Risorse Storage HTC

Corso di formazione per utenti - INFN DataCloud
5 Marzo 2025

Daniele Lattanzio
INFN - CNAF

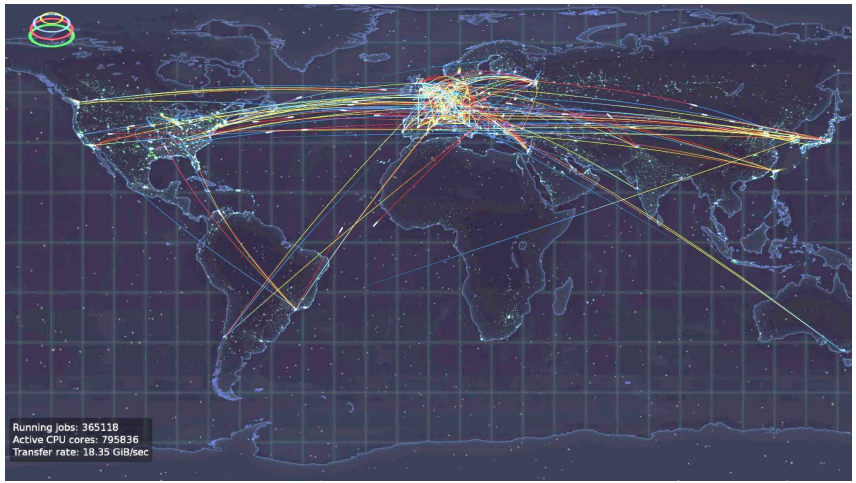
Outline

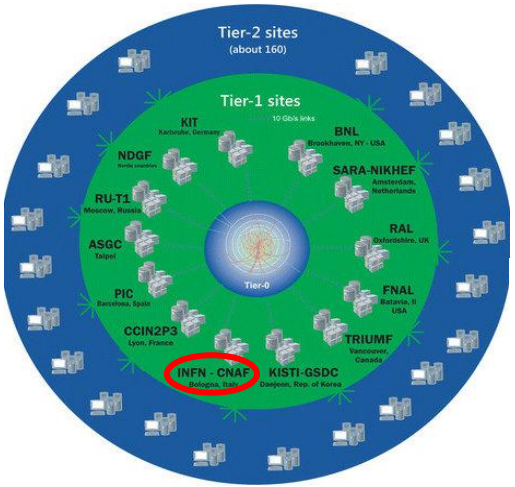
- Introduction
- Storage solutions
- File Systems
- Basic concepts
- Data transfer and data management
- Tape data management

Introduction

The WLCG collaboration

- The **Worldwide LHC Computing Grid** (**WLCG**)
- Involves around **170 computing centres** in more than **40 countries**
- Provides computational resources to store, distribute and analyse the **~200 PB** of data expected every year from the Large Hadron Collider (**LHC**) at CERN





WLCG
Worldwide LHC Computing Grid

The Italian WLCG Tier-1

- Since 2003 the Italian WLCG Tier-1 is located in **Bologna**
  - providing **resources**, **support** and **services** to **storage** activities, **distribution**, **processing** and **data analysis**
- managed by **INFN-CNAF**
- **60+ scientific communities** using the data centre
  - not only LHC and not only from the physics field

The Italian WLCG Tier-1

- ~**130 PB of disk space** shared among all nodes via a **distributed file system**
- ~**189 PB of tape space** used as the main **long-term storage** medium

Storage Solutions

Tiered Storage

- Tiered storage is a data storage environment consisting of two or more kinds of storage delineated by differences such as **price**, **performance**, **capacity**, **function**
- Any significant difference in one or more of the defining attributes can be sufficient to justify a separate storage tier
  - **Disk and tape**: two separate storage tiers identified by differences in all four defining attributes
  - Old technology disk and new technology disk
  - High performing disk storage and less expensive, slower disk of the same capacity and function.

Why Tape?

