

StoRM Tape status and plans

Enrico Vianello, Daniele Cesini
INFN-CNAF

Pre-GDB, November 7th, 2023
enrico.vianello@cnafe.infn.it, daniele.cesini@cnafe.infn.it,

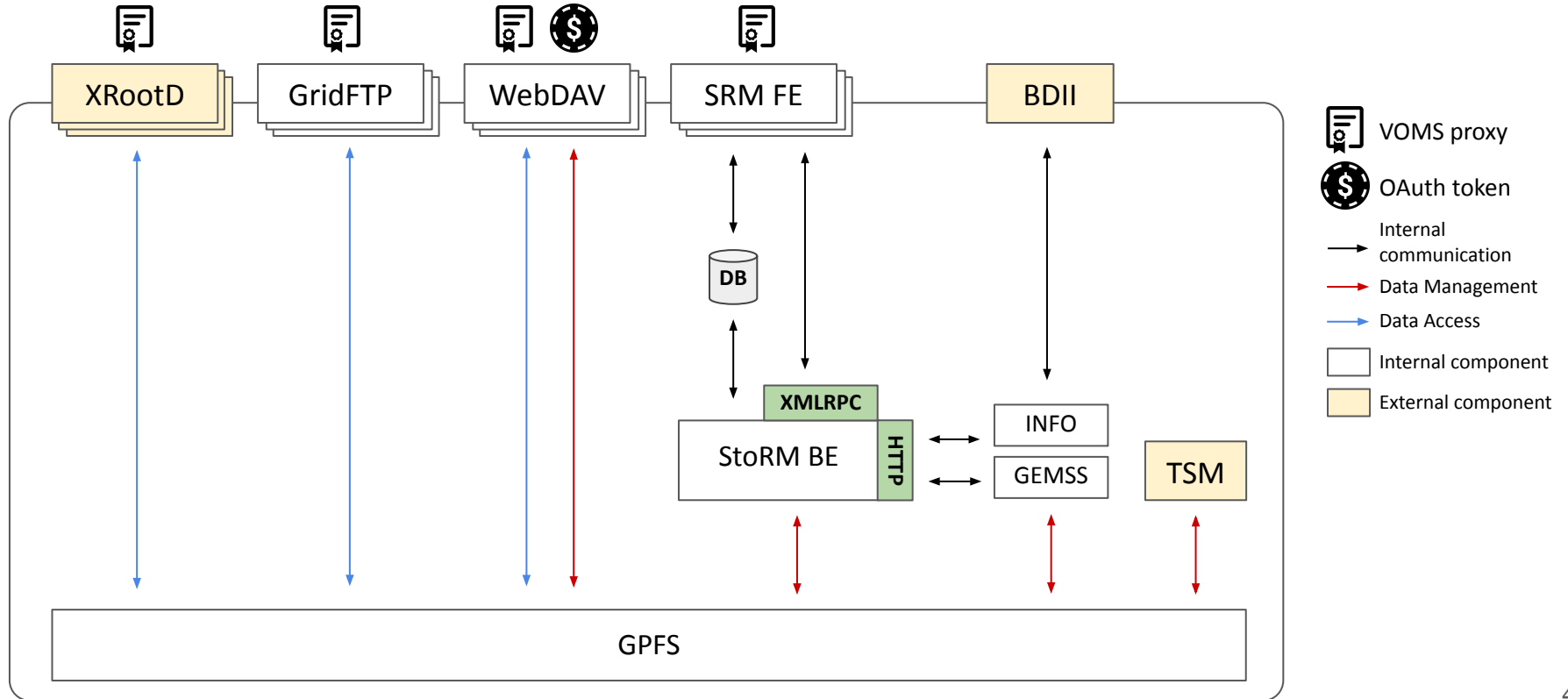


Topics

- StoRM Tape basics
 - GEMSS component
 - Current data life cycle within a tape-enabled storage area
- StoRM Tape REST API
 - The WLCG Tape REST API specification
 - NGINX and OPA deployment roles
 - OPA authorization example
 - Testing tools
 - Ongoing developments
- INFN-CNAF T1 Tape status
 - Tape acquisition
 - Tape libraries migration to the new datacenter

StoRM Tape basics

StoRM: a typical deployment architecture



GEMSS component

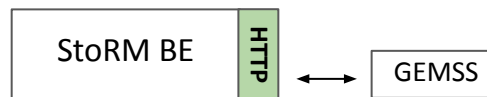
GEMSS is a full Hierarchical Storage Management (HSM) system, integrating:

- IBM General Parallel File System (**GPFS**)
- IBM Tivoli Storage Manager (**TSM**)
- **StoRM Backend**

StoRM Backend is not able to directly recall a file from tape: this operation is managed by GEMSS.

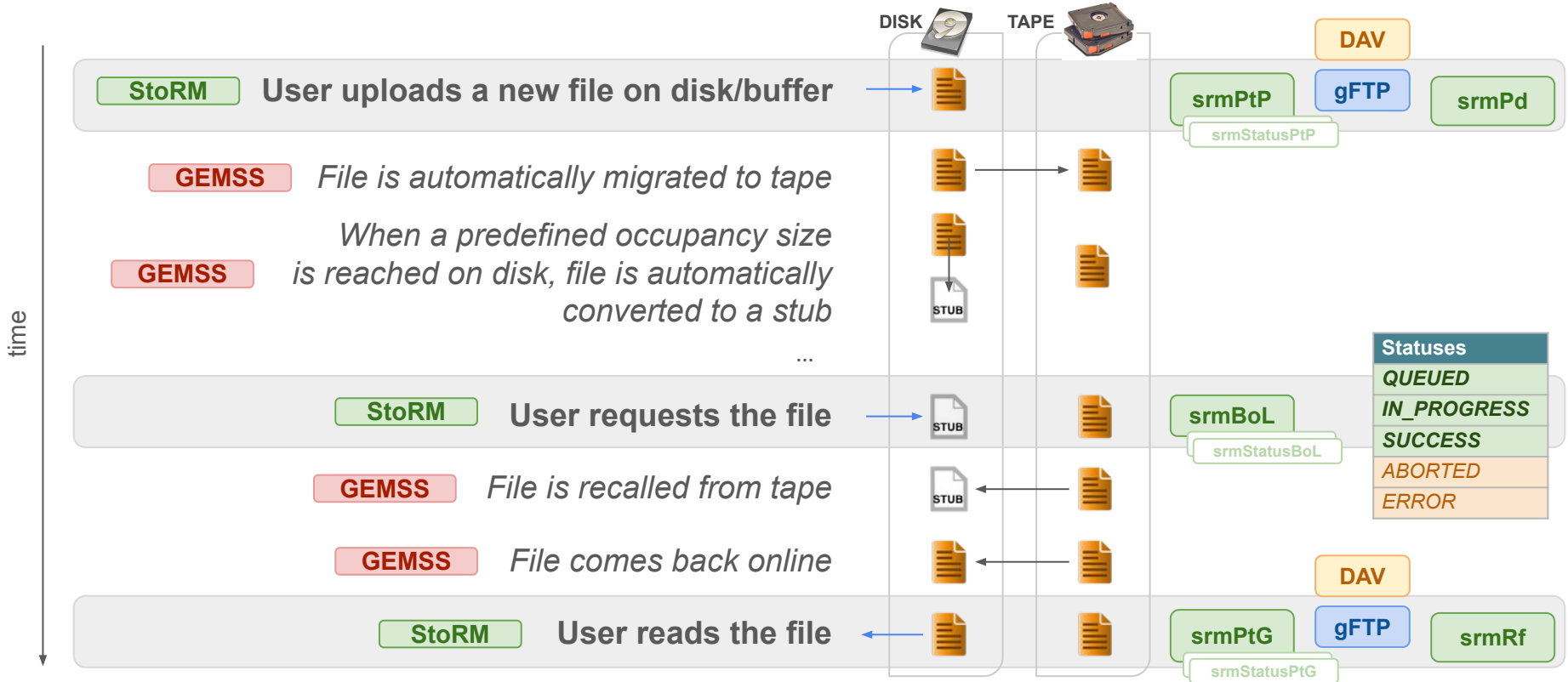
GEMSS advantages:

- high level of reliability
- minimum management effort needed for daily maintenance



GEMSS queries StoRM Backend in order to retrieve all the file recall requests.

