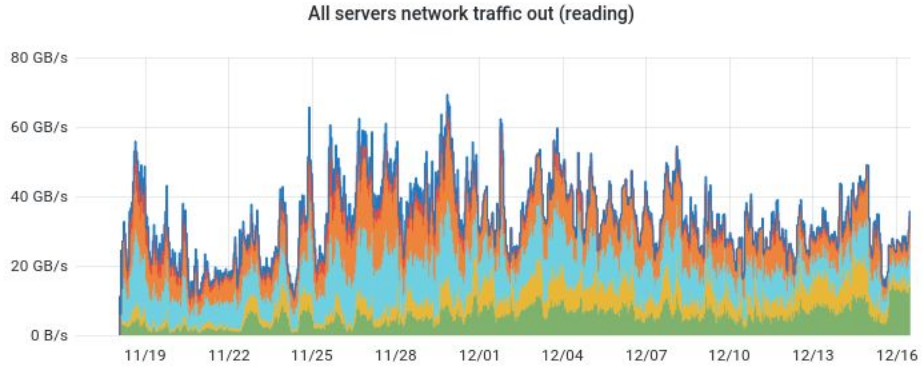
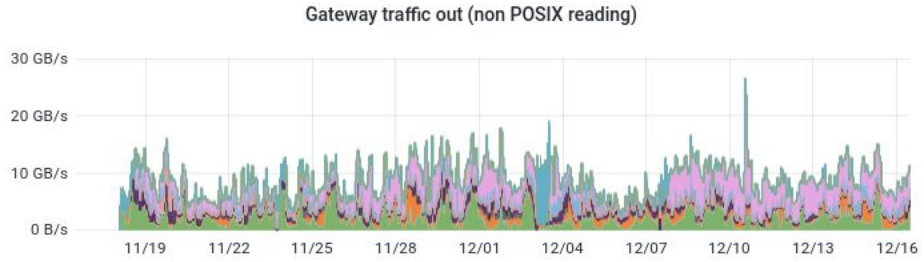
State of Storage

CdG 17 dicembre, 2021

Business as usual

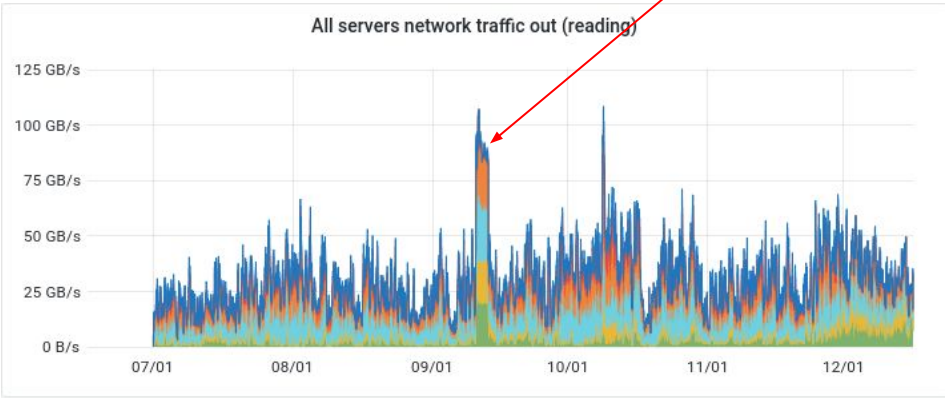
Last month

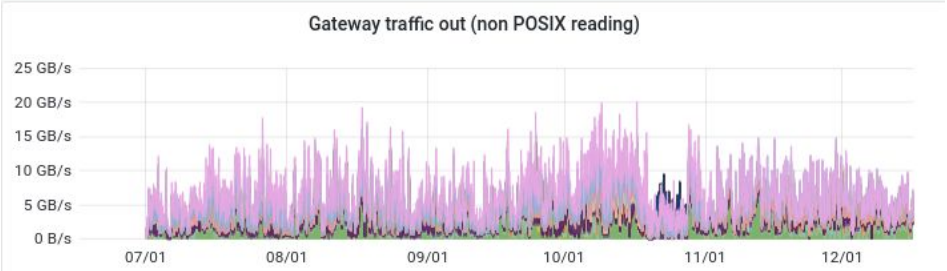




Last 6 months

CEPH Commissioning





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	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-7	Dell MD3820f	20	metadati, home, SW	12/2022
md-5, md-6	Dell MD3820f	8	metadati	06/2021 → 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

