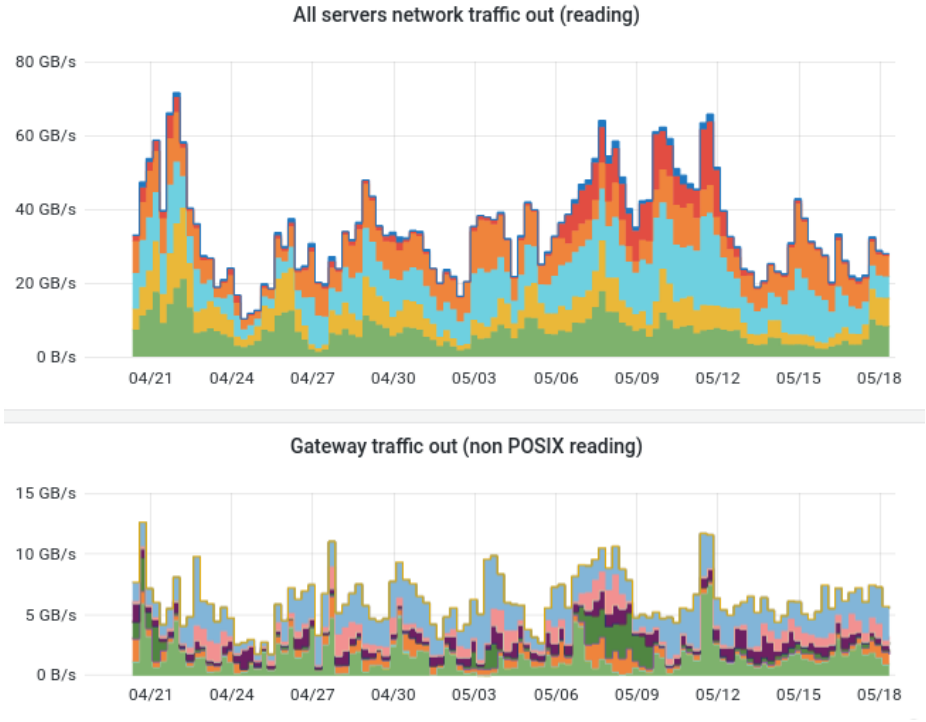
State of Storage

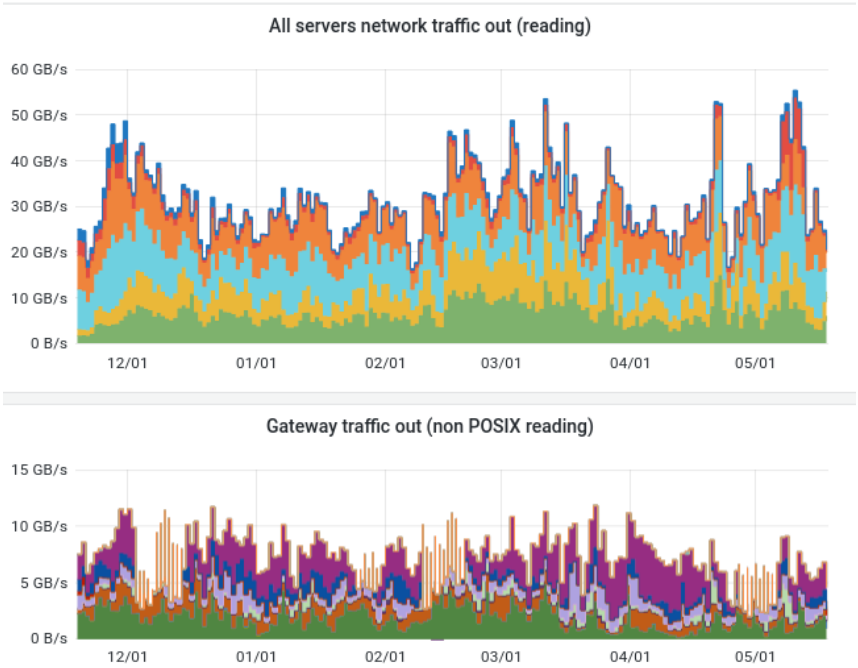
CdG 20 Maggio, 2022

Business as usual

Last month



Last 6 months



Disk storage in produzione

Installed: 50.07 PB, Pledge 2022: 59.1 PB, Used: 41.4 PB

Sistema	Modello	Capacita' netta, TB	Esperimenti	Scadenza
ddn-10, ddn-11	DDN SFA12k	10752	ALICE, AMS	03/2021→ 06/2023
os6k8	Huawei OS6800v3	3400	GR2, Virgo	06/2022
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	11/2021 → 12/2022
md-5, md-6, md-7	Dell MD3820f	28	metadati, home, SW	12/2022
os18k1, os18k2	Huawei OS18000v5	7800	LHCb	2023
os18k3, os18k5, os18k5	Huawei OS18000v5	11700	CMS	2024
ddn-12, ddn-13	DDN SFA 7990	5060	GR2,GR3	2025
ddn-14, ddn-15	DDN SFA 2000NV	24	metadati	2025
os5k8-1,os5k8-2	Huawei OS5800v5	8999	ATLAS	2027
Cluster CEPH	12xSupermicro SS6029	3400	ALICE, cloud, etc	2027

Prossimi acquisti

- Gara storage 2022 (14PB netti)
  - Capitolato e il resto della documentazione mandata a Frascati per l'approvazione
  - Installazione e messa in PROD verso fine dell'anno
- Acquisto di emergenza (solo dischi da inserire in sistemi già esistenti)
  - DDN - 76 slot liberi x14TB = 1064TB raw (850TB netti) → gpfs_data
    - In arrivo
- Ulteriore acquisto di tape (14PB per arrivare a pledge 2022) a breve

Current SW in PROD

- GPFS 5.0.5-9 (to be updated soon to 5.1.2-4)
