Large-scale Data Handling experience at INFN-CNAF Data Center

Lucia Morganti
on behalf of the storage team @INFN-CNAF

Outline

- Introduction
- Architectural choices
- Data access
- Future challenges and conclusions

Introduction

The CNAF Data Center

- CNAF, located in Bologna, is the INFN National Center dedicated to Research and Development on Information and Communication Technologies
- CNAF hosts the main INFN data center, the INFN Tier-1 in the WLCG e-infrastructure

SCIENTIFIC COMPUTING FOR WLCG EXPERIMENTS, ASTROPARTICLES AND GW EXPERIMENTS, THEORETICAL PHYSICS AND BEAM SIMULATIONS.

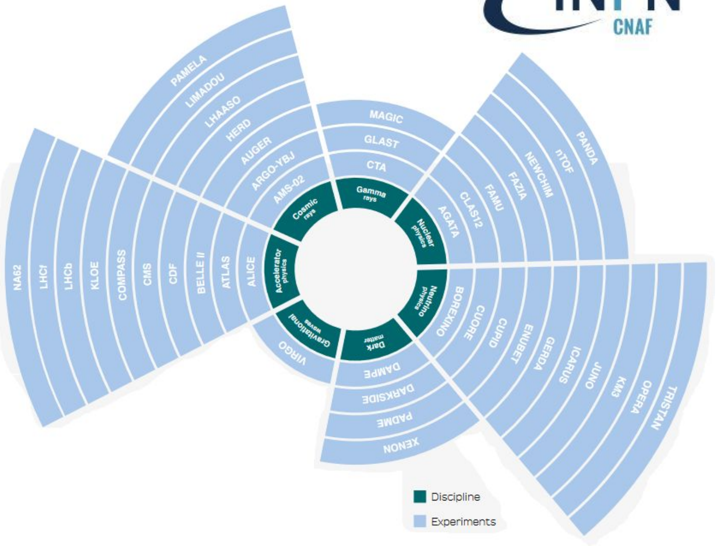
TECHNOLOGY TRANSFER TOWARDS INDUSTRY, PUBLIC ADMINISTRATION AND SOCIETY AT LARGE; SCOUTING FOR EXTERNAL PROJECTS.

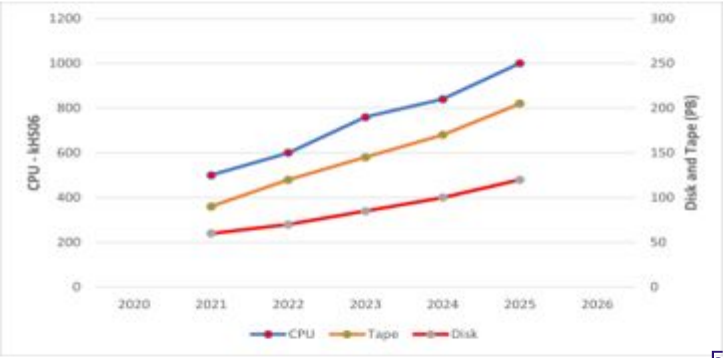
RESEARCH AND INNOVATION: DISTRIBUTED SYSTEMS (CLOUD AND GRID), SOFTWARE DEVELOPMENTS FOR EXPERIMENTS AND EXTERNAL PROJECTS, TRACKING OF THE NEW HARDWARE TECHNOLOGY.

ICT SERVICES FOR INFN TO DEVELOP, MANAGE AND SUPPORT GENERAL UTILITY SERVICES SUCH AS BOOKKEEPING, ENTERPRISE CONTENT MANAGEMENT, WEB SERVERS, ETC.

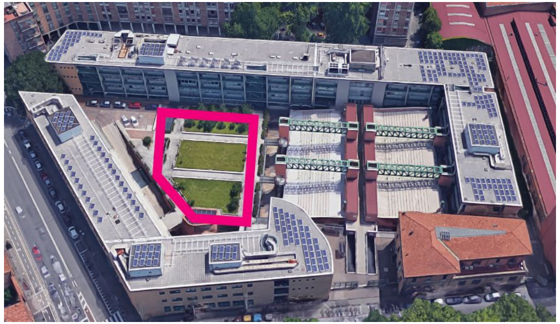
WLCG Tier-1

- Provides services and resources to more than 40 scientific collaborations
  - LHC experiments so far the more demanding
  - ~42k cores,~50 PB of disk, ~116 PB of tape
- Huge increase of resources foreseen in the coming years. By 2025:
  - ~130k cores, ~110 PB of disk, ~250 PB of tapes
  - and even more (x10) from 2027 (HL-LHC)
- The Data Center is moving to a new location. Three main drivers for the move to the Tecnopolo:
  - expected huge increase in IT resources
  - infrastructural problems at the current site
  - the opportunity offered by the new location.





