

# Data Management in Distributed systems - The CNAF services

D.Cesini - INFN-CNAF

L.Morganti - INFN-CNAF

E. Corni - INFN-CNAF



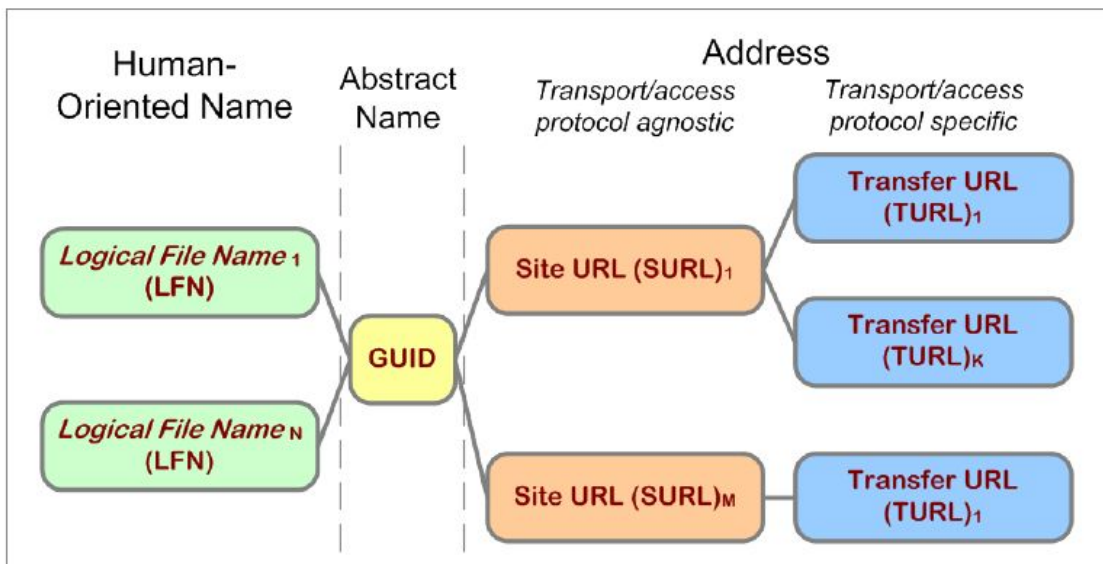
# Outline

- ❖ Identify a File
- ❖ Storage Systems at a Site
- ❖ Authentication methods
- ❖ Transfer protocols
  - SRM
  - SRM+gridFTP
  - Xrootd
  - HTTP/WebDav
  - Caching with XCache
- ❖ Tape Usage
- ❖ Orchestration Services
- ❖ DataLakes

# Identify a File

# 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.



# File Names - 2

- ❖ Logical File Name (LFN) : An alias created by a user to refer to some item of data
  - [lfn:cms/20030203/run2/track1](#)
- ❖ Grid Unique Identifier (GUID): A non-human-readable unique identifier for an item of data
  - [guid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6](#)
- ❖ Site URL (SURL) : The location of an actual piece of data on a storage system
  - [srm://storm.cnaf.infn.it:8444/infnguid/test.txt](#)
- ❖ Transport URL (TURL) : Temporary locator of a replica + access protocol understood by a SE
  - [gsiftp://storm.cnaf.infn.it:2811/storage/infnguid/test](#)

