Evolving storage services at INFN-T1

Andrea Rendina

Outline

- Introduction
- GridFTP transitioned to HTTPS
- X.509 certificates transitioning to JSON Web Tokens
- XrootD: status and issues
- Conclusions and future challenges

Introduction

CNAF Data Center

- CNAF hosts the main INFN data center, and the INFN Tier-1 in the WLCG e-infrastructure
- Provides services and resources to more than 40 scientific collaborations
  - LHC experiments so far the more demanding
  - ~46k cores, ~50 PB of disk, ~130 PB of tape
- Huge increase of resources foreseen in the coming years. By 2025:
  - ~130k cores, ~110 PB of disk, ~250 PB of tape
  - and even more (x10) from 2027 (HL-LHC)

Data access overview

- Several computing models to cope with
  - Experiment-driven (managed) vs user-driven (unmanaged)
  - Different storage usage, different requirements, different solutions
    - POSIX access (mainly read) from the WNs and the UIs
  - Heterogeneous protocols for data transfer
    - gsiftp (w/ and wo/ srm), https (w/ and wo/ srm), xrootd
  - Caches of various flavours
    - Xrootd proxy/caching proxy in support of the HPC datacenter integration: jobs running in Marconi (CINECA) access the full xrootd federation without external networking connectivity
    - StashCache for Virgo-Ligo, using CVMFS “external-data” feature
  - Different auth/z methods
    - Digital certificates, VOs and VOMS proxies, token-based

GridFTP transitioned to HTTPS

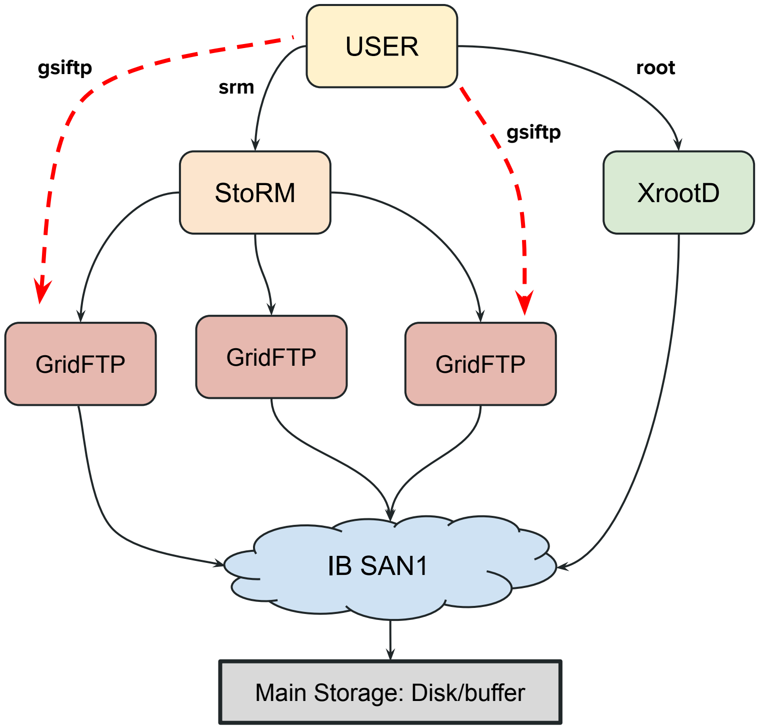
The Globus Toolkit retirement and its consequences

- In 2017, Globus announced they would stop supporting Globus Toolkit (end-of-life targeted for 2022)
    - WLCG uses two major features from the Globus Toolkit:
    - GridFTP as the primary third-party-copy transfer protocol for the WLCG infrastructure
        - DOMA (TPC) working group investigated alternatives for bulk transfers across WLCG sites
        - All storage elements to support WebDAV or XrootD-based TPCs
            - We provide support for HTTP-TPC with StoRM WebDAV
            - No plans to support XrootD-TPC at INFN-T1
    - GSI authentication, which is being transitioned to tokens

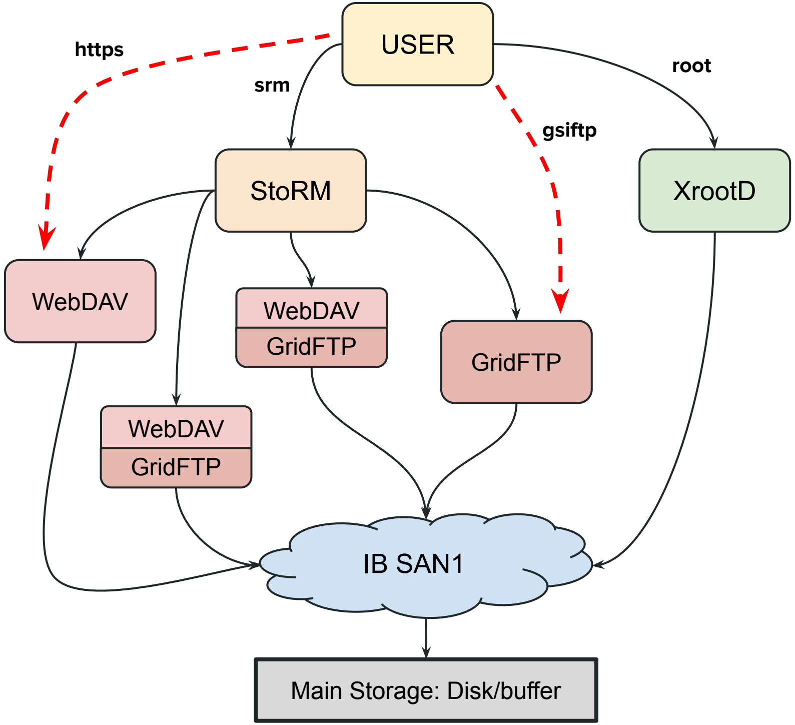
StoRM and GridFTP status at CNAF

- CNAF is a StoRM site (tape support via our HSM solution GEMSS)
  - A dedicated StoRM endpoint for each of ATLAS, CMS, LHCb
  - 2 endpoints shared among the (many) other VOs
- Each StoRM endpoint has a dedicated pool of StoRM WebDAV transfer nodes (14 in total)
  - 10 endpoints dedicated to the LHC experiments
  - StoRM WebDAV can also be used stand-alone, for data management
- 14 GridFTP transfer nodes
  - 6 endpoints dedicated to the LHC experiments
  - 4 for tape data management only
  - Before the transition they were 16

