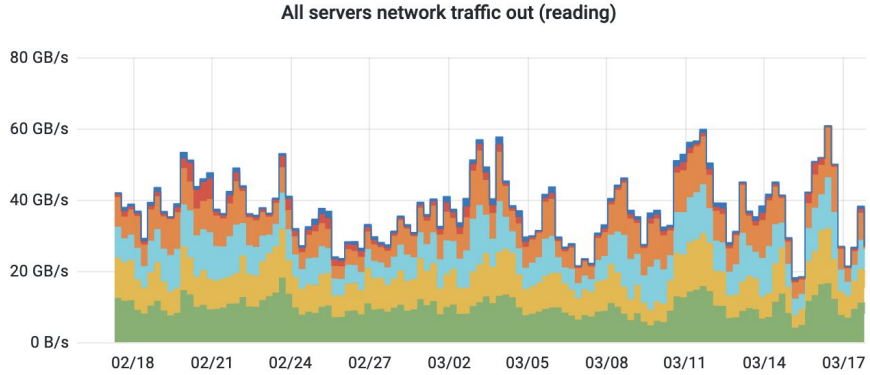
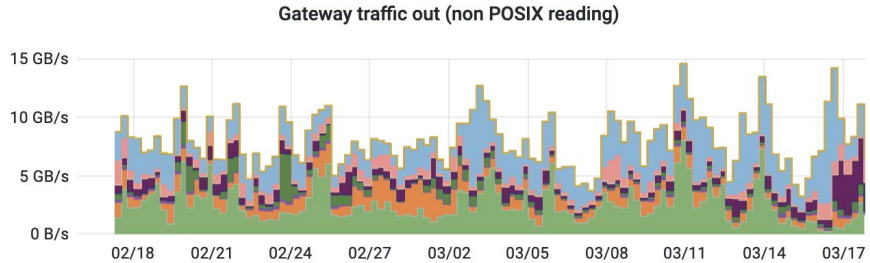
State of Storage

CdG 18 Marzo, 2022

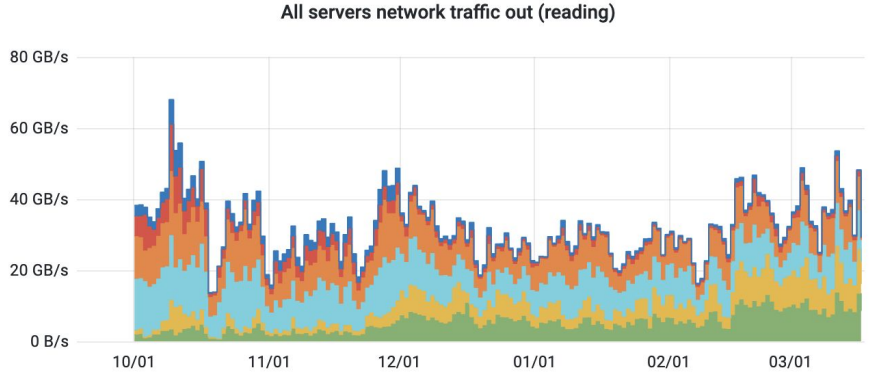
Business as usual

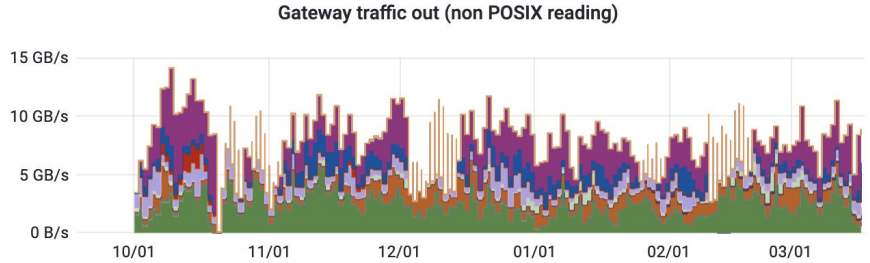
Last month





Last 6 months





Disk storage in produzione

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

Sistema	modello	Capacita', 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	5184(raw)	ALICE, cloud, etc	2027