CNAF Data Center

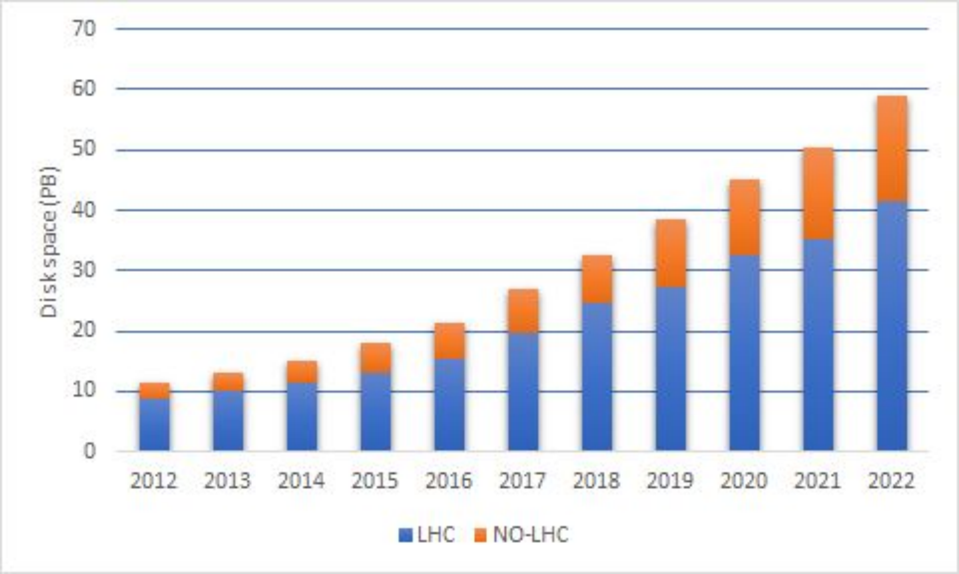


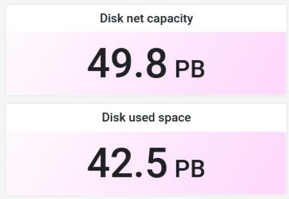
- Usable area of the data centre: 800 m^2;
- Maximum electrical power 1.4 MW
- With the current IT technology, we would be able to host IT resources to cover the requirements up to the end of LHC Run 3 (2024).

… moving to the Tecnopolo

- Usable area larger than 2000 m^2, electrical power from 3 MW in the first phase to 10 MW from 2027.
- A greener DC (targeting 1.08-1.10 PUE)
- Goal: meet the requirements for the data taking of the HL-LHC experiments up to 2035 and beyond, providing as well services for many other INFN experiments, projects, and activities.

Disk storage @CNAF





Disk net capacity

49.8 PB

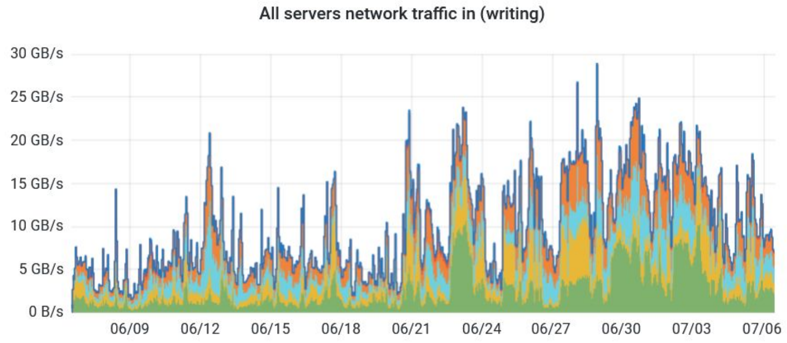
Disk used space

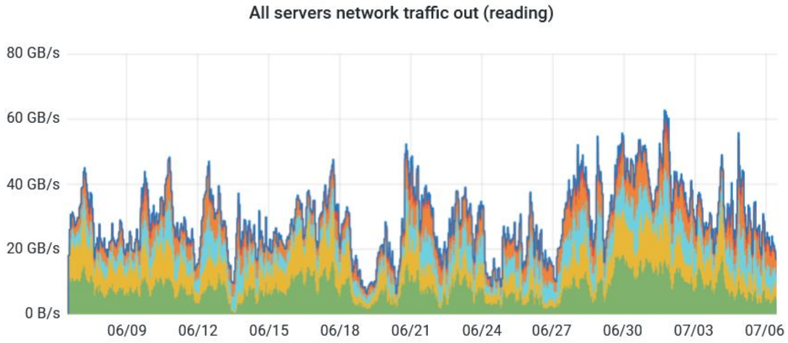
42.5 PB

Disk storage:

- DDN SFA12K (x2) and SFA7900 (x2)
- Huawei OS18800v5 (x5), 6800v5 and 5800v5 (x4)
- DELL MD3860F (x4)
- Some SSD and NVME disks for metadata

Disk storage @CNAF



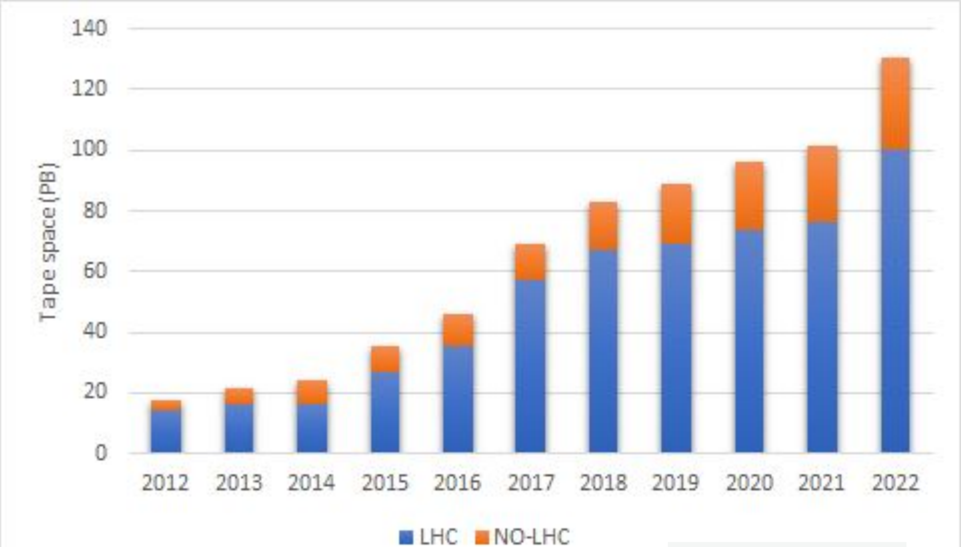


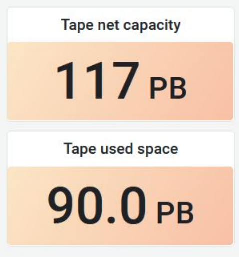
Disk storage:

- DDN SFA12K (x2) and SFA7900 (x2)
- Huawei OS18800v5 (x5), 6800v5 and 5800v5 (x4)
- DELL MD3860F (x4)
- Some SSD and NVME disks for metadata

- ~700 TB in, ~1.7 PB out per day
- ~300k transferred files per day

Tape storage @CNAF





Tape net capacity

117 PB

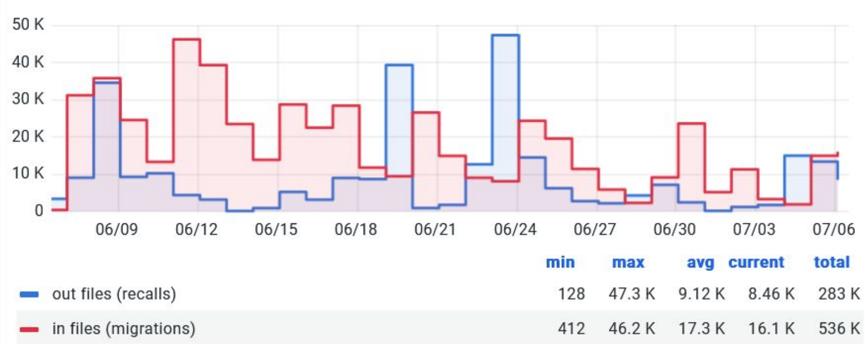
Tape used space

90.0 PB

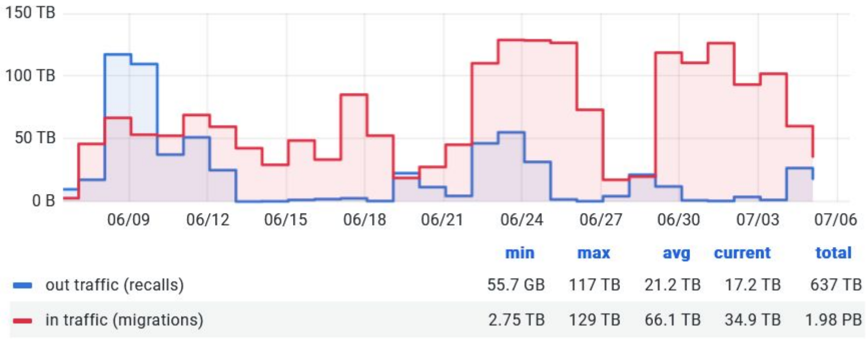
Two tape libraries:

- ORACLE SL8500
  - 10000 slots fully filled
  - 16 T10KD tape drives
  - 8.4 TB tape cartridges
  - 250 MB/s bandwidth per drive
- IBM TS4500
  - 6198 slots
  - 19 TS1160 tape drives
  - 20 TB tape cartridges
  - 400 MB/s bandwidth per drive

Tape storage @CNAF



	min	max	avg	current	total
🟦 out files (recalls)	128	47.3 K	9.12 K	8.46 K	283 K
🟥 in files (migrations)	412	46.2 K	17.3 K	16.1 K	536 K



	min	max	avg	current	total
🟦 out traffic (recalls)	55.7 GB	117 TB	21.2 TB	17.2 TB	637 TB
🟥 in traffic (migrations)	2.75 TB	129 TB	66.1 TB	34.9 TB	1.98 PB

Two tape libraries:

- ORACLE SL8500
  - 10000 slots fully filled
  - 16 T10KD tape drives
  - 8.4 TB tape cartridges
  - 250 MB/s bandwidth per drive
- IBM TS4500
  - 6198 slots
  - 19 TS1160 tape drives
  - 20 TB tape cartridges
  - 400 MB/s bandwidth per drive
- 10k recalled files, 20k migrated files per day
- 20 TB recalled, 70 TB migrated per day

Team

- Mission: provide and operate storage solutions and tools for data management and data transfer to experiments and users
- 8 people (1 group coordinator + 5 staff members + 2 research fellows)
- At least 2 people share operational know-how of each task/service/tool

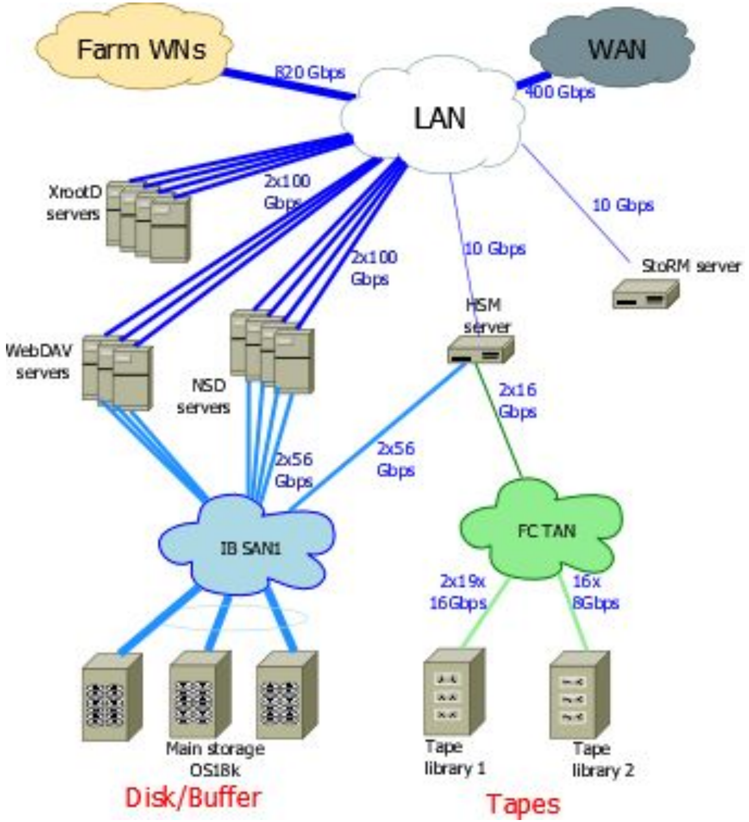
Architectural choices

Architectural choice

- Solution well consolidated over the years
- Storage servers:
    - SAN-based solution
    - Backend: Infiniband 56Gbps (FDR) and 100Gbps (EDR) and FiberChannel 16Gbps
    - Frontend: 2x100 GbE, 2x25 GbE and 4x10 GbE
- Software:
    - Parallel file system IBM Spectrum Scale (aka GPFS) as POSIX interface and backend for all data management and data transfer services
    - Interface to tape: IBM Spectrum Protect (aka TSM) + in-house optimization layer
    - Advantages:
        - performance
        - relying on stable and well supported sw
        - minimizing support effort