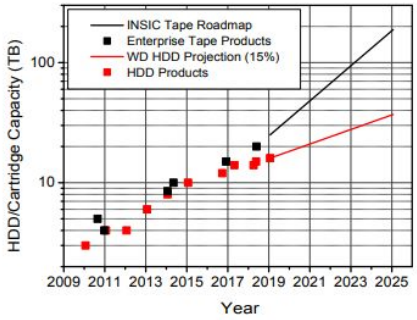
- **Cost** lower than disk
- **Longevity**: lifetime may last decades
  - long term backup
- **Large capability**
  - e.g: Tape@CNAF: 20 TB/cartridge (50 TB coming soon)
- **Easily scalable**

Why Tape?

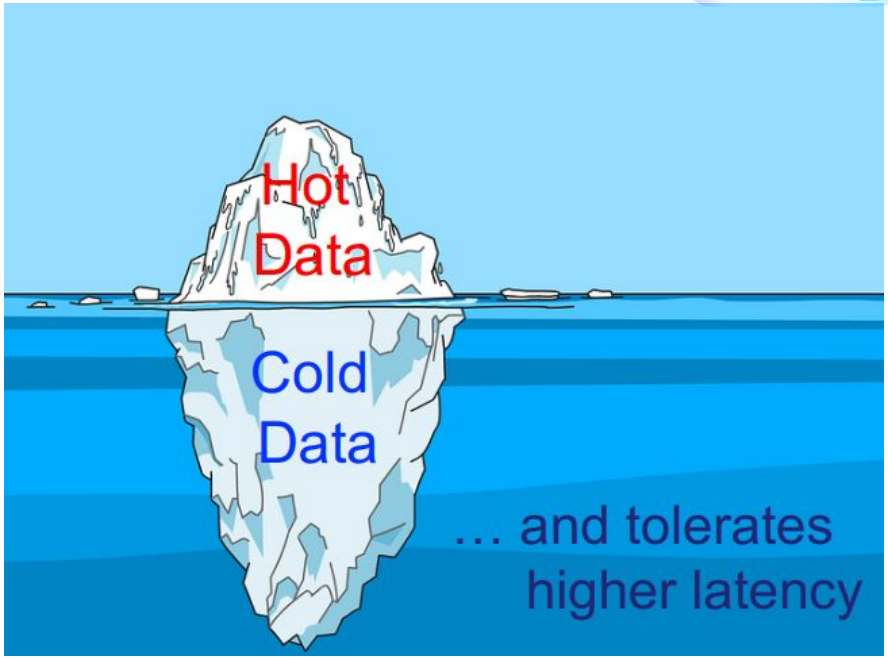
Cost

@ hyperscale
HDD Cost 3.7x Tape

Scaling



INSIC Tape Capacity
Scaling 34% CAGR to 2029



Storage Resources

Management of storage and access to experiment data:

- "Hot" data, on **disk**, accessible via distributed file systems **GPFS** (General Parallel File System) and **CEPH**
  - **LOCAL** (POSIX): via user-interface
  - **GRID**: via **protocols** requiring **authN** and **authZ** based either on **JWT** or **certificates**:
    - **StoRM** - srm
    - **StoRM WebDAV** - https/davs
    - XrootD - root
    - ~~GridFTP - gsiftp~~ (dismissed)