Life-cycle of a file hosted on a tape-enabled storage area



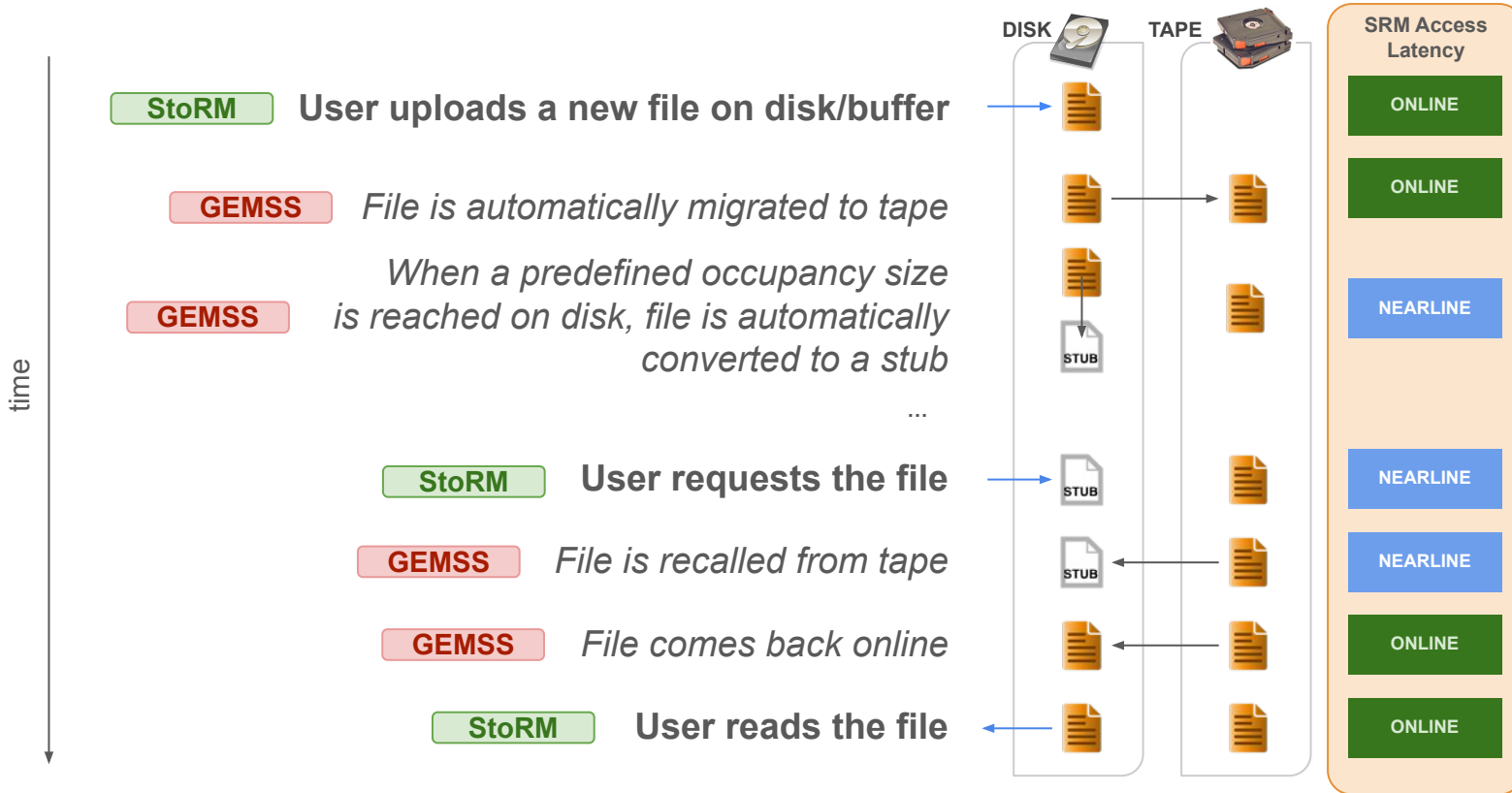
Data access latency

Data access latency can be described as the amount of time necessary for data to become available.

SRM latency values:

- **Online**
 - The lowest possible latency
 - *Online* files can be read/transferred by clients
- **Nearline**
 - Can be improved to a lower latency by staging the file on a disk buffer
 - Recalling a file from a tape storage means, in SRM terms, changing its latency from *Nearline* to *Online*.

Life-cycle of a file hosted on a tape-enabled storage area



Data Life Cycle - Stubbification

When a predefined occupancy size is reached on disk, files are automatically converted to stubs

stub → for StoRM this is a file which has been migrated/copied to tape and its content has been deleted from local disk buffer. File still exists with its altered metadata: **same size but zero block size**.

```
$ [root@storm-test ~]# ls -ls /storage/gemss_test1/tape/testfile
0 -rw-r--r-- 1 storm 5059 5242880 14 gen 2021
/storage/gemss_test1/tape/testfile
```

blocks=0

size=5242880

is $\text{blocks} * \text{BLOCK_SIZE}$ less than size ? If yes it's a stub.

Rely on File-System

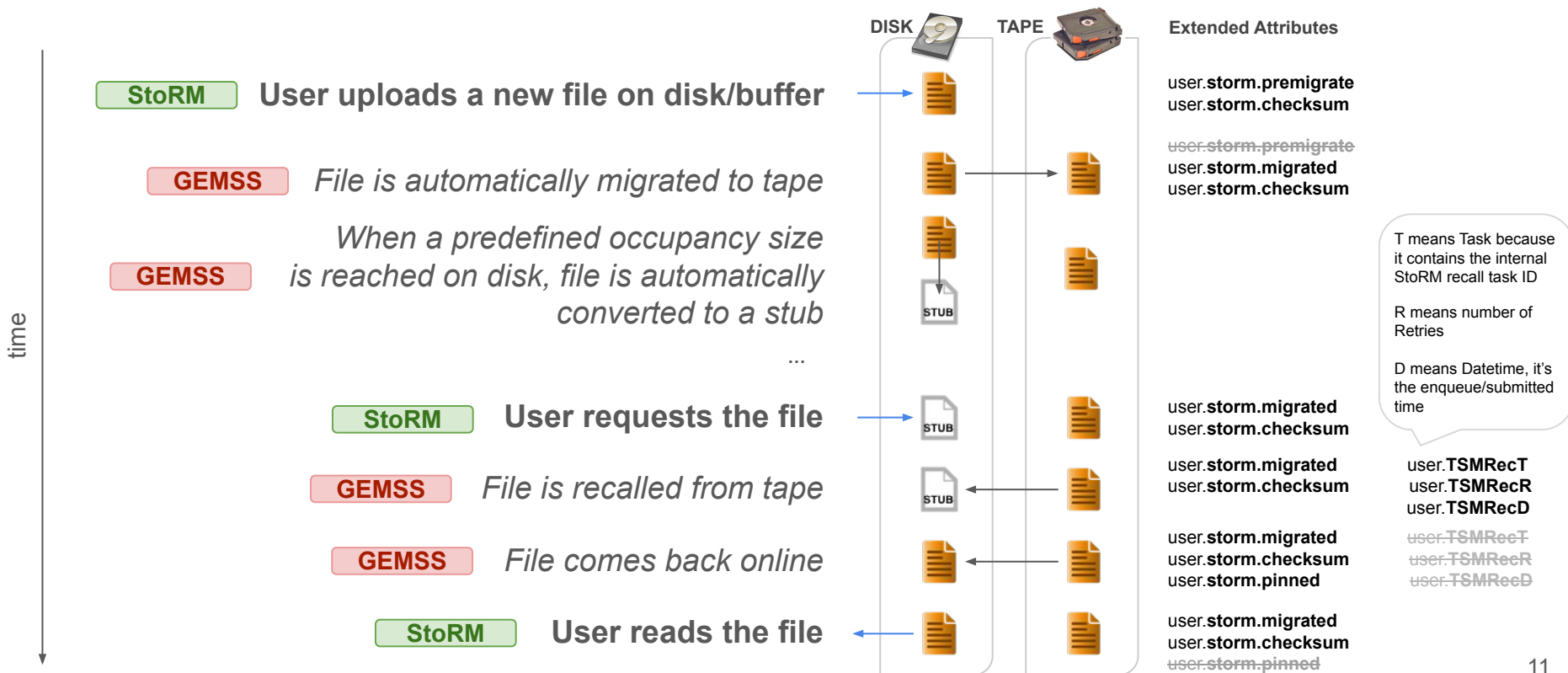
At StoRM side, all the information related to file latency or recall progress are computed from the underlying File-System, e.g.:

- file/directory exists
- permissions through ACLs
- file is available to be read (online) or is only on tape, etc.

StoRM uses the extended attributes mechanism to store extra metadata:

- checksum value
- file locality
- ongoing recall from tape: request id, num. retries and submitted timestamp

Life-cycle of a file hosted on a tape-enabled storage area



StoRM Tape REST API

StoRM Tape REST API - Introduction

- The StoRM implementation of [WLCG tape REST API specification](#)
 - a common HTTP interface, defined within a collaboration between different WLCG storage providers (StoRM, dCache, EOS+CTA) and clients (FTS)
- Supports bulk-request of tape-stored files
 - track progress of a previously staged bulk-request;
 - cancel a previously staged file replicas from disk;
 - retrieve information about the progress of file's staging.
- It can coexist with a typical SRM deployment (srmBringOnline commands)
- The API will be accessed via authentication mechanisms like X509 + VOMS (proxy-based) or token based (JWT).

