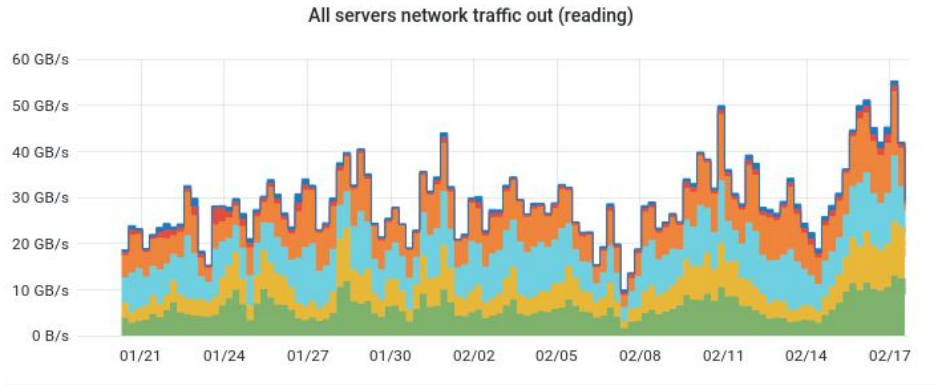
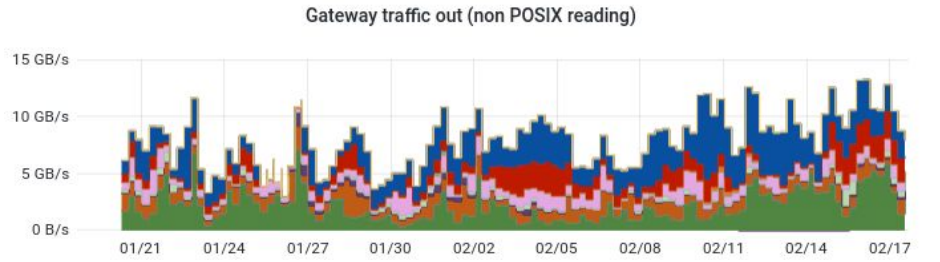
State of Storage

CdG 18 Febbraio, 2022

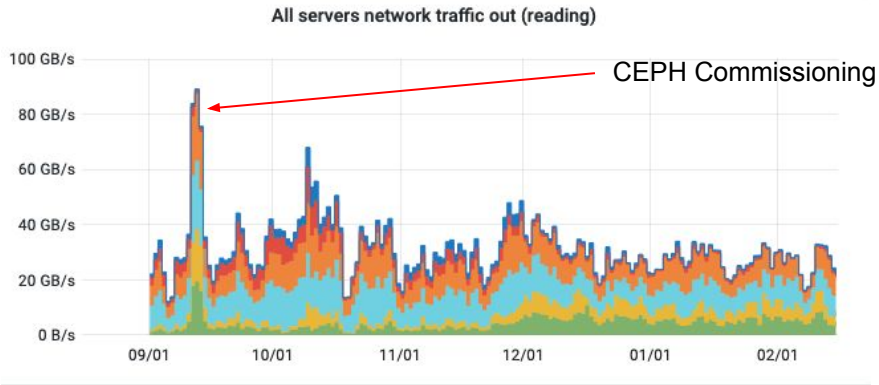
Business as usual

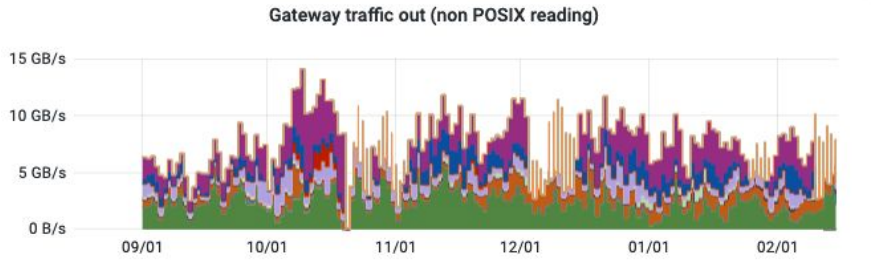
Last month





Last 6 months





Disk storage in produzione

Installed: 50.07 PB, Pledge 2022: 59.1 PB, Used: 40.8 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

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)

