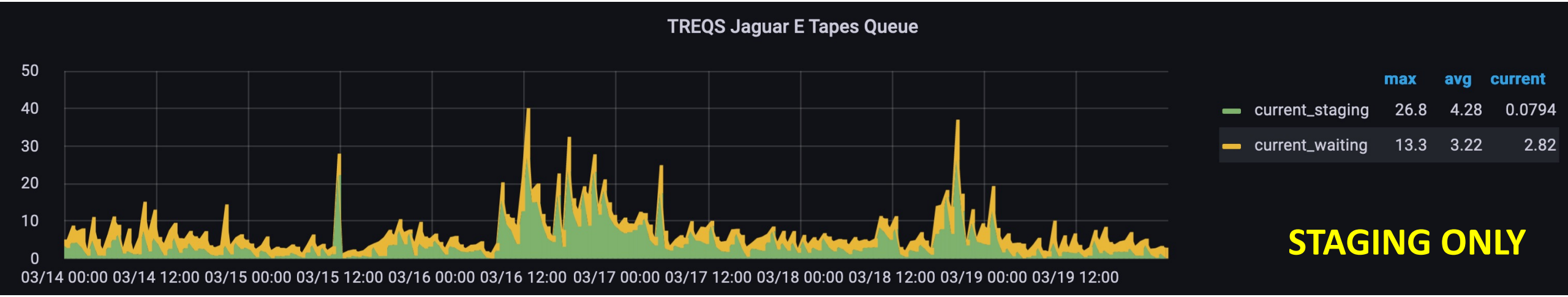
DT: LHC MIGRATIONS + GENERAL MIGRATIONS



DT: DRIVE USAGE (1/2)