Typical disk data flow before the transition

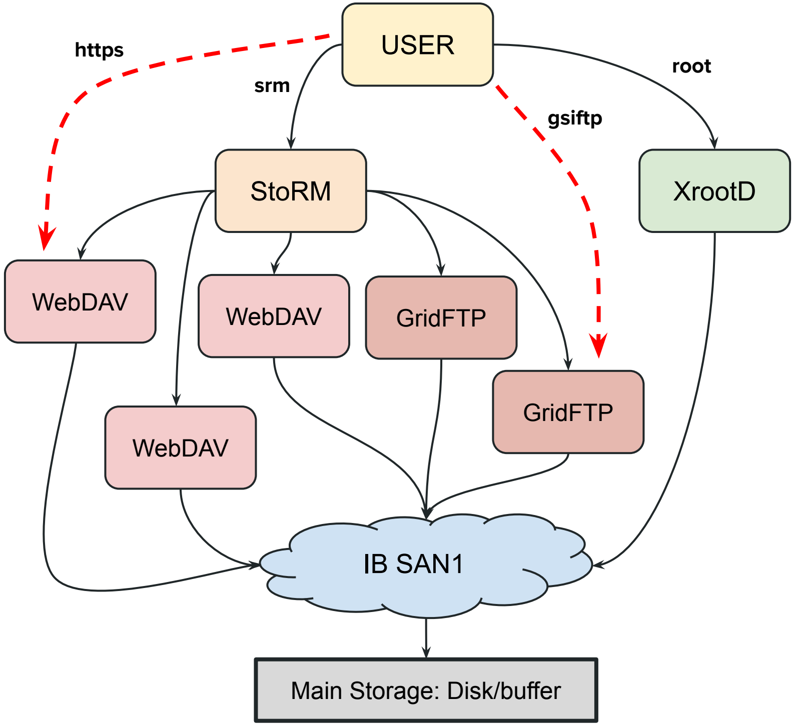


Typical disk data flow at the transition beginning



- Load balancing issue
  - Difficult to monitor the share of WebDAV and GridFTP in terms of load and connections;
  - splitting the services in dedicated machines seems to increase significantly the efficiency of the transfers

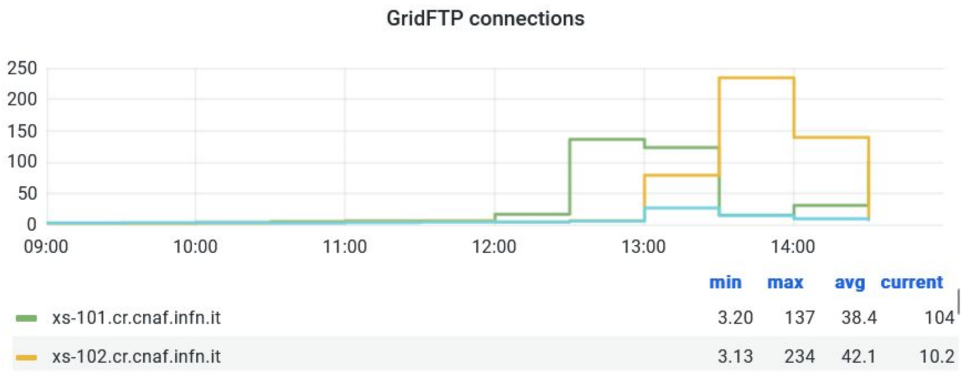
Typical disk data flow (now)



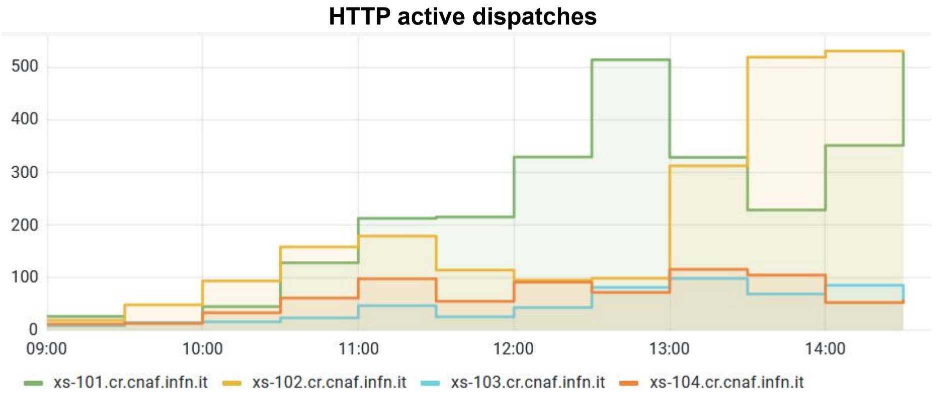
- Load balancing issue
  - Difficult to monitor the share of WebDAV and GridFTP in terms of load and connections
  - splitting the services in dedicated machines seems to increase significantly the efficiency of the transfers

	WebDAV	GridFTP
ATLAS	4	3
CMS	3	2
LHCb	3	1

Typical disk data flow (now)



	min	max	avg	current
🟩 xs-101.cr.cnaf.infn.it	3.20	137	38.4	104
🟨 xs-102.cr.cnaf.infn.it	3.13	234	42.1	10.2

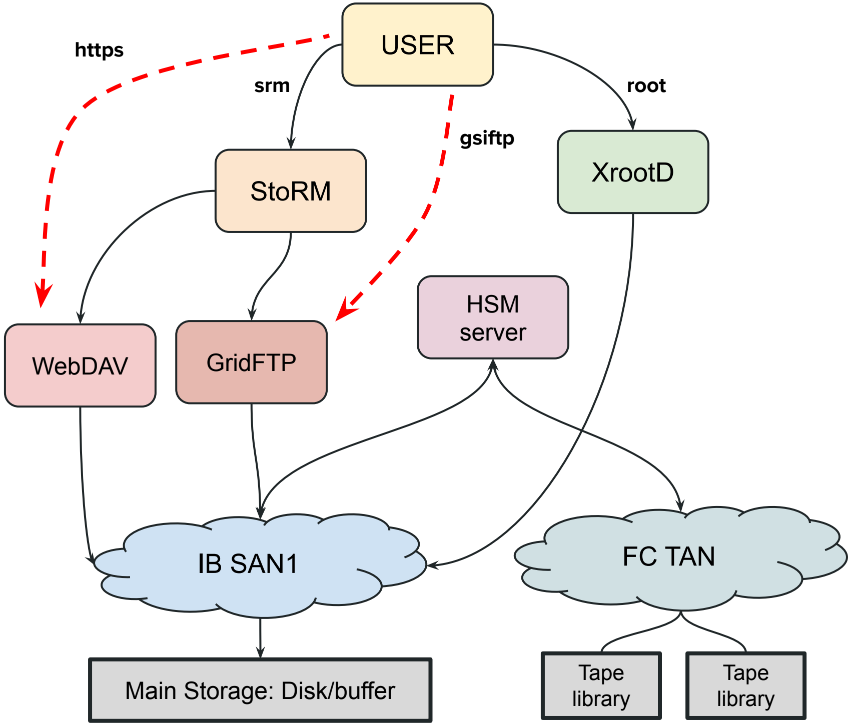


Load balancing issue

- Difficult to monitor the share of WebDAV and GridFTP in terms of load and connections
- splitting the services in dedicated machines seems to increase significantly the efficiency of the transfers

