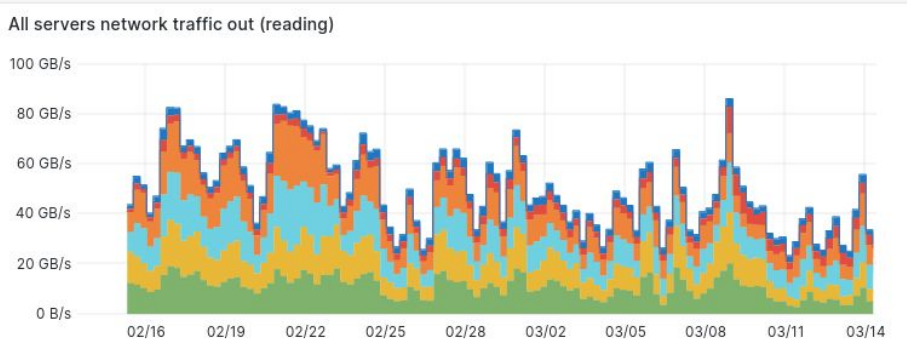
INFN CNAF

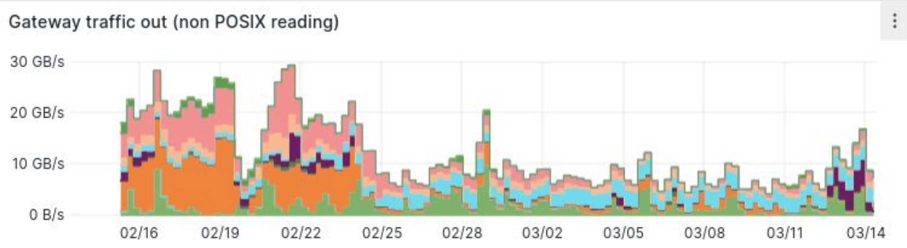
State of Storage

CdG 15 marzo, 2024

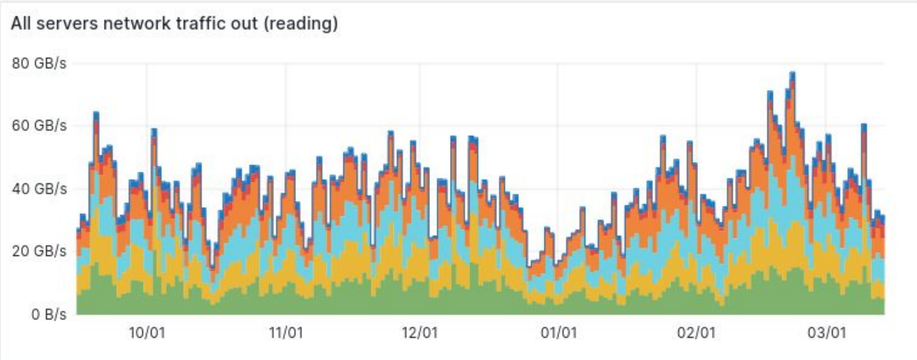
Business as usual

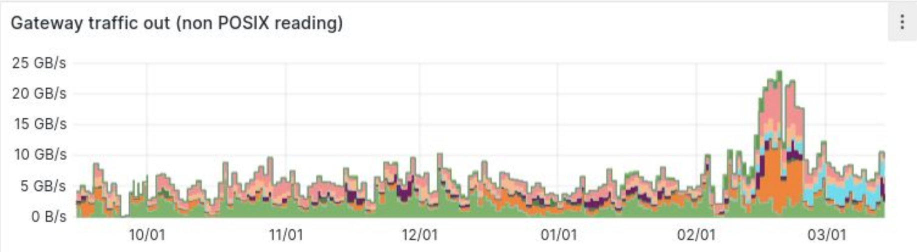
Last month


All servers network traffic out (reading)


Gateway traffic out (non POSIX reading)

Last 6 months


All servers network traffic out (reading)


Gateway traffic out (non POSIX reading)

Disk storage in produzione

Installed: 53.64 PB, Pledge 2023: 69.6 PB, Used: 48.8 PB

Storage system	Model	Net capacity, TB	Experiment	End of support
ddn-10, ddn-11	DDN SFA12k	**10120**	ALICE, AMS	**12/2022** (+10 spare hdd)
os6k8	Huawei OS6800v3	3400	GR2, Virgo	12/2024
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	05/2024
md-5, md-6 e md-7	Dell MD3820f	50	metadati, home, SW	11/2023 e 12/2024
os18k1, os18k2	Huawei OS18000v5	7800	LHCb	7/2024
os18k3, os18k5, os18k5	Huawei OS18000v5	11700	CMS	6/2024
ddn-12, ddn-13	DDN SFA 7990	5840	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