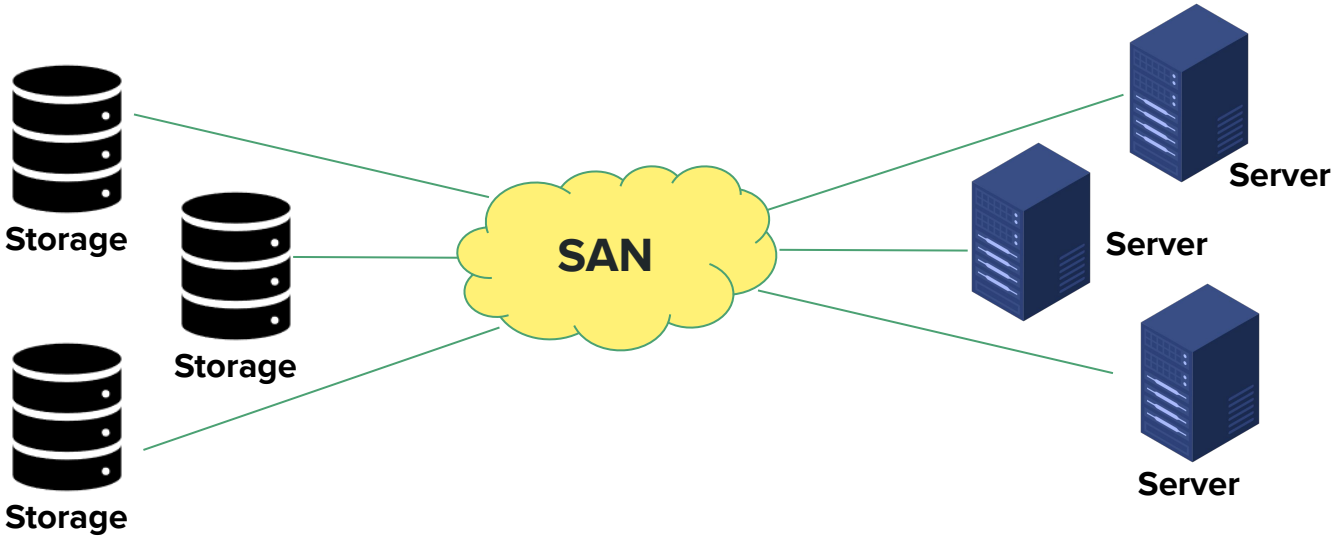
  Developed at CNAF
- "Cold" data, on **tape**, available via a complex software stack

IBM Spectrum Scale (GPFS)

ceph

Storage Area Network (SAN)

- A dedicated network interconnecting a shared pool of storage devices to multiple servers, so that the devices appear as a direct attached storage
- Can use different **protocols** to transport data (FC, SCSI, etc...)



Tape Area Network (TAN)

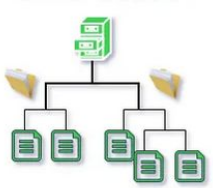
- The part of the SAN dedicated to the interconnection among servers, libraries and tape drives
- Tape drives can be installed in a central array and attached to the SAN, making them accessible to tape servers on the network

Tape libraries @CNAF

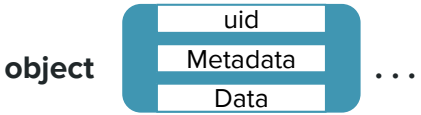
- **2 x IBM TS4500**
  - **1 tape library with 19 tape drives TS1160** (20TB/cartridge)
    - 102 PB installed, ~50 PB used
    - This library has been moved to the new data center
  - **1 tape library with 18 tape drives TS1170** (50TB/cartridge)
    - Acquired and installed at the new datacenter in 2024

File systems

Storage architectures



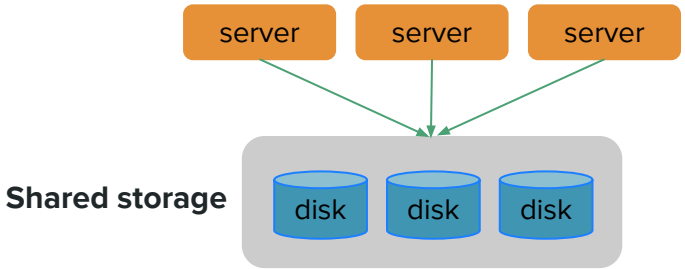
- **File storage**
  - POSIX read/write access
  - File systems which manage data as a **file hierarchy** with directories and subdirectories
  - In case of cluster file system, the worker nodes belong to the fs cluster and see the storage as a local file system. Jobs perform a direct access to the files and directories
- **Object storage**
  - **Flat structure**
  - Each object is a self-contained repository that owns the **data**, a **unique identifier** that allows the object to be found over a distributed system, and the **metadata** that describes the data.
  - More effective for handling large amounts of unstructured data
  - Highly **scalable**