WLCG Tape REST API specification

STAGE Requests that tape-stored files are made available on disk

Generated with Swagger, source [here](#)



POST /api/v1/stage Bulk-request of files to be transferred from tape to disk



GET /api/v1/stage/{id} Track progression of a previously staged bulk-request



DELETE /api/v1/stage/{id} Deletion of a previously submitted STAGE bulk-request



POST /api/v1/stage/{id}/cancel Indicates that the targeted subset of files are no longer needed



RELEASE Indicate that previously staged files through STAGE are no longer required on disk



POST /api/v1/release/{id}



ARCHIVEINFO Requests information about file locality



POST /api/v1/archiveinfo



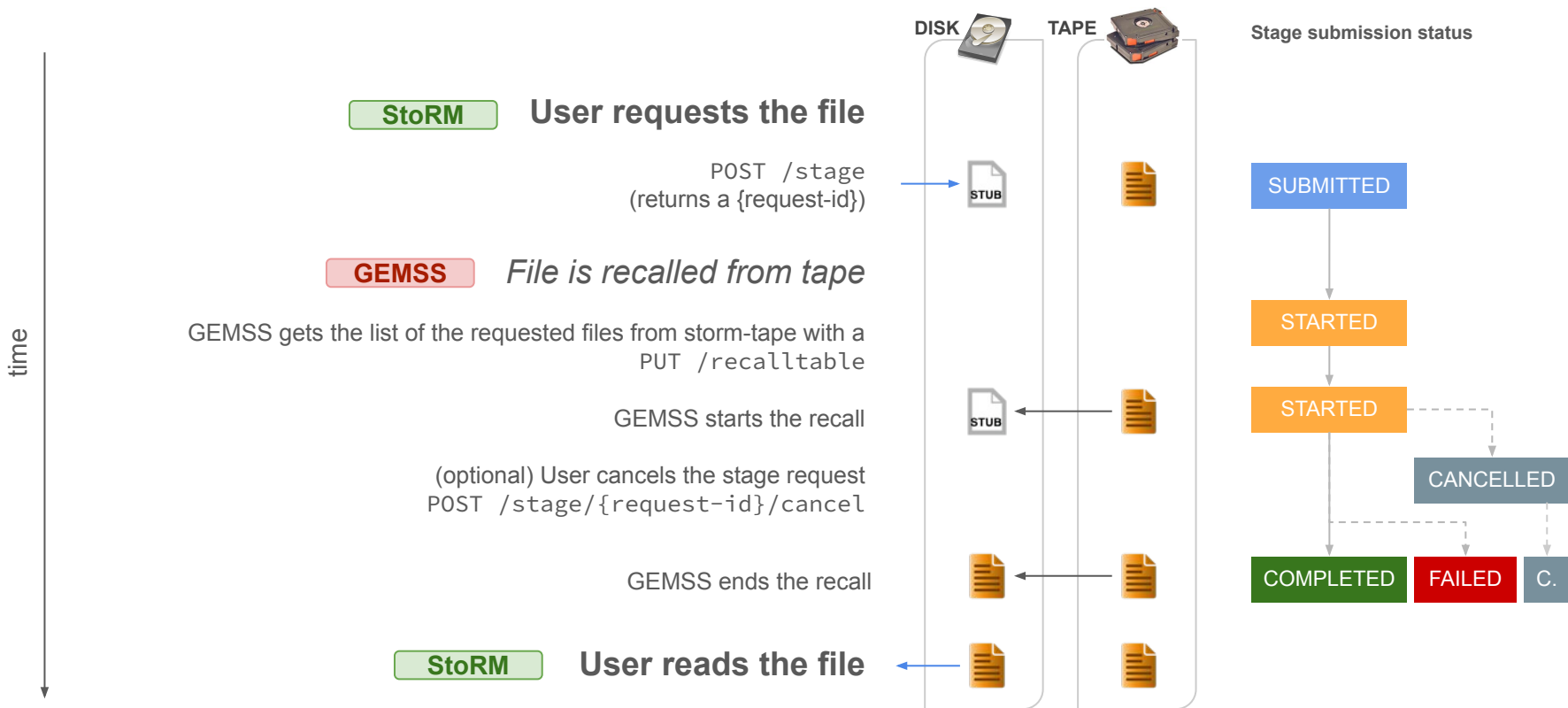
StoRM Tape REST API - Development details

- The [StoRM Tape REST API](#) service (storm-tape) is written in C++, based on the [Crow](#) framework and uses [SOCL](#) library as abstraction layer over the [SQLite](#) database engine
- The service checks file locality directly from the underlying storage system (GPFS). Information on the files are handled using extended attributes:
 - `user.storm.migrated`
 - `user.TSMReCT`
- It provides an additional endpoint for GEMSS to replicate the current interaction with StoRM
 - GET `https://<storm-tape-host>/recalltable/cardinality/tasks/readyTakeOver`
 - PUT `https://<storm-tape-host>/recalltable/tasks`
- Packaged as a [Docker image](#) or as RPM
- AuthN/Z is handled by external services (see later)

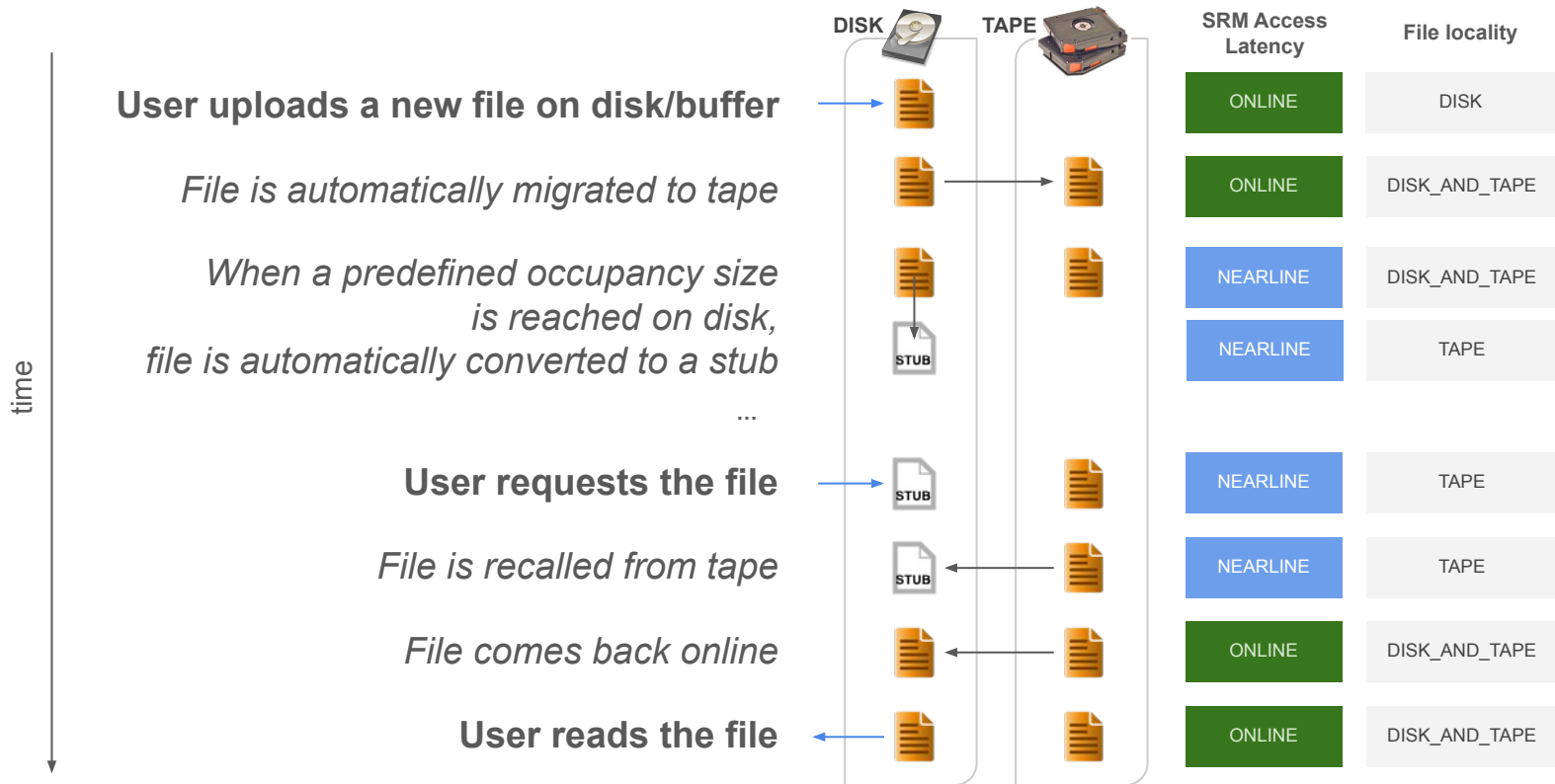
StoRM Tape REST API at CNAF-T1

- Currently v.0.4.0 *beta* version is available in production at INFN-CNAF
 - Installed via RPM and deployed as a standalone component
 - Deployed three separated instances, one for each WLCG experiment
 - <https://tape-atlas.cr.cnaf.infn.it:8443>
 - <https://tape-cms.cr.cnaf.infn.it:8443>
 - <https://tape-lhcb.cr.cnaf.infn.it:8443>
 - Direct access from remote users
 - not deployed within StoRM WebDAV
 - Configured to authorize users via VOMS proxies (ATLAS, CMS, LHCb)
 - ATLAS and CMS authZ via WLCG tokens is also available
- not used yet ⇒ ready for tests and production