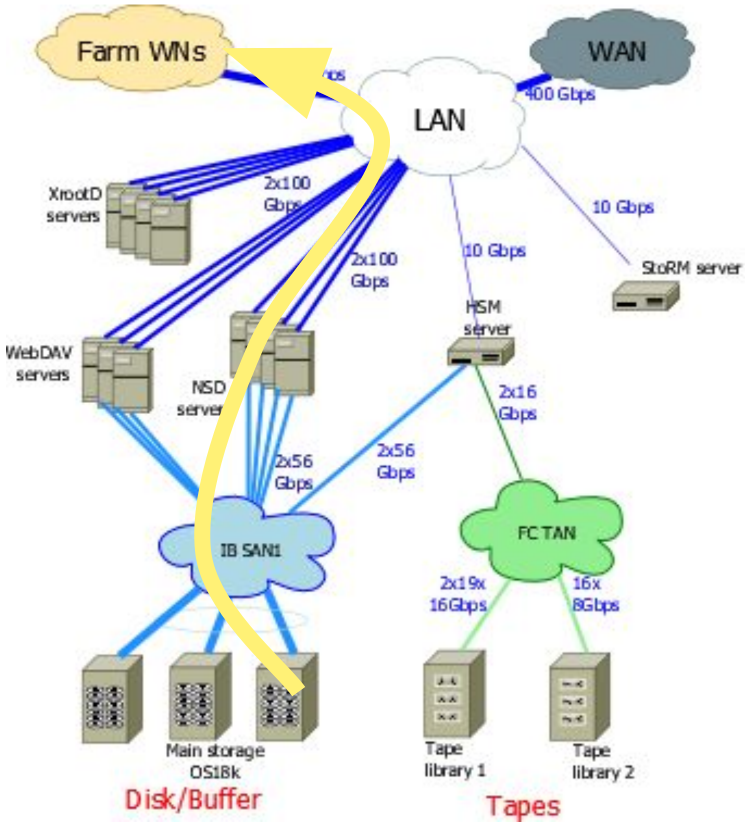
Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



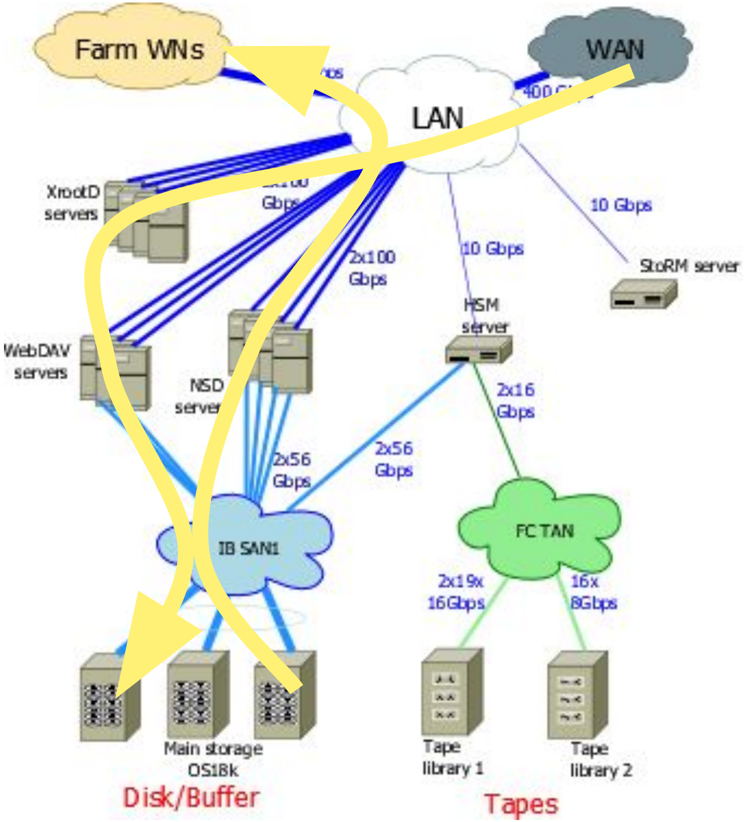
Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



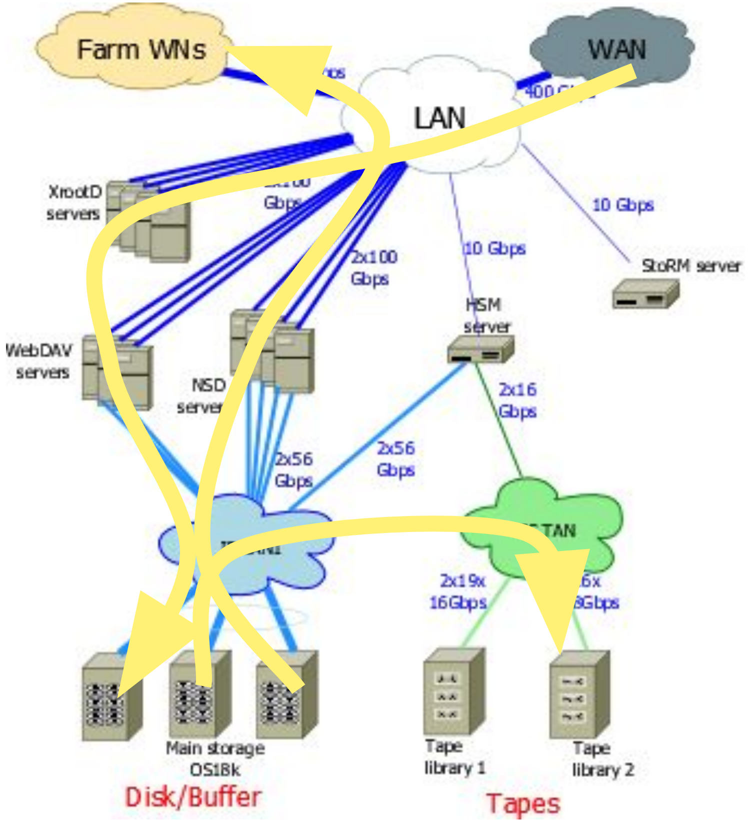
Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