Acquisti recenti e futuri

- Gara storage 2022 (14PB netti)
  - LENOVO DE6600: Collaudo non superato
    - Nuova proposta con apparati DDN SFA7990X
- AQ storage 2023-2024
  - Il vincitore è Huawei con sistemi OceanStore Micro 1500/1600
  - Richiesta fornitura di 64PB nel 2023
  - Installazione al Tecnopolo iniziata
- Gara Tape Library
  - Contratto fermo in AC, possibile ritardo di qualche mese
- Gare nastri
  - Ulteriori 30 PB di nastro per repack dei dati dalla libreria Oracle
    - Gara in preparazione

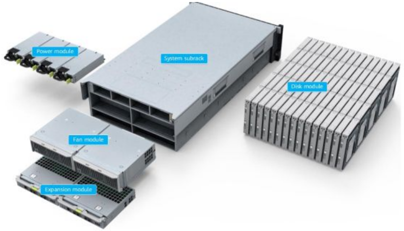


Figure 6 - Structure of the 4U high density disk enclosure

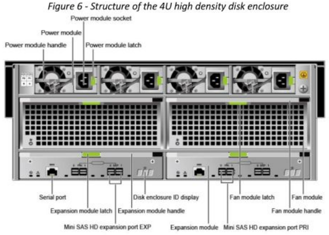


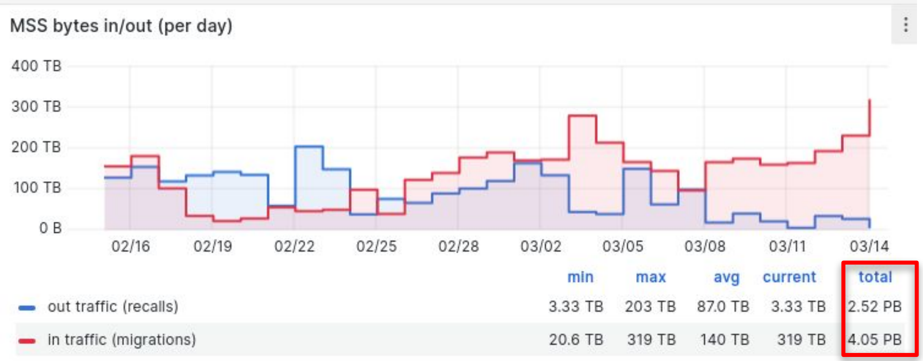
Figure 7 – Rear view of the 4U high density disk enclosure

Current SW in PROD

- GPFS 5.1.2-13
- StoRM BackEnd 1.11.22 (latest)
- StoRM FrontEnd 1.8.15 (latest)
- StoRM WebDAV 1.4.2 (latest)
- StoRM globus gridftp 1.2.4
- XrootD 5.5.4-1
    - ALICE CEPH updated to 5.5.5-1.el8
- Ceph 16.2.6 (Pacific)

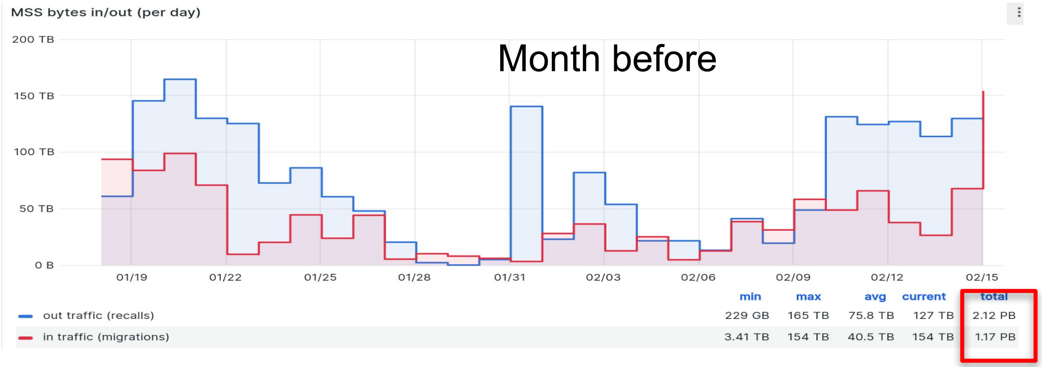
Stato tape

Last month



MSS bytes in/out (per day)

	min	max	avg	current	total
out traffic (recalls)	3.33 TB	203 TB	87.0 TB	3.33 TB	2.52 PB
in traffic (migrations)	20.6 TB	319 TB	140 TB	319 TB	4.05 PB