STAGING & MIGRATION

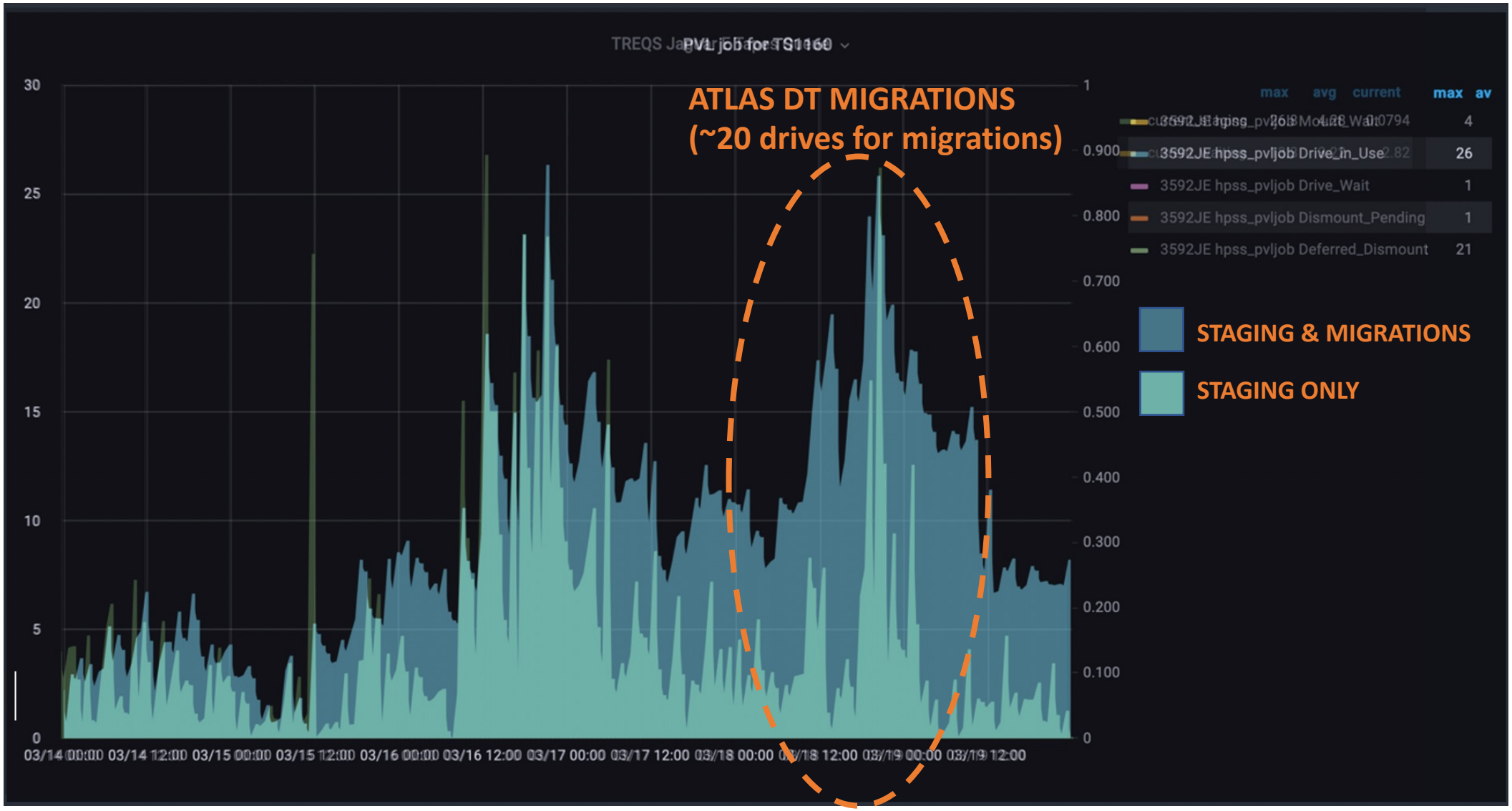




STAGING ONLY

DT migrations could get up to 20 drives (the min was 10) while DT stagings got max 26

DT: DRIVE USAGE (2/2)



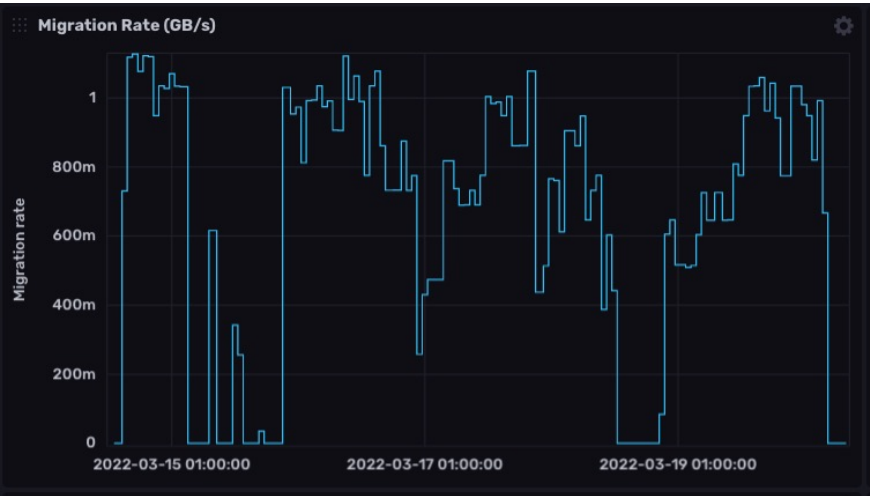
DT: ATLAS MIGRATIONS

CC VIEW

VO VIEW	DT (migrations)	
	MAX	AVG
ATLAS	4GB/s	1.8GB/s

- ATLAS requests less scattered and less delayed wrt other VOs => VO/CC Rate discrepancies
- Likely more drives wrt LHCb & CMS due to less competition from GENERAL (but unlikely from COS FF since AVG file size > 2GB)

DT: CMS MIGRATIONS



AVG Rate

0.65 GB/s

MAX Rate

1.13 GB/s

CC VIEW

Total written Size

325,528 GB

Mean written file size (Clone)