Data management services: StoRM

- SW (middleware) mainly in the hands of experiments or small groups of developers (3-5 people)
    - Very specific, far from being industry-standard
- CNAF is a StoRM site (tape support via our HSM solution GEMSS)
    - A dedicated StoRM endpoint for each of ATLAS, CMS, LHCb; two endpoints shared among the (many) other VOs
    - Each StoRM endpoint has a dedicated pool of StoRM WebDAV transfer nodes (14 in total)
    - GridFTP transfer nodes are still there (14 in total) ⚠
    - 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)

Data access

Data access

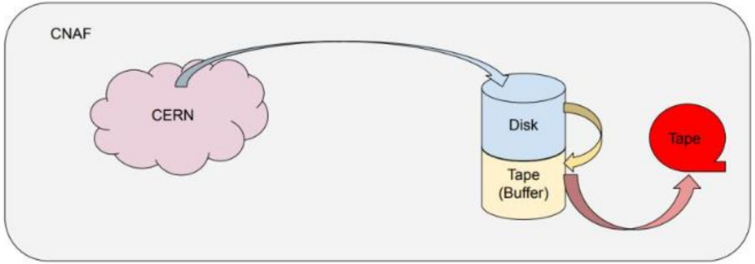
- Several computing models to cope with
  - Experiment-driven (managed) vs user-driven (unmanaged)
  - Different storage usage, different requirements, different solutions
    - POSIX access (mainly read) from the WNs and the UIs
    - Heterogeneous protocols for data transfer
      - gridftp (w/ and wo/ srm), xrootd, https (w/ and wo/ srm)
    - Caches of various flavours
      - Xrootd proxy/caching proxy in support of the HPC datacenter integration: jobs running in Marconi (CINECA) access the full xroot 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

Data transfer protocols: GridFTP protocol replacement

- 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, and the DOMA working group (DOMA TPC) investigated alternatives for bulk transfers across WLCG sites
    - All storage elements to support WebDAV- or XrootD-based TPCs
    - No plans to support XrootD-TPC at INFN-T1, we provide support for HTTP-TPC with StoRM WebDAV
  - GSI authentication, which is being transitioned to tokens.
- The HTTP-TPC transition is most advanced, and should be completed “before Run3” (quoting DOMA BDT 16/2/2022)

GridFTP transitioned to HTTPS

- The 2021 Network Data Challenge was carried out using HTTP-TPC (disk) as the final step of the commissioning process
  - INFN-T1 performed very well
- The 2022 Tape Data Challenge used srm+http
  - INFN-T1 achieved target rates
  - @INFN-T1, LHCb disk and buffer share hw and file system, thus LHCb workflow saturated StoRM WebDAV threads
    - Known bug on FTS management of DNS cache
    - Probably need load-balancing strategy for StoRM WebDAV endpoints

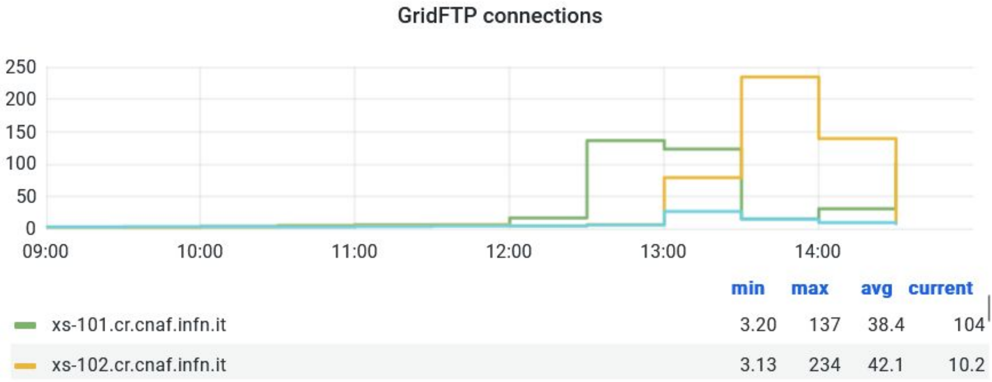


LHCb workflow (https://l.infn.it/n3)