# Storage Systems at a Site

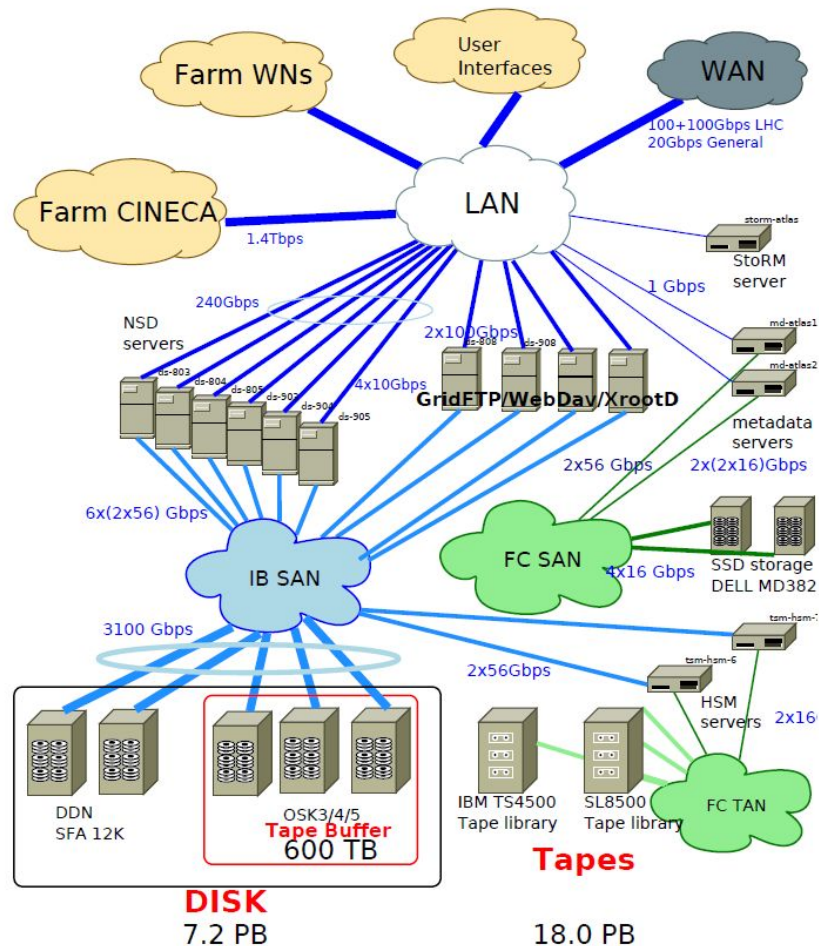
# Storage systems at a site - CNAF example

## ❖ SAN

- A dedicated network to store and access data
- Storage devices (disk arrays or tape libraries) are accessible to servers as **block level** data storage
- The interconnection is made by distinct protocols, such as Fibre Channel, iSCSI or Infiniband
- Storage and network devices are dedicated to the SAN and can be heterogeneous
- On top of the SAN can be created **a file system to access data at file level**

## ❖ TAN

- **Tape Area Network**



# Data Management Services





# Data Management and Data Transfer Services

## ❖ Deployment @CNAF:

- One or more StoRM frontends for each major experiment
- “Smaller” experiments share a single StoRM FE
- Dedicated pool of GridFTP transfer nodes to LHC and shared for the other exps
  - VOMS based authentication
- Easy interface with tape via GEMMS
- “plain” GridFTP transfer nodes available to allow plain proxy auth
  - need manual intervention for recall
  - VO not needed;
- Several XrootD servers (and redirectors) dedicated to the experiments
- XCache service also available
  - VOMS authentication and authorization required;
- dedicated StoRM WebDAV endpoints
  - **Third Party Copy support**
  - VOMS and token-based authentication
    - ◆ group-based authorization coming soon
- dedicated Apache server (token-based authorization with group support) capable of browserability

# Protocols vs AAI matrix @CNAF

	POSIX	GridFTP Plain	SRM+GridFTP	HTTP/WebDav	HTTP Browser	XrootD
Local	x					x
Grid Proxy		x	x			x
VOMS Proxy		x	x	x		x
IAM Token				x	x	