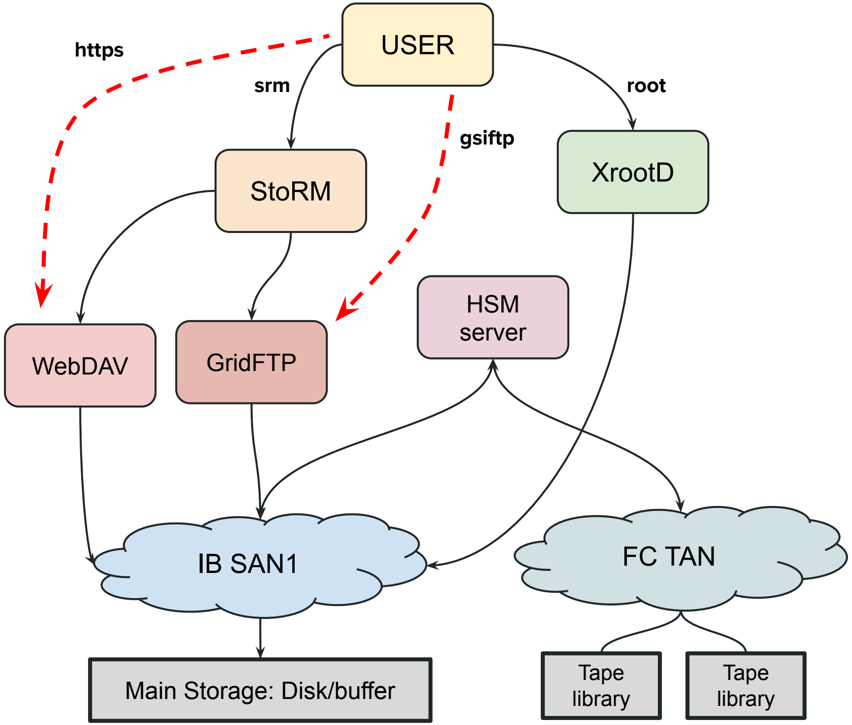
	WebDAV	GridFTP
ATLAS	4	3
CMS	3	2
LHCb	3	1

Typical tape data flow



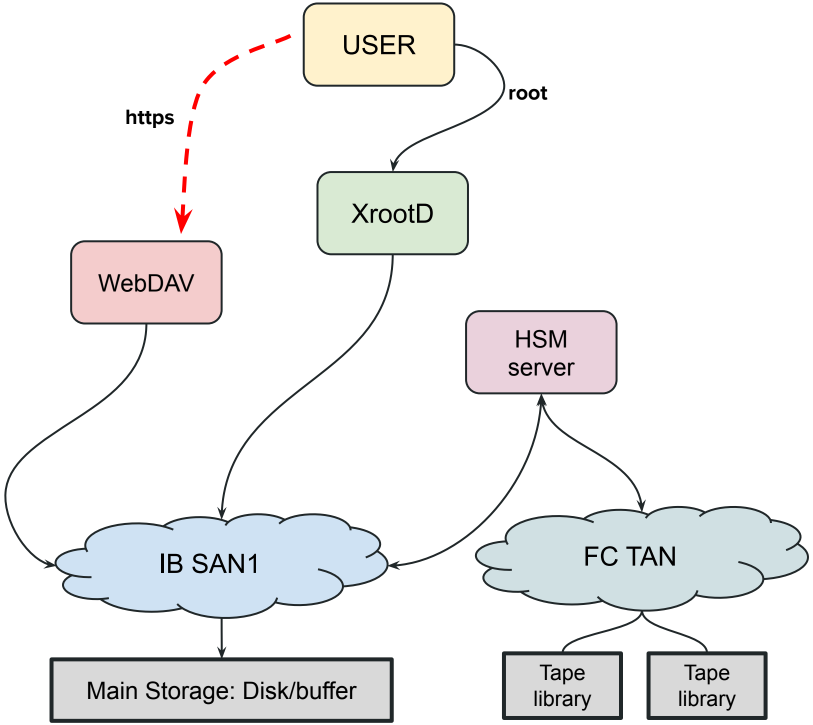
- The writing on tape and the related migrations from the disk buffer to tape still work fine both using StoRM WebDAV and GridFTP, but to recall data it is only possible via srm+gsiftp
- ALICE uses xrootd for tape as well
  - a specific plugin was developed @CNAF to manage tape recalls with XrootD

Typical tape data flow



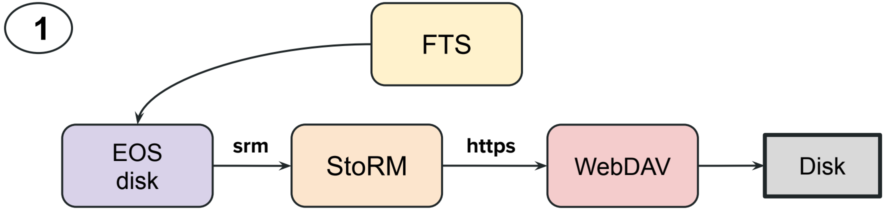
- The writing on tape and the related migrations from the disk buffer to tape still work fine both using StoRM WebDAV and GridFTP, but to recall data it is only possible via srm+gsiftp
- ALICE uses xrootd for tape as well
  - a specific plugin was developed @CNAF to manage tape recalls with XrootD

Typical data flow (near future)