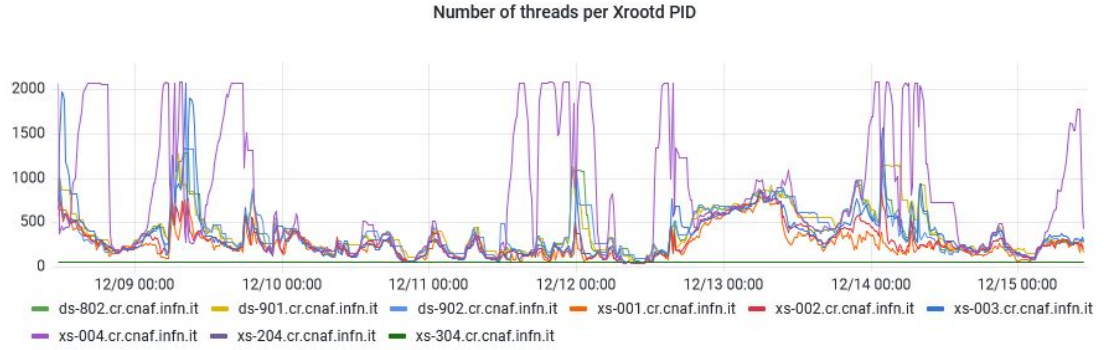
The neverending story seems ended.

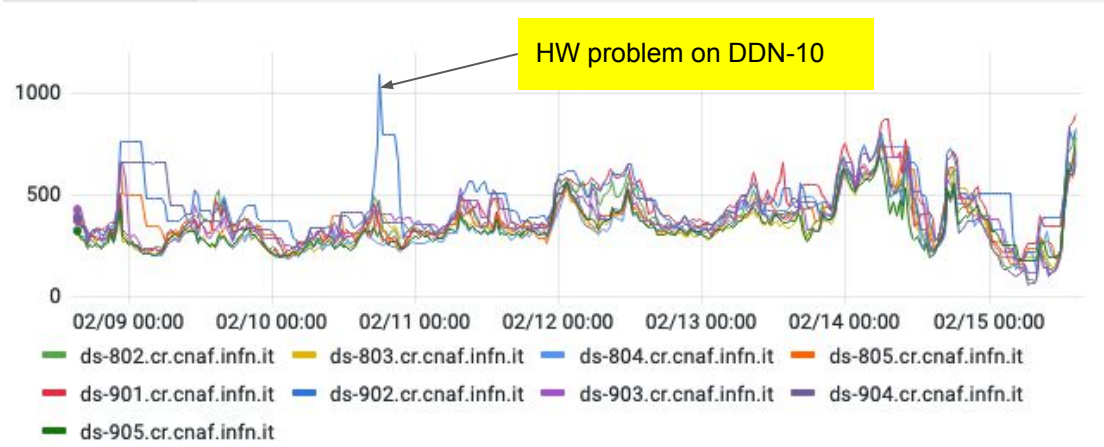
- Disabled *sendfile()* for read requests setting *xrootd.async nosf ()*. This greatly alleviated the load issues, and allowed us to remove limitation on *max threads*
- Manually set the default value *max threads* (2048)
  - Not needed according to the configuration reference (confirmed as an error in documentation by developers)
- Specified “**s**” for seconds after which to recycle an unused thread:
  xrd.sched mint 16 maxt 2048 avlt 8 idle 60s
  - “Optional” according to the configuration reference, additional request for the developers
- On GPFS side
  - Increased gpfs *pagepool* (for all XrootD servers) to 16GB
  - Separated NSD from XrootD servers

The neverending story: XrootD :-)

For ALICE, we notice an uneven distribution of threads among servers (interference between GPFS NSD and XrootD)

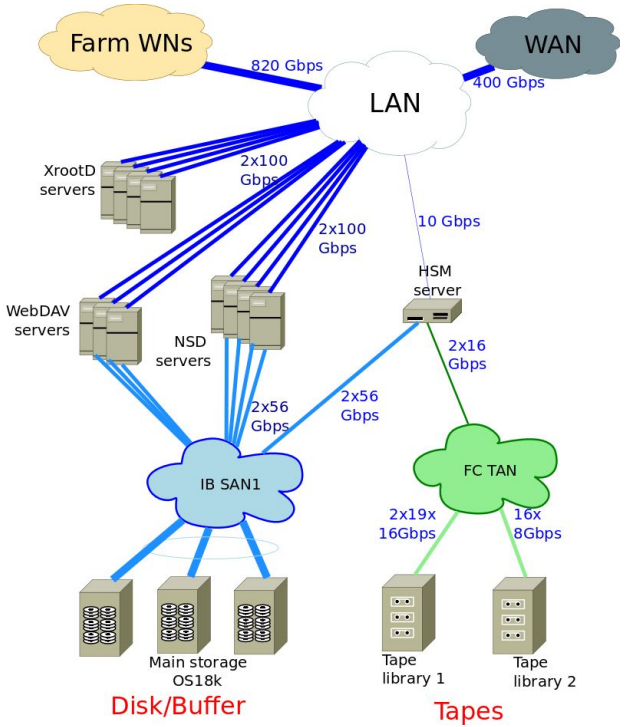
Now it does not happen any more





CMS cluster layout (new)