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)

DONE

~ 16 PB of data moved (in 3 weeks)

Current SW in PROD

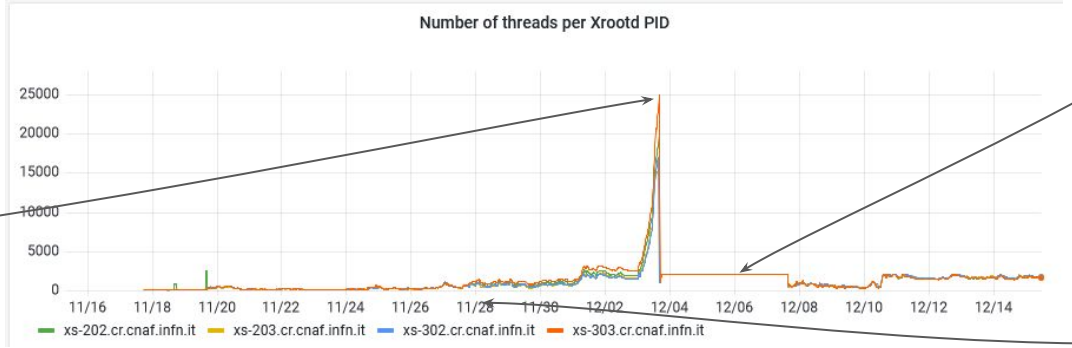
- GPFS 5.0.5-2 → being updated to 5.0.5-9 (to fix security vulnerability)
  - Done on farm nodes, UI and on the majority of servers
- StoRM BackEnd 1.11.21 (latest)
- StoRM FrontEnd 1.8.15 (latest)
- StoRM WebDAV 1.4.0 (latest)
- StoRM globus gridftp 1.2.4
- XrootD 4.11.2
  - updated to 4.12.4 in the 4 CMS servers
  - 5.3.1-1 on CMS redirectors (local and EU/IT/FR)

Recent problems

- CMS
  - Corrupted copies to be manually deleted on tape (GGUS ticket 155255)
  - XrootdD redirector version upgraded 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 (GGUS ticket 154249)
- ATLAS
  - XrootD not reading in some folders, missing read ACL for Atlas group, to be fixed
    - Contrary to StoRM backend, StoRM WebDAV does not allow to set special ACLs, so we need to enforce ACL (e.g. atlas r) at the fs level
    - An upgraded version of storm-native-libs is needed to fix this issue permanently, tests are ongoing (https://issues.infn.it/jira/browse/STOR-1502)
  - One stuck staging request (pin expired), fixed (GGUS ticket 155029)
  - Transfer failures as destination, ATLAS started to use http for tape while tape storage areas were not configured in our StoRM WebDAV endpoints, fixed. Checksum issues escalated to DDM ops, solved from the site point of view (GGUS ticket 155012)
  - Staging errors, problems with a tape drive, fixed (GGUS ticket 155200, 155201)