- StoRM BackEnd 1.11.21 (latest)
- StoRM FrontEnd 1.8.15 (latest)
- StoRM WebDAV 1.4.1 (latest)
- StoRM globus gridftp 1.2.4
- XrootD 4.11.2
  - updated to 4.12.4 in the 4 CMS servers
  - 5.3.1-1 on CMS redirectors (local and EU/IT/FR)

Recent problems

- ATLAS
  - Transfer and deletion errors; we suspect CERN Grid CA started using its new certificate in advance wrt what announced (GGUS 156979)
  - Transfer failures to INFN-T1_MCTAPE; buffer full (GGUS 156762)

- CMS
  - SAM warning cert expiring in 6 days (GGUS 157255)
  - Migration of JINR_Tape files to T1_IT_CNAF_Tape (GGUS 157161)
  - WebDAV SAM tests failing (GGUS 156967)
    - Increased connections in pool; gridftp moved to different servers
  - Destination Overwrite error at T1_IT_CNAF_Tape; files removed (GGUS 156487)
  - Deletion of files from previous tape challenge (GGUS 156236)

Recent problems

- LHCb
  - Server side credential failures; mistake in the upgrade of server cert (GGUS 157208); proposing a change in LHCb folder structure
    - Following this activity together with Lucio
    - Also, tape challenge reprise next week with new tests

- ALICE
  - Need to upgrade kernel (reboot) for ds-801, ALICE XrootD redirector (currently a single point of failure)
    - Switch to redirector in each server mode on hold
      - Already in production in the xrootd-ceph cluster