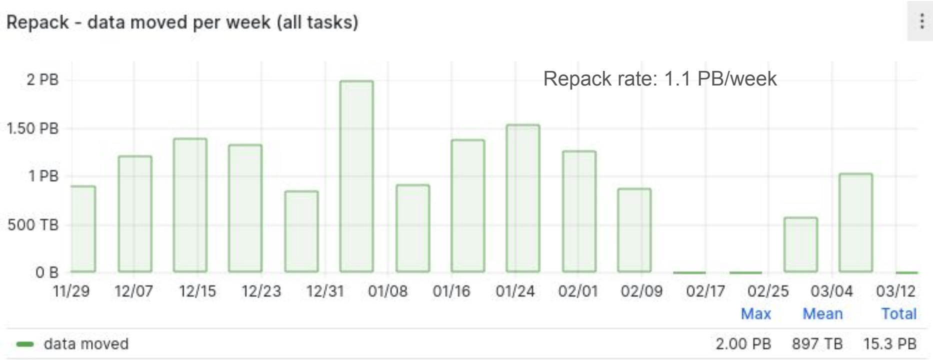


MSS bytes in/out (per day)

Month before

	min	max	avg	current	total
out traffic (recalls)	229 GB	165 TB	75.8 TB	127 TB	2.12 PB
in traffic (migrations)	3.41 TB	154 TB	40.5 TB	154 TB	1.17 PB

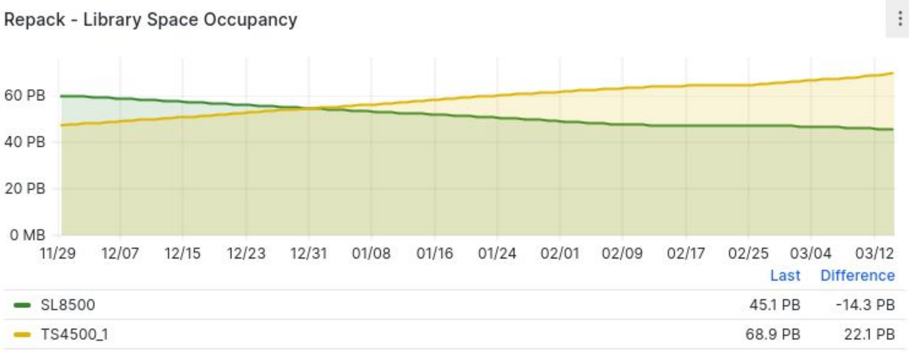
Tapes: Migration from Oracle to IBM library



Repack - data moved per week (all tasks)