Stage submission statuses



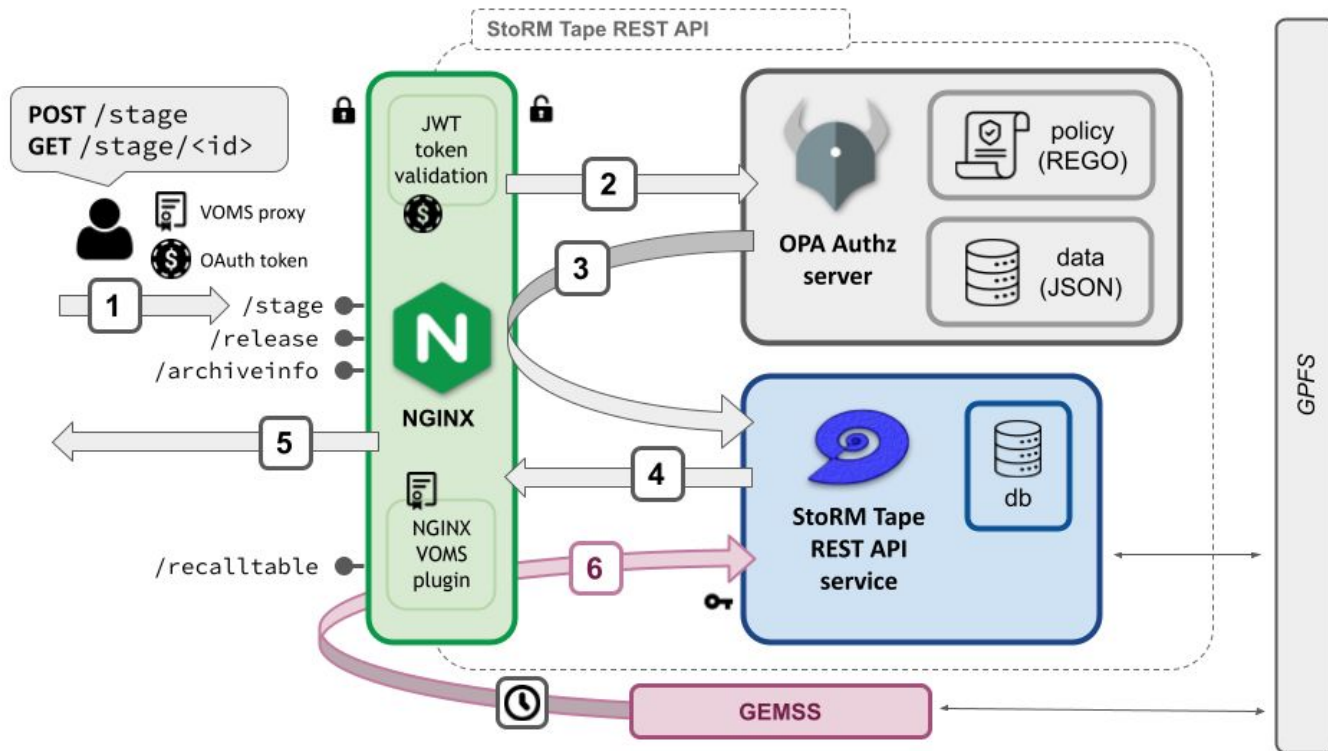
SRM access latency and Tape REST API file locality



StoRM Tape REST API: deployment

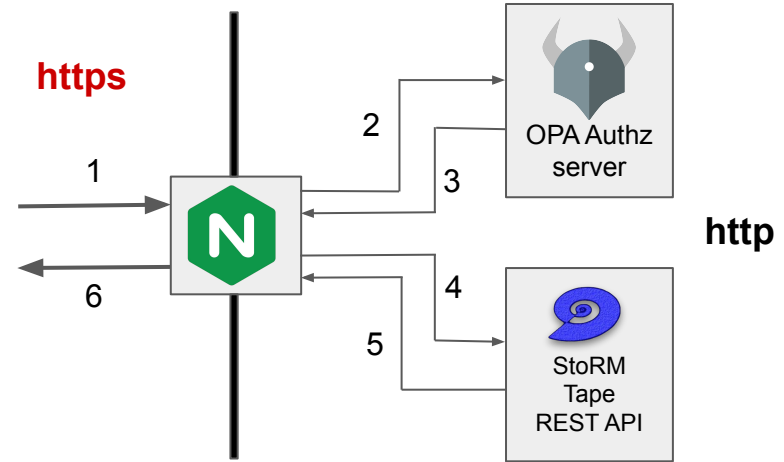
The StoRM Tape REST API relies on external components for authN/Z

- NGINX
 - authentication
- Open Policy Agent
 - authorization



NGINX + OPA Authorization flow

1. The user submits an API request, which is VOMS/TLS terminated by NGINX
2. NGINX sends the request to the Open Policy Agent (OPA) engine
3. OPA makes the authZ decision using its rules and data and sends it back to NGINX
 - o in case of negative authZ, 403 is returned
4. In case of successful authZ, the request is forwarded to the StoRM Tape REST API service
5. (and 6.) The response from the service is relayed to the client via NGINX





NGINX deployment role

- [NGINX](#) is an open-source HTTP server and reverse proxy, known for:
 - high performance
 - high stability
 - rich feature set
 - simple configuration
 - low resource consumption
- NGINX has been chosen as part of this deployment for
 - **TLS termination**
 - Authentication with **JWT**
 - Authentication with **VOMS/X509**



AuthN with NGINX

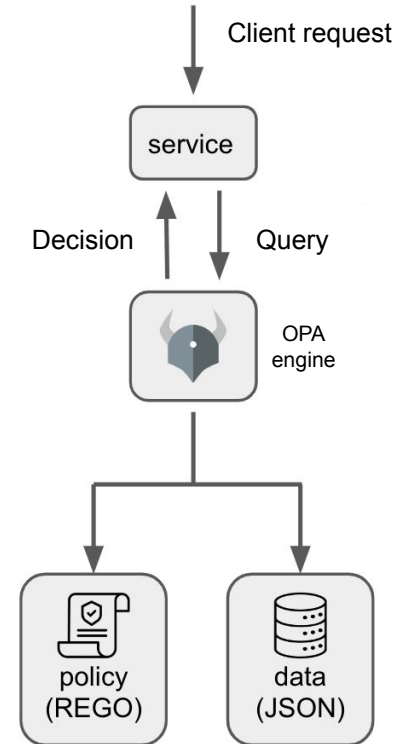
- [ngx_http_voms_module](#) is an NGINX module (developed at CNAF) which:
 - enables client-side authentication based on X.509 proxy certificates
 - defines a set of embedded variables whose values are extracted from the Attribute Certificate, e.g. the **voms_fqans**
- A custom NJS script (developed at CNAF) is used to:
 - check the presence of a JWT in the HTTP Header and, in case, validate it
 - or else, check the presence of X.509/VOMS variables parsed by the VOMS module above (*voms_fqans*, *ssl_client_s_dn*)
 - pass the above data to OPA and handle its response

```
subject      : /DC=org/DC=terena/DC=cn=  
issuer       : /DC=org/DC=terena/DC=cn=  
identity     : /DC=org/DC=terena/DC=cn=  
type         : RFC3820 compliant  
strength    : 2048  
path         : /tmp/x509up_u1000  
timeleft    : 00:59:35  
key usage   : Digital Signature, Encryp  
=== VO wlcg extension informatic  
VO          : wlcg  
subject     : /DC=org/DC=terena/DC=cn=  
issuer      : /DC=org/DC=terena/DC=cn=  
attribute   : /wlcg  
attribute   : /wlcg/mc  
attribute   : /wlcg/pilots  
attribute   : /wlcg/xfers  
timeleft    : 11:59:53  
uri         : wlcg-voms.cloud.cri
```



OPA deployment role

- [Open Policy Agent](#) (OPA) is an open-source authorization engine that:
 - unifies policy enforcement across the stack
 - is based on an high-level declarative language
 - allows the definition of policies as code
- OPA has been chosen as part of this deployment to implement **authorization** policies based on X.509/VOMS proxies or JWT tokens.
- It seems flexible enough to replace other authorization engines
 - e.g. Argus





OPA AuthZ (simple) example

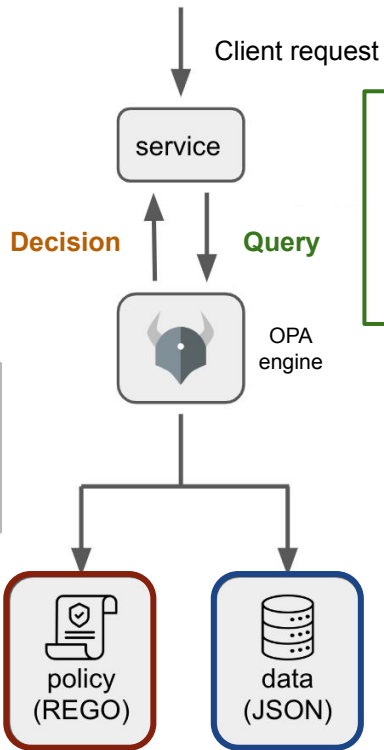
OPA AuthZ example of a stage bulk-request submission done with an allowed JWT (i.e. based on write scopes)

```
{  
  "allow": "true"  
}
```

This policy has been defined for POST requests to the stage endpoint

```
# POST /api/v1/stage  
allow if {  
  input.method == "POST"  
  input.path == "/api/v1/stage"  
  
  true in [  
    write_scopes_allowed,  
    voms_fqans_allowed  
  ]  
}
```

has allowed WLCG scopes? **OR** has allowed FQANs?



```
{  
  "method": "POST",  
  "path": "/api/v1/stage",  
  "access_token":  
  "eyJraWQ1OjJyc2ExIiwiaWwiYXNlIjoiaUlMyNTYifQ..."  
}
```

The JWT contains within its scopes a storage.stage/

```
{  
  "scope": {  
    "read": [  
      "storage.stage/*",  
      "storage.read/*"  
    ],  
    "write": [  
      "storage.stage/*"  
    ]  
  },  
  ...  
}
```