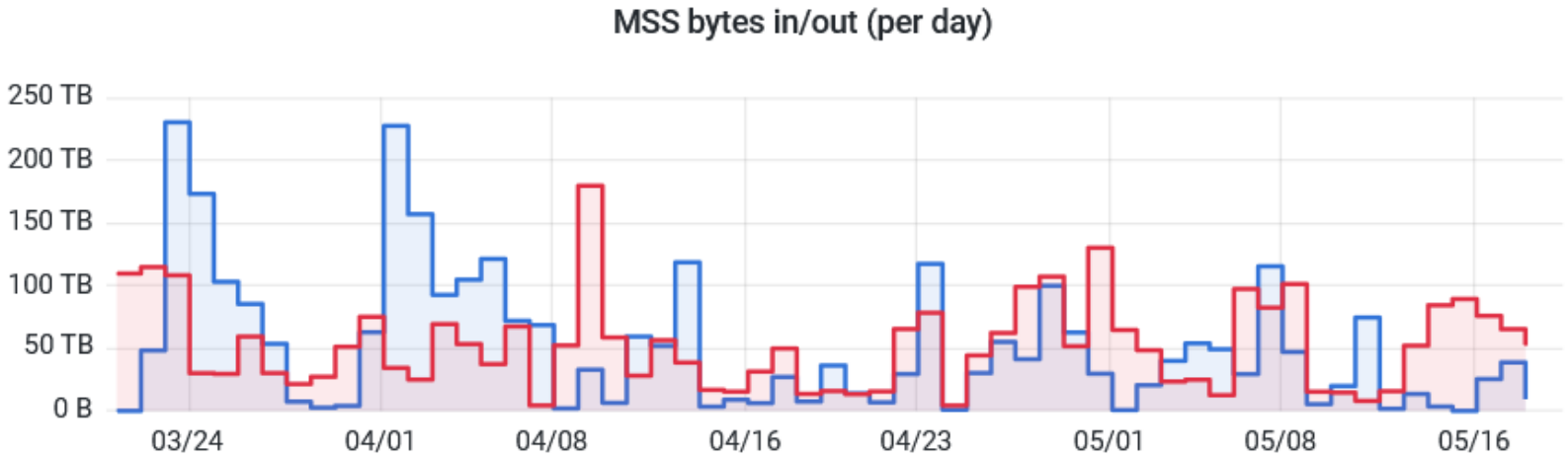
Actions

1. **Need to align authz policies for ATLAS, CMS and LHCb**
   - Only atlasprd can write in /atlasdatadisk, /atlasdatatape, /atlasgrouptape, /atlasmctape
   - Every CMS user can r/w in /cmsdisk/store/temp/user and /cmsdisk/store/user; only cmsprd can write in /cmsdisk/store and /cmstape
   - Every LHCb user can r/w in /disk/lhcb/user /disk/lhcb/failover; only lhcbprd can write in /disk/lhcb and /tape/lhcb

1. **Need to schedule a downtime for expansion of DDN storage (June?)**
   - gpfs_data and gpfs_virgo not available

Stato tape

19 Mar 2022 - 18 May 2022



	min	max	avg	current	total
out traffic (recalls)	105 GB	230 TB	51.0 TB	9.38 TB	3.01 PB
in traffic (migrations)	4.39 TB	180 TB	52.6 TB	52.1 TB	3.10 PB

Stato tape

- Liberi 13.4 PB (su cassette vuote, complessivamente sulle 2 librerie). Usati 96.6 PB.
  - Pledge 2022: 130.5 PB
  - Installato attuale: 116.5 PB
  - Spazio teoricamente ricavabile da reclamation: 4.7 PB
    - di cui ~1 PB sarà liberato entro qualche settimana
    - ~3.5 PB difficilmente recuperabili (inefficienza fisiologica del 3%)
  - Spazio libero su cassette semipiene: 1 PB

Library	Tape drives	Max data rate/drive, MB/s	Max slots	Max tape capacity, TB	Installed cartridges	Used capacity, PB
SL8500 (Oracle)	16*T10KD	250	10000	8.4	~10000	75.2
TS4500 (IBM)	19*TS1160	400	6198	20	1750	21.4

Prossimo acquisto tape

- Cancellazioni
  - ATLAS ha appena rimosso 2 PB
    - Space reclamation in corso
  - Usato CMS oltre il pledge (0.9 PB)
    - Prevista a breve campagna di cancellazione. Quando? Quanti dati?
  - Cancellazioni LHCb a breve?
- Scritture LHC intense non prima di
  - Agosto/settembre (CMS/LHCb)
  - Dicembre (ALICE)
- Con info più precise sulle prossime cancellazioni può valere la pena attendere qualche settimana per acquisto nuove tape
  - Per avere un dato di spazio libero più attendibile
  - Nuova tecnologia disponibile nel corso del 2023