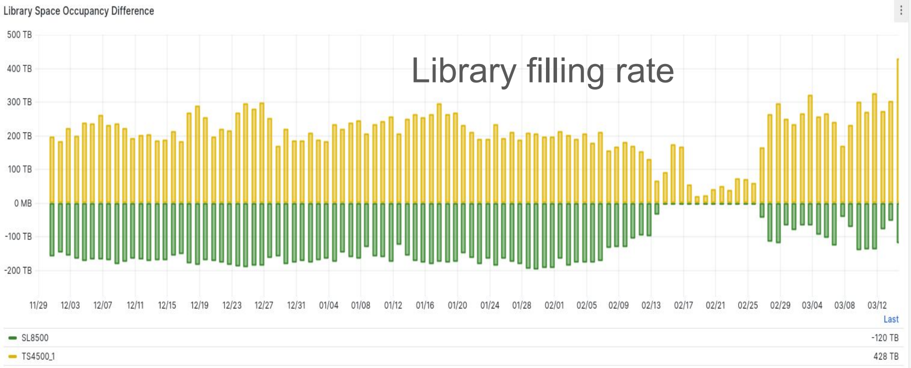
Repack rate: 1.1 PB/week

	Max	Mean	Total
data moved	2.00 PB	897 TB	15.3 PB



Repack - Library Space Occupancy

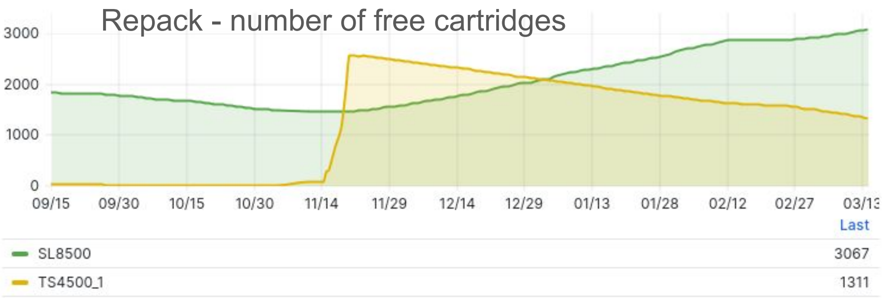
	Last	Difference
SL8500	45.1 PB	-14.3 PB
TS4500_1	68.9 PB	22.1 PB



Library Space Occupancy Difference

Library filling rate

	Last
SL8500	-120 TB
TS4500_1	428 TB



Repack - number of free cartridges

	Last
SL8500	3067
TS4500_1	1311

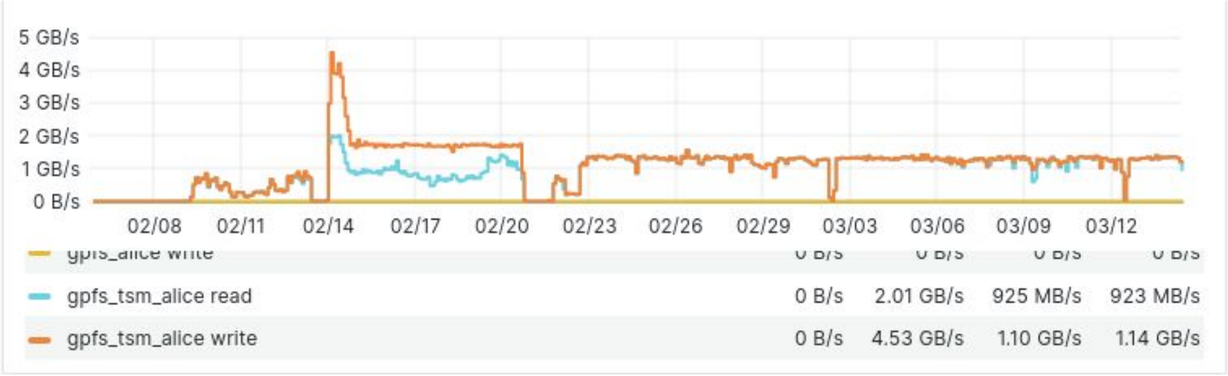
Stato tape

- Liberi ~26 PB (Scratch tape sulla libreria IBM).
- Usati ~119 PB.
    - In preparazione gara per altri 30 PB

Library	Tape drives	Max data rate/drive, MB/s	Max slots	Max tape capacity, TB	Installed cartridges	Used space, PB	Free space, PB
SL8500 (Oracle)	16*T10KD	250	10000	8.4	~10000	**45**	-
TS4500 (IBM)	19*TS1160	400	6198	20	5104	**68.9**	**25.6**

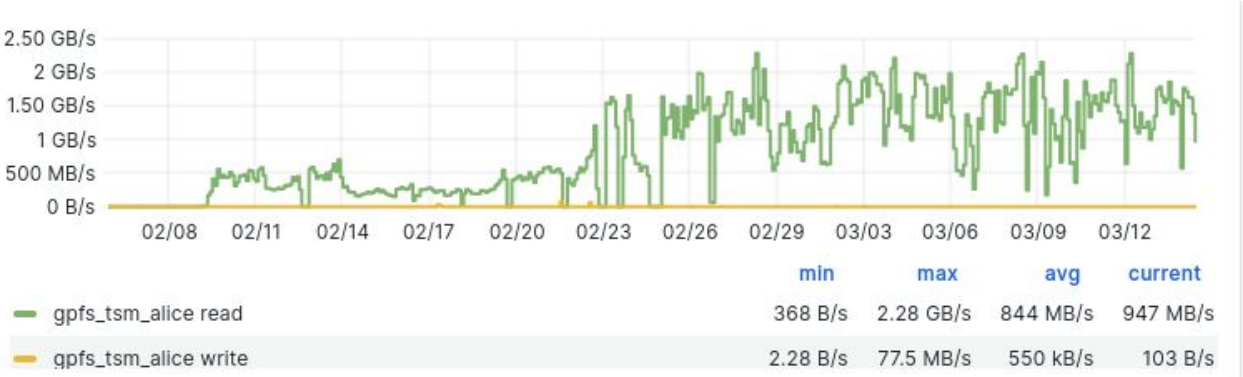
ALICE: CERN→CNAF tape

CERN→disk buffer



gpfs_alice write	0 B/s	0 B/s	0 B/s	0 B/s
gpfs_tsm_alice read	0 B/s	2.01 GB/s	925 MB/s	923 MB/s
gpfs_tsm_alice write	0 B/s	4.53 GB/s	1.10 GB/s	1.14 GB/s