10.03 GB

Total written files

32,449

VO VIEW	DT (migrations)	
	MAX	AVG
CMS	7GB/s	0.9GB/s

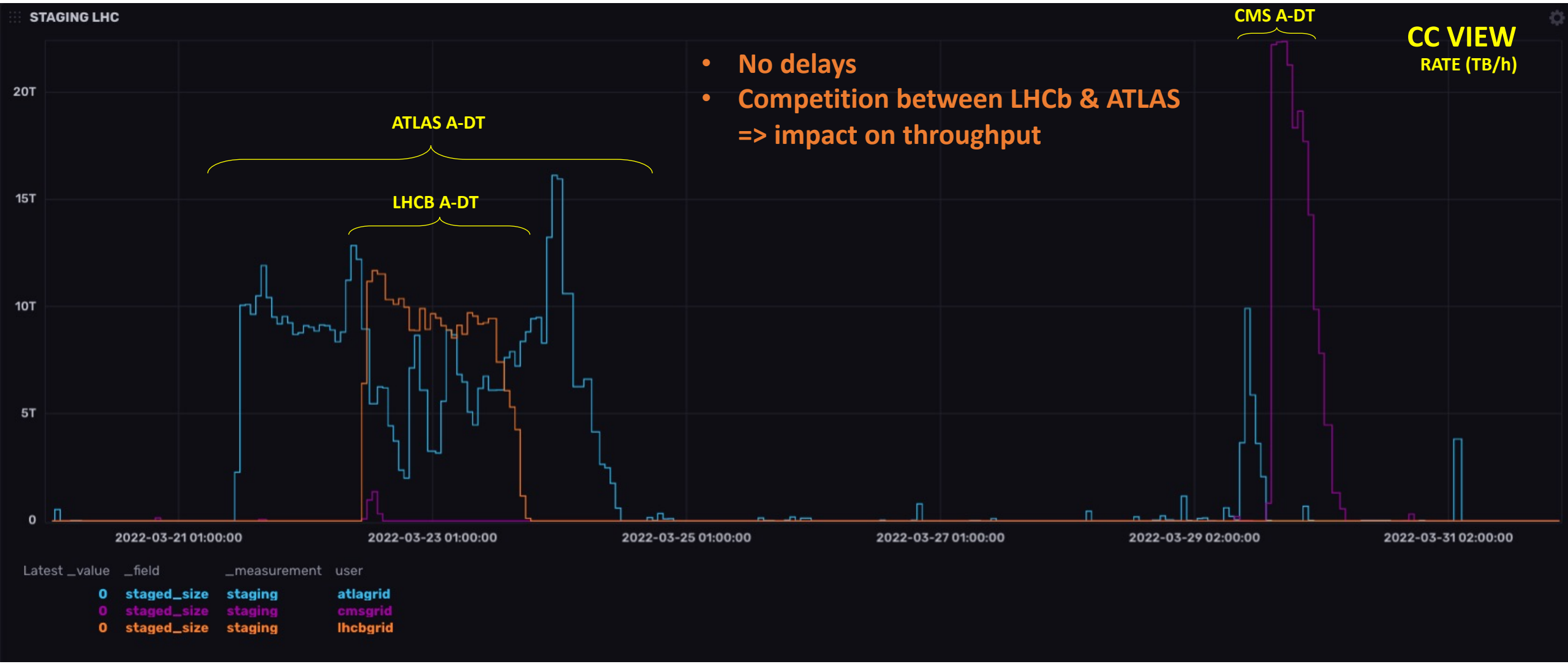
- CMS activity most scattered wrt ATLAS & LHCb and more delayed then ATLAS => VO/CC Rate discrepancies
- Only max 2 drives
- More competition from GENERAL

DT: LHCB MIGRATIONS

VO VIEW	DT (migrations)	
	MAX	AVG
LHCB	4GB/s	1.4GB/s

- LHCb activity more scattered than ATLAS but less than CMS, and more delayed then ATLAS => VO/CC Rate discrepancies
- More competition from GENERAL then ATLAS