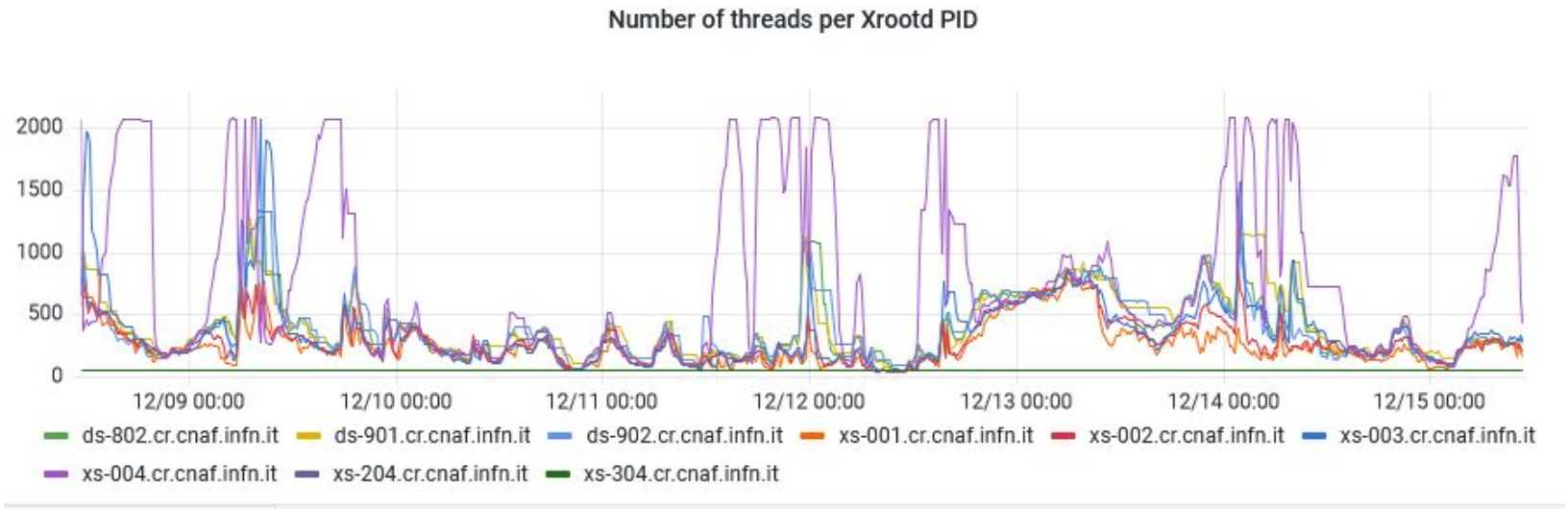
The neverending story: XrootD :-)

- Following debug with CERN people, we disabled sendfile() for read requests setting xrootd.async nosf (). This greatly alleviated the load issues, and allowed us to remove limitation on max threads
  - For both CMS and ALICE
- After a while, we realized that the default value for max threads (2048) was not actually used, and threads reached up to 17k
- We then manually set the default value
  - Not needed according to the configuration reference, request for the developers
- Also, we specified "**s**" for seconds after which to recycle an unused thread: xrd.sched mint 16 maxt 2048 avlt 8 idle 60**s**
  - "Optional" according to the configuration reference, additional request for the developers



Plateau of unrecycled threads

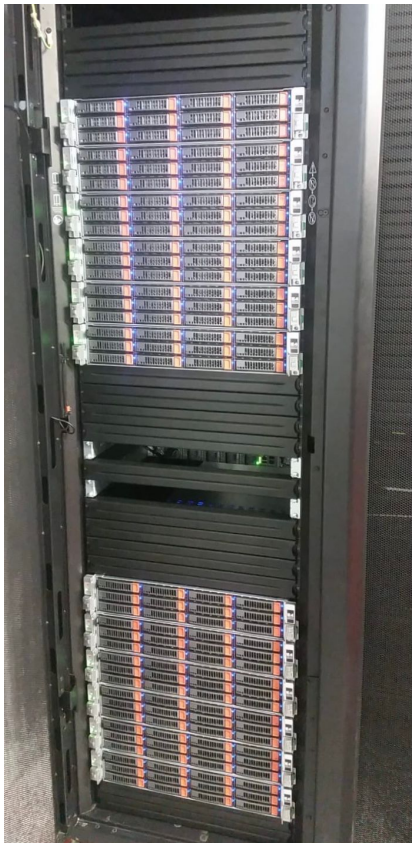
The neverending story: XrootD :-)



- For CMS, debug still ongoing with CERN people
- For ALICE, we notice an uneven distribution of threads among servers (redirector?)