Uno sguardo al 2023

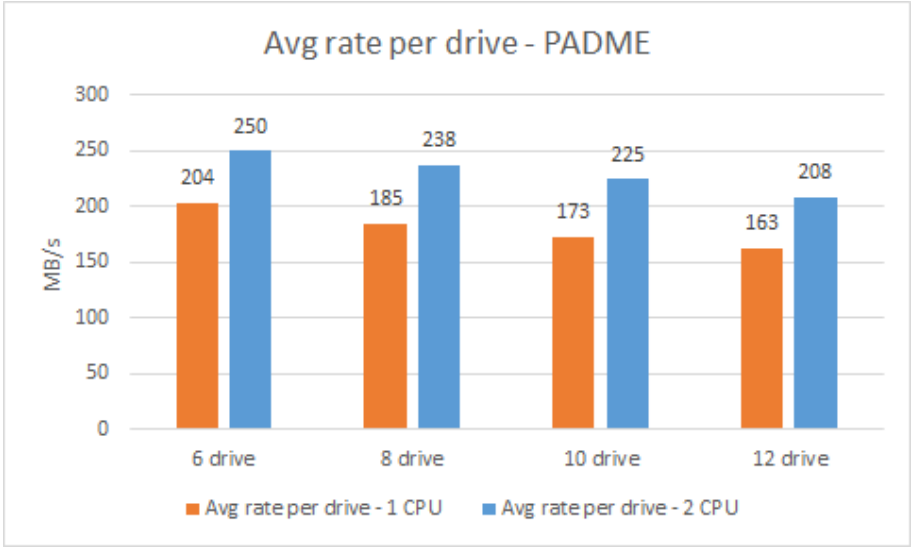
	ALICE	ATLAS	CMS	LHCb	No LHC	Totale
Pledge 2023 (PB)	23.7	32	41	26.6	38.1*	161.4
Δ 2023 da 2022 (PB)	4.8	7.5	7.2	3.6	7.8*	**30.9**

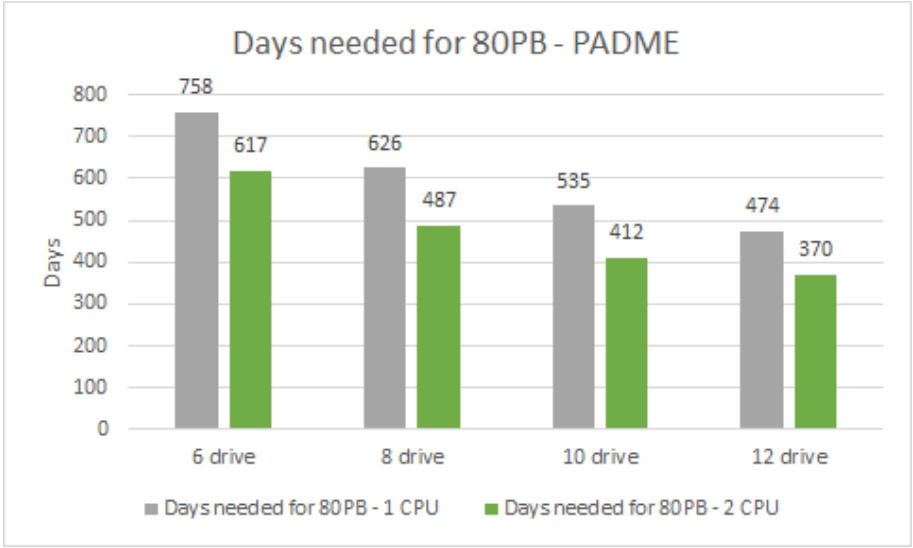
* Considerando lo stesso Δ del 2022 rispetto al 2021

- Su libreria IBM abbiamo 70 PB potenziali disponibili
  - Una volta esaurito il pledge 2022
  - Non è possibile acquistare ulteriori moduli della libreria prima dello spostamento al Tecnopolo (probabilmente fine 2023)
  - Spazio sufficiente per pledge 2023
  - Nuova tecnologia disponibile nel corso del 2023
    - Cassette più capacitive -> ulteriore spazio potenziale su libreria