General Parallel File System - GPFS

General Parallel File System / Spectrum Scale

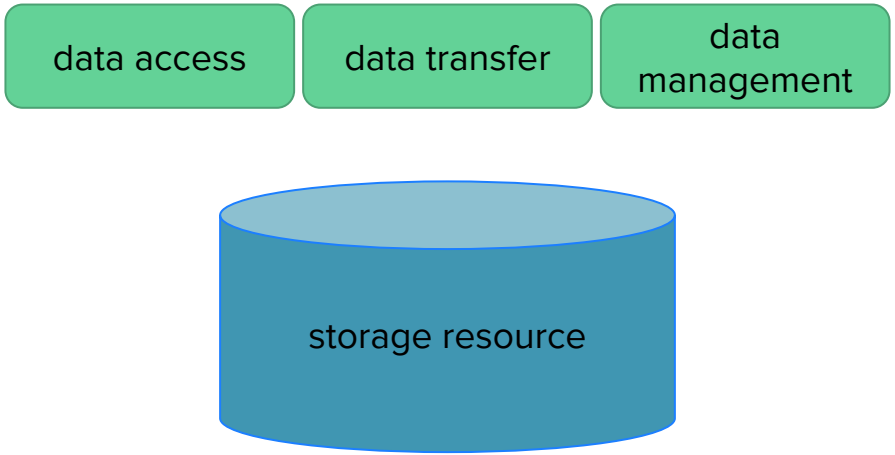
- Software licensed by IBM
- **Cluster** File System: provides **concurrent access** to a single file system or set of file system from multiple nodes
- **Shared storage** structure through collection of multiple disks connected to the cluster nodes
- **Redundant**: GPFS cluster allows failure of up to 50% of the servers



Basic concepts

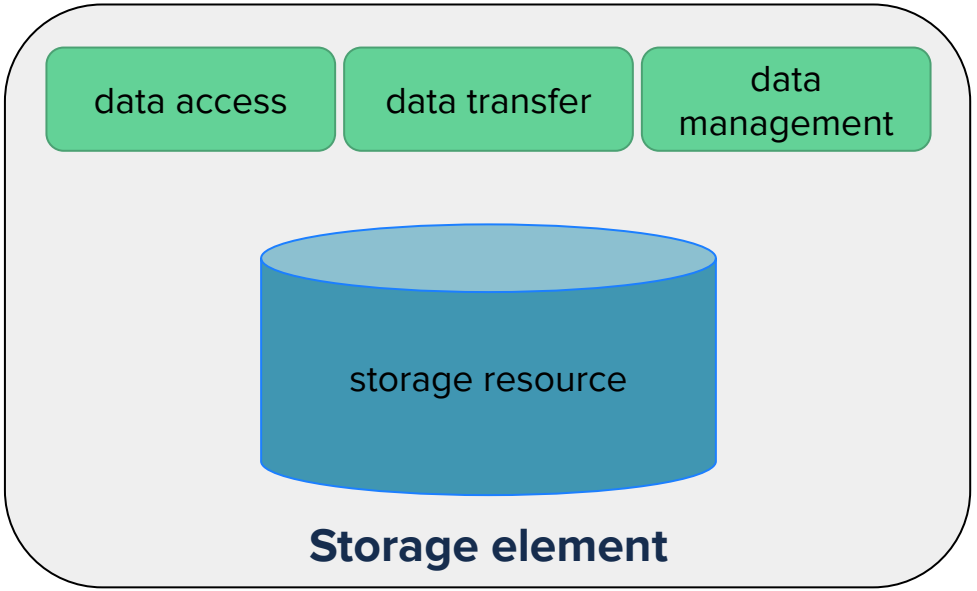
Storage element

- All data are stored on a data storage resource
- The storage resource is wrapped by storage services exposed to users so to make the data available
- Three categories of interfaces exist:
  - **Data access**
  - **Data transfer**
  - **Data management**



Storage element

- All data are stored on a data storage resource
- The storage resource is wrapped by storage services exposed to users so to make the data available
- Three categories of interfaces exist:
  - **Data access**
  - **Data transfer**
  - **Data management**