# Exp vs protocol vs AAI matrix @CNAF

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Experiment	Data Management protocols				Data Transfer protocols				Technologies					AAI				Specific needs/problems
2		POSIX	SRM	WebDAV	XrootD	POSIX	GridFTP	HTTP	XrootD	StoRM	GEMMS	XrootD	GridFTP	dataclient	VOMS	GRID	dataclient	token-based	
3	ALICE	no	no	no	yes	no	no	no	yes	no	no	yes	no	no	no	no	no	no	
4	ATLAS	no	yes	yes	yes	yes	yes	soon	yes	yes	yes	yes	yes	no	yes	no	no	interested	
5	CMS	no	yes	no	yes	yes	yes	no	yes	yes	yes	yes	yes	no	yes	no	no	interested	
6	LHCb	yes	yes	no	yes	no	yes	no	soon	yes	yes	yes	yes	no	yes	no	no	have to be interested	
7																			
8	white: no answer																		
9	AGATA	yes	tape	no	no	yes	tape	no	no	tape	yes	no	no	no	yes	no	no	no	
10	AMS	no	yes	no	no	no	yes	no	yes	yes	yes	yes	no	no	yes	no	no	interested (see DODAS)	
11	ARGO	yes	yes	no	no	yes	yes	no	no	tape	yes	no	no	no	yes	no	no	no	closing
12	AUGER	maybe local users	yes	no	no	no	yes	no	no	yes	no	no	no	no	yes	no	no	no	no tape, use Dirac
13	BELLE	no	yes	yes	no	no	yes	yes	no	yes	no	no	no	no	yes	no	no	interested	tape not used
14	BOREXINO	yes	no	no	no	yes	no	no	no	no	no	no	no	no more	no	no	no more	no	tape not used, closing
15	COMPASS	no	yes	interested	interested	no	yes	interested	interested	yes	no	no	no	no	yes	no	no	interested	no tape
16	CORELIB	yes	no	no	no	no	yes	no	no	no	no	no	yes	no	no	yes	no	no	no tape
17	COSMO_WNEXT	yes	no	no	no	no	yes	no	no	no	no	no	yes	no	no	yes	no	no	tape not used
18	CTA	no	yes	no	future	no	yes	future	no	yes	yes	no	no	no	yes	no	no	interested	valido per MC e DIRAC, non per utenti locali
19	CUORE	yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no tape, rsync (100G/day)
20	CUPID	yes	no	no	no	no	yes	no	no	no	no	no	yes	no more	no	yes	no more	no	tape not used
21	DAMPE	yes	tape	no	no	no	yes	no	no	tape	yes	no	yes	no	yes	yes	no	interested	
22	DARKSIDE	yes	no	no	no	no	yes	no	no	no	no	no	yes	no	not used	yes	no		tape?
23	ENUBET	yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	interested	no tape
24	FAMU	yes	no	no	no	yes	yes	no	no	no	no	no	yes	no	no	yes	no	interested	no tape
25	GERDA	yes	tape	no	no	no	tape	no	no	tape	yes	no	no	no	yes	no	no	interested	
26	GLAST	no	yes	no	interested	no	yes	no	no	interested	yes	no	no	no	yes	no	no		tape not used
27	ILDG	no	yes	no	no	no	yes	no	no	yes	no	no	no	no	yes	no	no	no	no tape
28	ICARUS	no	yes	no	no	no	yes	no	no	yes	yes	no	no	no	yes	no	no		
29	JUNO	yes	yes	yes	no	yes	yes	yes	no	yes	yes	no	yes	no	yes	no	no	maybe in the future	no tape
30	KM3NET	yes	yes	no	interested	no	yes	no	no	yes	no	interested	yes	yes	yes	yes	yes	interested	tape not used
31	LHAASO	yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no tape, closing
32	LHCf	yes	no	no	no	yes	no	no	no	no	no	no	no	no	no	no	no	no	no tape
33	LIMADOU	yes	no	interested	no	no	yes	interested	no	no	no	no	yes	no	no	yes	no	interested	no tape
34	MAGIC	no	tape	no	no	yes	tape	no	no	tape	yes	no	no	no	yes	no	no		convergono con CTA
35	NA62	no	yes	no	no	no	yes	no	no	yes	no	no	no	no	yes	no	no	only if it becomes DIRAC standard	tape not used; use xrootd (in other sites)
36	NEWCHIM	yes	no	no	no	no	yes	no	no	no	no	no	yes	no more	no	yes	no more		tape: write to buffer with guc. Recall?
37	PADME	no	yes	no	interested	no	yes	no	no	interested	yes	yes	no	no	yes	no	no	interested	
38	PAMELA	no	tape	no	no	yes	tape	no	no	tape	yes	no	no	no	yes	no	no	no	
39	THEOPHYS	no	tape	no	no	yes	tape	no	no	tape	yes	no	no	no	yes	no	no	no	
40	VIRGO	yes	yes	yes	no	yes	yes	yes	no	yes	yes	no	yes	no	yes	yes	no	interested	also a storage area for storm no voms
41	XENON	no	yes	no	no	no	yes	no	no	yes	yes	no	no	no	yes	no	no		

# X509 Authentication

# Obtaining a personal X509 Certificate

- ❖ A personal certificate can be obtained from:
  - <https://www.digicert.com/sso>
  - select "Grid Premium" in the drop-down menu named "Product"
  - Then click on "Request Certificate" and you will install the certificate on your browser.