Test di repack

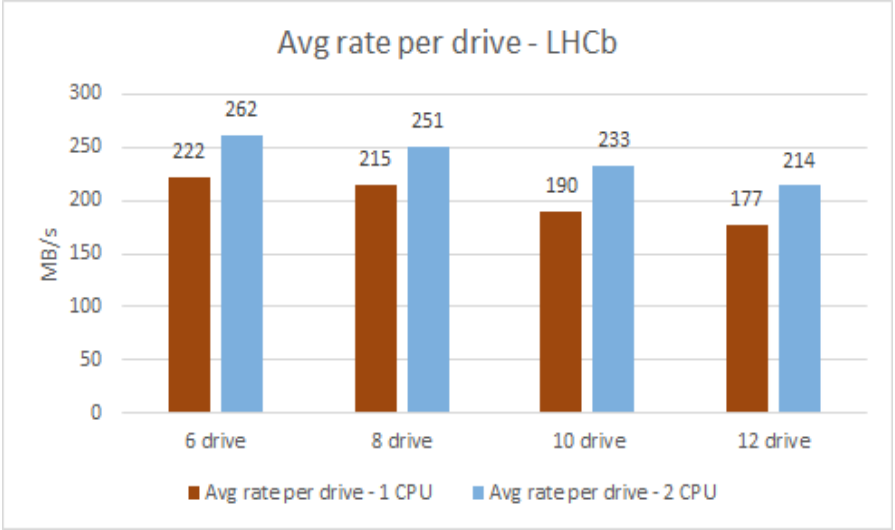
- In vista della possibile dismissione della libreria Oracle
  - A un drive Oracle che legge corrisponde un drive IBM che scrive sulla nuova libreria
  - Massimo rate per drive: 250 MB/s
  - Sofferenza della CPU del TSM server in operazioni amministrative con diversi drive
  - Test via TSM server con 1 CPU (setup attuale) o 2 CPU
  - PADME: dimensione media file di 670 MB

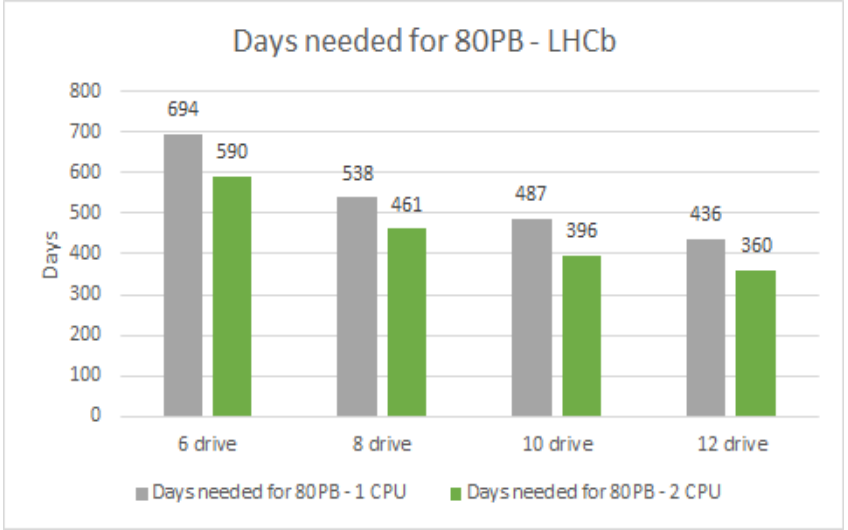




Test di repack

- LHCb: dimensione media file di 1.65 GB





- A parità di tempo necessario per effettuare l'intero repack di 80 PB, la seconda CPU permette di farlo con ~2 drive in meno
- Da confrontare costi tape drive con costi licenze TSM per seconda CPU

