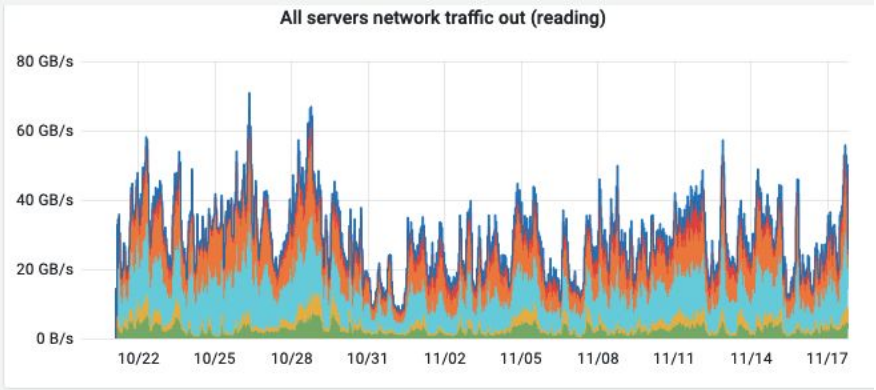
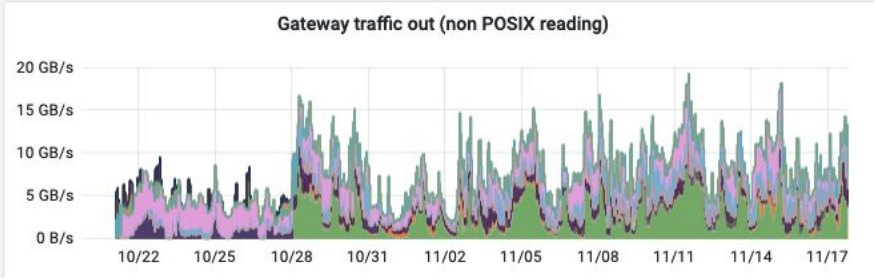
State of Storage

CdG 19 novembre, 2021

Business as usual

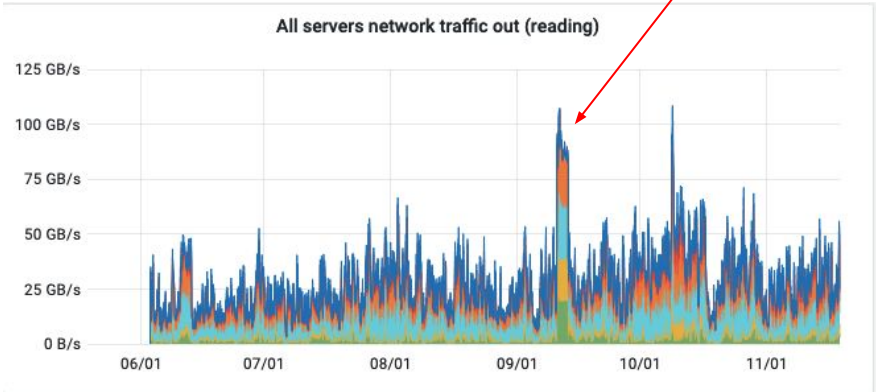
Last month

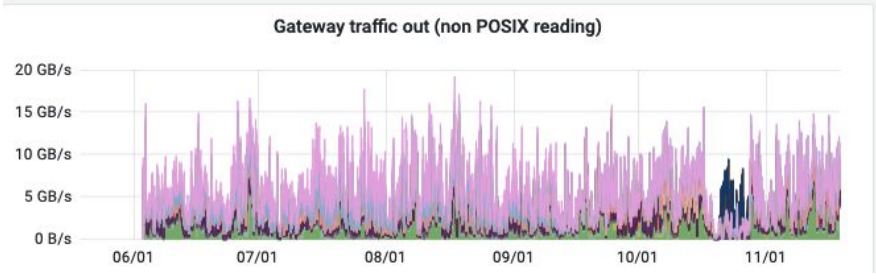




Last 6 months

CEPH Commisioning





Disk storage in produzione

Installed: 50.07 PB Pledge 2021: 50.4 PB Used: 36.1 PB

Sistema	modello	Capacita', TB	esperimenti	scadenza
ddn-10, ddn-11	DDN SFA12k	10752	Atlas, Alice, AMS	03/2021→ 06/2023
os6k8	Huawei OS6800v3	3400	ALICE, GR2	2022
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	11/2021
md-7	Dell MD3820f	20	metadati, home, SW	2022
md-5, md-6	Dell MD3820f	8	metadati	06/2021
os18k1, os18k2	Huawei OS18000v5	7800	LHCb, ALICE	2023
os18k3, os18k5, os18k5	Huawei OS18000v5	11700	ATLAS, 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

