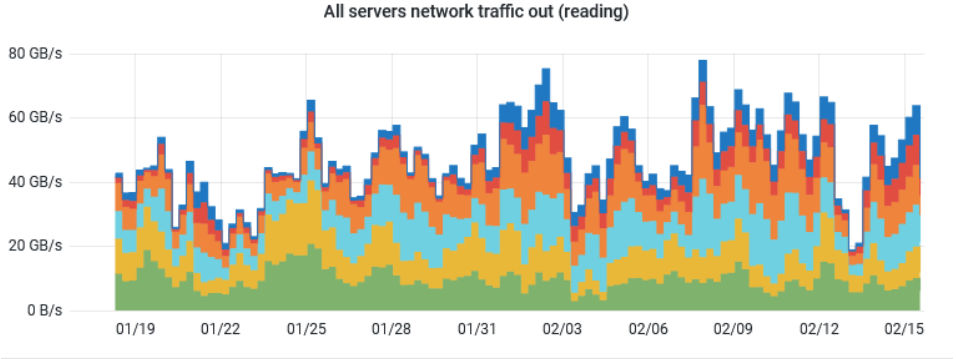
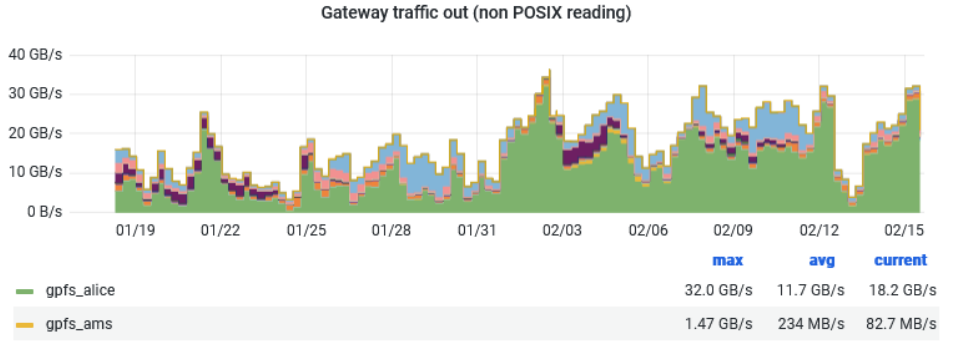
State of Storage