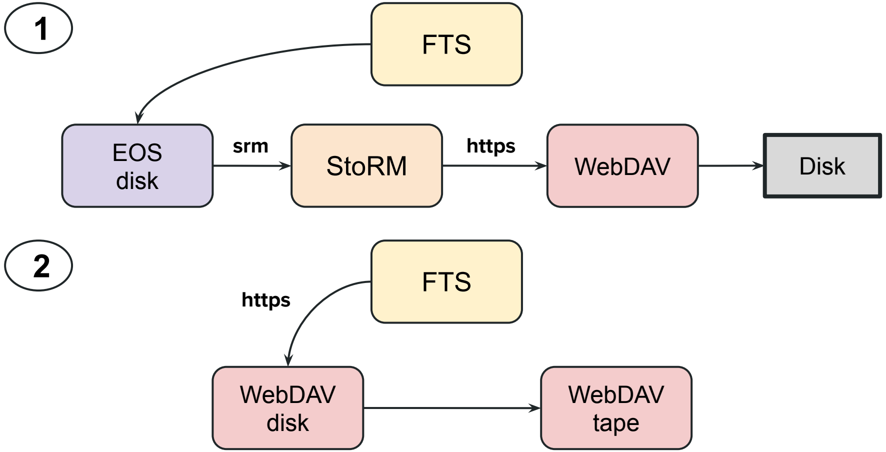
- StoRM developers are working at the WLCG Tape REST API, a common http rest interface allowing clients to manage access to files stored on tape (and to ultimately replace the SRM protocol)

The LHCb tape data challenge



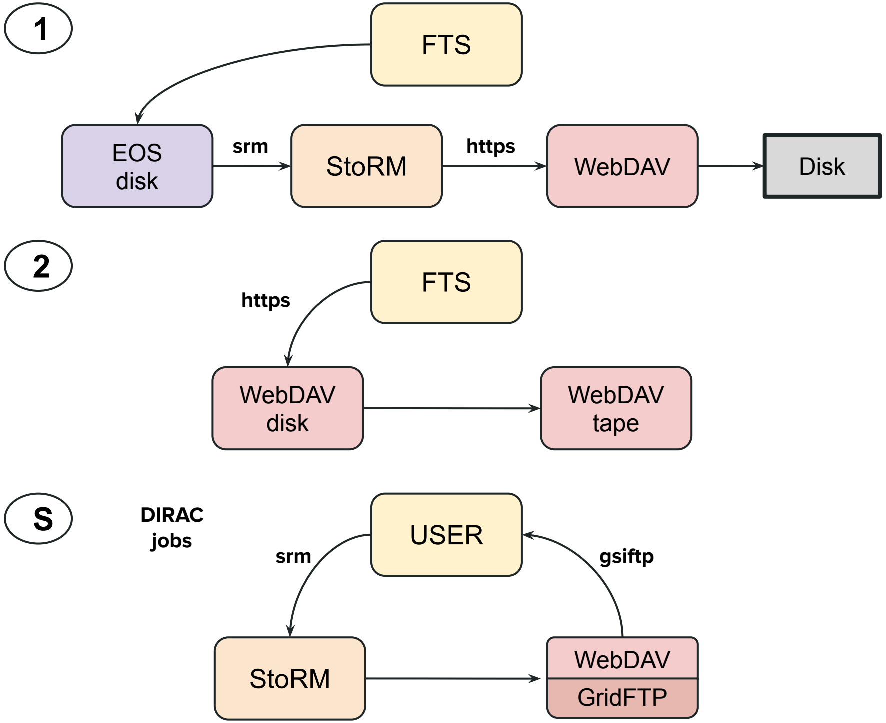
1. The data are copied from EOS disk to WebDAV disk via srm+https