- XrootD servers accessing filesystem via dedicated NSD servers via 100GbE
- Considering connect XrootD servers disks directly via IB

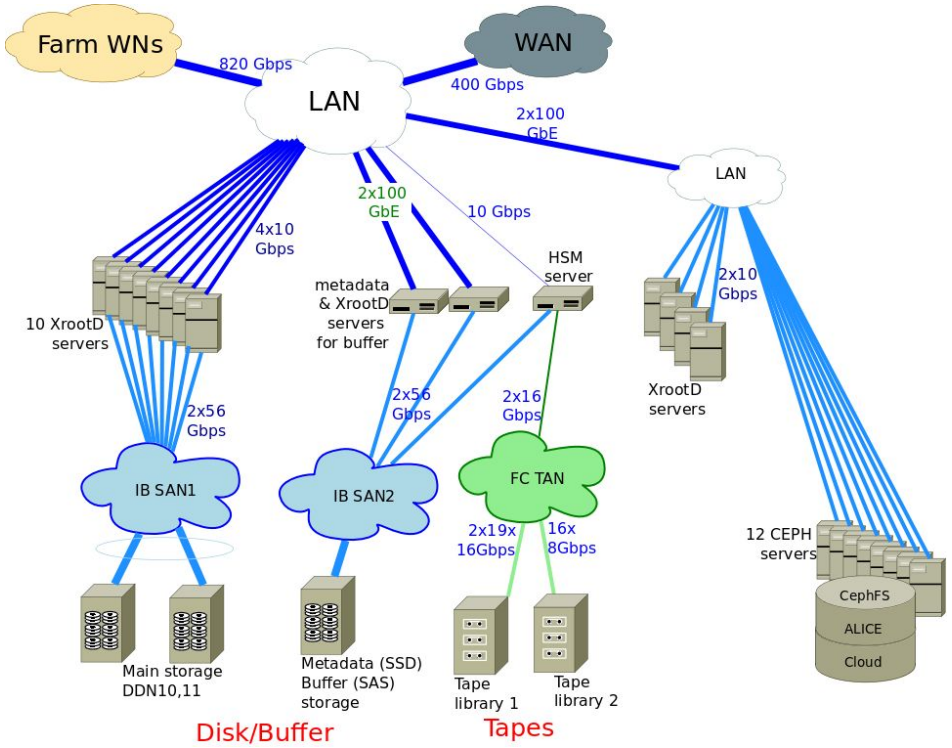


Test XrootD ALICE su CEPH

- 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 + 3 server su server separati
  - Scritti 400TB → Peak throughput 8GB/s!!!
  - Candidate for **pledged disk**

ALICE cluster layout (new)

- Consolidated all ALICE disk-only storage space on 2 DDN systems
- Added 6 new XrootD servers with direct access to DDN disks
- Proposing to use 1.6PB (increase of 2022) on CephFS



