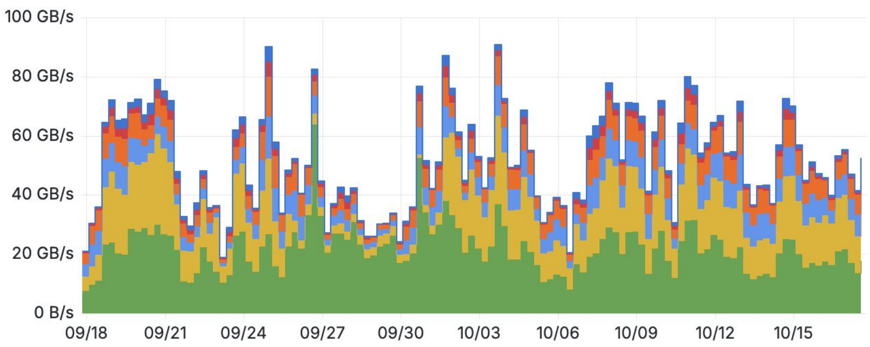
State of Storage

CdG 18 Ottobre, 2024

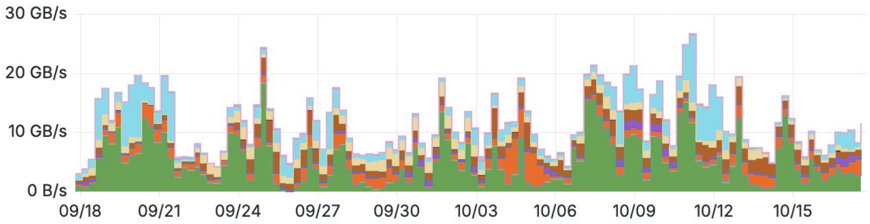
Business as usual + migration to TP

Last month

All servers network traffic out (reading)

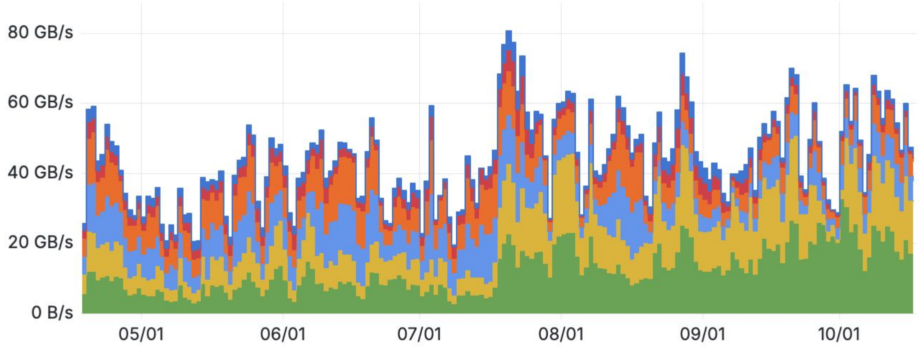


Gateway traffic out (non POSIX reading)

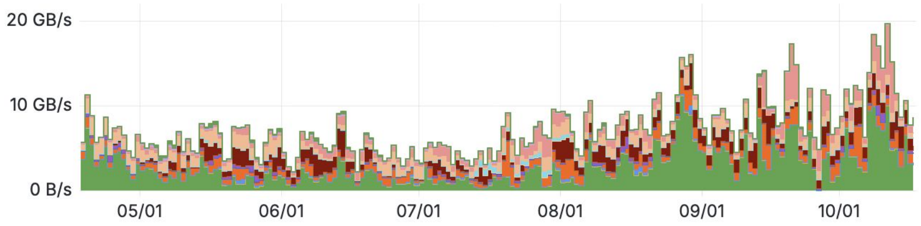


Last 6 months

All servers network traffic out (reading)



Gateway traffic out (non POSIX reading)



Disk storage in produzione

Installed: 113PB - 33PB (in dismissione)=80.6PB Pledge 2024: 82.08PB, Used: 48.8PB

Storage system	Model	Net capacity, TB	Experiment	End of support
os6k8	Huawei OS6800v3	3400	GR2, Virgo	07/2024
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	12/2024
md-5, md-6 e md-7	Dell MD3820f	50	metadati, home, SW	11/2023 e 12/2024
os18k1,	Huawei OS18000v5	960	LHCb (buffer tape)	7/2024
os18k3, os18k5, os18k5	Huawei OS18000v5	1200	ATLAS,ALICE (buffer tape)	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	Moving to TecnoPolo	2027
od1k6-1,2,3,4,5,6	Huawei OD1600	60000	ALICE,ATLAS,LHCb, CMS	2031
od1k5-1,2	Huawei OD1500(NVMe)	400	Metadati, LHCb hotadata	2031