The LHCb tape data challenge



1. The data are copied from EOS disk to WebDAV disk via srm+https
2. Then from WebDAV disk to WebDAV tape via https

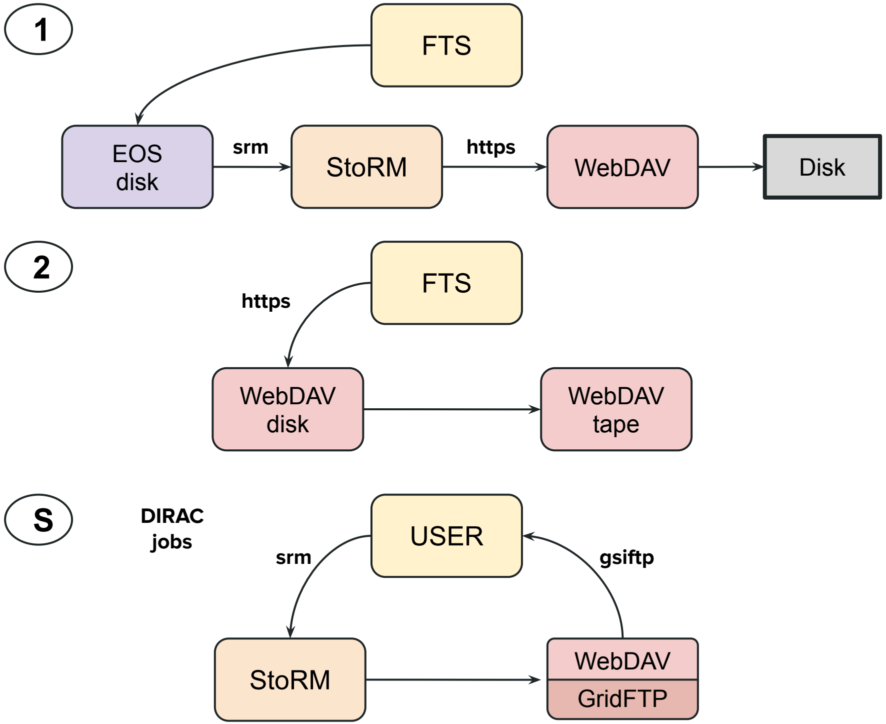
The LHCb tape data challenge



1. The data are copied from EOS disk to WebDAV disk via srm+https
2. Then from WebDAV disk to WebDAV tape via https

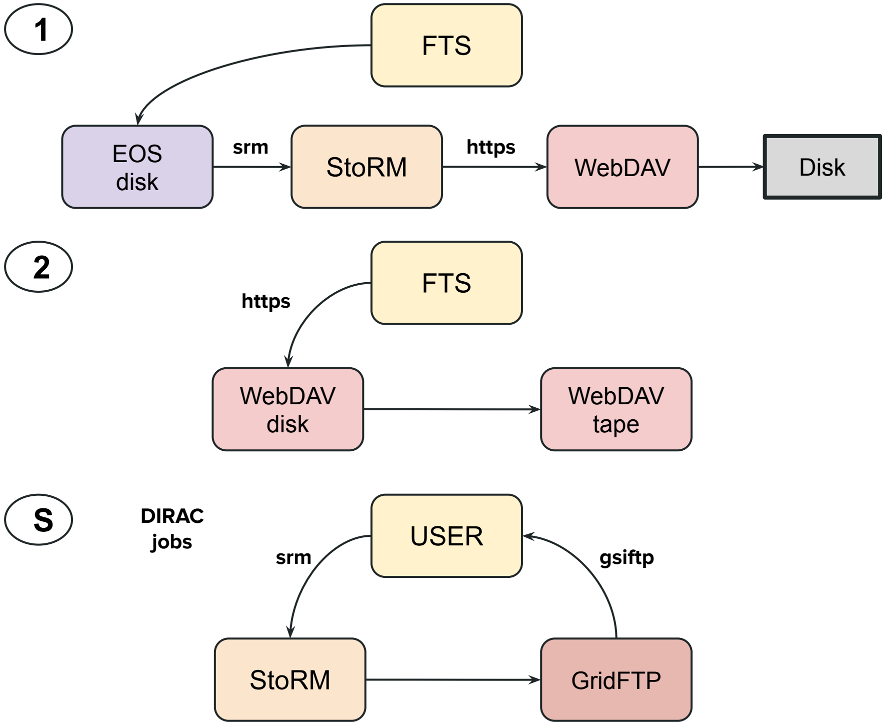
S. Simultaneously, DIRAC job download use srm+gsiftp

The LHCb tape data challenge



- High rate failure was observed:
  - LHCb disk and buffer share hw and file system, thus LHCb workflow saturated StoRM WebDAV threads
  - A known bug on FTS management of DNS cache implied a not balanced load among all the StoRM WebDAV servers

The LHCb tape data challenge



- The solution strategy:
  - Reserving one endpoint to GridFTP and the other to https seems to increase significantly the efficiency of the transfers
  - Probably need load-balancing strategy for StoRM WebDAV endpoints as StoRM already does by the *srm* protocol

X.509 certificates transitioning to JSON Web Tokens

Token-based authentication transition

- The GSI authentication is being transitioned to OAuth2.0 token-based authorization in all relevant WLCG workflows, following the WLCG Authorization Working Group initiative
  - A realistic timeline for the full transition from X.509 certificates to JSON Web Tokens seems to be slowly advancing towards 2024
  - In our storage services, we provide support for token-based authentication/authorization with StoRM WebDAV
  - Token-based authN/Z is already enabled for 15 No-LHC experiments (none of the LHC ones)
    - There are 4 IAM instances as token-issuers deployed by the Software Development group at CNAF
  - Some experiments are already accessing part of their data with X.509 VOMS proxies and JWT tokens at the same time within the same directory structure

The StoRM WebDAV integration

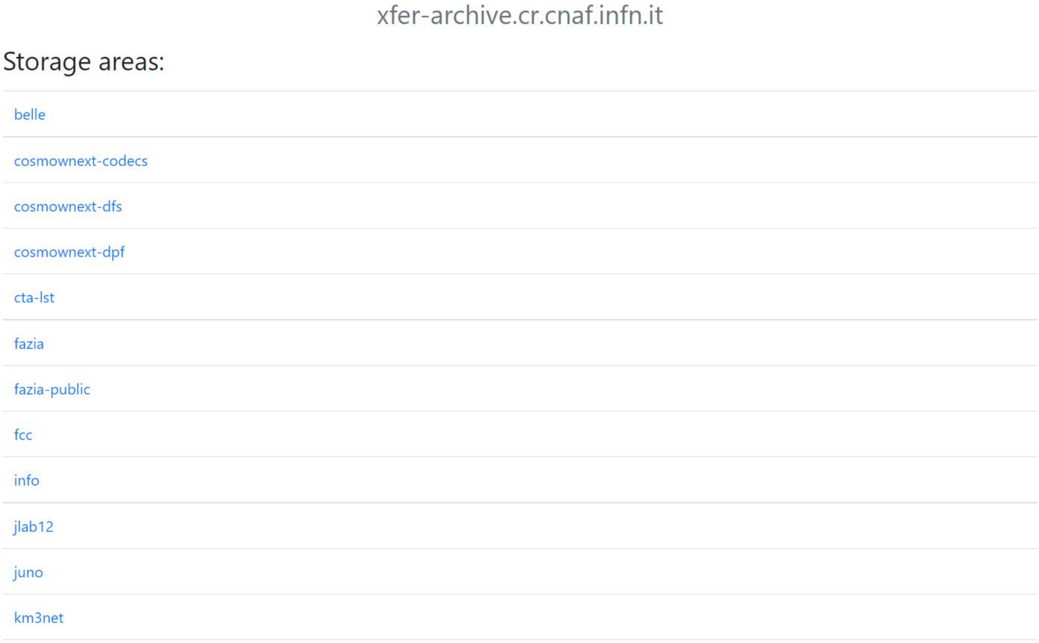
- StoRM WebDAV supports OpenID connect authentication and authorization on storage areas, so tokens can be used instead of proxies
    - The *group* claim allows to set the permissions for a user
        - As the voms-proxy roles
    - Also, the *scope* claim will allow to set the proper authorization for the operations
        - *storage.read*, *storage.create*, *storage.modify…*
    - It is possible to navigate the SA via web browser
    - The third-party copies avoid the macaroons usage
- The WLCG Tape REST API will be integrated with the token-based authentication

The StoRM WebDAV integration - the Belle II case

- Belle II uses StoRM WebDAV in production to access data stored on disk
- Authentication and authorization are both via voms-proxy and IAM-token, in production
  - Authorization on storage areas is NOT flat: StoRM WebDAV allows fine-grained authorization for both X509 proxies and IAM tokens
    - Specific access policies are defined for given folders in the storage area, and targeting specific authenticated group of users (e.g. "prod")

The StoRM WebDAV integration - the Belle II case

- It is possible to navigate the SA via **web browser** pointing at https://xfer-archive.cr.cnaf.infn.it:8443/