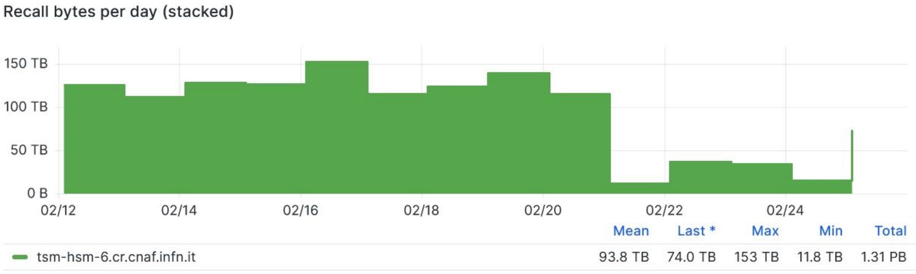
Disk buffer →tape



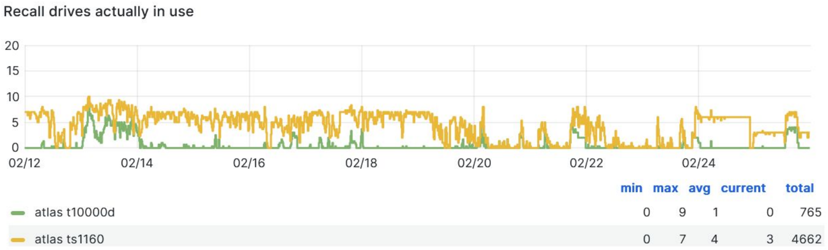
	min	max	avg	current
gpfs_tsm_alice read	368 B/s	2.28 GB/s	844 MB/s	947 MB/s
gpfs_tsm_alice write	2.28 B/s	77.5 MB/s	550 kB/s	103 B/s

Data Challenge 2024

- Yesterday we presented our impressions at the DOMA Retrospective (talk)
  - The DC24 time-range should be excluded from A/R computation, see e.g. https://ggus.eu/index.php?mode=ticket_info&ticket_id=165509
  - FTS optimizer reducing transfers when failures arise would be very welcome
  - Very difficult to distinguish between DC and production load; very difficult for the sites to debug when production activity have an heavy impact



	Mean	Last *	Max	Min	Total
🟩 tsm-hsm-6.cr.cnaf.infn.it	93.8 TB	74.0 TB	153 TB	11.8 TB	1.31 PB



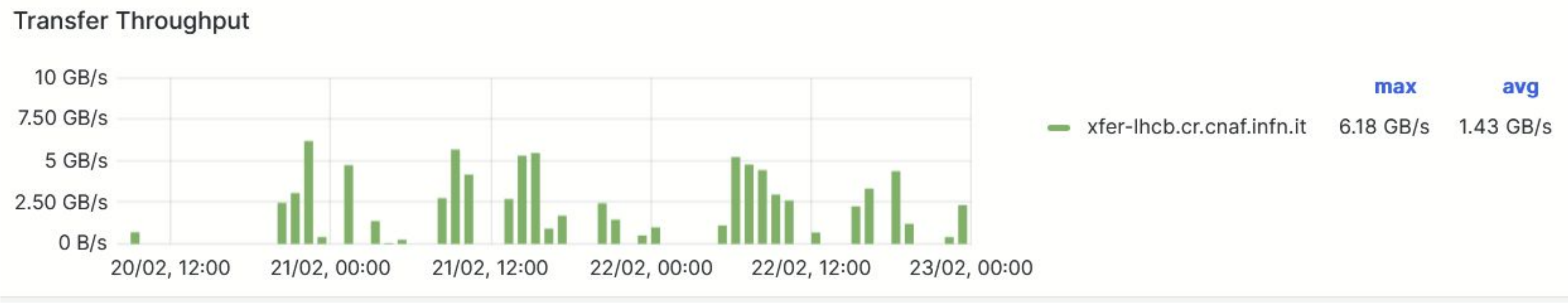
	min	max	avg	current	total
🟩 atlas t10000d	0	9	1	0	765
🟨 atlas ts1160	0	7	4	3	4662