Here we're defining that in order to write or read you need to have one of this lists of allowed scopes



```

1 package server_rules
2
3 import future.keywords.contains
4 import future.keywords.if
5 import future.keywords.in
6
7 default allow := false
8
9 default certificate_dn_allowed := false
10
11 default voms_fqans_allowed := false
12
13 default read_scopes_allowed := false
14
15 default write_scopes_allowed := false
16
17 certificate_dn_allowed if {
18     not input.access_token
19     not input.voms_fqans
20     input.ssl_client_s_dn == data.allowed_dn[_]
21 }
22
23 voms_fqans_allowed if {
24     not input.access_token
25     input.voms_fqans[_] == data.fqan[_]
26 }
27
28 read_scopes_allowed if {
29     not input.voms_fqans
30     allowed_scopes(data.scope.read, jwt_scopes)
31 }
32
33 write_scopes_allowed if {
34     not input.voms_fqans
35     allowed_scopes(data.scope.write, jwt_scopes)
36 }
37
38 allowed_scopes(patterns, scopes) if {
39     glob.match(patterns[_], [], scopes[_])
40 }

```

INPUT

```

1 ▾ {
2     "access_token": "eyJrawQ10iJyc2ExIiwiaWxnbG9kaXkiOiU1MyNTYifQ.eyJ3bG9nLnZlciI6IjEuMCI5InN1YiI6IjBmZDc2YjNjLWMzZjEtNDI4MCI5
3     "method": "GET",
4     "path": "/api/v1/stage/24035c6f-f092-49c1-8401-211681e2c568"
5 }

```

DATA

```

9     /wcy/stage
10 ],
11 ▾ "scope": {
12 ▾ "read": [
13     "storage.stage:/rendina",
14     "storage.read:/*"
15 ],
16 ▾ "write": [
17     "storage.stage:/*"
18 ]
19 }
20 }

```

OUTPUT

Found 1 result in 499µs.

```

1 {
2     "allow": true,
3     "certificate_dn_allowed": false,
4     "jwt_scopes": [
5         "storage.stage:/rendina"
6     ],
7     "read_scopes_allowed": true,
8     "voms_fqans_allowed": false,
9     "write_scopes_allowed": true
10 }

```

LINT

```

1 package server_rules
2
3 import future.keywords.contains
4 import future.keywords.if
5 import future.keywords.in
6
7 default allow := false
8
9 default certificate_dn_allowed := false
10
11 default voms_fqans_allowed := false
12
13 default read_scopes_allowed := false
14
15 default write_scopes_allowed := false
16
17 certificate_dn_allowed if {
18     not input.access_token
19     not input.voms_fqans
20     input.ssl_client_s_dn == data.allowed_dn[_]
21 }
22
23 voms_fqans_allowed if {
24     not input.access_token
25     input.voms_fqans[_] == data.fqan[_]
26 }
27
28 read_scopes_allowed if {
29     not input.voms_fqans
30     allowed_scopes(data.scope.read, jwt_scopes)
31 }
32
33 write_scopes_allowed if {
34     not input.voms_fqans
35     allowed_scopes(data.scope.write, jwt_scopes)
36 }
37
38 allowed_scopes(patterns, scopes) if {
39     glob.match(patterns[_], [], scopes[_])
40 }

```

Share

NEW

<https://play.openpolicyagent.org/p/QrEywa9cUj>

Copy

Install OPA

v0.58.0

[OPA installation docs](#)

Linux

macOS

Windows

```

curl -L -o opa \
https://openpolicyagent.org/downloads/v0.58.0/opa_darwin_amd64; \
chmod 755 ./opa

```

Copy

Run OPA with playground policy

Heads up! The Rego playground is intended for development. Don't rely on it for your production deployments.

```

./opa run --server \
--log-format text \
--set decision_logs.console=true \
--set bundles.play.polling.long_polling_timeout_seconds=45 \
--set services.play.url=https://play.openpolicyagent.org \
--set bundles.play.resource=bundles/f7rdlv6uMo

```

Copy

Query OPA with playground input

Test by piping your playground's JSON input into your OPA served playground policy

```

curl https://play.openpolicyagent.org/v1/input/f7rdlv6uMo \
| curl localhost:8181/v1/data -d @-

```

Copy

StoRM Tape REST API - Development status

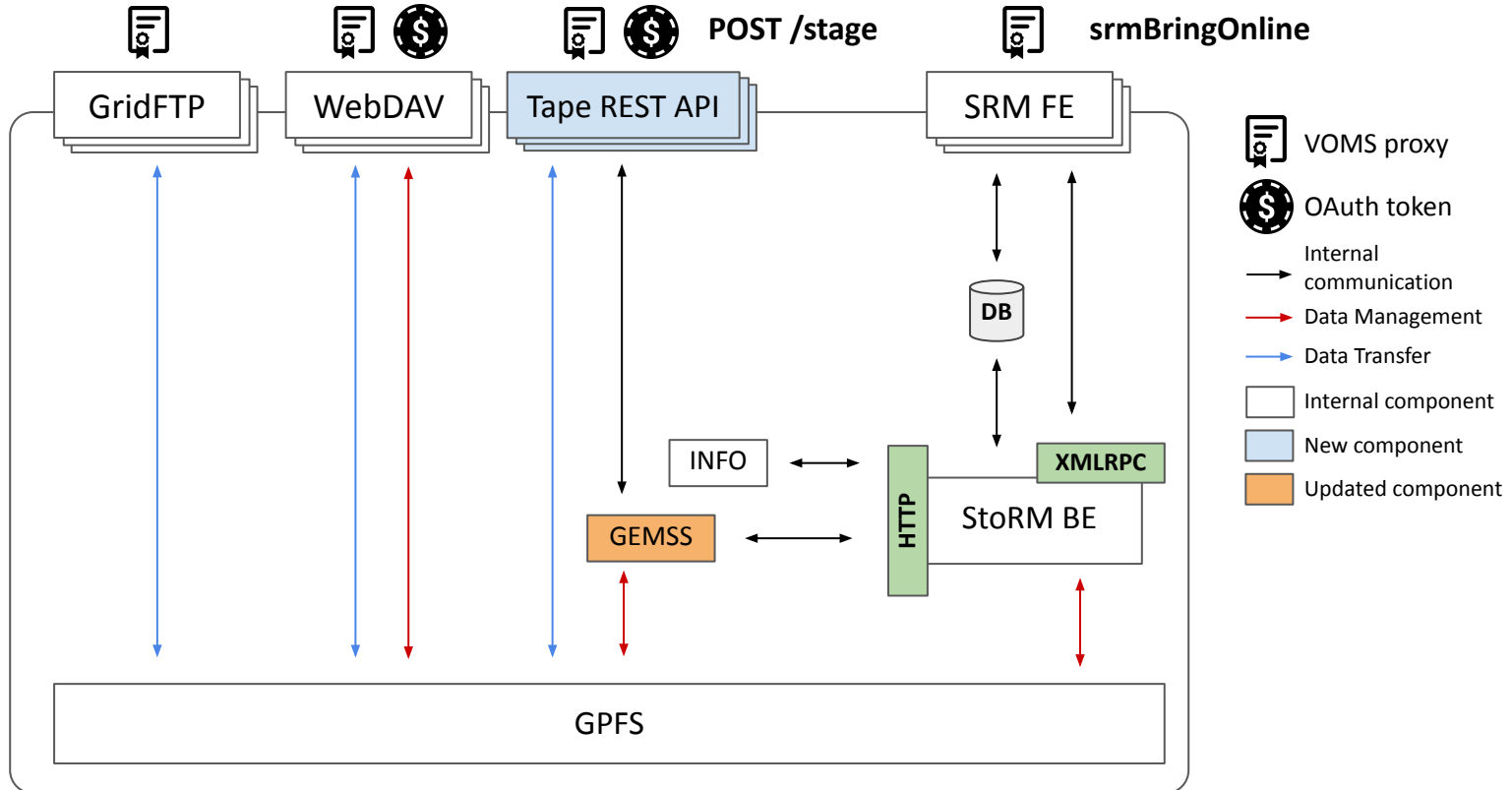
Done:

- Implemented HTTP request/response for all the API endpoints defined in the WLCG specification
- Packed the entire service in a Docker image (also as RPM)
- Implemented the authZ workflow using NGINX and OPA
- Persistency with SQLite
- Interaction with GEMSS for tape recall
 - Replicated the same interaction between GEMSS and StoRM Backend
 - Added an “in-progress” endpoint to return to GEMSS the list of STARTED requests
- Extensions to the WLCG API definition:
 - ARCHIVEINFO endpoint accepts also a request object equal to the stage submission object

In progress:

- Cache JWT validation keys within NGINX
- Proper scope-based authZ
 - the parametric path of storage.* scopes is not yet evaluated at OPA level

StoRM deployment: the Tape REST API scenario



Testing: deployment tests with Robot Framework

Deployment tests use [Robot Framework](#).