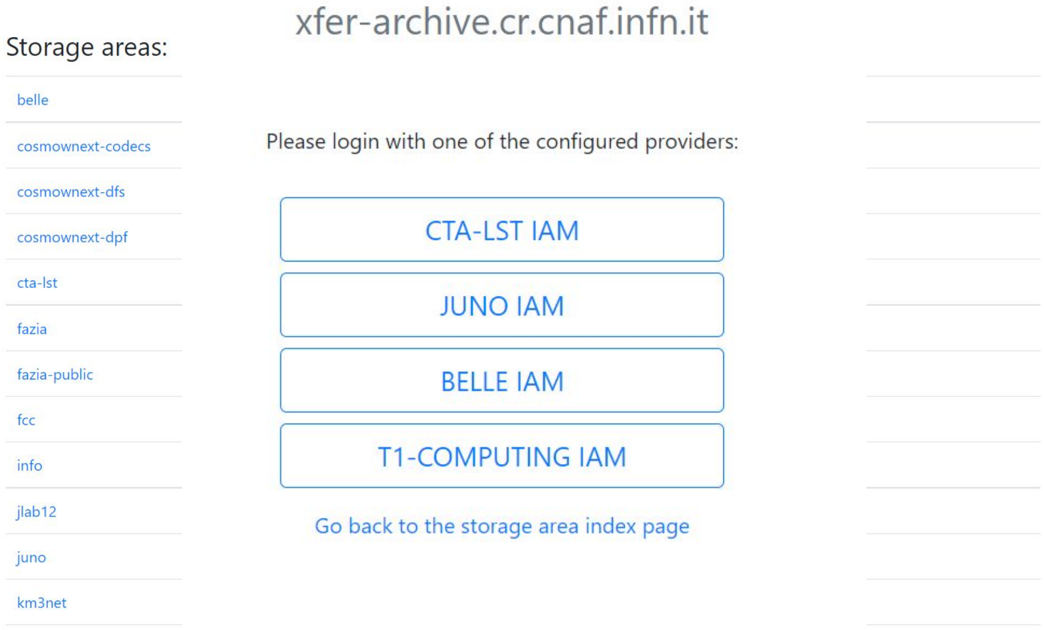
xfer-archive.cr.cnaf.infn.it

Storage areas:

- belle
- cosmownext-codecs
- cosmownext-dfs
- cosmownext-dpf
- cta-lst
- fazia
- fazia-public
- fcc
- info
- jlab12
- juno
- km3net

The StoRM WebDAV integration - the Belle II case

- It is possible to navigate the SA via **web browser** pointing at https://xfer-archive.cr.cnaf.infn.it:8443/



xfer-archive.cr.cnaf.infn.it

Storage areas:

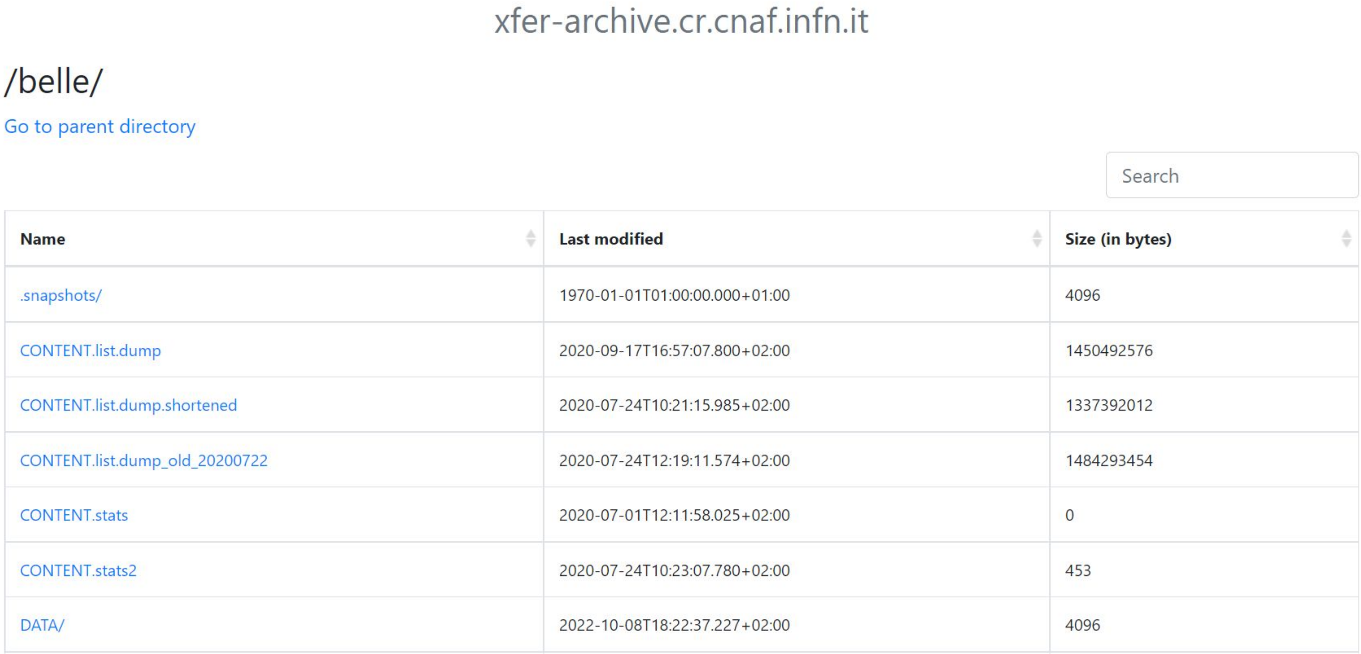
- belle
- cosmownext-codecs
- cosmownext-dfs
- cosmownext-dpf
- cta-lst
- fazia
- fazia-public
- fcc
- info
- jlab12
- juno
- km3net

Please login with one of the configured providers:

Go back to the storage area index page

The StoRM WebDAV integration - the Belle II case

- It is possible to navigate the SA via **web browser** pointing at https://xfer-archive.cr.cnaf.infn.it:8443/



xfer-archive.cr.cnaf.infn.it

/belle/

Go to parent directory

Search

Name	Last modified	Size (in bytes)
.snapshots/	1970-01-01T01:00:00.000+01:00	4096
CONTENT.list.dump	2020-09-17T16:57:07.800+02:00	1450492576
CONTENT.list.dump.shortened	2020-07-24T10:21:15.985+02:00	1337392012
CONTENT.list.dump_old_20200722	2020-07-24T12:19:11.574+02:00	1484293454
CONTENT.stats	2020-07-01T12:11:58.025+02:00	0
CONTENT.stats2	2020-07-24T10:23:07.780+02:00	453
DATA/	2022-10-08T18:22:37.227+02:00	4096