Staging activity from ATLAS during the DC

Data Challenge 2024

We have difficulties interpreting FTS monitoring

- Rates averaged over no-transfers periods



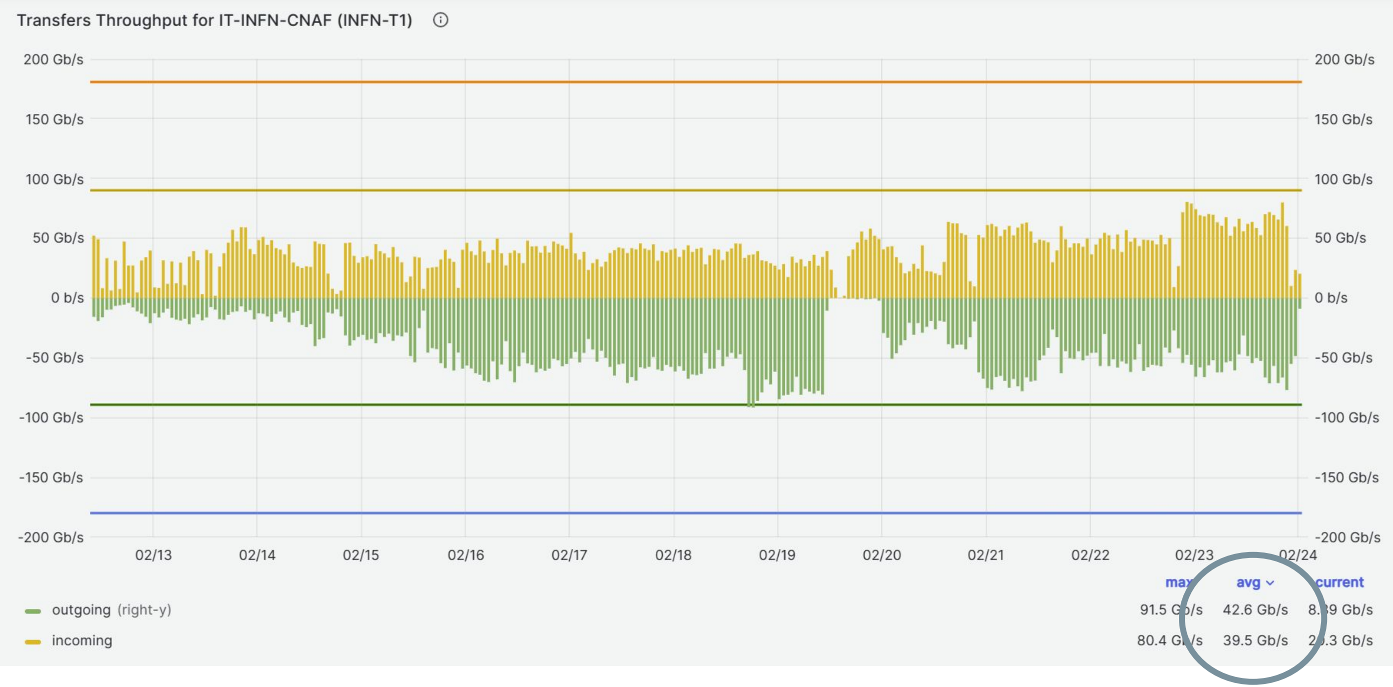
LHCb, Tape-Disk, which is actually Disk_buffer-Disk
(FTS plot provided by A. Rogovskiy)

Data Challenge 2024

We have difficulties interpreting FTS monitoring

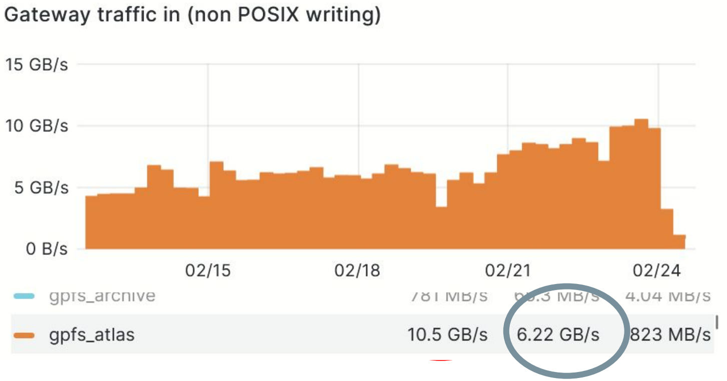
- Throughputs reported for our site are much lower than what we observe
    - Are we measuring an important contribution from production load?
        - Unfortunately, we cannot disentangle.
    - Is FTS throughput computed and reported only for successful transfers?
        - Again, unfortunately we cannot disentangle in the traffic we measure.
        - Shouldn’t success rate be reported together with throughput?