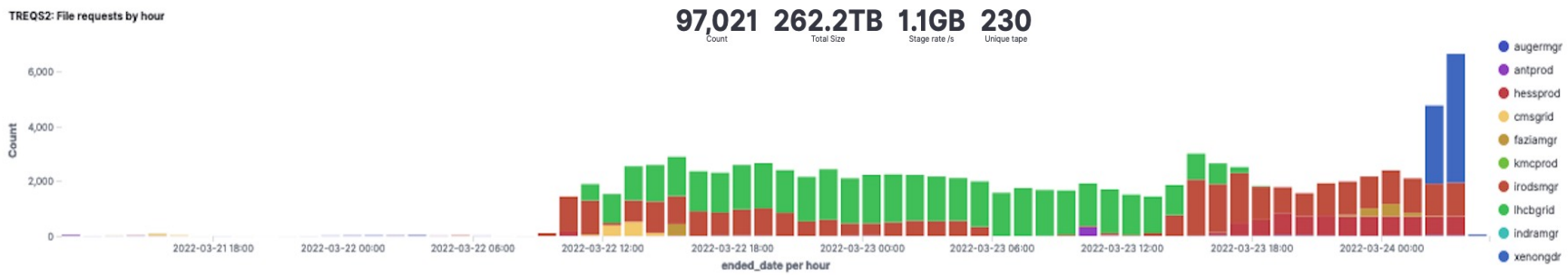
A-DT: LHC STAGINGS



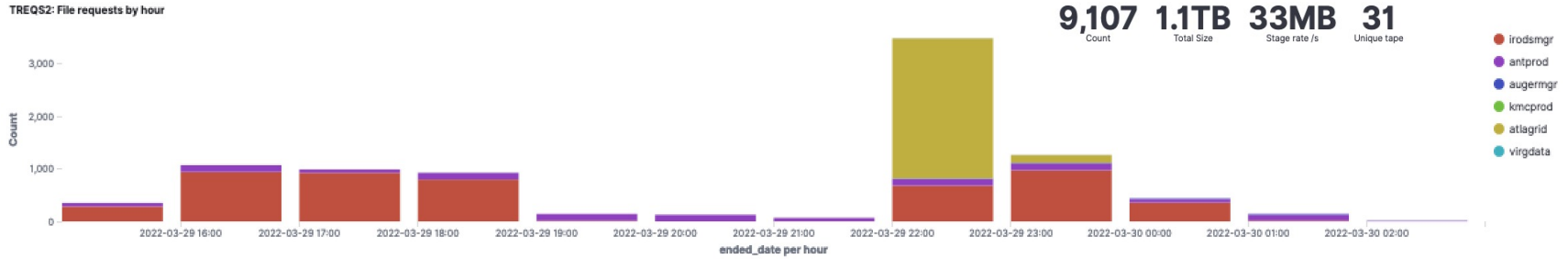
- No delays
- Competition between LHCb & ATLAS
  => impact on throughput