Prossimi acquisti

- Gara storage 2022 (14PB)
  - 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 usable) → gpfs_data
  - Huawei - 96 slot liberi x10TB =960TB raw (768TB usable) → gpfs_atlas
  - Spesa
    - ~67KE (offerta) → DDN
    - ~56KE (estimate) → Huawei
  - Tempi di consegna ~2 mesi
- Ulteriore acquisto di tape (14PB per arrivare alla pledge 2022) da fare a maggio

Current SW in PROD

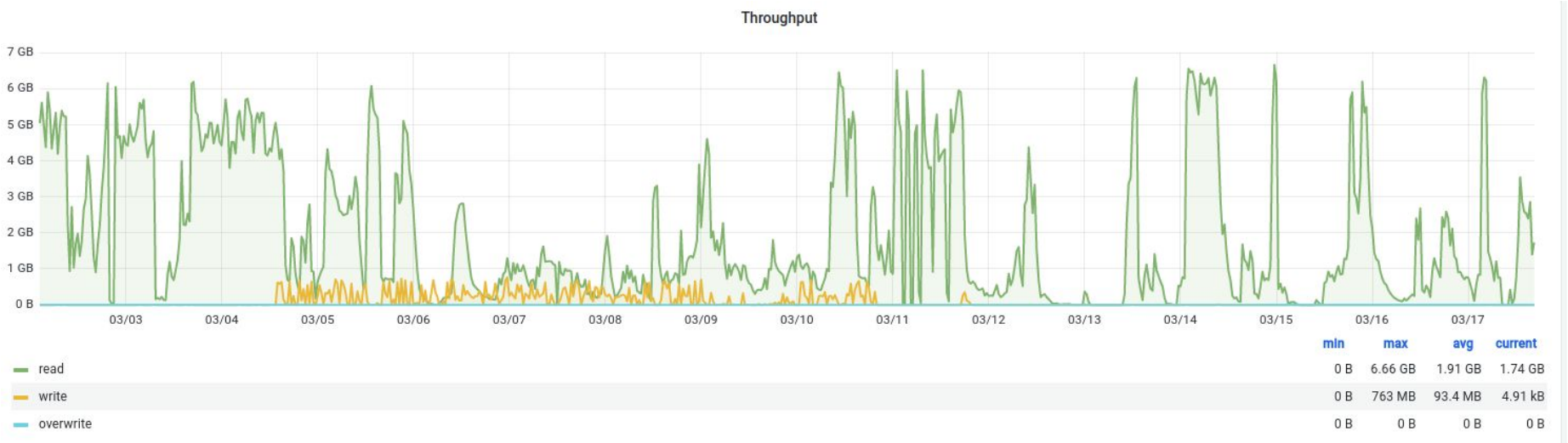
- GPFS 5.0.5-9
- StoRM BackEnd 1.11.21 (latest)
- StoRM FrontEnd 1.8.15 (latest)
- StoRM WebDAV 1.4.1 (latest)
- StoRM globus gridftp 1.2.4
- XrootD 4.11.2
  - updated to 4.12.4 in the 4 CMS servers
  - 5.3.1-1 on CMS redirectors (local and EU/IT/FR)

Test XrootD ALICE on CEPH

- Storage with CEPH (5 PB raw)
- 1.5PB Pledge for ALICE
- 5 server with Xrootd v4.8.4 (Redirector + Server)
- Many thanks to Francesco Noferini for the setup

	CEPH_TEST												
	AliEn SE		Catalogue statistics						Storage-provided information				
SE Name	AliEn name	Tier	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version
1. CNAF - CEPH_TEST	ALICE::CNAF::CEPH_TEST	1	1.332 PB	400.4 TB	963.6 TB	29.35%	563,756	FILE	1.361 PB	400.4 TB	993 TB	28.74%	Xrootd v4.8.4
Total			**1.332 PB**	**400.4 TB**	**963.6 TB**		**563,756**		**1.361 PB**	**400.4 TB**	**993 TB**		