XrootD: status and issues

Overview

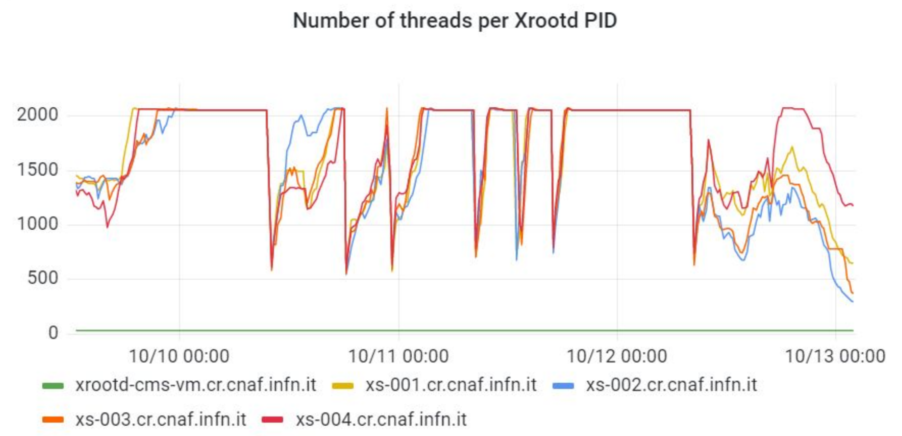
- ALICE has always performed data access using XrootD
  - Alice XrootD installation at INFN-T1 is specific and optimized to work on top of General Parallel File System (GPFS, by IBM).
  - A specific plugin was developed @CNAF to manage tape recalls
- CMS uses an XrootD federation
  - INFN-T1 hosts national and local redirectors, plus several servers
- ATLAS and LHCb use it sparingly for streaming data access
- Other experiments use dedicated XrootD instances, e.g. AMS, DAMPE, JUNO, PADME
- VIRGO uses a Stashcache instance to read data from /cvmfs
- They all add up to 37 XrootD instances
  - Only 33 make traffic

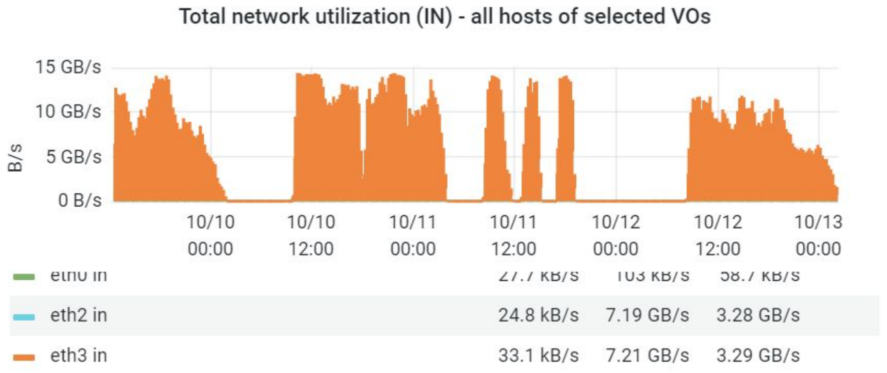
XrootD thread saturation

- We have been experiencing for some months an issue with the XrootD instances dedicated to the CMS experiment
- Threads saturated unevenly among servers with load increasing
    - We have already debugged it with the help of colleagues @CERN

XrootD thread saturation

- Suggested solutions:
  - Disable sendfile() for read requests setting xrootd.async nosf () greatly alleviated the load issues, and allowed us to remove limitation on max threads
  - Need to set the default value max threads (2048), otherwise threads rise up to 17k, with too high load
  - On GPFS side
    - Increase pagepool to 16GB
    - Separate NSD from XrootD servers

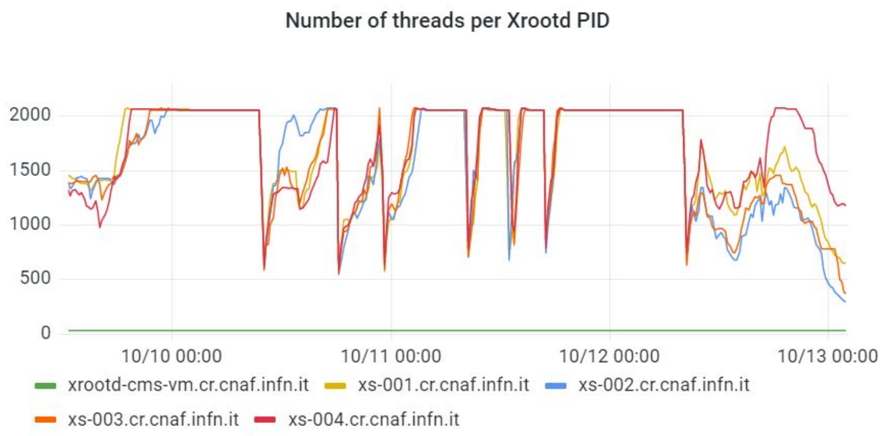


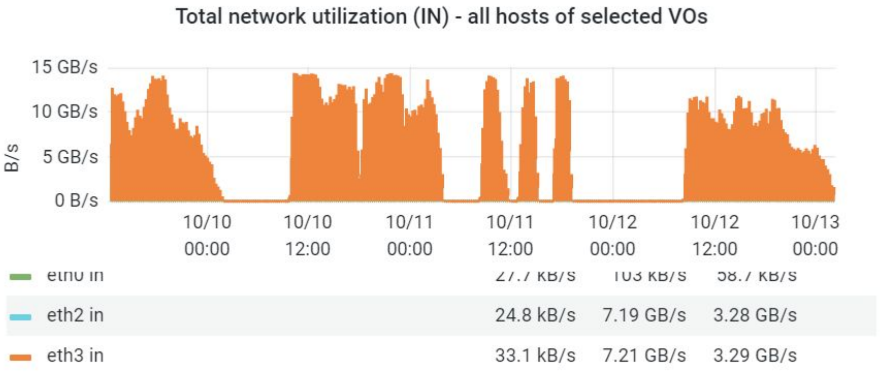


eth0 in	27.7 kB/s	103 kB/s	58.7 kB/s
eth2 in	24.8 kB/s	7.19 GB/s	3.28 GB/s
eth3 in	33.1 kB/s	7.21 GB/s	3.29 GB/s

XrootD thread saturation

- Still, it happens that threads saturate to the maximum value while
  - throughput goes to 0
  - the service stops logging in its log file
  - the last lines of the log files show a long sequence of **broken pipe** and **send failure** errors
  - a restart of the service solves the issue
  - at the peak load, each server is doing 4GB/s (40Gbps) of I/O (in and out) via network, serving ~2k connections
  - on average, more than 400 files are transferred per minute
- Currently investigating this with the help of the XrootD developers
  - double max_threads up to 4k