A-DT COMPETING STAGING ACTIVITES PER VO

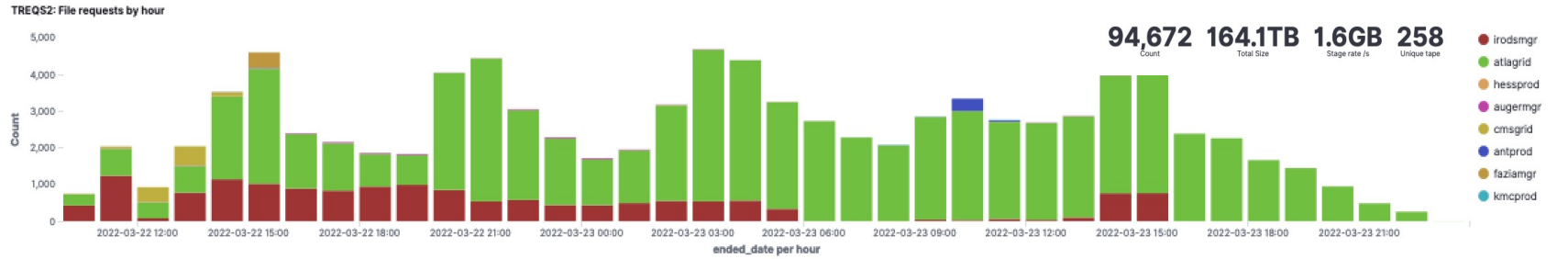
LHC & NON-LHC COMPETITORS



vs ATLAS



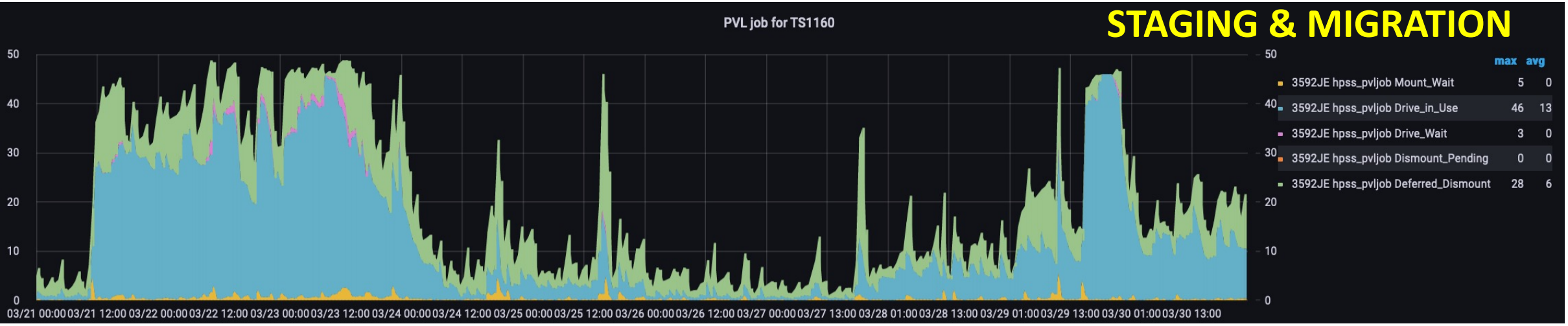
vs CMS

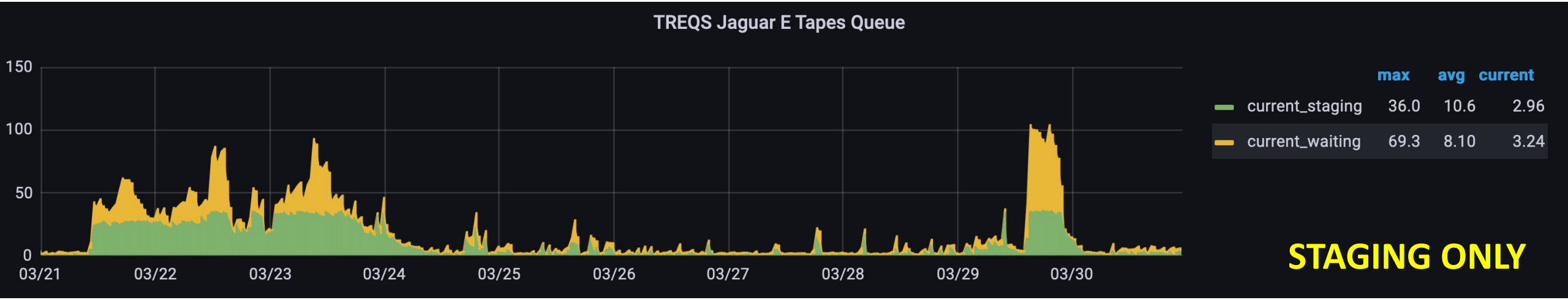


vs LHCB

A-DT: DRIVE USAGE (1/2)