# X509 format conversion

Once obtained the pk12 certificate (cert.p12), it is necessary to split it in a public and private keys and put them in the .globus folder inside user home directory in the UI:

```
cd $HOME
mkdir .globus
cd .globus
openssl pkcs12 -clcerts -nokeys -in cert.p12 -out usercert.pem
openssl pkcs12 -nocerts -in cert.p12 -out userkey.pem
chmod 600 usercert.pem
chmod 400 userkey.pem
```

# VOMS Proxy Creation

- ❖ To transfer files using VO, first we have to generate a proxy with voms extensions using the command:

```
voms-proxy-init --voms <vo name>
```

- ❖ To check the proxy:

```
voms-proxy-info --all
```



# Long Lived Proxy

- ❖ The maximum lifetime of myproxy is one week
- ❖ Store a password for the token retrieval:

```
myproxy-init --voms virgo:/virgo/virgo -s myproxy.cnaf.infn.it -d
```

- ❖ Then, every two days using the following:

```
echo "<password>" hidden_file
```

```
myproxy-logon --proxy_lifetime 12 -d -S < hidden_file
```

# Data Transfer

- SRM
- GridFTP
- GFAL utils
- WebDav & IAM+Apache
- XRootD
  - Caching with XCache

# SRM (Storage resource Manager) protocol:

- ❖ The Storage Resource Manager Interface defined in Specification v2.2
  - Open Grid Forum GFD-R-P-129
  - <http://www.ogf.org/documents/GFD.129.pdf>
- ❖ Specifies a common control interface to storage resource management systems
- ❖ Provides a common interface to heterogeneous storage and file systems

- Examples
  - `srmPing`
  - `srmMkdir`
  - `srmRmdir`
  - `srmRm`
  - `srmLs`
  - `srmMv`
  - `srmReserveSpace`
  - `srmGetSpaceMetadata`
  - `srmPrepareToPut`
  - `srmPutDone`
  - `srmStatusOfPrepareToPutRequest`
  - `srmPrepareToGet`
  - `srmStatusOfGetRequest`
  - `srmCopy`
  - `srmStatusOfCopy`

# SRM PtP and PtG

To request the storage space in the destination file system two commands are available:

- Transfer from local to remote:
  - `clientSRM PTP` (*Prepare To Put*):
- Transfer from remote to local:
  - `clientSRM PTG` (*Prepare To Get*):

# SRM PtP and PtG

```
clientSRM PTP -v NIG -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444 -s \  
srm://storm-fe-archive.cr.cnaf.infn.it:8444/srm/managerv2?SFN=/virgo4/test.mt.002
```

```
clientSRM PTG -v NIG -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444 -s \  
srm://storm-fe-archive.cr.cnaf.infn.it:8444/srm/managerv2?SFN=/virgo4/test.mt.002
```

To check the status of the request, use “clientSRM SPTP” (*Status of prepare To Put*) or “clientSRM SPTG”

```
clientSRM SPTP -v -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444 \  
-t fd5256b6-5cb3-4fe2-a82d-3ba316f1f1f8
```

```
clientSRM SPTG -v -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444 \  
-t fd5256b6-5cb3-4fe2-a82d-3ba316f1f1f8
```

# Gridftp copy

- ❖ SPTG and SPTP provides, when completed the TURL
- ❖ To be used for the transfer with “globus-url-copy”

```
#Local to remote:  
globus-url-copy file:///<local_path>/file \  
gsiftp://gridftp-storm-archive.cr.cnaf.infn.it:2811//storage/gpfs_virgo4/virgo4/test.mt.002  
  
#Remote to local:  
globus-url-copy \  
gsiftp://gridftp-storm-archive.cr.cnaf.infn.it:2811//storage/gpfs_virgo4/virgo4/test.mt.002\  
file:///<local_path>/file
```

TURL

Third party is also possible

# gridFTP transfers

```
globus-url-copy file:///<local_path>/file \
gsiftp://gridftp-storm-archive.cr.cnaf.infn.it:2811/<remote_path>/file
```

## Interesting options:

**-p** <n>: multiple streams enable

**-tcp-bs** SIZE, -tcp-buffer-size SIZE

**-bs** block SIZE, -block-size block SIZE

**-r** : recursive

**-sync**

**-sync-level** number

Criteria for determining if files differ when performing a sync transfer. The default sync level is 2. The available levels are:

- Level 0 will only transfer if the destination does not exist.
- Level 1 will transfer if the size of the destination does not match the size of the source.
- Level 2 will transfer if the time stamp of the destination is older than the time stamp of the source.
- Level 3 will perform a checksum of the source and destination and transfer if the checksums do not match. The default

algorithm used for this checksum is MD5, but other algorithms can be specified with the -algo parameter.