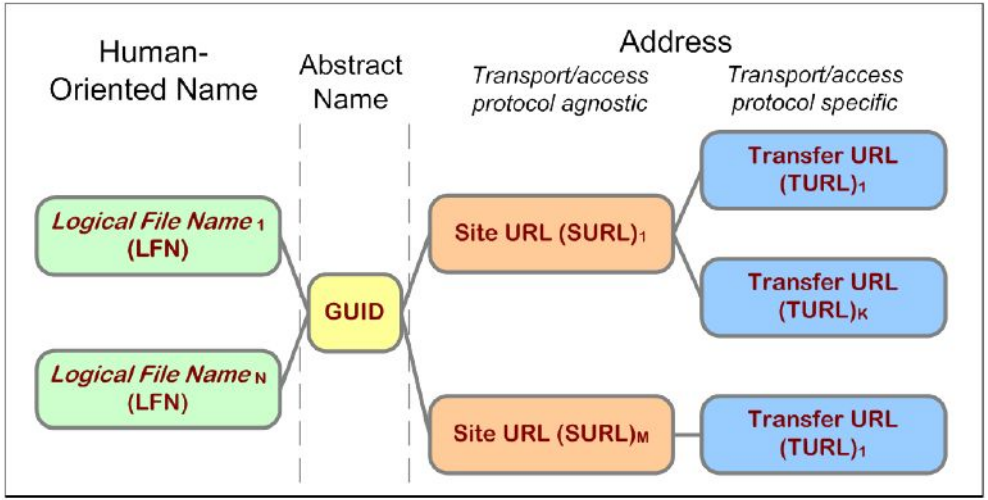
File Names

Files in the Grid can be referred by different names:

- **Logical File Name (LFN)**: an alias created by a user to refer to some data
- **Grid Unique IDentifier (GUID)**: a non-human-readable unique identifier for an item of data
- **Site URL (SURL)**: the location of an actual piece of data on a storage system
- **Transport URL (TURL)**: Temporary locator of a replica + access protocol understood by a SE

File Names

- While the GUIDs and LFNs identify a file irrespective of its location, the SURLs and TURLs contain information about where a physical replica is located, and how it can be accessed



Storage Areas

- Storage Areas (**SA**)
  - Many storage servers, with different transfer protocols
    - **StoRM**, **StoRM WebDAV**, **XrootD**
  - Different types of **authN/authZ**
    - X509 certificates - **VOMS proxy**
    - OIDC - **JWT**
  - A complete and exhaustive list of SA available @CNAF can be found at: https://www.cnaf.infn.it/~usersupport/

StoRM WebDAV storage areas with JWT authentication

aa.wp6

StoRM WebDAV endpoint	Access point	Root path
xfer.cr.cnaf.infn.it	/DataCloud-TB	/storage/gpfs_escape/datacloud-tb

belle

StoRM WebDAV endpoint	Access point	Root path
xfer-archive.cr.cnaf.infn.it	/belle	/storage/gpfs_data/belle

cta-lst

StoRM WebDAV endpoint	Access point	Root path
xfer-archive.cr.cnaf.infn.it	/cta-lst	/storage/gpfs_data/ctadisk/cta-lst

Data transfer and data management

Data transfer tools

- Common file transfer tools between hosts: FTP, scp/sftp, rsync
  - Great compatibility, widely available, convenient and familiar to many users
- These tools work fine in a local environment, for small and quick, not so-frequent transfers... BUT...
- They are unsuitable for large bulk data transfers and unreliable connections and hosts
- Authentication/authorization mechanisms are difficult to integrate in a distributed environment for large communities
- They perform poorly on a WAN, not making efficient use of available bandwidth for wide area data movement

GridFTP - a data transfer protocol

- Extension of the **File Transfer Protocol** (FTP) for **grid** computing
- GridFTP and the corresponding Grid Security Infrastructure (**GSI**)-based authentication and authorization system have been data transfer pillars of WLCG for many years
- The Globus Alliance supplies both the protocol and a reference implementation for server and client (**globus-url-copy**)

BUT...