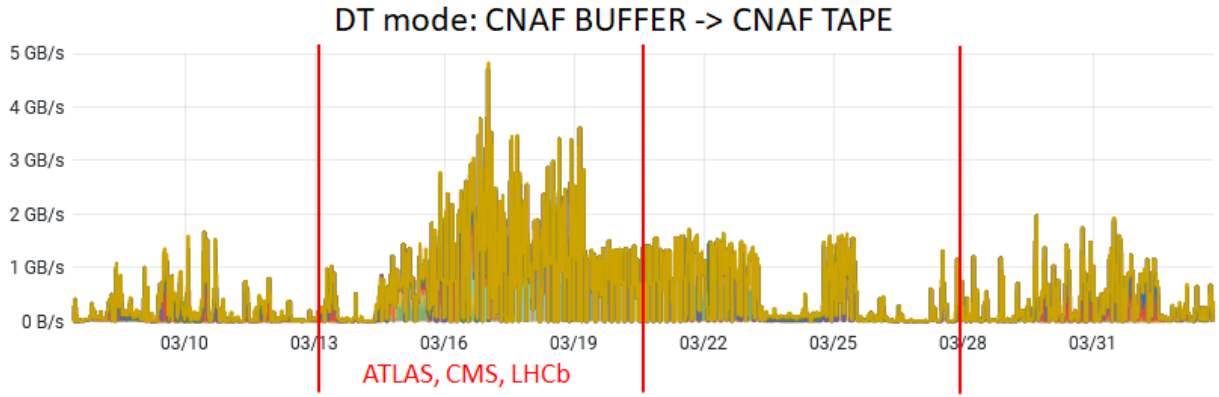
Tape challenge: risultati

	Writes (DT)			Reads (A-DT)		
VO	Expected GB/s	Result avg GB/s	Target achieved?	Expected GB/s	Result avg GB/s	Target achieved?
ATLAS	0.9	1	YES	0.8	1.9	YES
CMS	0.37	1.1	YES	NA	1.46	YES
LHCb	1.72	1.52/1.72*	~YES*	1.35	1.8	YES

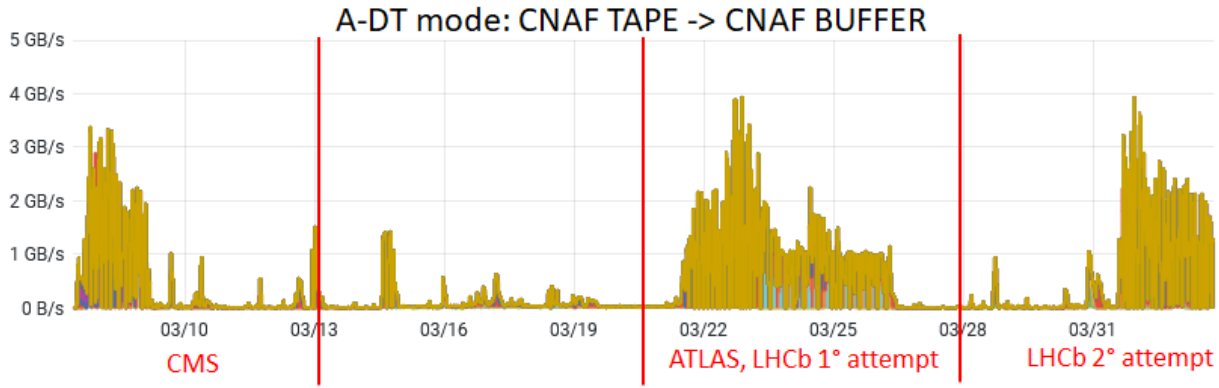
* Migration rate reaches the target if buffer has enough data to migrate