Different kinds of functionality are tested:

- authN/Z → **opa**, **token authz** and **voms-authz**
- compliance with the specification → **stage** (including **cancel**), **archiveinfo** and **release**
- integration with GEMSS → **gemss**

Total Statistics	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
All Tests	85	82	3	0	00:02:00	

Statistics by Tag	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
archiveinfo	18	18	0	0	00:00:16	
cancel	15	14	1	0	00:00:20	
extension	6	6	0	0	00:00:04	
gemss	13	12	1	0	00:00:14	
opa	33	33	0	0	00:01:18	
release	13	12	1	0	00:00:18	
stage	50	48	2	0	00:01:24	
token-authz	61	58	3	0	00:00:59	
voms-authz	17	17	0	0	00:00:59	

Statistics by Suite	Total	Pass	Fail	Skip	Elapsed	Pass / Fail / Skip
Test	85	82	3	0	00:02:00	
Test.Archiveinfo	12	12	0	0	00:00:07	
Test.Authorization	33	33	0	0	00:01:18	
Test.Gemss	10	9	1	0	00:00:10	
Test.General	1	1	0	0	00:00:02	
Test.Release	8	7	1	0	00:00:06	
Test.Stage	21	20	1	0	00:00:16	

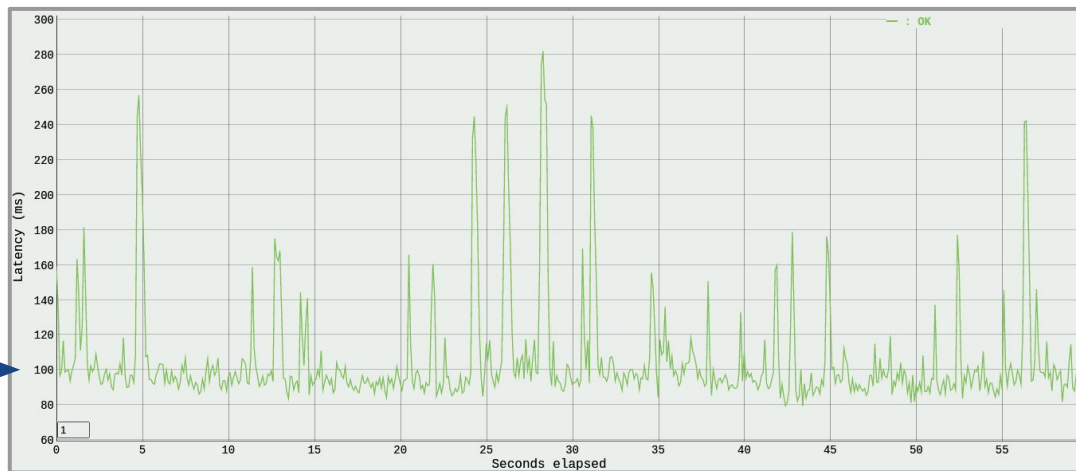
Fixes of failures are in progress!



Testing: load tests with Vegeta

Vegeta is a versatile HTTP load testing tool built out of a need to drill HTTP services with a constant request rate

```
shtimmerman@shtimmerman:~$ echo "POST https://storage-tape-rest.cr.cnaf.infn.it/api/v1/stage" | vegeta attack -duration=40s -rate=20 -body stage.json -insecure -header "Authorization: Bearer $AT" | vegeta report
Requests      [total, rate, throughput]    800, 20.03, 0.00
Duration      [total, attack, wait]        40s, 39.95s, 50.592ms
Latencies     [min, mean, 50, 90, 95, 99, max]  43.07ms, 59.358ms, 53.637ms, 69.843ms, 106.512ms, 141.565ms, 155.793ms
Bytes In      [total, mean]                 0, 0.00
Bytes Out     [total, mean]                 0, 0.00
Success       [ratio]                       0.00%
Status Codes  [code:count]                  0:800
Error Set:
Post "https://storage-tape-rest.cr.cnaf.infn.it/api/v1/stage": EOF
```



Average latency is **~100 ms**.

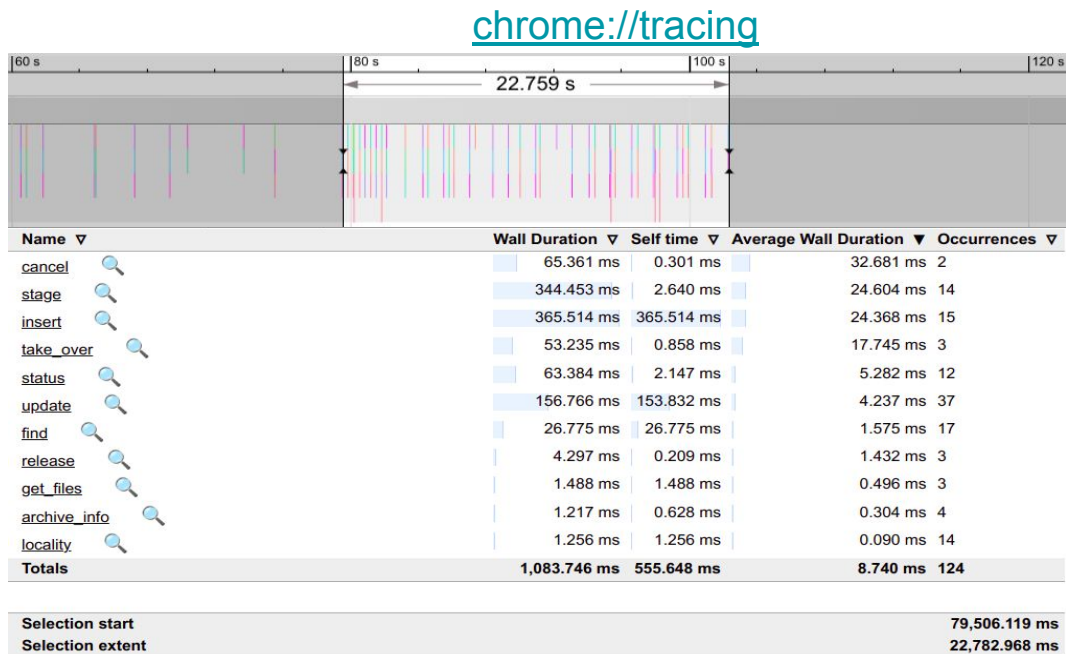
It includes the full interaction with NGINX, OPA and Tape service

Testing: basic profiling

The functions of the StoRM Tape REST API source code have been instrumented to contain tracing informations.

At the end of the service run, a JSON file `results.json` is stored locally.

Average latency is **~10 ms**.







Ongoing developments: integration with StoRM WebDAV

StoRM WebDAV now allows users to use a browser to navigate storage area files.

Ongoing developments have the aim to provide users the ability to trigger a bulk stage request through the folder view ⇒

/test/
[Go to parent directory](#)

Shopping cart icon (1) Cloud icon (2) Search

Name	Last modified	Size (in bytes)	Actions
 bosons.dat	2023-09-21T22:31:34.547+02:00	27240	
 example	2023-09-21T22:31:34.547+02:00		+ Add to queue
 fermions.dat		3560	
 super-asymmetry.dat		3080	+ Add to queue (1)

Showing 1 to 4 of 4 rows

Selected files

Path	Actions
/test/super-asymmetry.dat	x Remove
/test/bosons.dat	x Remove

[x Remove all files](#) [> Submit](#) (3)