- In 2017, Globus announced the **retirement** of its open source Globus Toolkit, which provides the reference implementation for the GridFTP protocol
- Work ongoing in DOMA TPC WG to **phase out the GridFTP** protocol in favor of alternative approaches such as **HTTP**

XRootD: exTended Root Daemon

- High-performance, fault-tolerant, and secure solution for handling massive amounts of data distributed across multiple storage resources
- Originally developed at SLAC for BaBar experiment and later extended to meet the needs of the LHC experiments at CERN
- **Scalable** to hundreds of servers
- X509 VOMS auth/authz

Data transfer with XRootD

- **Listing directory**:

```
-bash-4.2$ xrdfs root://xrootd-ams.cr.cnaf.infn.it:8082// ls /
```

- **Download:**

```
-bash-4.2$ xrdcp root://xrootd-ams.cr.cnaf.infn.it:8082//test_1906 copia_locale
```

- **Upload:**

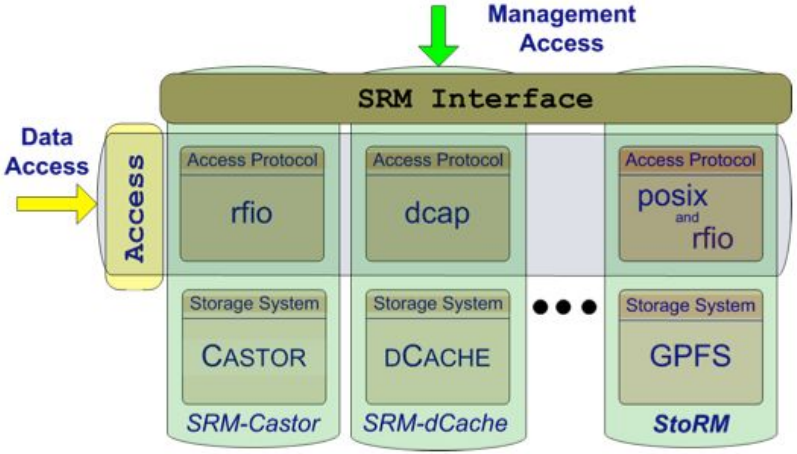
```
-bash-4.2$ xrdcp copia_locale root://xrootd-ams.cr.cnaf.infn.it:8082//test_0809
```

- **Removing a file:**

```
-bash-4.2$ xrdfs root://xrootd-ams.cr.cnaf.infn.it:8083// rm /test_0809
```

SRM (Storage Resource Manager)

- A **Storage Resource**: a storage system in the Grid. The basic logical entities of a storage resource are space and file.
- **Storage Resource Managers** (SRMs): middleware services that provide dynamic space allocation and file management of shared storage components.
- SRMs services agree on a **standard interface** to hide storage dependent characteristics and to allow interoperability between heterogeneous resources.
- The SRM solution developed and used @CNAF is **StoRM**.



N.B.: SRMs do not perform file transfers, but can invoke middleware components that perform this job

GFAL - Grid File Access Library

- A simple set of generic and useful **command line tools** to perform data transfers operations
- **POSIX-like** API
  - gfal-ls, gfal-mkdir, gfal-copy, gfal-rm, gfal-sum...
- Work with ANY protocol for both metadata operations and remote I/O: file://, gsiftp://, srm://, http://, root://...
- Supports *Third Party Copies* (**TPC**)

GFAL in practice

- **Listing directory:**

```
-bash-4.2$ gfal-ls srm://storm-test.cr.cnafinfn.it:8444/demo
```

- **Download:**

```
-bash-4.2$ gfal-copy srm://storm-test.cr.cnaf.infn.it:8444/demo/myfile.txt myfile_local.txt
```

- **Upload:**

```
-bash-4.2$ gfal-copy myfile_local.txt srm://storm-test.cr.cnaf.infn.it:8444/demo/myfile.txt
```