Acquisti recenti e futuri

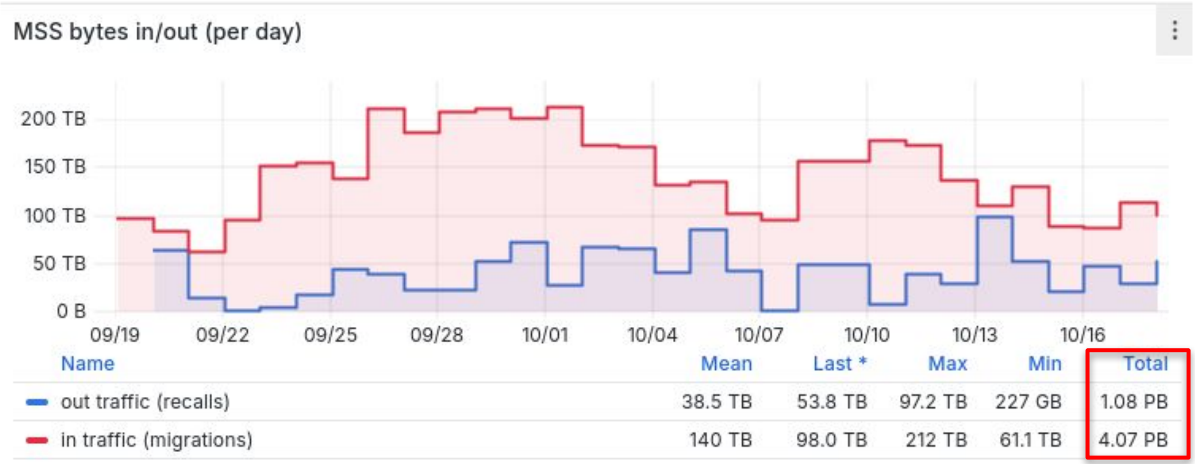
- Gara storage 2022 (14PB netti)
  - Nuova proposta con apparati DDN SFA7990X
  - In fase di installazione a TP
- Tape Library
  - Installata, collaudo completato
  - Le cassette JF da 50TB sono state inserite nella libreria (7.8PB)
  - In fase di configurazione per la PROD
- Gare nastri
  - Nuova gara di acquisto tape JF (96PB)

Problemi relativi a “I/O intensive workflow” di LHCb

- Per diminuire stress del work flow di LHCb abbiamo migrato il buffer tape su HW separato
- Abbiamo considerato la possibilità di creare un “buffer disco” per i dati “hot”
    - 200 TB NVMe per *.dst files in /storage/gpfs_lhcb/disk/lhcb (Oct 2nd)
        - Riempito immediatamente di file non più acceduti
    - Il path giusto sarebbe stato /storage/gpfs_lhcb/disk/lhcb/buffer/ (830 TB), che non è un fileset
        - rsync dei dati su un nuovo fileset buffer è molto lento (traffico di produzione?)
        - La placement policy dovrebbe essere basata sul filename, e cambiare quotidianamente

Stato tape

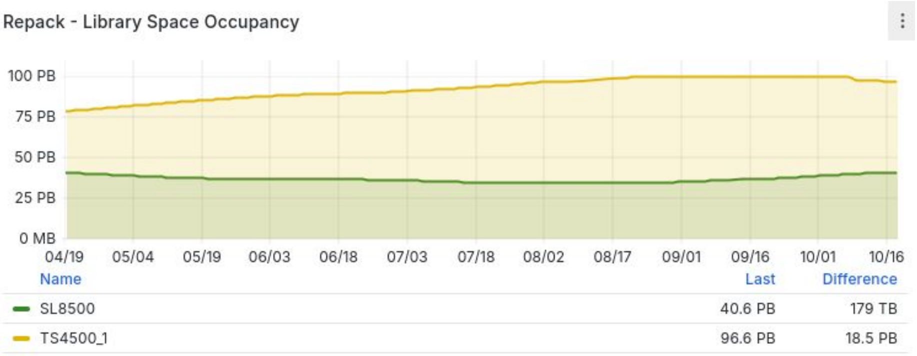
Last 2 months



Name	Mean	Last *	Max	Min	Total
🟦 out traffic (recalls)	38.5 TB	53.8 TB	97.2 TB	227 GB	1.08 PB
🟥 in traffic (migrations)	140 TB	98.0 TB	212 TB	61.1 TB	4.07 PB