Stato installazione storage - lotto GPFS

- Con GPFS vediamo 8.9 PB netti (~200TB in piu' rispetto la richiesta del capitolato)
- Collaudo completato
- Le prestazioni in linea a quanto richiesto

	Scrittura sequenziale da GPFS (MB/s)	Lettura sequenziale su GPFS (MB/s)	Random read su GPFS (IOPS)
Target	31500	31500	9000
Risultato	38707	32295	13943

- Rimane da fare:
  - montare slitte per i server
  - Sostituire 4 cavi IB (FDR->EDR)

Piano spostamenti e ampliamenti

- Spostare dati di ATLAS su 2xOS5k8 (9 PB) lasciando il buffer tape su 3xOS18k (570 TB)
  - Vengono liberati 4344 TB su DDN10/11 e 2400 TB su OS18k
- +2400 TB su OS18k → CMS (per arrivare alla pledge 2021)
- +3350 TB su DDN10/11 → ALICE - 2067TB(OS6k8) = pledge 2021
- +1400 TB su OS6k8 → gpfs_data (ora al 99%)
  - Per tamponare l'emergenza +560TB (presi da ATLAS Su OS5k8 per il tempo di migrazione)
- +627 TB su OS6k8 → DarkSide (gpfs_ds50)
- +154 TB su DDN10/11 → AMS (gpfs_ams)

Stato installazioni storage - lotto CEPH

- Lotto CEPH (5 PB raw) consegnato e installato in rack
- Collaudo eseguito con successo
- Installato RHEL 8.4 e Ceph Pacific
  - Proposto test di pre-produzione ad ALICE
  - Il test sarà fatto in lettura e scrittura in collaborazione con Francesco Noferini
  - L’attività ha attirato l’attenzione dei colleghi di RAL con cui condivideremo l’esperienza

Stato installazioni storage - lotto CEPH

- Eseguiti benchmark estensivi e prove di failover
- Pronto per essere utilizzato con servizi di trasferimento

Current SW in PROD

- GPFS 5.0.5-2 → being updated to 5.0.5-9 (to fix security vulnerability)
  - Done on farm nodes, UI and on the majority of servers
- StoRM BackEnd 1.11.21 (latest)
- StoRM FrontEnd 1.8.15 (latest)
- StoRM WebDAV 1.4.0 (latest)
- StoRM globus gridftp 1.2.4
- XrootD 4.11.2; updated to 4.12.4 in the 4 CMS servers

Recent problems

- CMS
  - Tests ongoing in collaboration with CERN to understand the need for maximum thread in xrootd; xrootd upgraded to 4.12.4 in the 4 servers
  - Plan for upgrading XrootdD redirector version 4.11.2 -> 5.3.1 (GGUS ticket 154790)
  - Tests for tape data challenge; the experiment is currently arranging the deletion of files written during the challenge (154249)
  - Support for testing tape access with srm+https (GGUS ticket 153570)

Recent problems (continued)