- **Removing a file:**

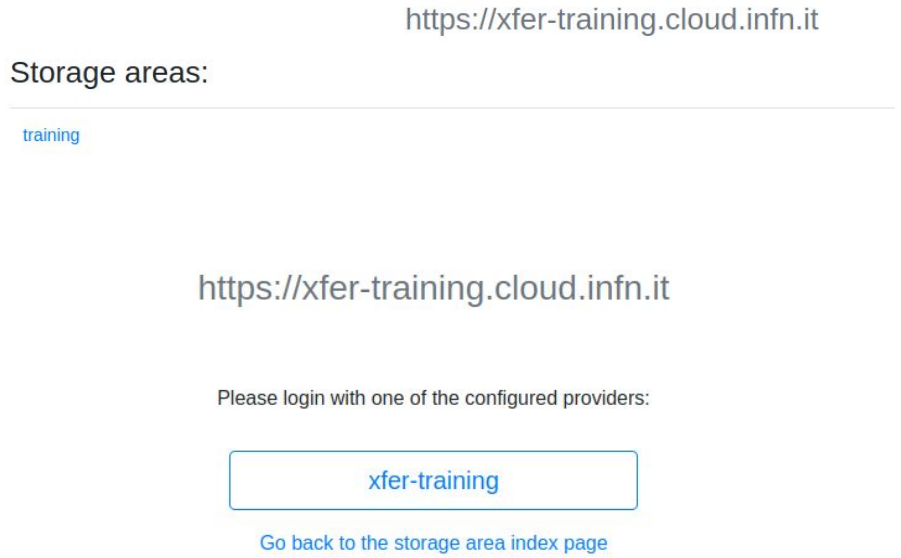
```
-bash-4.2$ gfal-rm srm://storm-test.cr.cnaf.infn.it:8444/demo/myfile.txt
```

HTTP/WebDAV

- The Hypertext Transfer Protocol (**HTTP**): a well-known and widely adopted standard protocol, underpinning the World Wide Web
- **WebDAV** stands for Web-based Distributed Authoring and Versioning
- **Extension of the http** protocol which allows users to collaboratively edit and manage files on remote web servers
- It supports file sharing, editing, and versioning directly through a web interface

WebDAV Clients

- The most common WebDAV clients to navigate into the storage area content are **web browsers**.
  - https://xfer-training.cloud.infn.it:8443/
- **HTTP endpoint**: anonymous read-only access enabled
- **HTTPS endpoint**: users access through their X509 certificate is enabled

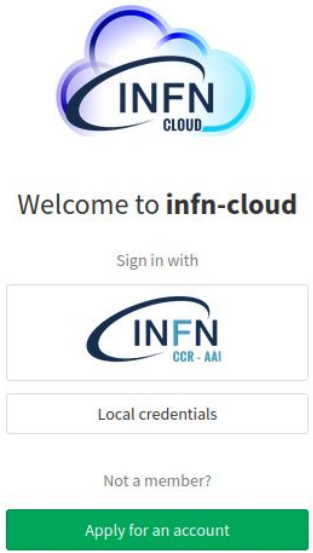


StoRM WebDAV

- **StoRM WebDAV** is the StoRM service that provides valid WebDAV endpoints for the experiments' storage areas
- AuthN/Authz possible both with:
  - **VOMS proxies**
  - **OIDC tokens** with **group-based authorization**
- Widely used by many communities
- **StoRM WebDAV** + **token authN/authZ** suggested when onboarding new experiments at Tier-1

StoRM WebDAV with tokens

- StoRM WebDAV supports **OpenID connect** authentication and group-based authorization on storage areas.
- Multiple **IAM (Identity and Access Management)** instances can be configured; once registered within IAM, an access token can be retrieved with **oidc-agent**



StoRM WebDAV with tokens

- StoRM WebDAV supports **OpenID connect** authentication and group-based authorization on storage areas.
- Multiple **IAM (Identity and Access Management)** instances can be configured; once registered within IAM, an access token can be retrieved with **oidc-agent**


```
{
  "sub": "b17691a1-b6a1-4aa0-a4b8-43bc6b94a65f",
  "iss": "https://iam-t1-computing.cloud.cnaf.infn.it/",
  "groups": [
    "eupraxia",
    "dampe",
    "litebird",
    "Cygno",
    "foot",
    "newchim",
    "user-support",
    "luna",
    "km3net",
    "muoncoll",
    "gamma",
    "borexino"
  ],
  "preferred_username": "dlattanzio",
  "organisation_name": "t1-computing",
  "client_id": "05e11c51-16fa-4b53-a960-dc1d3fa1e846",
  "scope": "address phone openid offline_access profile
  "name": "Daniele Lattanzio",
  "exp": 1721119262,
  "iat": 1721115662,
  "jti": "cfbc91da-7d44-4753-b2c3-4727bc428d64",
  "email": "daniele.lattanzio@cnaf.infn.it"
}
```