Conclusions and future challenges

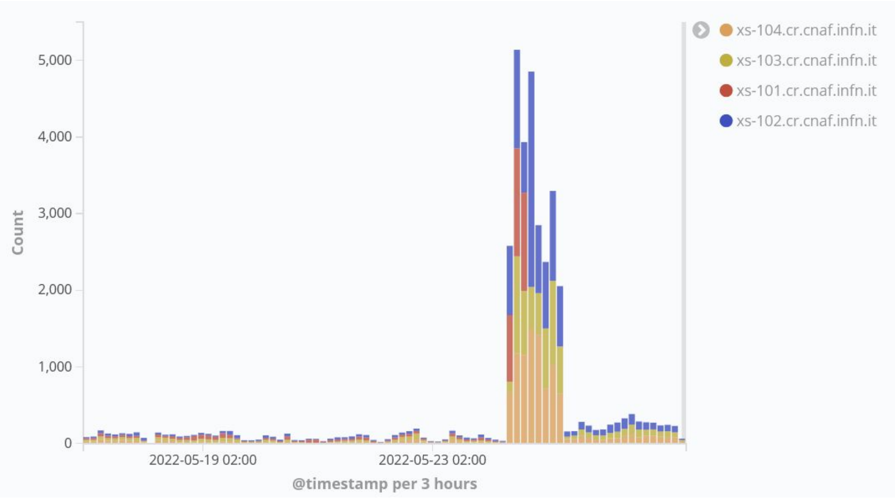
Conclusions and future challenges

- The transition from GridFTP to HTTP is still ongoing
  - Investigate and fix the open issues
- The transition towards token-based auth/z is ongoing, following DOMA
  - "By March 2022 all storage services to provide support for tokens including operations for which currently SRM is used (tape)"
  - Our storage services support token-based auth/z with StoRM WebDAV
    - Used by several no-LHC experiments

Thanks

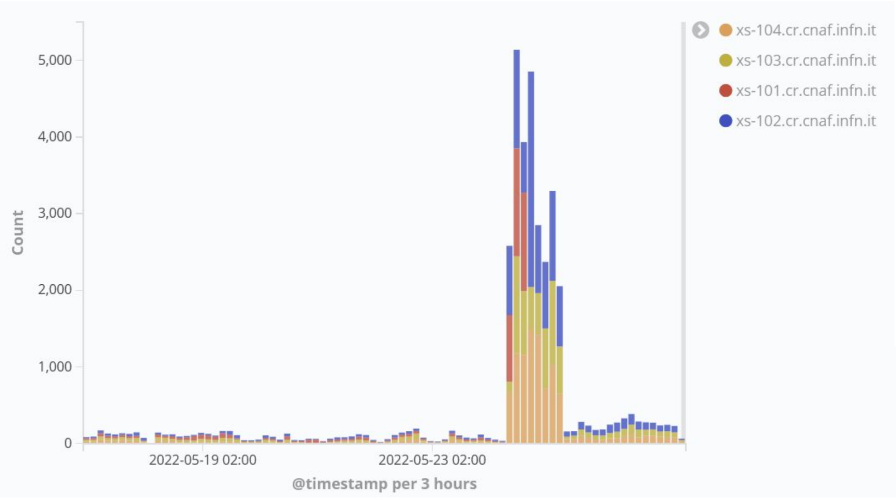
Backup slides

The LHCb tape data challenge



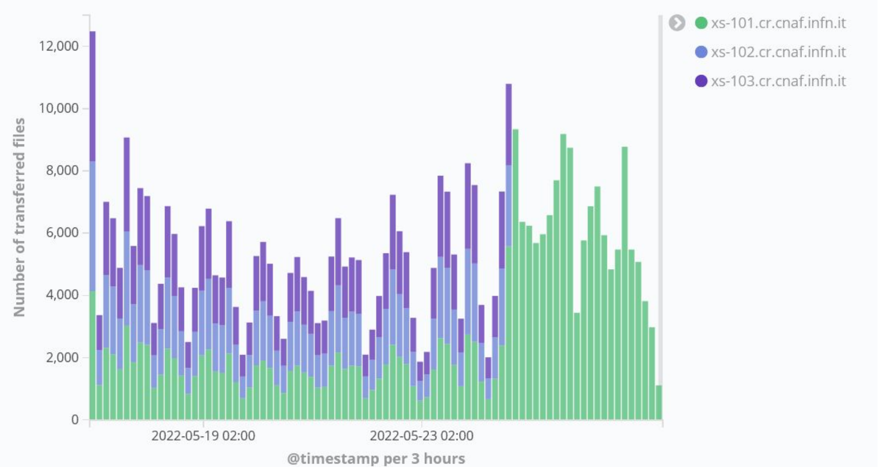
- The number of TPC is shown to peak during the tape challenge

The LHCb tape data challenge

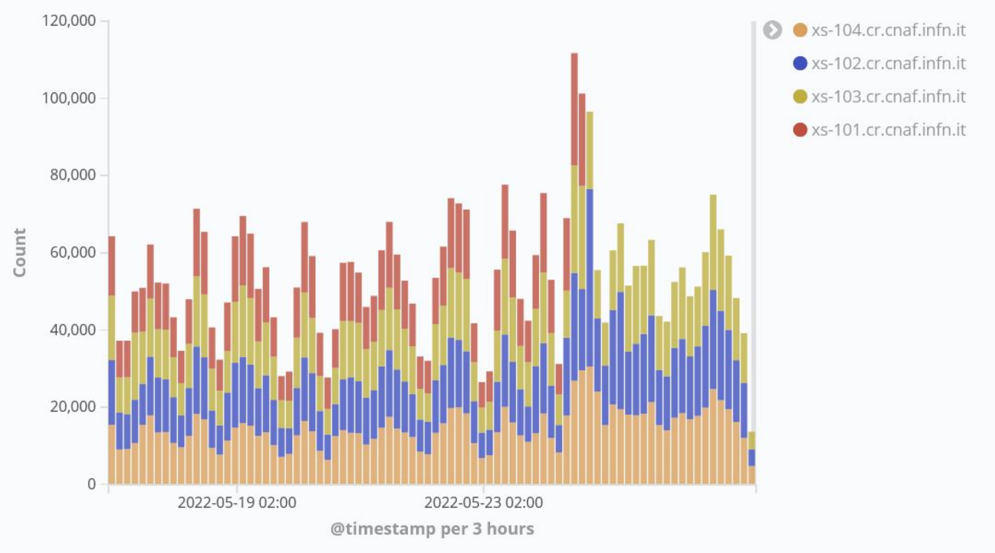


- The number of TPC is shown to peak during the tape challenge

- The number of GridFTP connections used to download files from the DISK started to involve only one server (xs-101)



The LHCb tape data challenge



- The plot shows the HTTP connections
- The balancing is better, but not perfect