Recent problems

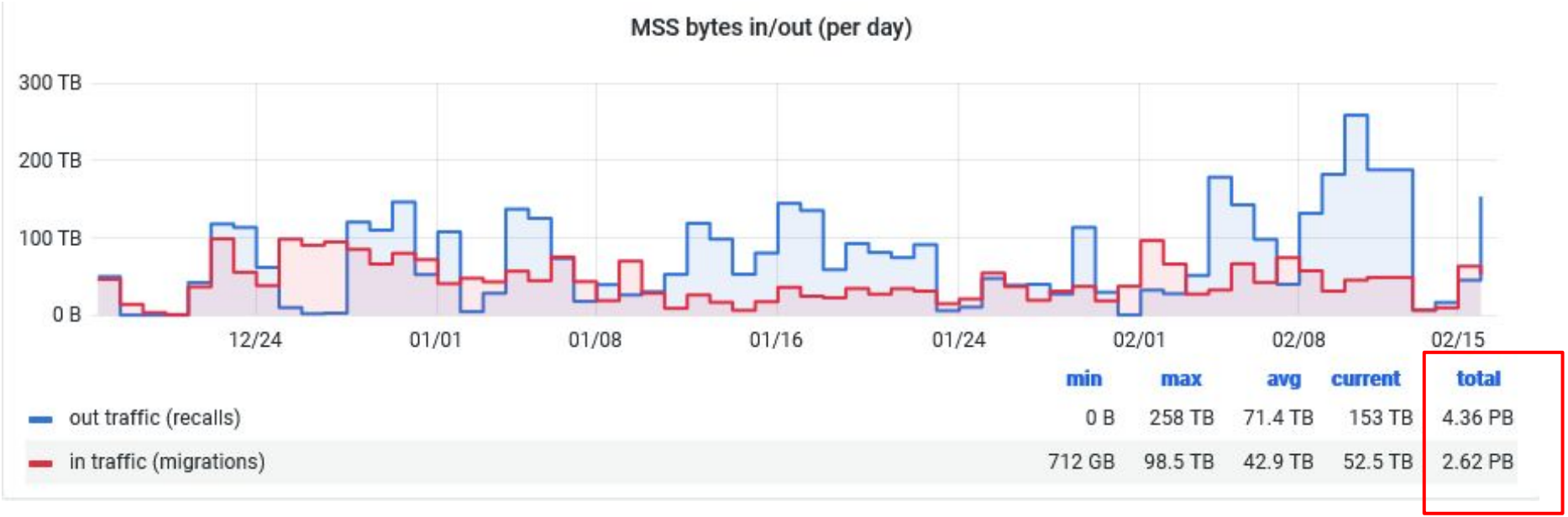
- CMS
  - Consistency check (cc) scans failing at T1_IT_CNAF_Disk; contacting the wrong xrootd endpoint from k8s pods; **waiting for reply** (GGUS 155890)
  - “File exists and overwrite not enabled”; folders were not configured to migrate to tape, fixed (GGUS 155872)
  - “File exists and overwrite not enabled”; files missing from our fs (GGUS 155679)
- ATLAS
  - gridftp jobs to RAL-LCG fail; fs not mounted after temporary fs down (GGUS 155705)
  - Staging errors; issues with a tape drive (GGUS 155595)
  - Transfer errors; ipv6 not configured (GGUS 155587)
- LHCb
  - User cannot list his directory; fs ACL to be enforced when performing DM with StoRM WebDAV (GGUS 155608)
  - Storm SRM does not return an https turl; LHCb started to use http for tape while tape storage areas were not configured in our StoRM WebDAV endpoints, fixed (GGUS 155481)

Recent activities

- ATLAS
  - Monthly dumps of gpfs_atlas fs available to be processed (?)
  - 4 new StoRM WebDAV endpoints, which are not NSD servers
    - Need to switch off the old ones to see TPCs in the new endpoints (even after 20 days); probably due to caching effects in FTS at the glibc level
  - We deleted empty directories
- CMS
  - 4 new XrootD servers, which are not NSD servers
    - Daniele S. reported no failures in a Release Validation production run at M100 :-)
  - XrootD proxy is back in xs-404 for jobs from M100
  - Writing on subdirs of /cmstape/store/ that not actually migrate on tape. Why?
  - Issue of files corrupted. Any news on fix in RUCIO?
- 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
- OTHER
  - Support to network team in preparation of the Juno data challenge

Stato tape

17 Dec 2021 - 16 Feb 2022



Stato tape

- 3 PB liberi (su cassette vuote, complessivamente sulle 2 librerie). Usati 94 PB.
  - Gran parte delle scritture su nuova libreria
    - Tutti LHC
    - Xenon, CTA, Virgo, ARGO, Juno, Icarus, Auger
  - Pledge 2022: 130,5 PB
  - Nuova gara: in arrivo 14,8 PB a inizio marzo

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	75.4
TS4500 (IBM)	19*TS1160	400	6198	20	1010	18.3

Tape challenge 2022

- CMS : 2nd week of March for A-DT test
- ATLAS/CMS/LHCb : 3rd week of March for DT test
- ATLAS/LHCb : 4th week of March for A-DT test
- ALICE won't be in this challenge due to conflict with other commissioning activities
- Particular attention on LHCb DT writes
  - Expected 2.24 GB/s rate
  - We will evaluate if a second HSM server is needed

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	0.86	
Total	0.3	5.14	3.86	1.5