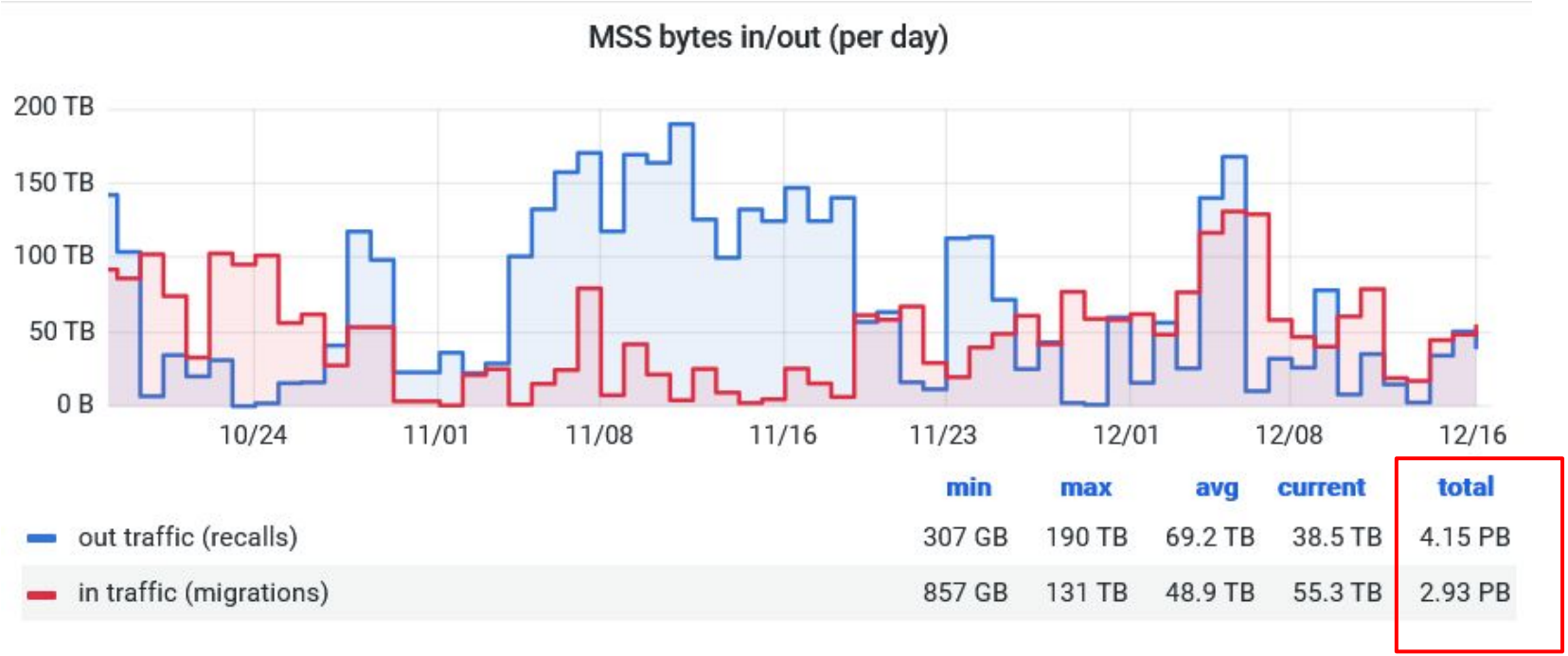
Test XrootD ALICE su CEPH

- Storage con CEPH (5 PB raw) up and running
- Installato RHEL 8.4 e Ceph Pacific
- 500 TB per test con XrootD di ALICE
  - Scritture/letture con 1 redirector + 1 server configurati su 1 server Ceph
  - Endpoint diverso dalla produzione
  - Test iniziato, per ora scritti 11 TB



Stato tape

17 Oct - 16 Dec 2021



Stato tape

- 10 PB liberi (complessivamente sulle 2 librerie). Usati 91 PB.
  - Gran parte delle scritture su nuova libreria
    - Tutti LHC
    - Xenon, CTA, Virgo, ARGO, Juno, Icarus, Auger
  - Pledge 101 PB

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	75.3
TS4500 (IBM)	19*TS1160	400	6198	20	1010	15.7

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	0.86	
Total	0.3	5.14	3.86	1.5

- Rate di scrittura di ATLAS comprende picco di 3.5 GB/s di 10 ore ogni 5 giorni
  - Buffer disco non si riempe se il trasferimento verso tape è fatto a una media di 0.9 GB/s