Monit plot provided by A. Forti for ATLAS+CMS



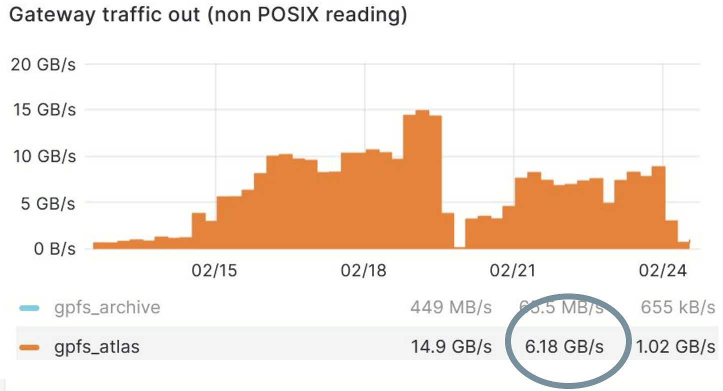
	max	avg	current
🟩 outgoing (right-y)	91.5 Gb/s	42.6 Gb/s	8.39 Gb/s
🟨 incoming	80.4 Gb/s	39.5 Gb/s	2[illegible].3 Gb/s

Gateway traffic in (non POSIX writing)



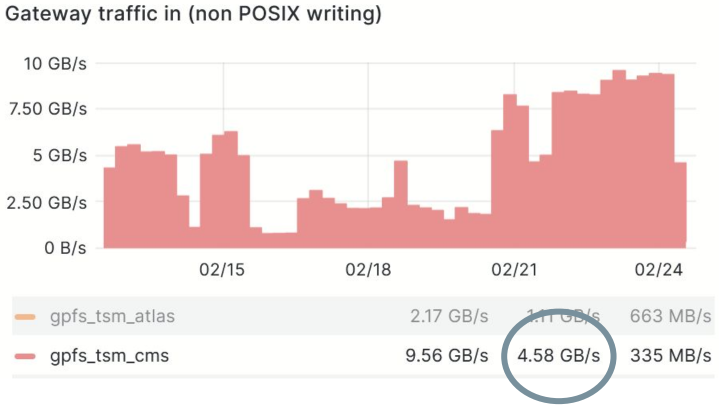
gpfs_archive	781 MB/s	69.3 MB/s	4.04 MB/s
gpfs_atlas	10.5 GB/s	6.22 GB/s	823 MB/s

Gateway traffic out (non POSIX reading)



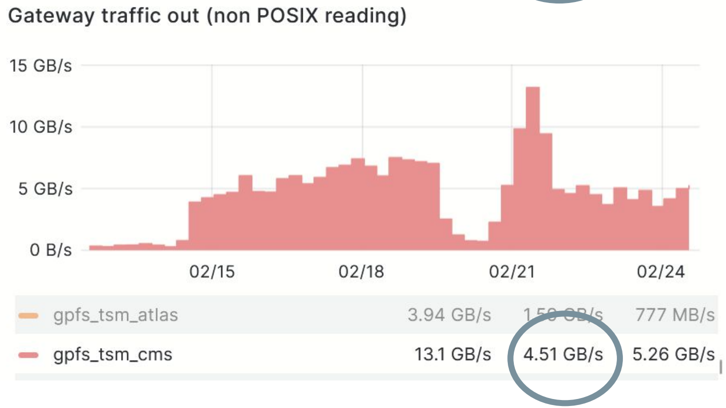
gpfs_archive	449 MB/s	65.5 MB/s	655 kB/s
gpfs_atlas	14.9 GB/s	6.18 GB/s	1.02 GB/s

Gateway traffic in (non POSIX writing)



gpfs_tsm_atlas	2.17 GB/s	1.11 GB/s	663 MB/s
gpfs_tsm_cms	9.56 GB/s	4.58 GB/s	335 MB/s

Gateway traffic out (non POSIX reading)



gpfs_tsm_atlas	3.94 GB/s	1.50 GB/s	777 MB/s
gpfs_tsm_cms	13.1 GB/s	4.51 GB/s	5.26 GB/s