Test XrootD ALICE su CEPH



Recent problems

- CMS
  - Check for buffer to be cleaned before tape challenge (GGUS 156271)
  - Discussing deletion of files from previous tape challenge (GGUS 156236)
  - Checksum mismatch; a single file on tape was invalidated (GGUS 156044)
  - Consistency check (cc) scans failing at T1_IT_CNAF_Disk; contacting the wrong xrootd endpoint from k8s pods (GGUS 155890)
- LHCb
  - Tape data challenge: LHCb reports many “Timeout waiting for connection from pool”; investigation ongoing (GGUS 155973)

Recent activities

- 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
      - Need to check efficiency of DNS load balancing in xrootd-ceph cluster
- JUNO: support to network team in preparation of the Juno data challenge
- BELLE: alignment of authorization policies based on VOMS FQAN between StoRM backend and StoRM WebDAV

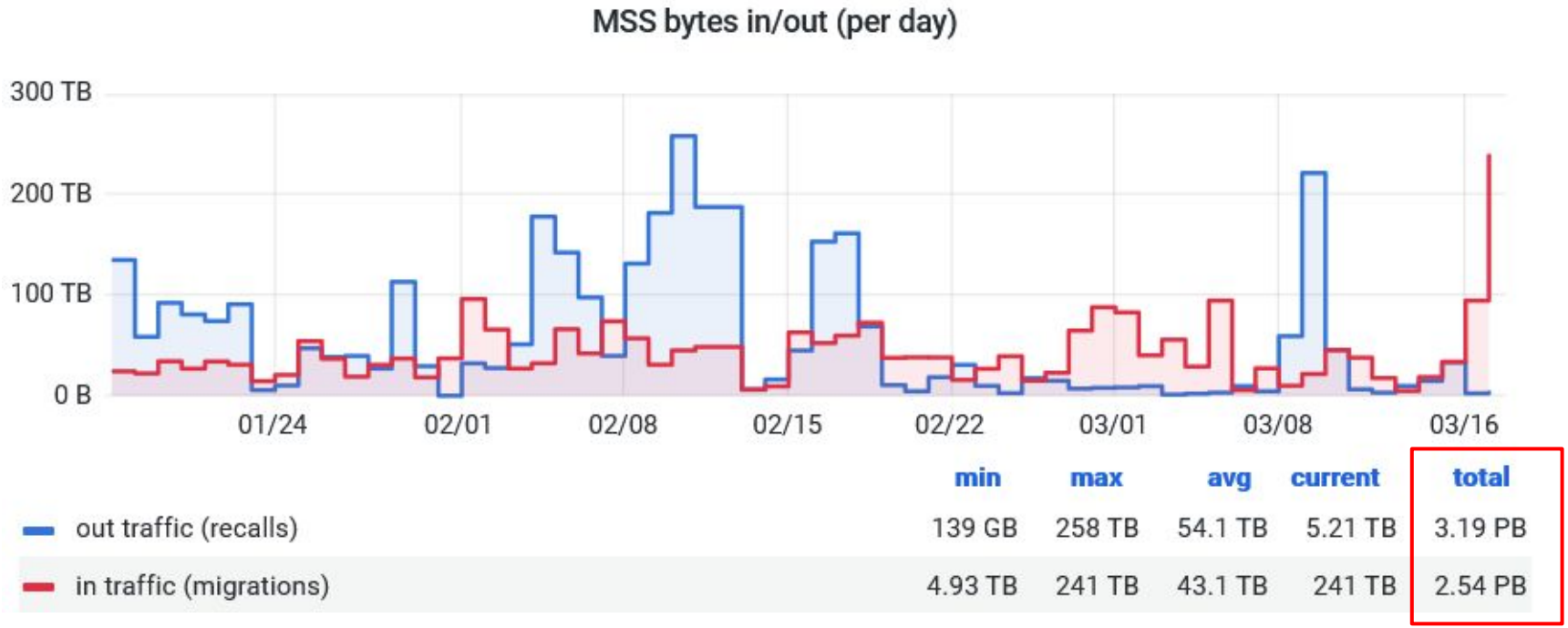
Actions

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

2. **Need to reboot all StoRM endpoints for security issues (kernel upgrade)**
   - Service not reachable for max 5 minutes
   - We will schedule an “at risk” in GOCDB for next week