CdG 17 febbraio, 2023

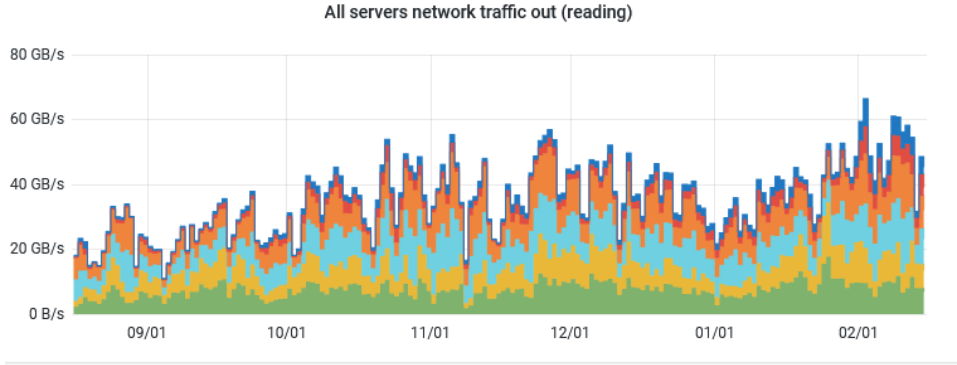
Business as usual

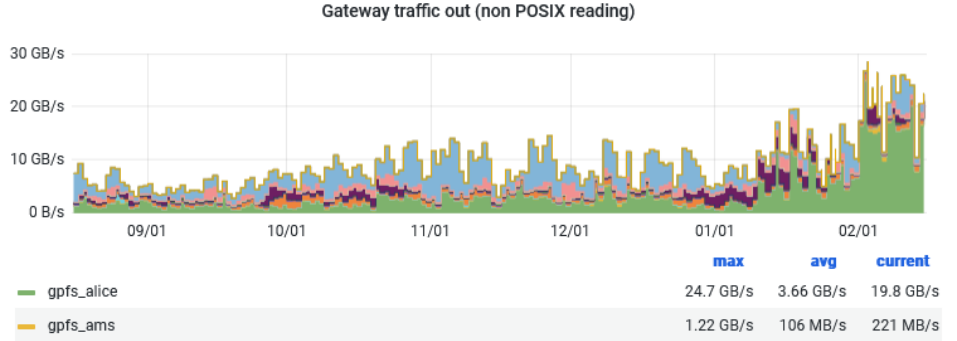
Last month





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	12/2022
os6k8	Huawei OS6800v3	3400	GR2, Virgo	12/2023
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	11/2023
md-5, md-6 e **md-7**	Dell MD3820f	50	metadati, home, SW	**11/2023** e **12/2024**
os18k1, os18k2	Huawei OS18000v5	7800	LHCb	2023
os18k3, os18k5, os18k5	Huawei OS18000v5	11700	CMS	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

- Disco ulteriore per sistemi Huawei
  - 600TB netti
  - Arrivati a fine 2022, aggiunti ad ATLAS
- Gara storage 2022 (14PB netti)
  - Il vincitore è Lenovo con ThinkSystem DE6600
    (Derivato dal NetApp DE6600 con un enclosure con i dischi NVME)
  - Installazione e messa in produzione sperabilmente ad aprile/maggio 2023
- AQ storage 2023-2024
  - 64PB nel 2023 + 14PB nel 2024
  - Bando pubblicato

Current SW in PROD

- GPFS 5.0.5-13 (to be updated soon to 5.1.2-8)
  - Updated to 5.1.2-8 for CMS and LHCb
  - Will be updated everywhere together with the installation of a new kernel
- 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 5.4.2-1
  - 5.3.1-1 on CMS redirectors (local and EU/IT/FR)
  - To be upgraded for ATLAS and no-LHC (still at 4.11.2-1)
- Ceph 16.2.6 (Pacific)

Tickets and problems

- Ready to configure access with tokens for all the storage areas upon request

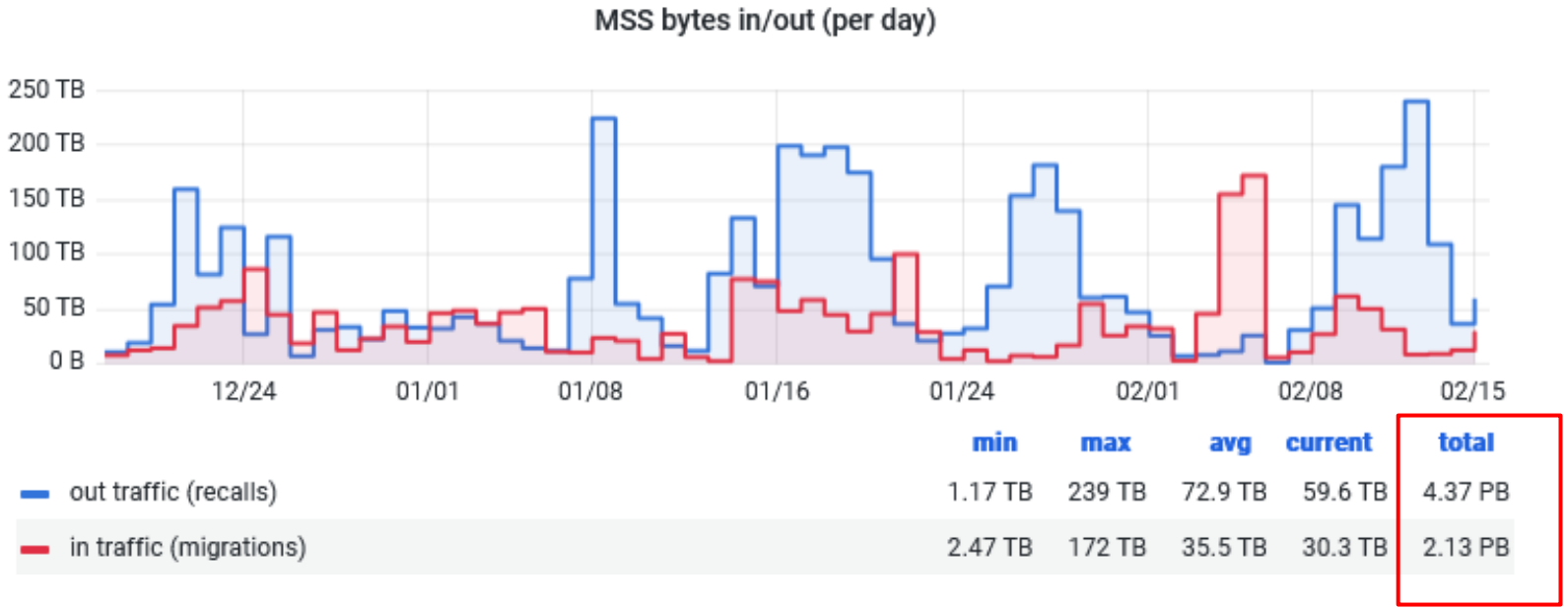
- ATLAS
  - 150 kB/s for BNL -> CNAF transfers; equipment failure in LHCOPN path from BNL
  - Switched off gridftp in December
  - TPCs in push-mode to GOOGLE_EU fail for all StoRM sites (GGUS 158487)
    - Issue for StoRM developers, on hold
- CMS
  - SAM tests fail given they keep testing gsiftp protocol, which had been removed in agreement with D. Spiga (GGUS 160530)
    - Gridftp restored, ready to switch off upon request
  - Xrootd cache is now authenticated (VOMS); tests are being executed from Vega
  - Quota increase by 700 TB in Rucio (GGUS 160141)
  - Issue opened to developers: thread saturation, log files stuck, need restart
    - still trying to verify the impact of xrootd.async off nosf