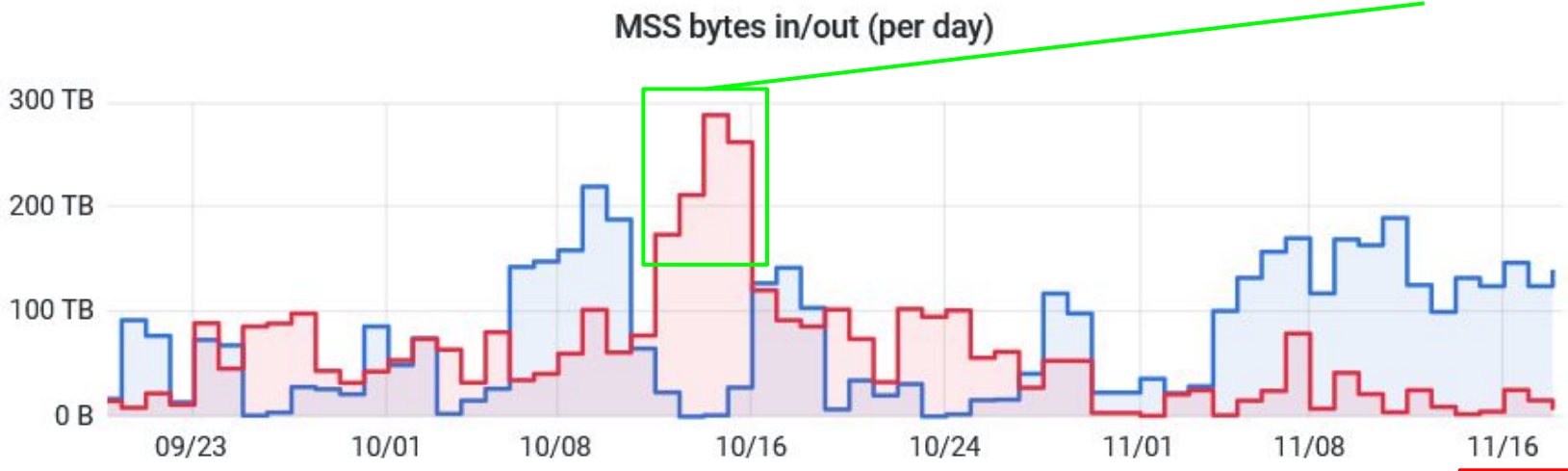
- ATLAS
  - Failed transfers due to unscheduled downtime published on GOCDB (GGUS ticket 154486)
  - Staging errors; a configuration issue prevented recall requests to reach our backend; solved (GGUS ticket 154419)

- LHCb
  - Failed FTS transfers, one tape was blocked in a failed robotic component; solved (GGUS ticket 154839)
  - xs-104 not reachable on IPv6; solved (GGUS ticket 154786)
  - Files not written; indeed they were removed by the job after being written; workflow to be checked/improved by the experiment (154332)

Stato tape

19 Sep - 18 Nov 2021

Tape challenge



	min	max	avg	current	total
out traffic (recalls)	114 GB	220 TB	76.9 TB	140 TB	4.62 PB
in traffic (migrations)	857 GB	288 TB	60.0 TB	6.52 TB	3.60 PB

Stato tape

- 11.3 PB liberi (complessivamente sulle 2 librerie). Usati 89.7 PB.
  - Gran parte delle scritture su nuova libreria
    - Tutti LHC
    - Xenon, CTA, Virgo, ARGO, Juno, Icarus
  - Pledge 101 PB. Appena installati 5.2 PB 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	74.9
TS4500 (IBM)	19*TS1160	400	6198	20	1010	14.8