Globus retirement

- In 2017, Globus announced they would stop supporting Globus Toolkit and focus on their closed-source cloud services.
- Final end-of-life targeted for 2022

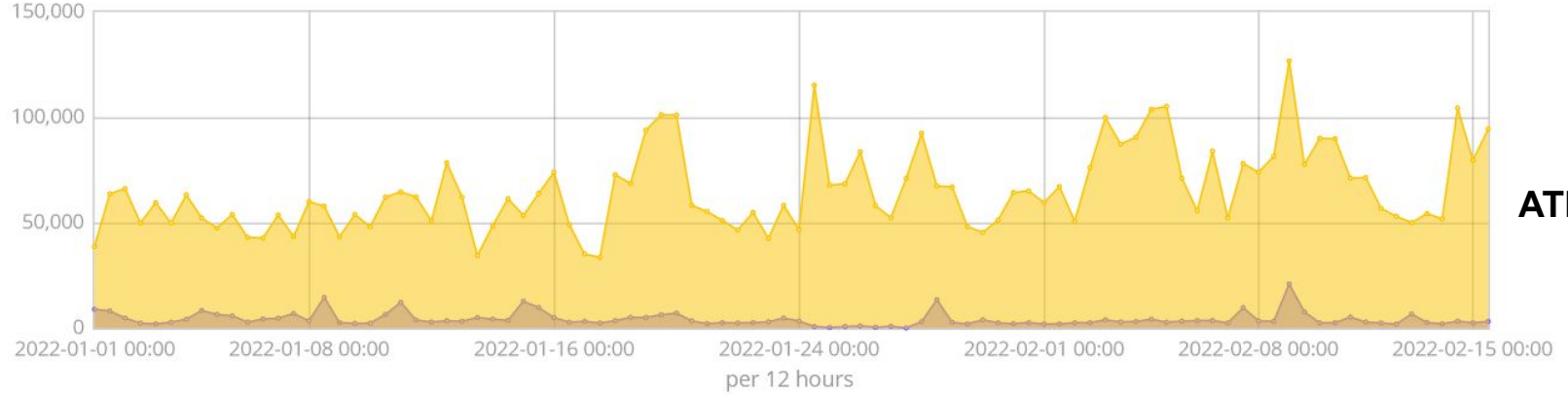
- WLCG uses two major features from the Globus toolkit:
  - GridFTP, which is being transitioned to HTTP-TPC
  - GSI authentication, which is being transitioned to tokens.

- The HTTP-TPC transition is most advanced, and should be completed “before Run3” (cit DOMA BDT 16/2/2022)