- Target raggiunti
- Ulteriori dettagli: https://indico.cern.ch/event/1145328/

Tape challenge: rate buffer - tape



- Avg on 30 mins
- Peak: **4.9 GB/s**

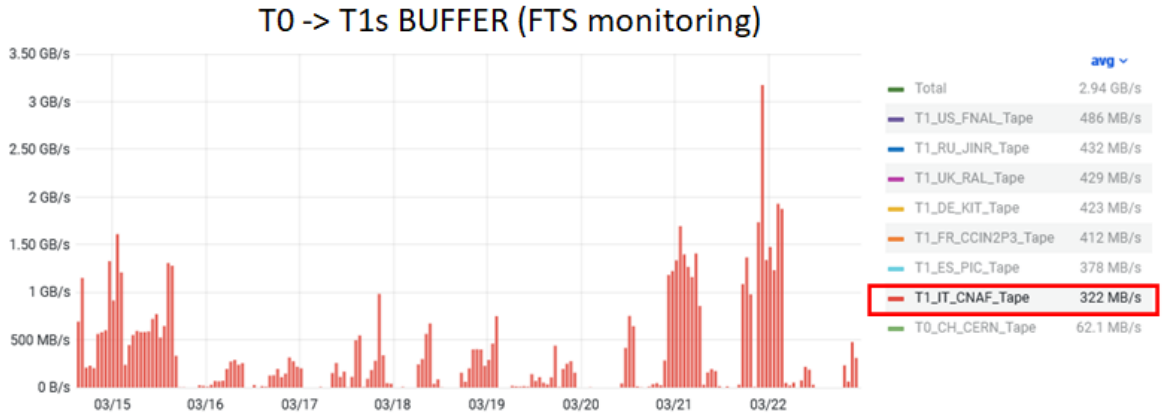


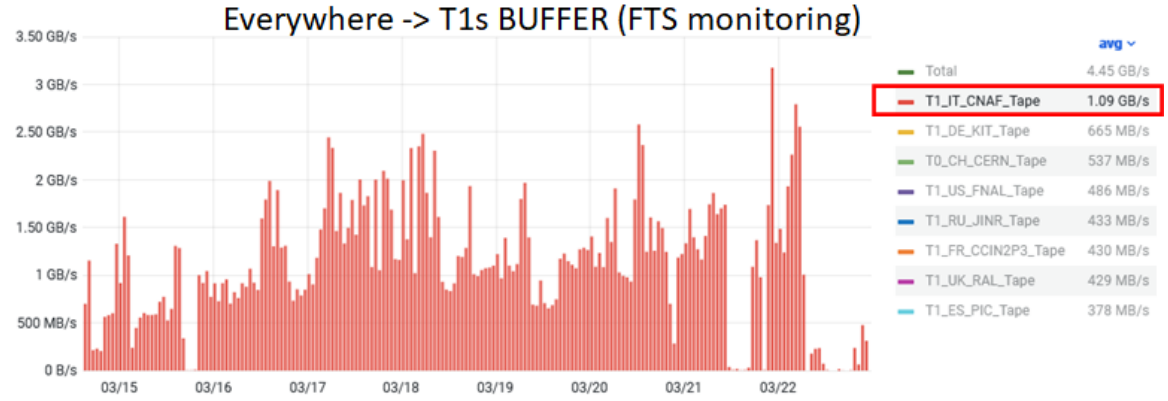
- Avg on 30 mins
- Peak: **4 GB/s**

Tape challenge: scritture LHCb

- Fallimenti FTS pre-trasferimento CNAF disco -> CNAF tape
  - Messaggio di errore: thread pool di webdav pieno
  - Strano failure rate del 75% (con 4 webdav server)
  - Non accade per CERN -> CNAF disco
  - In ogni caso, buon rate complessivo
  - Nuovi test settimana prossima
  - Scrittura su disco e poi su buffer ridondante
    - Stesso FS e stesso hardware

Tape challenge: scritture CMS





- Rate calcolato da CMS: 0.32 GB/s
  - Solo da CERN a CNAF buffer
- Rate complessivo stesso periodo: 1.09 GB/s
  - Da ovunque a CNAF buffer

Tape challenge: letture

- Rate medi per drive
  - ATLAS: 211 MB/s (85% rate nominale)
  - CMS: 146 MB/s (59% rate nominale)
  - LHCb: 360 MB/s (90% rate nominale)
- Una certa inefficienza è dovuta al carico sui server causato dallo scan del FS
  - Più importante per ATLAS e CMS (numero di file maggiore)
- Dati di CMS distribuiti in più cassette

	ATLAS	CMS	LHCb
Total files	73902	10761	61378
Total tapes	52	164	18
Total GiB	157943	164397	349895
Total GB	169590	176519,9	375696,9
Avg GiB per file	2,1	15,3	5,7
Avg GB per file	2,3	16,4	6,1
Avg files per tape	1421	66	3410
Avg GiB per tape	3037	1002	19439
Avg GB per tape	**3261**	**1076**	**20872**