4 PB of new data written to tapes in two month (since last CdG)

Tapes: Migration from Oracle to IBM library on hold

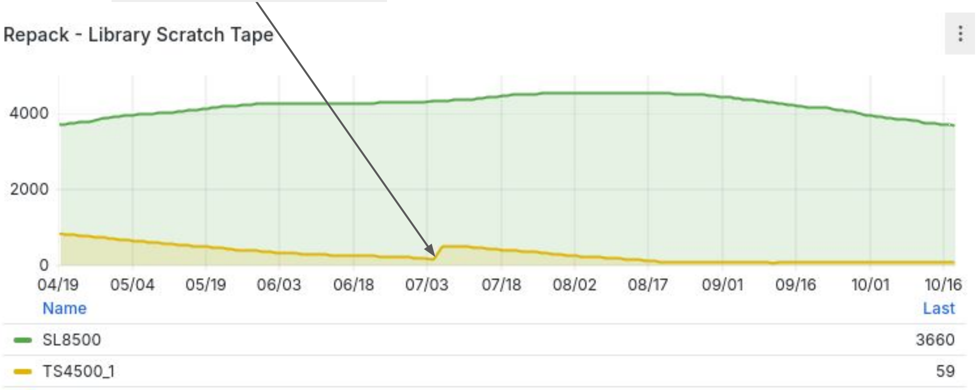


Repack - Library Space Occupancy

Name	Last	Difference
SL8500	40.6 PB	179 TB
TS4500_1	96.6 PB	18.5 PB

Repack - number of free cartridges

380 tapes inserted



Repack - Library Scratch Tape

Name	Last
SL8500	3660
TS4500_1	59

SL8500 has 36.6 PB to migrate

Stato tape

- Liberi 1.1 PB (Scratch tape sulla libreria IBM).
- Usati 136 PB.
- Spazio tape sulla libreria IBM praticamente esaurito
- La nuova libreria IBM non è ancora funzionante a causa di problemi di compatibilità con la versione TSM in produzione.

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	**36.7**	-
TS4500 (IBM)	19*TS1160	400	6198	20	5100+380	**99.6**	**1.1**
TS4500-2(IBM)	18*TS1170	400	7844	50	165	**0**	**8.2**

Current SW in PROD

- GPFS 5.1.2-15, in preparation to 5.1.9-6
  - RHEL 9 and ARM support
- StoRM BackEnd 1.11.22 (latest)
- StoRM FrontEnd 1.8.15 (latest)
- StoRM WebDAV 1.4.3 (latest)
- StoRM globus gridftp 1.2.4
- XrootD 5.5.4-1
  - LHCb updated to 5.5.5-1
- Ceph 16.2.6 (Pacific)
- GEMSS and tape drive orchestrator updated to support X tape libraries

Tickets and more

- ALICE
  - Open action: finalize the configuration for the XrootD tape cluster (xs-204, xs-304)
    - Waiting for the migration of servers to EL9 to install and test rpm for interaction xrootd-tape
- ATLAS
  - Found *one* corrupted file on tape following net problems on Sep 25th (12347 files checked on disk, 1092 files checked on tape)
    - declared as bad
  - Ongoing staging activity (650 TB)
    - Misha added a 80% limit of buffer filling based on information reported on *report.json*
  - GGUS 168445 (waiting for reply): failed transfers due to “tape buffer full”
    - *We highly recommend not to exceed the **mean daily** writing rate limit (recalls included) of 1.0GB/s*
    - We involved Lorenzo who investigated buffer status with ATLAS colleagues
  - GGUS 167957 (on hold): StoRM WebDAV does not permit the creation of non-existent parent directory even if the scope does it,
    - Waiting for a fix from StoRM developers

Tickets and more

- CMS
  - GGUS 167634 (waiting for reply): SAM tests failing
    - Thread limit reached, need to tune both FTS and StoRM WebDAV parameters
  - GGUS 168610 (solved): same issue of GGUS 167634
  - https://its.cern.ch/jira/browse/CMSDM-220: enabling overwrite-when-only-on-disk feature on CNAF tape
    - Error “Destination file exists and is on tape” (missing user.storm.checksum.adler32) and error “Could not check destination file locality” (missing user.storm.migrated)
      - CMS deleted those files, which had been transferred one year ago
    - We do not have access to/we cannot monitor CMS internal ticketing system
      - common WLCG GGUS ticketing system should be used to get prompt support
  - GGUS 167995 (on hold): StoRM WebDAV does not permit the creation of non-existent parent directory even if the scope does it
    - Waiting for a fix from StoRM developers