Tickets and problems

- LHCb
  - Tape data challenge (see below)
  - Switched off gridftp
  - Unable to use XrootD from Rhel 9(GGUS 160273)
    - Xrootd upgraded to 5.4.2-1
- storm-ams.cr.cnaf.infn.it (Ams, Dampe, Darkside, Juno) now uses the pool of gridftp servers of storm-archive.cr.cnaf.infn.it
- Belle II
  - Transfers failures due to configuration problem with KEK SE (GGUS 159984)
  - Refactoring of ACLs and default ACLs for the whole fileset
  - Waiting for S.Pardi to support transition srm+gsiftp->srm+https
- Virgo
  - Ongoing Stash-cache upgrade (OSG 3.5-> OSG 3.6)
- Configured StoRM WebDAV storage areas for Herd and Luna
- In a joint effort with User support, enforcing a transition towards StoRM WebDAV for all no-LHC experiments: Corelib, Dampe, Darkside, Km3, Limadou, Newchim

Stato tape

17 Dec 2022 - 15 Feb 2023



	min	max	avg	current	total
out traffic (recalls)	1.17 TB	239 TB	72.9 TB	59.6 TB	4.37 PB
in traffic (migrations)	2.47 TB	172 TB	35.5 TB	30.3 TB	2.13 PB

Stato tape

- Liberi 25 PB (su cassette vuote, complessivamente sulle 2 librerie). Usati 96 PB.
  - Pledge 2022/23: 134.3 PB (130.5 PB pledge 2022)
  - Installato attuale: 130.5 PB
  - Gara da 53 PB in uscita, per arrivare a 183.5 PB (pledge 2023)