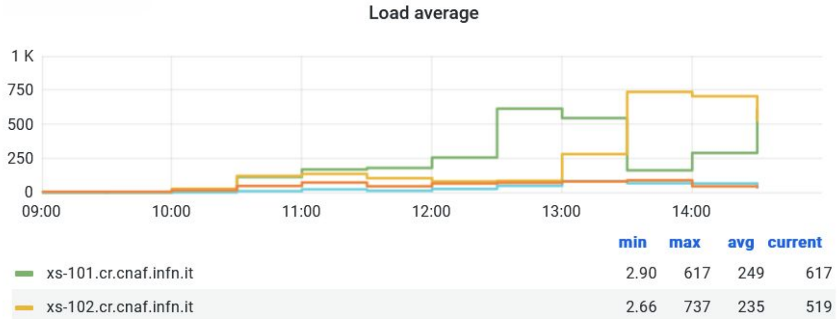
However, GridFTP still here :-(

LHC experiments are not gsiftp-free yet:

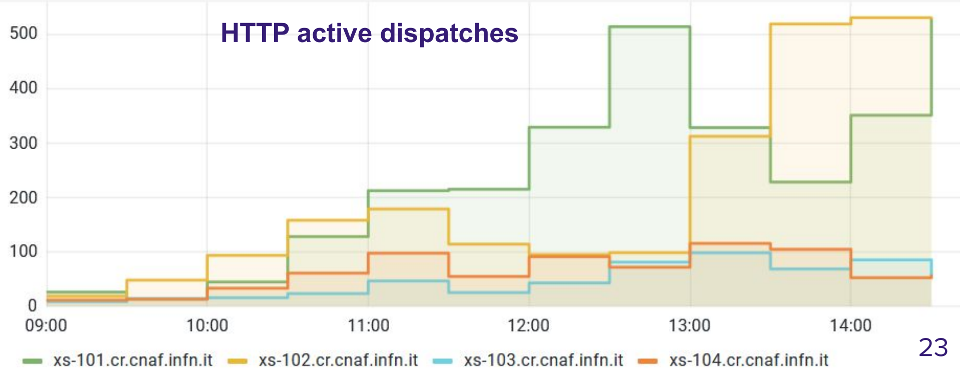
- Mainly traffic from T2s
- But also from WNs (LHCb)
- Reserving one endpoint to GridFTP and the other to https seems to increase significantly the efficiency of the transfers


GridFTP connections

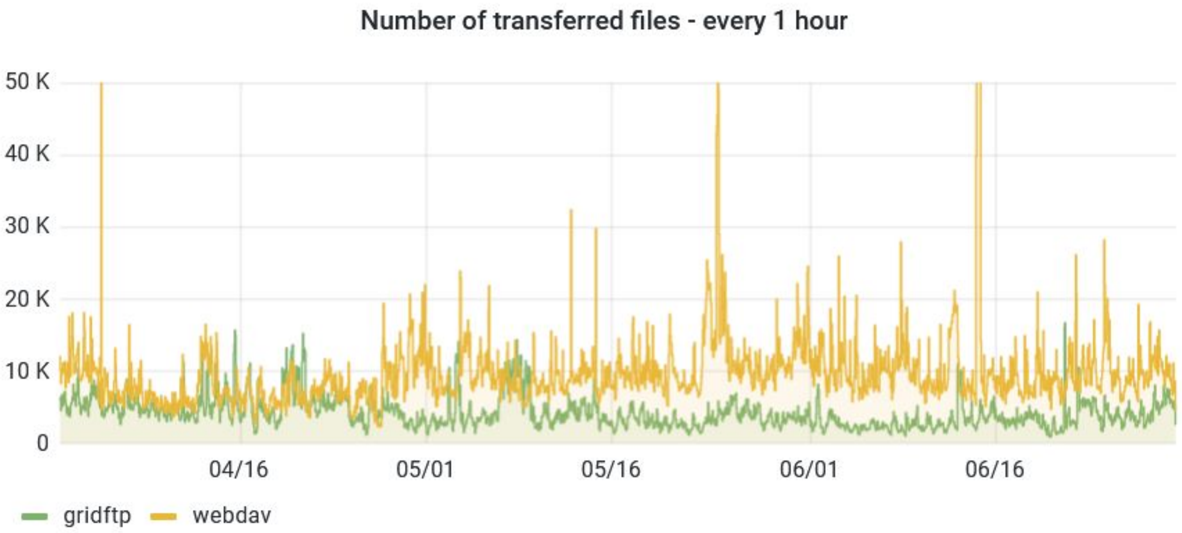
	min	max	avg	current
xs-101.cr.cnaf.infn.it	3.20	137	38.4	104
xs-102.cr.cnaf.infn.it	3.13	234	42.1	10.2