Tickets and more

- LHCb
  - GGUS 168542 (in progress): failed data transfers
    - The restarting of sprucing and merge jobs increased traffic, two out of 6 StoRM WebDAV servers reported thread saturation
      - Configuration problem fixed
  - GGUS 168495 (waiting for reply): corrupted files
    - Due to network problems occurred on Sep 25th, some communication messages have been lost when StoRM WebDAV was writing data to disks
    - Checksums were recalculated for more than 80k files on disk and 3.5k files on tape
      - Minimal impact - found only 7 corrupted files

Tickets and more

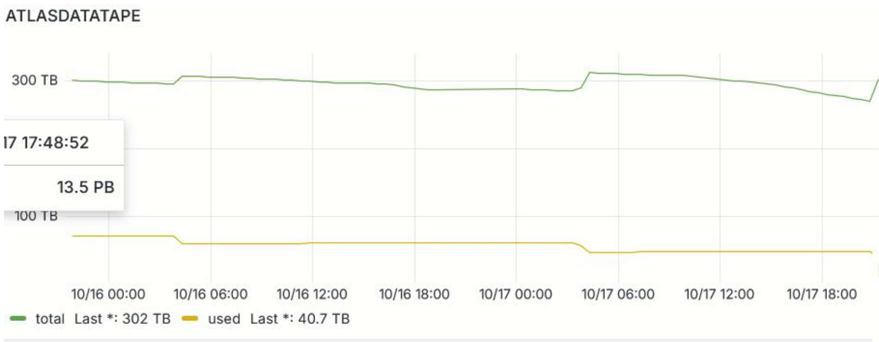
- LHCb
  - GGUS 167716 (in progress): low transfer efficiency with new storage HW installed at TP
    - Performance decreases with the file system occupancy and the pressure of the experiment data flow
      - 6 StoRM WebDAV servers separated from the NSD ones
      - Dedicated HW for tape buffer
      - Following closely the situation via weekly reports to WLCG management board & operations coordination since Aug 30th
      - The MB asked for a service incident report before Nov 15th

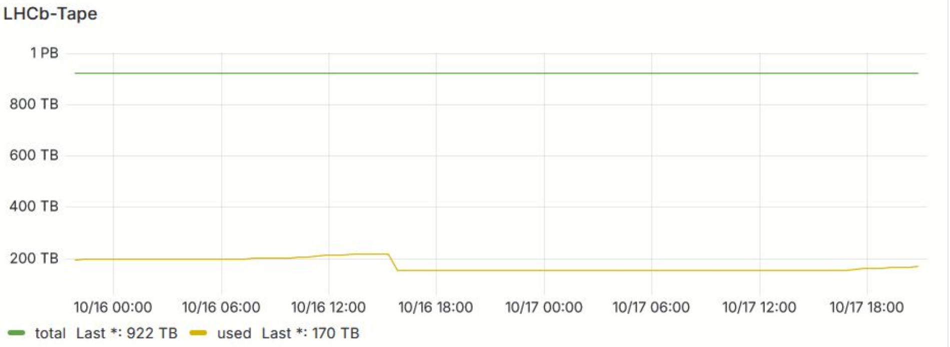
Tickets and more

- Gsiftp protocol via StoRM backend is still available for two experiments
  - New StoRM release should finally allow to switch GridFTP off (Xenon, CTA-LST)
- CTA
  - Local ‘cta’ users can now read data in /cta-lst posix. Grid tools strongly encouraged.
- Dampe
  - GridFTP “plain” still used
    - TPCs between XrootD server at IHEP and CNAF are working well
    - Rucio+FTS (https) should replace the current gsiftp transfers (WP6-DataCloud)
- DUNE
  - Data exposed in read mode also via XrootD (xrootd-archive); VOMS and scitokens authnz

Storage Site Report (/info/report.json)

New dashboard at t1metria-storage-site-report to visualize space used and assigned reported in /info/report.json for all ATLAS, CMS and LHCb storage areas


ATLASDATATAPE
300 TB
17 17:48:52
13.5 PB
100 TB
10/16 00:00
10/16 06:00
10/16 12:00
10/16 18:00
10/17 00:00
10/17 06:00
10/17 12:00
10/17 18:00
total Last *: 302 TB
used Last *: 40.7 TB


LHCb-Tape
1 PB
800 TB
600 TB
400 TB
200 TB
10/16 00:00
10/16 06:00
10/16 12:00
10/16 18:00
10/17 00:00
10/17 06:00
10/17 12:00
10/17 18:00
total Last *: 922 TB
used Last *: 170 TB