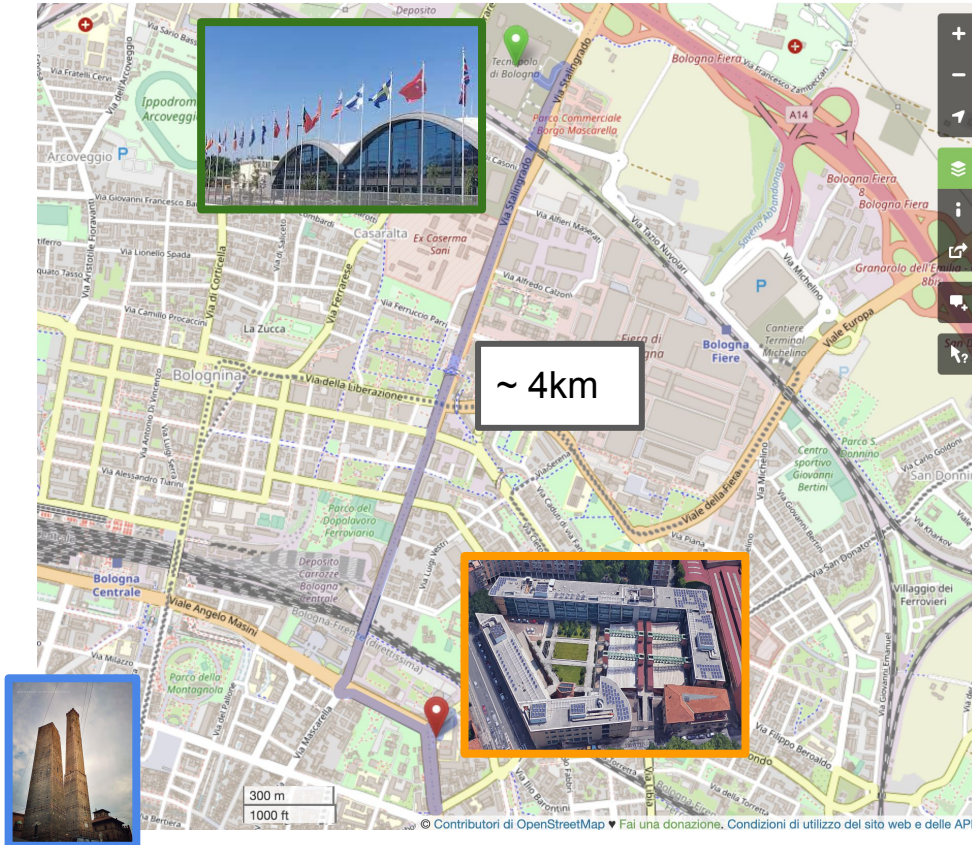
progress

INFN-CNAF T1 Tape status



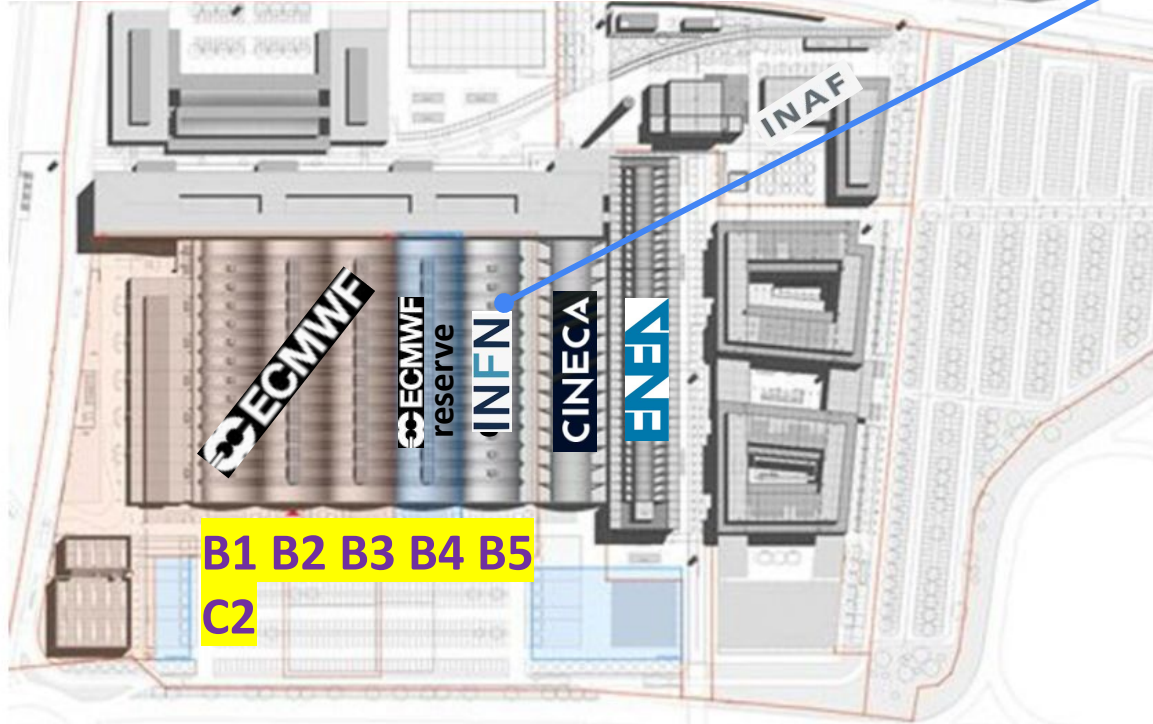
The new INFN Data Center at Bologna Tecnopolo

The new INFN-CNAF Data Center at Bologna Tecnopolo



What can the Tecnopolo host?

The computing infrastructures

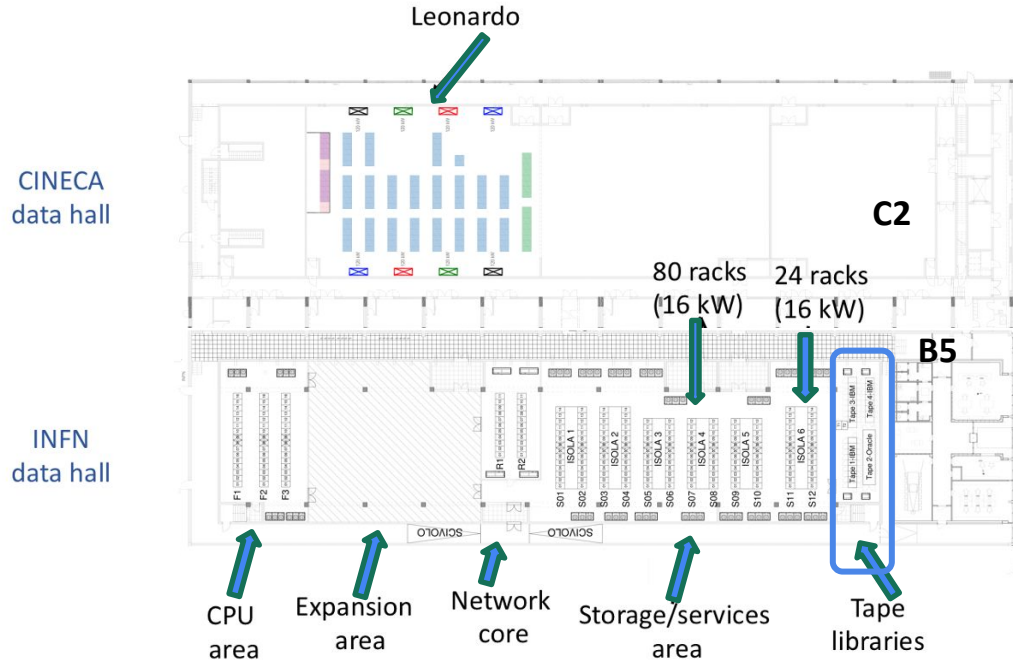


Each of the 6 “botti” (barrels) is
~5000m² of usable IT space



Same architect and design of the
“Sala Nervi” in the Vatican

CNAF and CINECA data halls



DLC 80kW



The new CNAF Datacenter will feature the following main areas

- High Density – 2-3 rows for 80kW racks
- Low density – 80+24 16kW racks
- Expansion area
- **Tape libraries areas**
 - **Up to 4 libraries**

The CPU area can host up to 3MW of CPUs via 42 DLC high density racks

The low-density area will be used to host

- Storage systems
- CNAF Cloud Infrastructures
- ISO certified Cloud racks

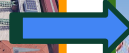
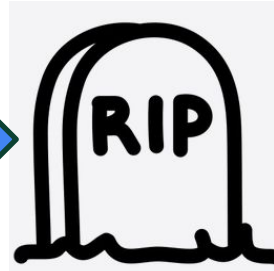
Cooling

- Air cooled Cold Corridor aisles
- Direct Liquid in High Density

3+1 redundancy in all the infrastructure facilities

CNAF Tape libraries and drives

- **1 x Oracle SL8500**
 - **1 tape library with 16 tape drives T10000D** (8.5TB/cartridge)
 - 80PB installed, 64PB USED
 - Repack on the other libraries needed
 - After completion of repack this library will be dismissed
- **2 x IBM TS4500**
 - **1 tape library with 19 tape drives TS1160** (20TB/cartridge)
 - 102 PB Installed, 50PB USED
 - cannot be further extended due to physical constraints in the current room
 - This library will be moved to the new data center
 - **1 tape library with 18 tape drives TS1170** (50TB/cartridge) acquired and will be installed at new data center Q1 2024





Metropolitan Tape Area Network

- 2 libraries at CNAF
- 1 new library at the Tecnopole
- About 7 km of fiber to connect the 2 datacenters
 - yellow + red paths
- 2 fiber pairs dedicated to extend the fiberchannel TAN
 - Brocade optics for 10km distance

BROCADE
A Broadcom Company

Product Brief

Brocade® 32Gb/s LWL
(10 km) SFP+

Optimized, Certified Optical Transceivers for
Extending Service Provider and Data Center
Networks

Overview

Today's enterprise data centers are undergoing an infrastructure transformation, requiring higher speeds, greater scalability, and higher levels of performance and reliability to better meet the demands of business. As speed and performance needs increase, optical transceivers—once considered a generic component of Fibre Channel switching technologies—have become an integral part of overall system design.

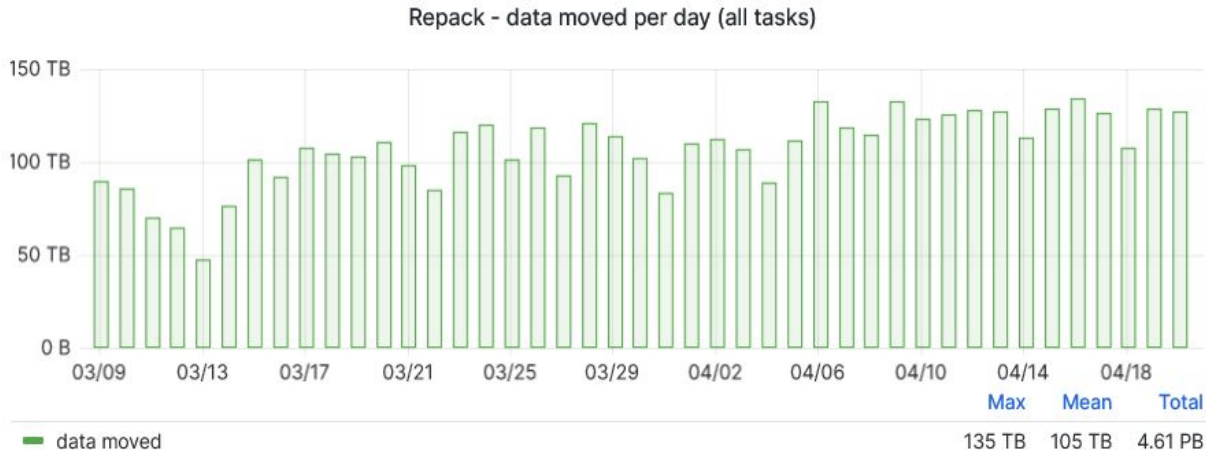
The Brocade® 32Gb/s Long Wavelength (LWL) 10 km SFP+ part of the

Highlights

- Provides high system reliability through rigorous qualification and certification processes.
- Leverages unique design parameters to provide the highest performance with industry-leading Brocade