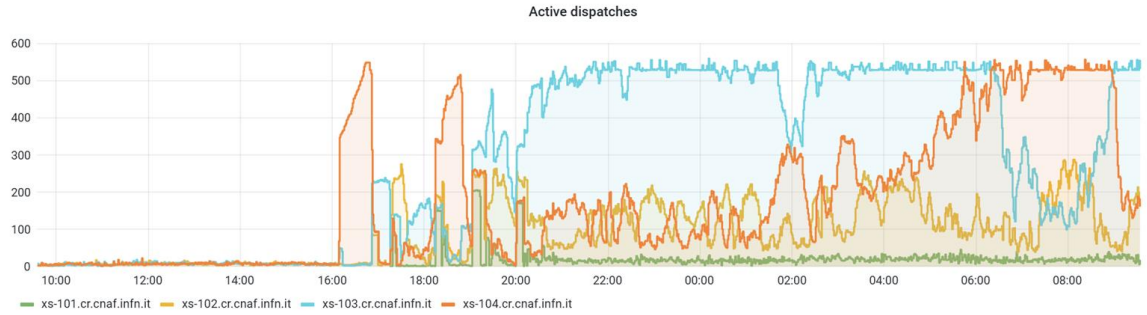
Library	Tape drives	Max data rate/drive, MB/s	Max slots	Max tape capacity, TB	Installed cartridges	Used capacity, PB
SL8500 (Oracle)	16*T10KD	250	10000	8.4	~10000	60
TS4500 (IBM)	19*TS1160	400	6198	20	2450	36

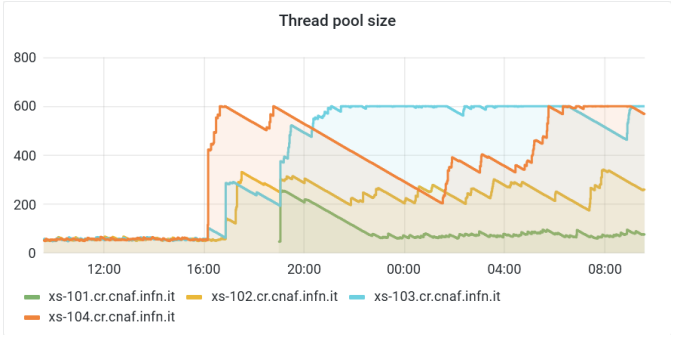
Evoluzione tape a medio termine

- Gara per acquisto nuova libreria in uscita
  - Libreria con supporto a drive IBM e LTO
    - In uscita nuova tecnologia tape drive e cassette
  - Da installare al Tecnopolo
- Repack dati da Oracle a IBM (60 PB)
  - Senza spostamento della libreria Oracle al Tecnopolo
  - Ulteriore acquisto di 30 PB di nastro nel 2023 e il resto nel 2024
  - Repack presumibilmente ultimato entro il 2024

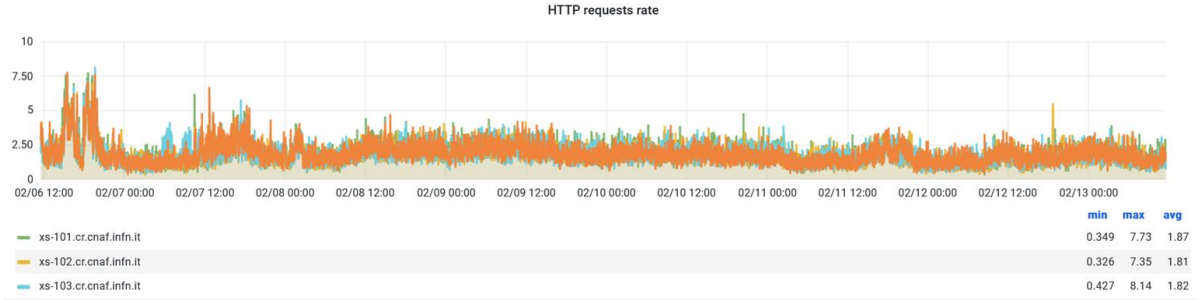
LHCb data challenge (load balancing)

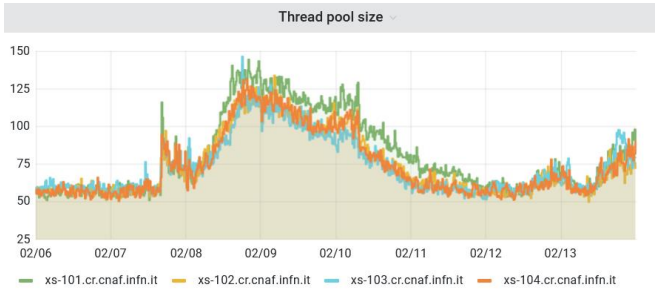
During writing EOS-> CNAF_DISK and CNAF_DISK->CNAF_Tape. Only the latter are balanced, thanks to storm endpoint





During recall, requests are much better balanced among hosts (FTS fix in place?)