Stato tape

17 Jan 2022 - 17 Mar 2022



Stato tape

- 16.5 PB liberi (su cassette vuote, complessivamente sulle 2 librerie). Usati 95 PB.
  - Nuova gara installata la scorsa settimana: 14.8 PB
  - Pledge 2022: 130.5 PB
  - Installato attuale: 116.5 PB
  - Ulteriore acquisto a maggio per arrivare a pledge

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	75.5
TS4500 (IBM)	19*TS1160	400	6198	20	1750	19.5

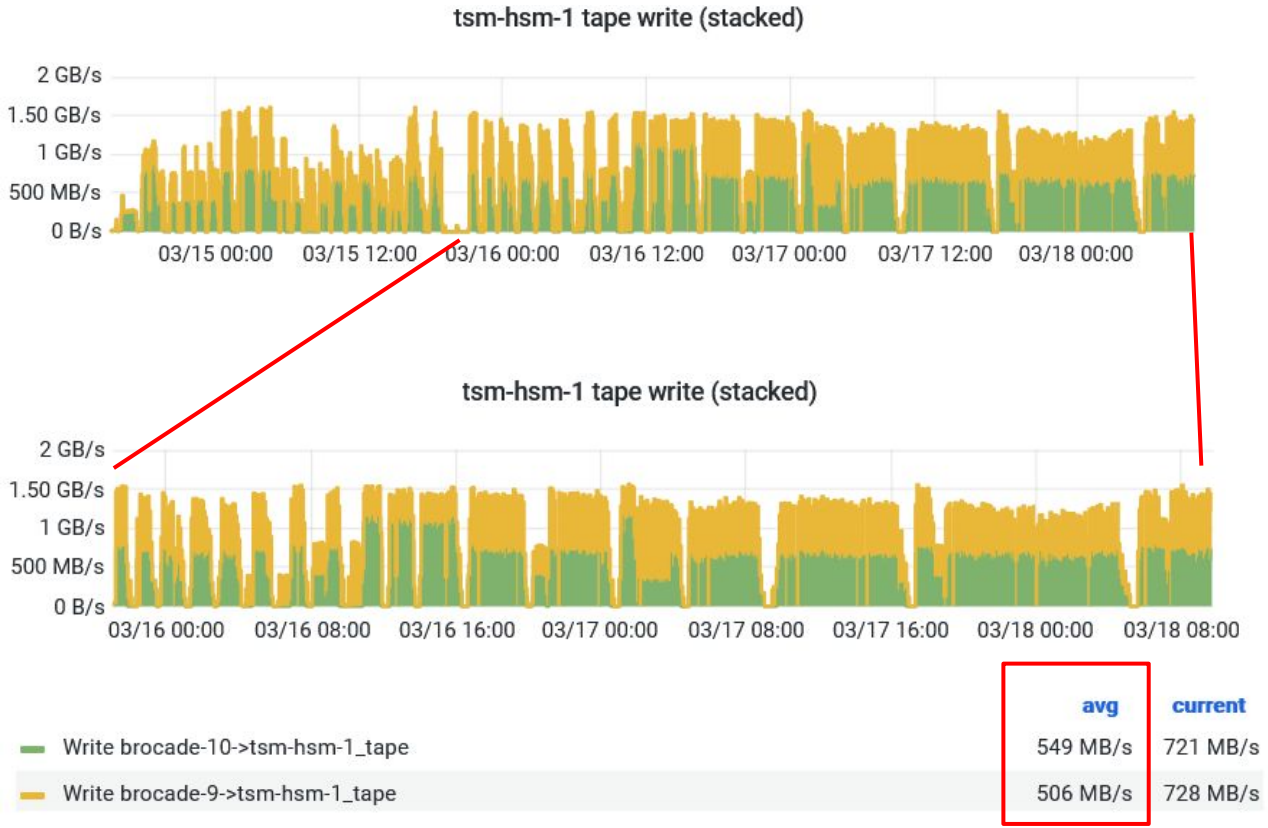
RUN3 targets

VO	Reads (DT) GB/s	Writes (DT) GB/s	Reads (A-DT) GB/s	Writes (A-DT) GB/s
ALICE		0.8	0.3	0.8
ATLAS	0.2	0.9	0.8	0.5
CMS	0.1	1.2	1.9	0.2
LHCb		~~2.24~~ 1.72	~~0.86~~ 1.35	
Total	0.3	4.62	4.35	1.5