Repack: migration + reclaim

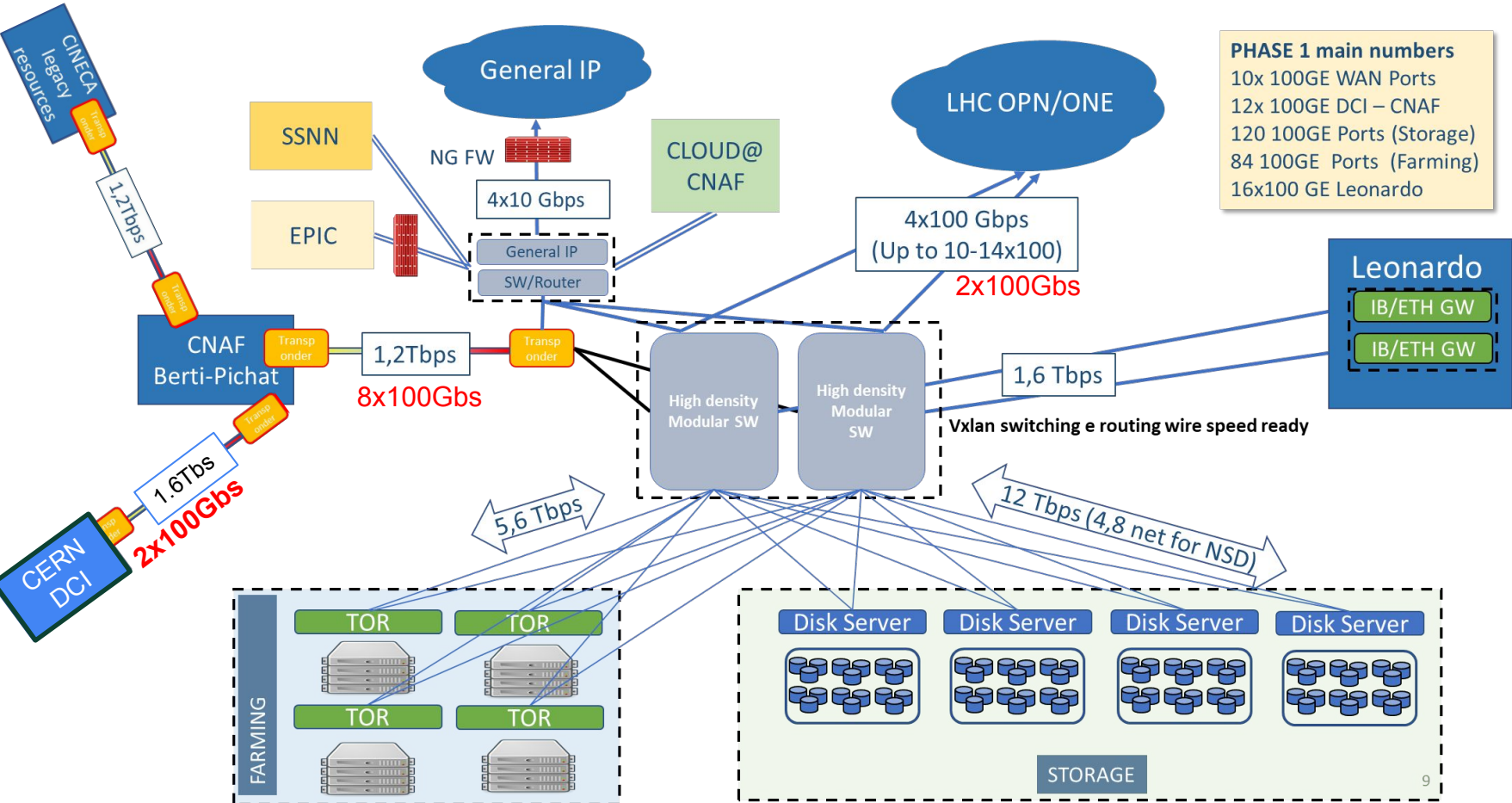
- Repack Oracle SL8500 ⇒ IBM TS4500 at ~100-120 TB / day will last ~18-20 months
 - 1 process per tape drive on the TSM server
 - 1 Socket Used ⇒ 2 sockets improve performances by 15% (need licenses)
 - During tests 7 Oracle drives used on average
 - Dynamic allocation of tape drives via Orchestrator based on production load
 - [E.Fattibene et al. - Dynamic sharing of tape drives accessing scientific data](#)
 - 174 MB/s per drive (lower than nominal: 250 MB/s)



CNAF tape summary

- Tape servers
 - 1 IBM Storage Protect server (TSM)
 - 1 HSM server for each LHC VO + 1 server for all the other experiments
 - All these servers will be replaced with new nodes at Tecnopole in Q1 2024

Networking Infrastructure



Live Relocation Timeline



- **Renovation work will terminate on November 19th**
 - After a long series of deferments
- 6 weeks to complete the network cabling inside the new datacenter
- New core switches (ARISTA) delivered in December 2023 ;
 - Smaller systems under evaluation during these days
- “live” migration except for:
 - IBM Tape Library
 - ISO27001 Certified Zone

} DISMANTLE and RE-ASSEMBLING

Thanks! Questions?

Contacts and references

StoRM source: <https://github.com/italiangrid/storm>

Documentation: <http://italiangrid.github.io/storm/>

Support mailing lists: storm-support@lists.infn.it storm-users@lists.infn.it

Developers mailing list: storm-devel@lists.infn.it

Other useful links:

- [WLCG Tape REST API](#) specification
- [StoRM Tape REST API](#) source code
- [A RESTful approach to tape management in StoRM](#), CHEP 2023
- [ngx_http_voms_module](#) source code
- [GEMSS](#) source code
- [StoRM Tape testsuite](#) source code

Backup slides



OPA AuthZ example

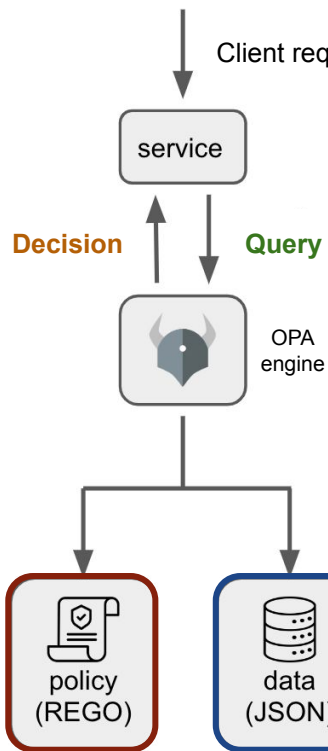
OPA AuthZ example of a stage status request for stage-id 9a8e34bd done by a user which authenticates with an X.509 personal certificate.

```
{
  "allow": "true"
}
```

```
# GET /api/v1/stage/<id>
allow if {
  input.method == "GET"
  glob.match("/api/v1/stage/*", ["/"],
input.path)

  true in [
    read_scopes_allowed,
    voms_fqans_allowed,
    certificate_dn_allowed
  ]
}
```

This policy has been defined for GET method on stage path.



```
{
  "method": "GET",
  "path": "/api/v1/stage/9a8e34bd",
  "client_s_dn":
    "CN=Enrico Vianello vianello@inf.n.it,
    O=Istituto Nazionale di Fisica Nucleare,
    C=IT,DC=tcs,DC=terena,DC=org"
}
```

```
{
  "allowed_dn": [
    "CN=Enrico Vianello
    vianello@inf.n.it,O=Istituto Nazionale di Fisica
    Nucleare,C=IT,DC=tcs,DC=terena,DC=org",
    "CN=test0,O=IGI,C=IT",
    ...
  ],
  ...
}
```

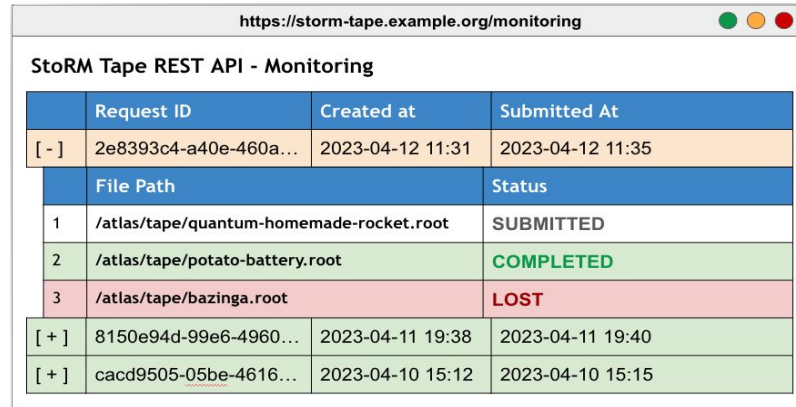
A static list of allowed DNs has been defined within internal OPA data

has allowed WLCG scopes? **X**
OR
 has allowed FQANs? **X**
OR
 has allowed DN? **✓**

StoRM Tape REST API: future looks

The idea is to provide to some privileged users a monitoring dashboard:

- monitor the requested file statuses per recall request
- cancel/delete/resume file requests



	Request ID	Created at	Submitted At
[-]	2e8393c4-a40e-460a...	2023-04-12 11:31	2023-04-12 11:35
	File Path	Status	
1	/atlas/tape/quantum-homemade-rocket.root	SUBMITTED	
2	/atlas/tape/potato-battery.root	COMPLETED	
3	/atlas/tape/bazinga.root	LOST	
[+]	8150e94d-99e6-4960...	2023-04-11 19:38	2023-04-11 19:40
[+]	cacd9505-05be-4616...	2023-04-10 15:12	2023-04-10 15:15