**-checksum-alg** CHECKSUM-ALGORITHM

Set the algorithm type to use for all checksum operations during the transfer.

**-verify-checksum**

Perform a checksum on the source and destination after each file transfer and compare the two. If they do not match, fail the transfer. The default algorithm used for this checksum is MD5, but other algorithms can be specified with the -checksum-alg parameter.

# Gfal (Grid File Access Library)

- ❖ To use this utilities a personal certificate is required in order to generate a valid proxy
- ❖ You can copy a file with “`gfal-copy`” or list the contents of a folder with “`gfal-ls`”
  - Automatically perform the two steps that we described in the previous slides

```
-bash-4.2$ gfal-copy srm://storm-fe-ams.cr.cnaf.infn.it:8444/darksidedisk/test .  
Copying srm://storm-fe-ams.cr.cnaf.infn.it:8444/darksidedisk/test [DONE] after 3s  
  
-bash-4.2$ gfal-ls -l srm://storm-fe-ams.cr.cnaf.infn.it:8444/darksidedisk  
test
```



# More on gfal utils

```
-bash-4.2$ gfal-  
gfal-cat          gfal-copy          gfal-legacy-register  
gfal-legacy-unregister gfal-mkdir        gfal-rm             gfal-stat  
gfal-xattr        gfal-chmod        gfal-legacy-bringonline  
gfal-legacy-replicas gfal-ls           gfal-rename         gfal-save  
gfal-sum
```

**Note: Missing support for gridftp multiple streams in streamed copy**

<https://its.cern.ch/jira/browse/DMC-1117>

# HTTP endpoints - StoRM WebDAV

- ❖ StoRM WebDAV is the StoRM service that provides valid WebDAV endpoints for the experiments' storage areas
- ❖ The most common WebDAV clients are browsers and command-line tools such as curl and davix.
- ❖ VOMS proxies are supported only by command-line tool, e.g.:

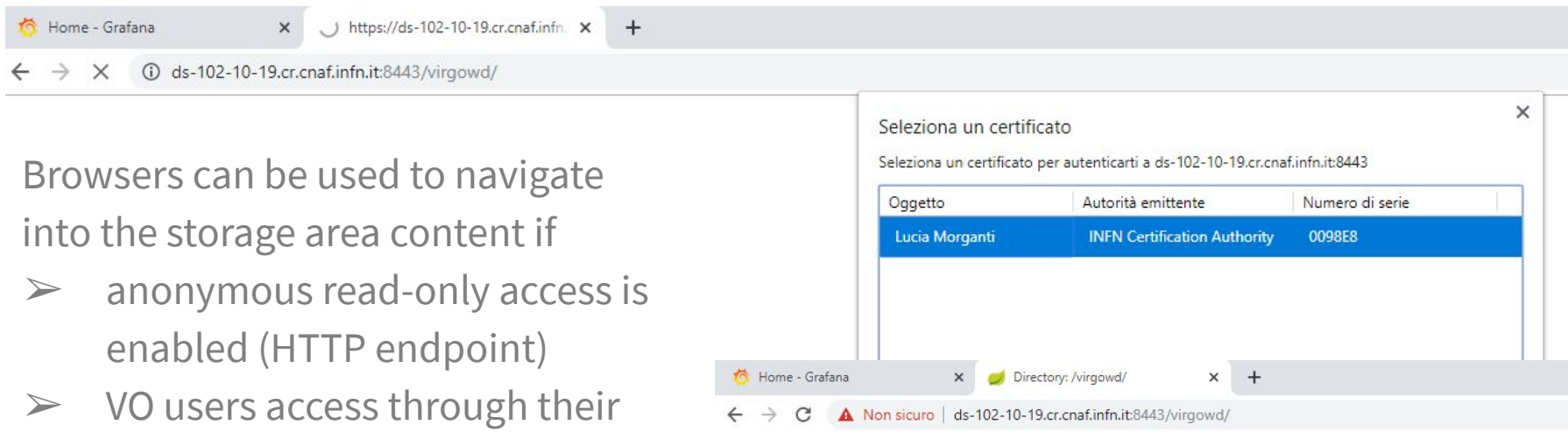
```
voms-proxy-init --voms dteam
```

```
Your proxy is valid until Thu Nov 21 21:57:20 CET 2019
```

```
davix-ls -P grid https://xfer-1.cr.cnaf.infn.it:8443/dteam  
smoke-test-celebrimbor-20038  
[...]
```