Tape challenge - letture CMS

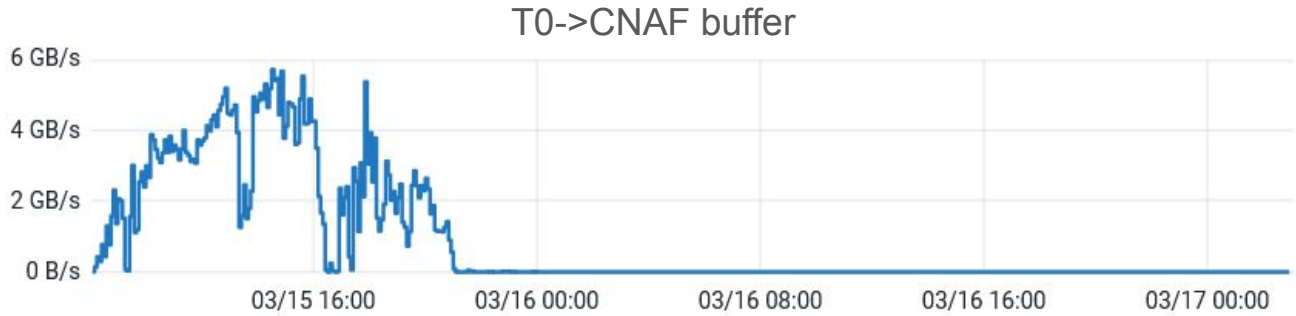
- Recall: 185 TB
- Tutti i dati su libreria Oracle
- Target rate: 1.9 GB/s
- Media rate complessiva: 1.45 GB/s
- Media rate in condizioni ottimali: 1.8 GB/s
- Per CMS lo scopo era stressare il sistema e verificare i rate rispetto al precedente test
- Siamo disponibili a fare test di dimensione equivalente con dati su entrambe le librerie
- Possibilità di aggiungere secondo HSM server

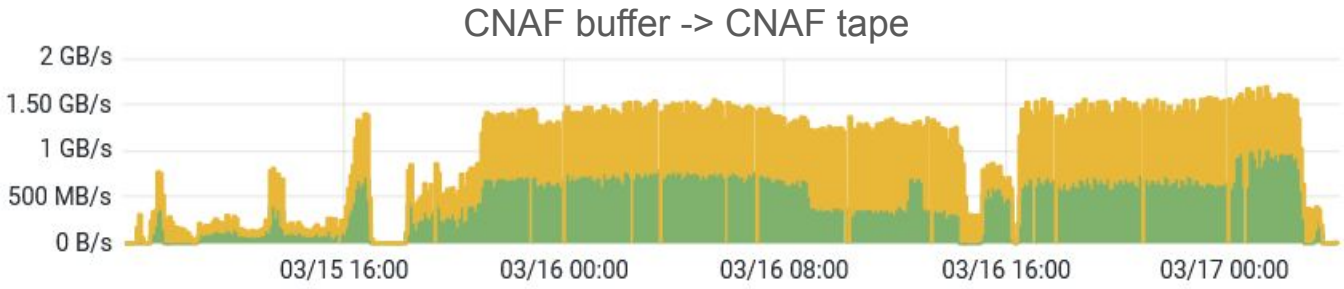
Tape challenge - scritture CMS



- Target rate: 0.9 GB/s
- Media rate: 1.05 GB/s
- Ampi periodi con rate superiori al target

Tape challenge - scritture ATLAS





- Scritture: 150 TB
- Target rate verso buffer: 3.5 GB/s per 12 ore
- Target rate buffer->tape: 0.9 GB/s
- Media rate verso buffer: 3 GB/s
- Media rate buffer->tape complessiva: 1 GB/s
- Media rate buffer->tape dalla fine scritture sul buffer in poi: 1.35 GB/s
- Da investigare se e come potenziare il buffer

Tape challenge - scritture LHCb

- Target rate: 1.72 GB/s
- Media rate: 1.52 GB/s
- Il target è superato quando sul buffer si sono acculumati dati, per cui, in caso di scritture continue sul buffer, non avremmo problemi di rate