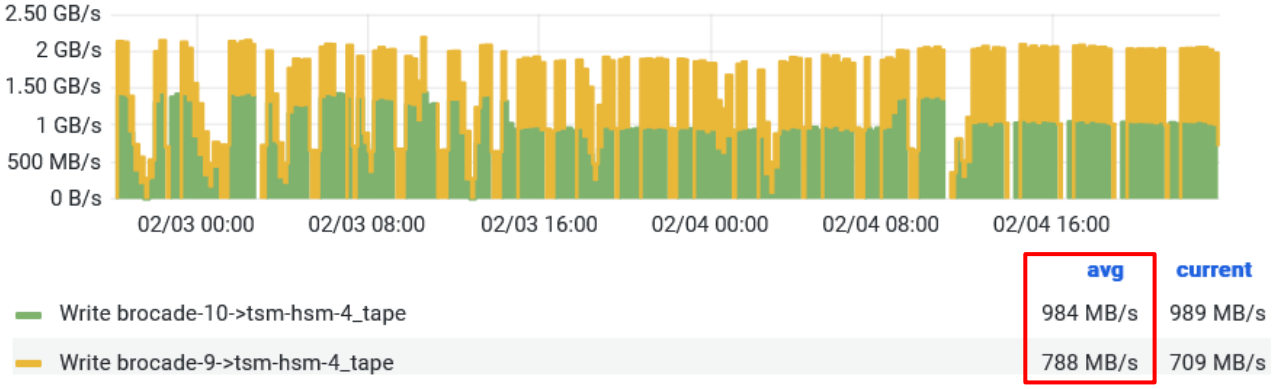
LHCb data challenge (scritture)

Target avg rate: **1.72 GB/s**

CERN -> CNAF DISK -> CNAF BUFFER

Overall avg rate: **2.2 GB/s**

CNAF BUFFER -> CNAF TAPE



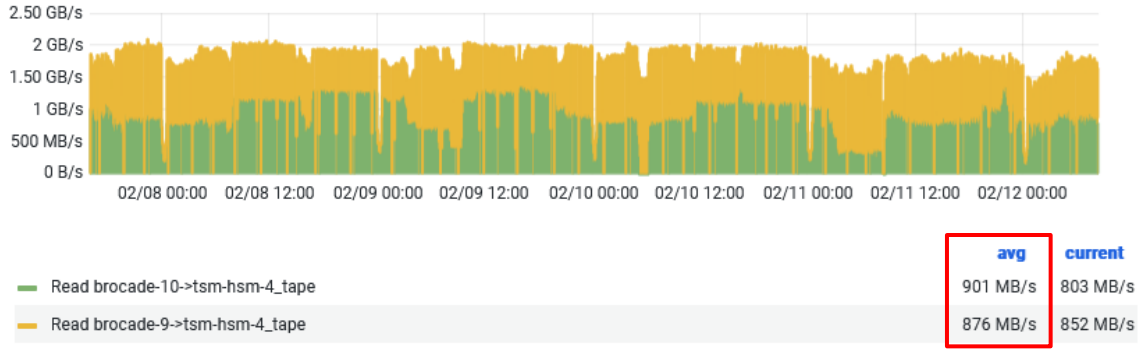
Overall avg rate: **1.77 GB/s**

Peak: > 2.1 GB/s

LHCb data challenge (letture)

Target avg rate: **1.35 GB/s**

CNAF TAPE -> CNAF



Overall avg rate: **1.76 GB/s**

Peak: > 2 GB/s

CNAF BUFFER -> CNAF DISK

Overall avg rate: **1.1 GB/s**

Considerazioni sul buffer

- Workflow LHCb prevede una permanenza dei dati sul buffer molto limitata
  - In scrittura migrazione automatica (scan tipicamente ogni ora)
  - In lettura dati copiati su disco non appena richiamati da tape sul buffer. E’ sempre cosi?
- Possibilità di dividere il file system tra disco e buffer
  - File system buffer piccolo (qualche decina di TB?) su storage veloce
  - Il disco non verrebbe penalizzato dalle attività sul buffer
- Stessa situazione per CMS
  - Workflow simile?