StoRM WebDAV with tokens

- An access token can be retrieved using oidc-agent:
  - **start** the oidc-agent service
  - **register a client** (only the first time)
  - **load the client** configuration
  - **get an access token** and store it in a environment variable

```
[dlattanzio@ui-tier1 ~]$ eval `oidc-agent-service use`
16626
[dlattanzio@ui-tier1 ~]$
[dlattanzio@ui-tier1 ~]$ oidc-add t1-computing
Enter decryption password for account config 't1-computing':
success
[dlattanzio@ui-tier1 ~]$
[dlattanzio@ui-tier1 ~]$ export BEARER_TOKEN=$(oidc-token t1-computing)
[dlattanzio@ui-tier1 ~]$
```

Data management with StoRM WebDAV

- **Listing directory:**

```
gfal-ls https://xfer-training.cloud.infn.it:8443/training
```

- **Download:**

```
gfal-copy https://xfer-training.cloud.infn.it:8443/training/file.txt file_local.txt
```

- **Upload:**

```
gfal-copy file_local.txt https://xfer-training.cloud.infn.it:8443/training/file.txt
```

- **Removing a file:**

```
gfal-rm https://xfer-training.cloud.infn.it:8443/training/file.txt
```

Tape data management

Tape data management

- Data on tape need to be copied on a **disk buffer** to be accessed
- The buffer is a disk (detached and generally different from the actual disk) that serves as a temporary platform for files that must be **migrated** or have been **recalled** from tape
- This is not a static disk but once it is full, the oldest and already migrated files are deleted by the **garbage collector**
- **Migration**: moving a file from disk buffer to tape
- **Recall**: moving a file from tape to disk buffer

Tape data management

- **Migrate files to tape**: when a file has to be moved to tape, one needs to put it into the buffer disk. From there, data will be automatically migrated to tape after a certain time
- **Recall files from tape**: to recall files from tape **using VO**, you can use “clientSRM bol” (**Bring On Line**) command
- For details on Tape usage @T1: https://confluence.infn.it/display/TD/Tape

```
clientSRM ls -l -v NIG -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444/ -s srm://storm-fe-archive.cr.cnaf.infn.it:8444/ams/${your_file}
```

Based on the information shown in the output, it is possible to locate the file by checking the value of the *fileLocality* line:

`-[0] fileLocality=0`	the file is on disk
`-[0] fileLocality=1`	the file is on tape
`-[0] fileLocality=2`	the file is both on disk and tape

Where is my file?

StoRM Tape REST API

- Data stored on tape can also be recalled with **HTTP protocol**, thanks to the development of a common **HTTP REST interface** within the WLCG community in support of the transition towards srm-less recalls
- Integrated with **GEMSS** (Grid-Enabled Mass Storage System)
- It supports authentication mechanisms based on VOMS proxies and tokens
- Developed at CNAF

Garbage collector

- GEMSS triggers a **periodic scan** of GPFS file system
- New files are migrated to tape through TSM (Tivoli Storage Manager)
- When the file system occupancy reaches a configured high threshold, the GPFS garbage collector starts to remove files from disk buffer
- Deletion ends when the file system occupancy reaches a configured low threshold

References

1. **Tier-1 User-GUIDE**: https://confluence.infn.it/display/TD
2. **StoRM-Webdav with tokens**: https://confluence.infn.it/display/TD/Data+transfers+using+http+endpoints#Datatransfersusinghttpendpoints-TokensWebDAV
3. Monitoring Tier-1 https://t1metria.cr.cnaf.infn.it/
   a. File systems quota and occupancy https://www.cnaf.infn.it/~vladimir/gpfs
4. StoRM https://italiangrid.github.io/storm/documentation/functional-description/1.11.2/
5. Indigo IAM https://github.com/indigo-iam/iam
6. OIDC-agent https://github.com/indigo-dc/oidc-agent
7. gfal2-utils https://github.com/cern-fts/gfal2-util

Thanks for the attention