STAGING & MIGRATION

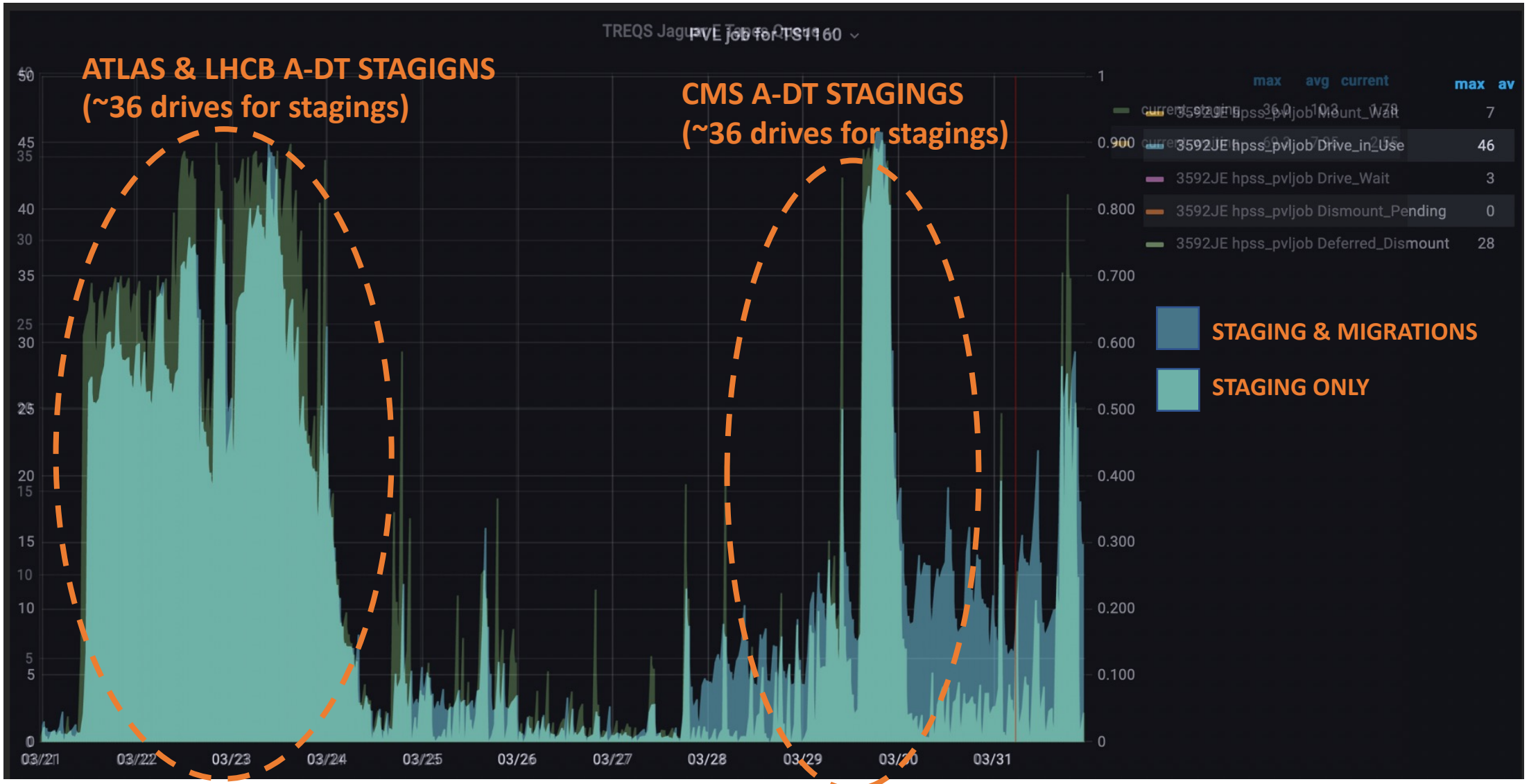




STAGING ONLY

A-DT stagings reached max 36 while A-DT migrations could get up to 10 drives

A-DT: DRIVE USAGE (2/2)



A-DT: ATLAS STAGINGS

CC VIEW

AVG Rate

2.03GB/s

MAX Rate

4.48 GB/s

Total Staged Size

554,388 GB

Mean Staged file size

2.55 GB

Request Count

217,821

VO VIEW	A-DT (stagings)	
	MAX	AVG
ATLAS	6.9GB/s	2.6GB/s

- ATLAS stagins more scattered in bunches of 15K FTS requests
- ATLAS smaller AVG file size wrt other VOs (bigger files => better performance)
- Competition with LHCb stagings
- VO/CC Rate discrepancies maybe due to files on tape buffer?

A-DT: CMS STAGINGS

VO VIEW	A-DT (stagings)	
	MAX	AVG
CMS	8.3GB/s	3.4GB/s

- CMS stagins more concentrated in one bunch of 14K FTS requests
- CMS bigger AVG file size wrt other VOs (bigger files => better performance)
- Less competition from other LHC VOs
- VO/CC Rate discrepancies maybe due to files on tape buffer?

A-DT: LHCB STAGINGS

VO VIEW	A-DT (stagings)	
	MAX	AVG
LHCB	4.5GB/s	2GB/s

- LHCb stagins in 1 bunch of 44k requests
- LHCb smaller AVG file size than CMS (bigger files => better performance)
- Competition from ATLAS
- VO/CC Rate discrepancies maybe due to files on tape buffer?

ISSUES PER VO

- **CMS**:
  - During DT (as scheduled): non-purged tape buffer => rescheduled

- **ATLAS**:
  - During DT: RFIO server overload due to CMS stagings (2h errors) => splitting RFIO server

- **LHCB**:
  - During DT: HPSS Tape mover down on 18/03 => +52TB (+25%) to tape buffer to digest LHCb writing activity
  - During A-DT: 4h timeout for bringonline requests => dCache configuration updated to 48h?

SUMMARY

- TAPE CHALLENGE @CC-IN2P3 successful both during DT and A-DT for all 3 VOs
- Some technical issues and configuration issues on HPSS and dCache side (fixed)
- Still improving storage system monitoring tools
- Future evolution of tape infrastructure during 2022 is expected to boost performance and reliability

Questions?