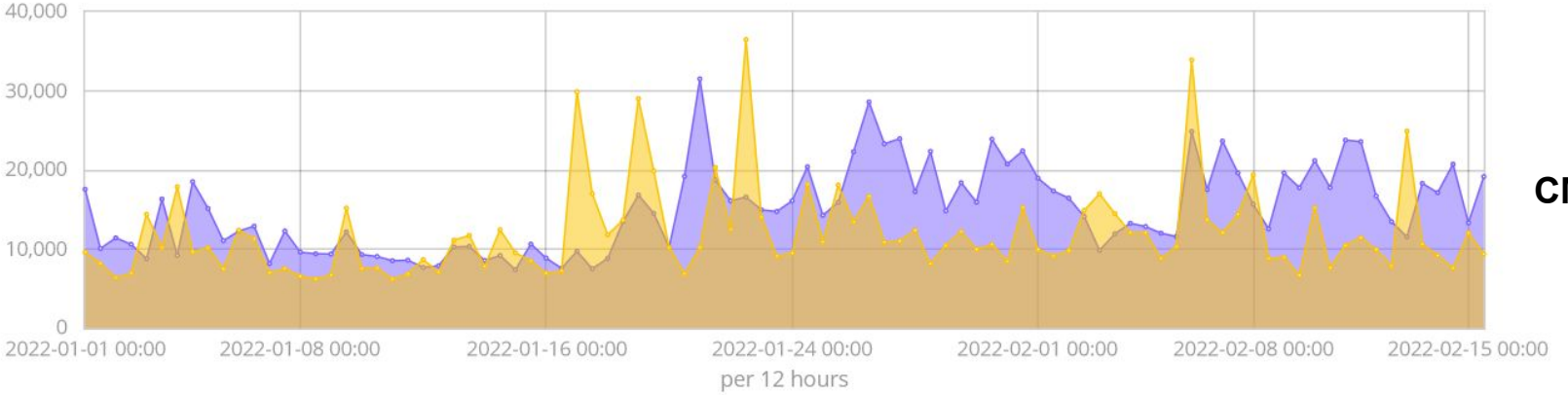
GridFTP protocol replacement

- The third-party-copy sub-group of the DOMA working group investigated alternatives to the GridFTP protocol for bulk transfers across WLCG sites
  - All storage elements to support WebDAV-based or XrootD-based TPCs
- Nov 2018: StoRM WebDAV was extended to support third-party transfers and become a viable alternative to GridFTP for bulk data transfers
- In 2019, we provided dedicated endpoints (and then ATLAS endpoints) for the stress tests, initially run with the DTEAM VO
- Over 2020, CMS and LHCb also started tests with StoRM WebDAV
- Nov 2020: HTTP TPC in production for ATLAS
- May 2021: stage-out also switched to HTTP for ATLAS; CMS and LHCb use HTTP in production
- Jul 2021: srm+http tests for tape storage areas (ATLAS)
- Oct 2021: srm+http tests for tape storage areas (CMS, LHCb)
- Dec 2021: srm+http in production for TAPE SAa (ATLAS)

Number of transferred files per protocol

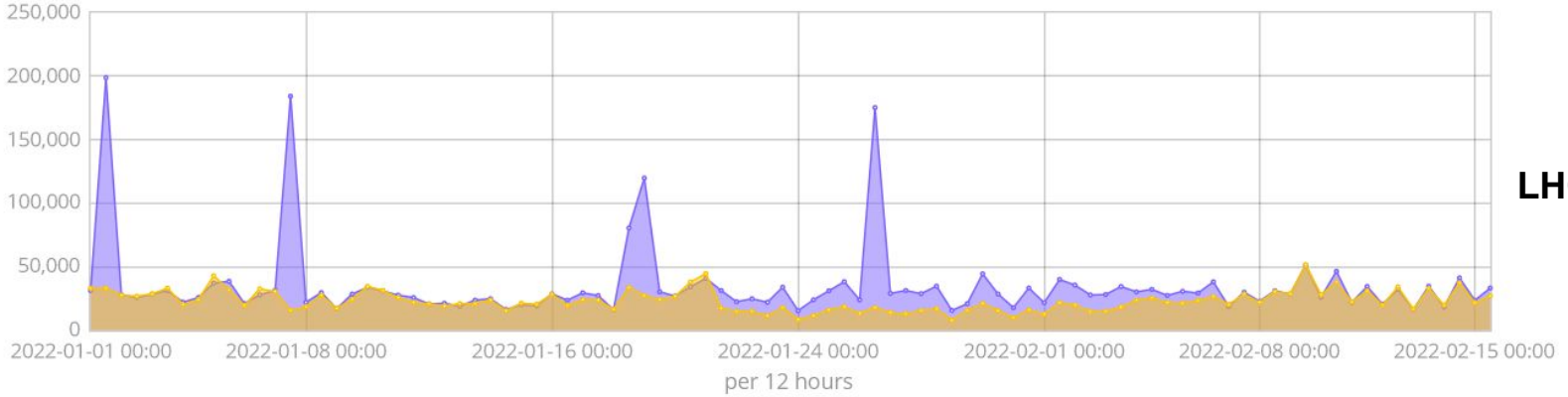


ATLAS



CMS

Number of transferred files per protocol



LHCb

- ATLAS, CMS and LHCb all use StoRM WebDAV in production (srm+http)
  - gsiftp still used, but the plan is to get rid of it before Run 3
  - StoRM WebDAV also provides a DM interface: fs ACLs enforced (ATLAS, LHCb)
- No-LHC experiments still rely on gsiftp
  - Exceptions (X509+StoRM WebDAV): Juno, Fcc, Belle
  - More exceptions (token+StoRM WebDAV): CTA-LST, nTof, JLab12, Km3Net, Fazia, Newsdm, Litebird, Belle

Transition to tokens

- Currently, our storage services support OAuth/OpenID Connect authentication and authorization mechanisms with StoRM WebDAV
- Several no-LHC experiments use its storage area browser application with tokens
- Timeline from WG for Transition to Tokens and Globus Retirement: “All storage services provide support for tokens in all relevant operations, including those for which currently SRM is still being used (tape)” by March 2022… quite optimistic :-)
- The StoRM developers are working at the WLCG Tape REST API, a common http rest interface allowing clients to manage access to files stored on tape (and to ultimately replace the SRM protocol)