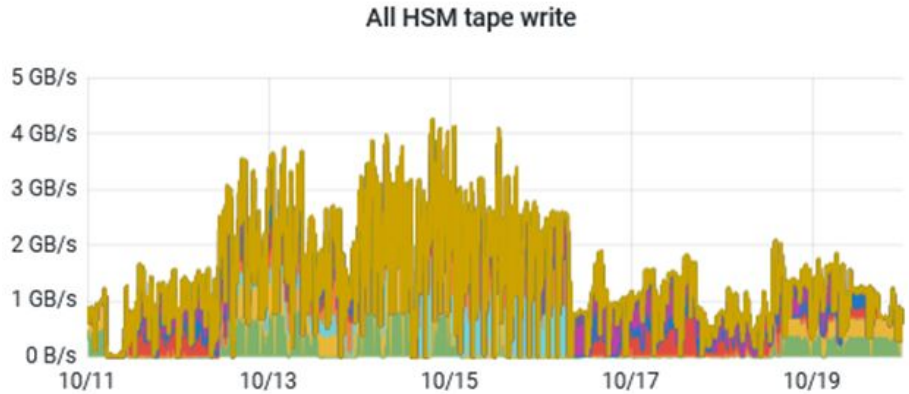
RUN 3: rate richiesti

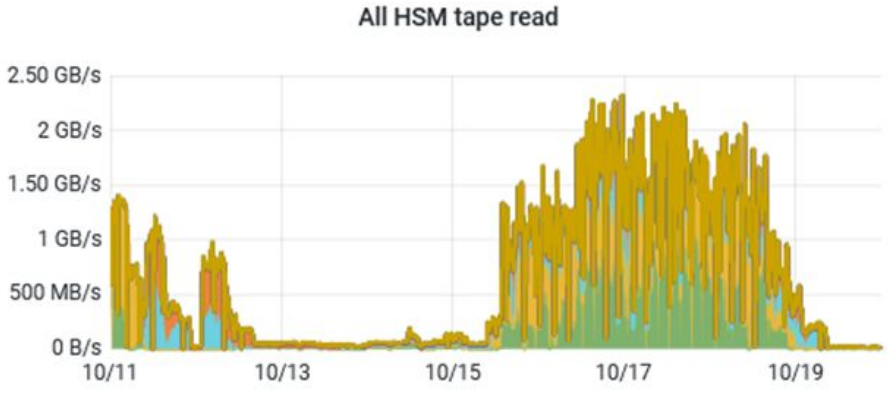
VO	Reads (DT) GB/s	Writes (DT) GB/s	Reads (A-DT) GB/s	Writes (A-DT) GB/s
ALICE		0.8	0.3	0.8
ATLAS	0.2	0.9	0.8	0.5
CMS	0.1	1.2	1.9	0.2
LHCb		2.24	1.72	
Total	0.3	5.14	5.72	1.5