Load average

	min	max	avg	current
xs-101.cr.cnaf.infn.it	2.90	617	249	617
xs-102.cr.cnaf.infn.it	2.66	737	235	519


HTTP active dispatches

However, GridFTP still here :-(



No-LHC experiments still rely heavily on gsiftp

Data transfer protocols: XrootD

- 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 40 XrootD servers

Future challenges and conclusions

Future challenges

- Transition gridftp ➔ http still ongoing
- Transition towards token-based auth/z 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
- CEPH is being considered as alternative for Disk-Only solution (i.e. without tape backend)
    - A dedicated file system was deployed and is used by ALICE
- CNAF data center has to cope with the next challenges of science
    - Resources x2 by 2025, and even more (x10) from 2027 (HL-LHC)

Conclusions

- Storage operations @INFN-T1 are proceeding smoothly
- We are supporting ongoing transitions to HTTP and token-based auth/z
    - We are currently deploying (too) many services, and we hope this ecosystem gets simpler
- We are actively planning and working for the transition (!) of the Data Center to the Tecnopolo

Thanks!

Backup slides

Data transfer services: issues with XrootD

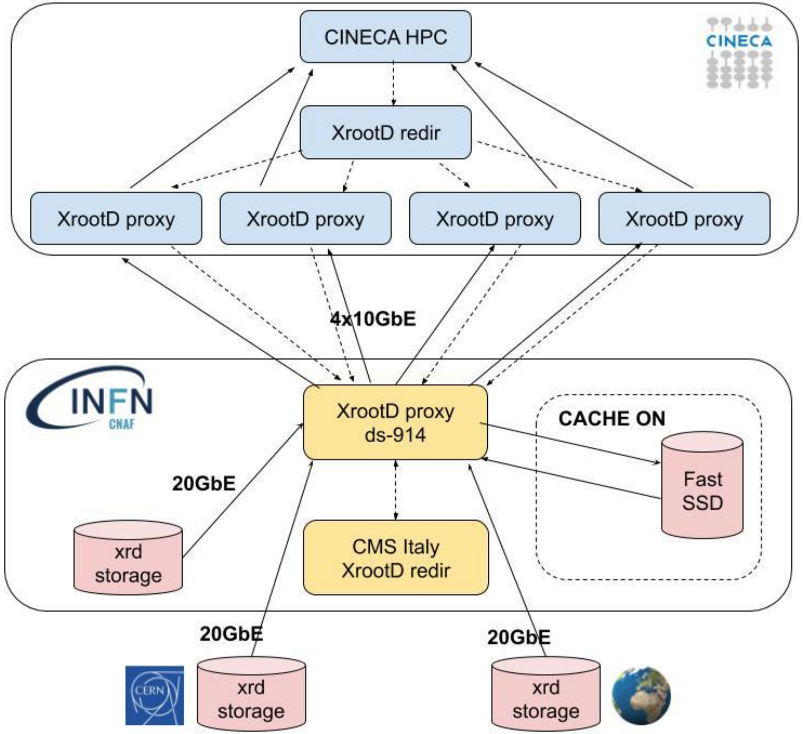
- Threads saturated unevenly among servers with load increasing
    - Disable *sendfile()* for read requests setting *xrootd.async nosf ()* greatly alleviated the load issues, and allowed us to remove limitation on *max threads*
    - Need to manually set the default value *max threads* (2048)
    - On GPFS side
        - Increase pagepool to 16GB
        - Separate NSD from XrootD servers
- Still, it happens that threads saturate to the maximum value while xrootd makes no traffic and server load is very low
    - A restart of the service solves the issue
    - Currently investigating this with the help of colleagues @CERN

Supporting INFN T1 extension to an HPC system

- Since 2015 CNAF has started a R&D program for the utilization of remote CPU resources to extend the data center beyond its premises
- PRACE Project Access to LHC Italy community for using the nodes from the CINECA Marconi KNL partition (3600 nodes, 68x4 cores per node, 96 GB RAM)
    - The nodes don’t have external network connectivity. The key issue is then remote access to data.
    - This limitation has been solved by enabling external networking to CNAF and CERN.

Supporting INFN T1 extension to an HPC system

- An XrootD proxy installed at CNAF makes the full XrootD federation visible to CMS, via an Xcache setup (caching also possible).
- With such a setup (+ CVMFS and Singularity on the nodes), the experiment workflows can be executed on KNL Marconi with only limitation in uplink between A2 and CNAF Storage



CINECA HPC
CINECA
XrootD redir
XrootD proxy
XrootD proxy
XrootD proxy
XrootD proxy
4x10GbE
INFN CNAF
XrootD proxy ds-914
CACHE ON
Fast SSD
20GbE
xrd storage
CMS Italy XrootD redir
20GbE
20GbE
CERN
xrd storage
xrd storage