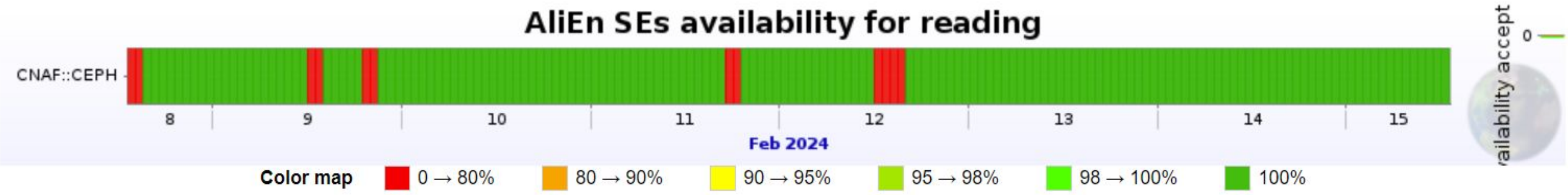
We measure 85 Gb/s OUT and 86 Gb/s IN (FTS says 42 Gb/s OUT and 39.5 Gb/s IN)

Data Challenge 2024 - and now?

- Currently re-running the DC24 for LHCb with 50 instead of 200 FTS max transfers
  - Also gpfs pagepool increased
- Currently re-running T0 export for ATLAS
  - We'll repeat the test with a new rpm for StoRM WebDAV, improving efficiency (https://github.com/italiangrid/storm-webdav/pull/40)
- We'll align the StoRM WebDAV instances dedicated to CMS since we observed higher load and higher failures in those servers having lower number of CPU cores
- We'll re-think LHCb hardware configuration so to accommodate their workflow, given @INFN-T1 tape buffer and disk are on the same filesystem, managed by the same endpoints

Tickets and more

- ALICE
  - Finishing XrootD configuration restyling of GPFS cluster:
    - Manage configuration files with Puppet
    - Revert to original configuration with all servers acting as managers as well
    - Upgrade to latest version in production (5.5.4-1.el7)
    - Check on the status of the service has been included within sensu framework of check and remediation
    - Waiting for the end of the Pb - Pb data transfer (approximately 2 months) to finalize the configuration with the tape cluster (xs-204, xs-304)
  - Misalignment between MonaLisa and our alerting system
    - We contacted Mario and Francesco to replicate MLsensor behaviour on our cluster



Tickets and more

- ATLAS
  - GGUS 165526: DC24 T0-T1 test repetition
  - GGUS 165355: 'SSL-connect' transfer errors due to overloading of the endpoints during the DC24
- CMS
  - GridFTP still used, only for SAM tests
  - StoRM Tape REST installed and configured; no tests yet and it did not help in getting rid of GridFTP
  - GGUS 165479: 12 recall requests stuck in FTS
    - The original BOL request was processed by StoRM backend during a restart of the service, on Feb 5th, generating an intermediate situation in the StoRM database, which was not purged by the garbage collector due to a known issue.
  - GGUS 165276, GGUS 165183: staging not working due to wrong configuration

Tickets and more

- LHCB
  - GGUS 165648 (in progress): new VOMS configuration to be added before April 10th but not now
  - GGUS 165048 (in progress): LHCb token authentication for disk storage
    - WLCG-scope-based token AuthZ implemented for disk storage area
    - StoRM WebDAV does not support full path scope
      - Access point/root path cannot be part of the scope path
    - Involved StoRM developers
      - Discussion ongoing
  - GGUS 164032
    - Assigned to StoRM
    - StoRM WebDAV provides storage tokens (macaroons) via the oauth/token endpoint, but FTS retrieves macaroons from the resource path. This should not be allowed by StoRM WebDAV (STOR-1602), no matter permissions of the storage area.

Tickets and more

- Gsiftp protocol via StoRM backend is still available for a few experiments
  - Tests srm+https and feedback very welcome (Belle, Xenon)
  - Goal: remove gsiftp protocol and switch off GridFTP
- CTA-LST
  - GridFTP switched on for CTA-LST storage areas to allow Third-party copies from PIC (gsiftp)
- Dampe
  - GridFTP “plain” still used
    - Testing XrootD server at IHEP to perform the transfers to CNAF (WP6-Datacloud)
- Virgo
  - Intensive usage of stashcache (downtime added in OSG, waiting for their feedback)
  - Hit the 2x10Gb=2.5GB/s limit traffic OUT