Total throughput Data Taking (DT): 5.44 GB/s

Total throughput After Data Taking (A-DT): 7.22 GB/s

Tape challenge: rate complessivi



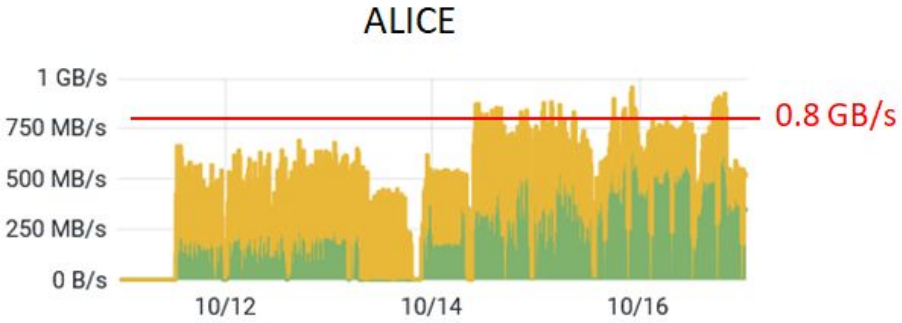


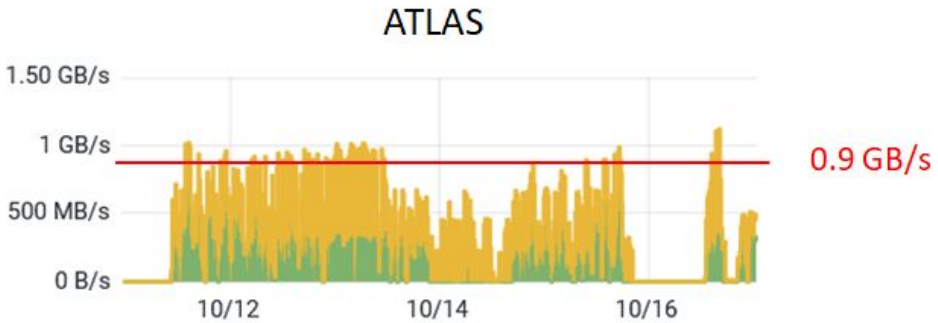
- Max (avg on 15 mins): 4.2 GB/s

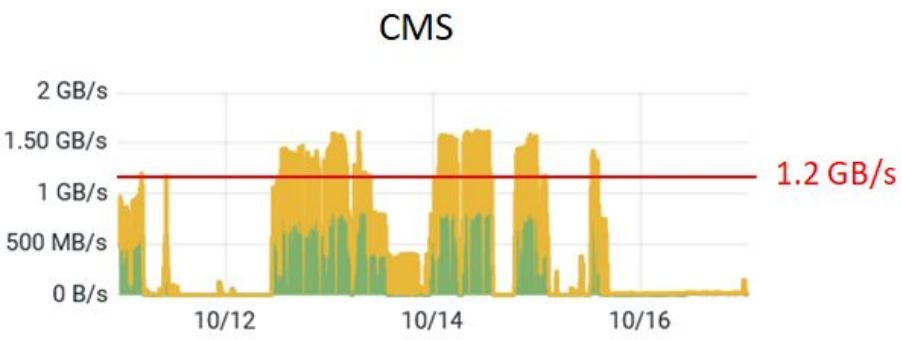
- Max (avg on 15 mins): 2.3 GB/s

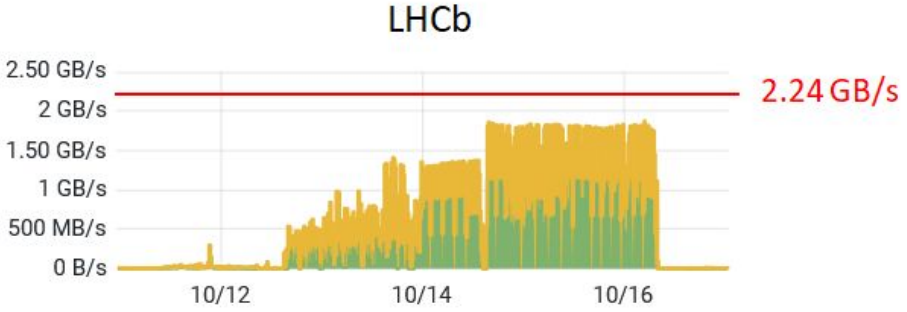
- Rate complessivi non raggiungono i numeri previsti
  - Test di VO diverse in momenti diversi

Tape challenge: scritture

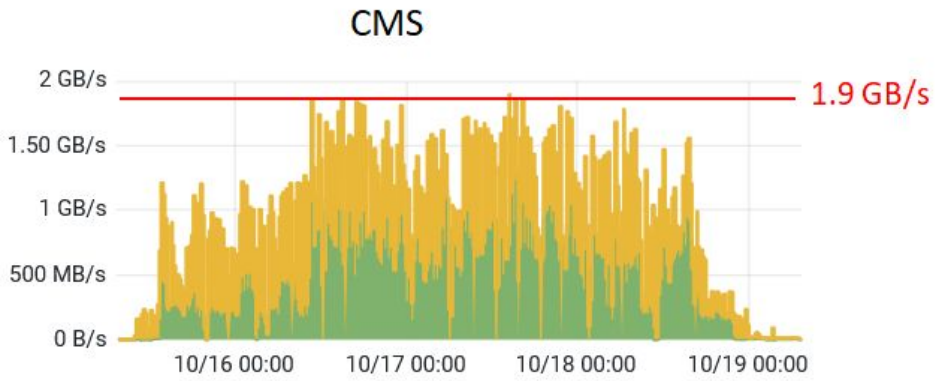


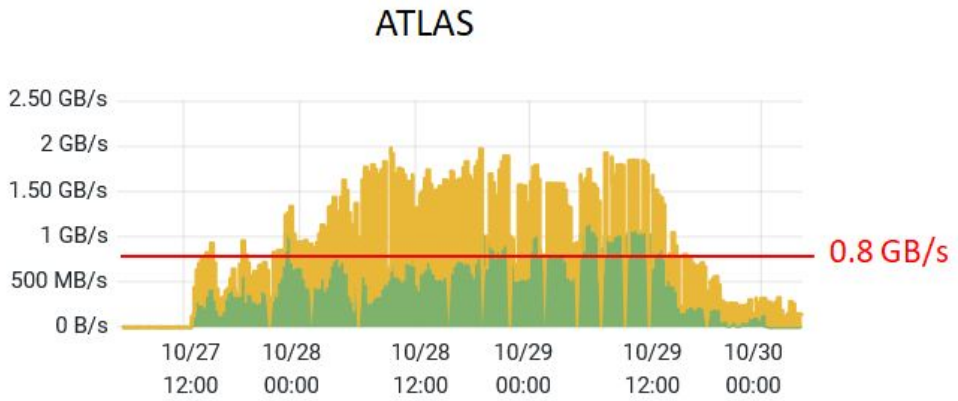






Tape challenge: letture





Tape challenge: considerazioni

- In generale organizzazione migliorabile
  - No tempistiche chiare
  - No coordinamento tra le VO
  - Difficoltà nella cancellazione dei dati di test
  - Punto di partenza per futuro test (forse a inizio 2022, probabilmente di 2 settimane)
- Scritture LHCb con rate basso per motivi legati all'esperimento
- Importante capire per quanto tempo è richiesto che i rate massimi siano raggiunti
  - Problema del riempimento del buffer