# StoRM WebDAV

- ❖ Browsers can be used to navigate into the storage area content if
  - anonymous read-only access is enabled (HTTP endpoint)
  - VO users access through their X509 certificate is enabled (HTTPS endpoint)



## Directory: /virgowd/

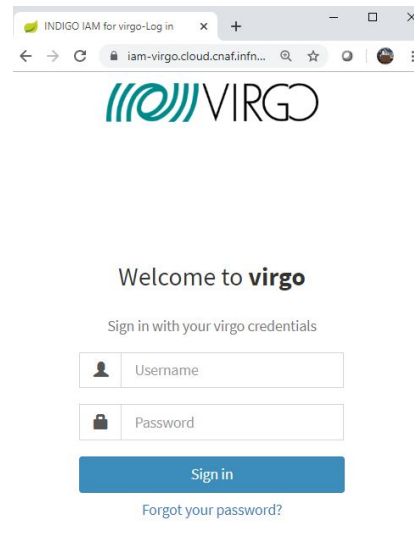
[Parent Directory](#)

[V-raw-1212474700-100.gwf](#) 4585583989 bytes Jun 8, 2018 9:07:40 AM

[V-raw-1212476800-100.gwf](#) 4447108735 bytes Jun 8, 2018 9:11:10 AM

# StoRM WebDAV with tokens

- ❖ StoRM WebDAV also supports OpenID connect authentication and authorization on storage areas.
- ❖ Dedicated IAM (Identity and Access Management) instances can be configured; once registered within IAM, an access token can be retrieved via browser (or dedicated script)



# StoRM WebDAV with tokens

- ❖ The access token, exported in the variable `$AT`, can be used instead of the VOMS proxy to access the storage area with http clients (curl, davix...):

```
$ davix-ls --capath /etc/grid-security/certificates/ -H "Authorization: Bearer ${AT}"  
https://ds-102-10-19.cr.cnaf.infn.it:8443/virgowd  
V-raw-1212476800-100.gwf  
V-raw-1212474700-100.gwf
```

# StoRM WebDAV and Third-Party-Copy

- ❖ StoRM WebDAV also supports Third-Party-Copy, as required by the activities of the DOMA TPC working group  
(<https://twiki.cern.ch/twiki/bin/view/LCG/ThirdPartyCopy>)
- ❖ In order to implement a form of delegated authorization in support of third-party transfers, an OAuth authorization server can be used by clients such as FTS to obtain an OAuth access token that grants the same privileges as a VOMS credential.



# Apache with tokens for group-based authorization

- ❖ Currently, StoRM WebDAV does not support group-based authorization
  - to be implemented soon
- ❖ A dedicated Apache server is configured for this use-case @CNAF
- ❖ A catch-all IAM instance available at iam-computing.cloud.cnaf.infn.it
  - registered users are assigned to specific groups
- ❖ Once registered within IAM, an access token can be retrieved
  - i.e. via browser contacting a dedicated “client”

The screenshot shows a web browser window with the URL `iam-t1-computing.cloud.cnaf.infn.it/iam-test-client/`. The page title is "IAM Login Service Test Client". Below the title, it says "You're now logged in as: Lucia Morganti" and "This application has received the following information:". Under this, there is a section for "access\_token (JWT):" which contains a long JWT token. An arrow points from this token to a separate box on the right titled "access\_token (decoded):". This box displays the decoded JSON of the token, where the "groups" array is circled. The groups listed are "asfin", "fazia", and "ntof".

iam-t1-computing.cloud.cnaf.infn.it/iam-test-client/

## IAM Login Service Test Client

You're now logged in as: Lucia Morganti

This application has received the following information:

- access\_token (JWT):  
`eyJraWQiOiJyc2ExIiwiaWVwXnIjoIUMyNTYifQ.eyJzdWIiOiJhM2I3MTljMS1hNzM0LTQzNDYtOGUzY1IiLCJm`

• access\_token (decoded):

```
{
  "sub": "a3b719c1-a734-4346-8e3b-d2d4a2172f3c",
  "iss": "https://iam-t1-computing.cloud.cnaf.infn.it/",
  "name": "Lucia Morganti",
  "groups": [
    "asfin",
    "fazia",
    "ntof"
  ],
}
```

# WebDAV/Apache with tokens

- ❖ The storage areas can be accessed with http clients (curl, davix...) using the access token (e.g. exported in the variable \$AT):

- `davix-ls --capath /etc/grid-security/certificates/ -H "Authorization: Bearer ${AT}"`  
<https://gridftp-storm-archive.cr.cnaf.infn.it:8443/virgowd/>

- ❖ Group-based token authorization is supported:

```
bash-4.2$ davix-ls -H "Authorization: Bearer ${AT}" https://ds-814.cr.cnaf.infn.it:8443/ntof  
test_lucia_folder
```

```
bash-4.2$ davix-ls -H "Authorization: Bearer ${AT}" https://ds-814.cr.cnaf.infn.it:8443/newsdm
```

```
(Davix::HttpRequest) Error: Authentication failed on server
```



## Xrootd (extended ROOT daemon):

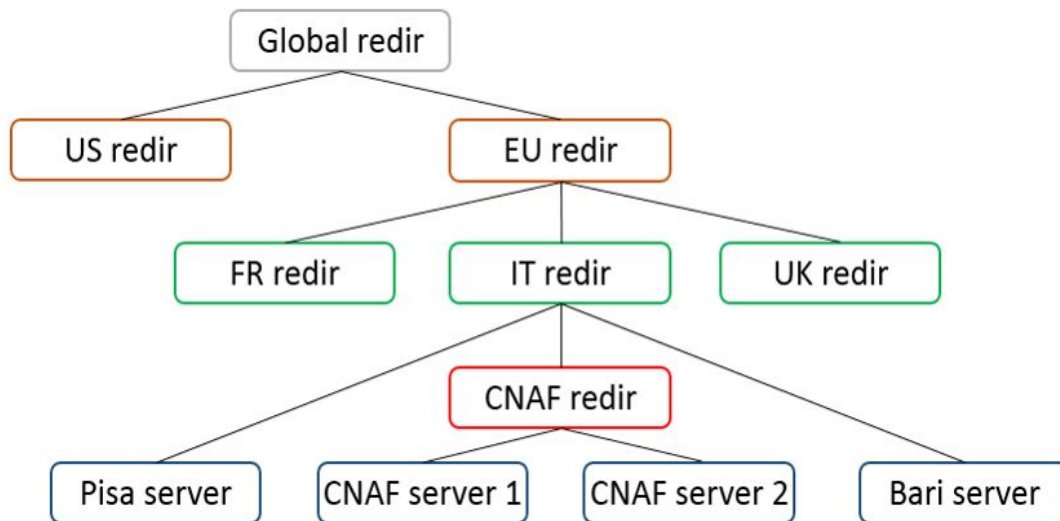
- ❖ To use this protocol a personal certificate is required in order to generate a valid proxy...
- ❖ ...other methods allowed
  - it depends on the configuration
- ❖ Before file transfer, it is necessary to generate a valid proxy with the command:  
`voms-proxy-init --voms <V0>`

# What is a XRootD redirector?

- ❖ A redirector is a service that allows to query the site to search for the requested file
- ❖ The redirection structure is a tree structure where the trunk is the global redirector and the branches are the regional redirector up to the local redirectors.

# XRootD Federations

- ❖ FEDERATION - When the global redirector is contacted, it searches for the file in the lower levels (European redirector and then Italian one)



# How to use Xrootd to transfer a file?

- ❖ In case of a federation - When you attempt to open a file, your application must query a redirector to find the file.
- ❖ You must specify the redirector to the application. Which redirector you use depends on your region, to minimize the distance over which the data must travel and thus minimize the reading latency.
- ❖ These "regional" redirectors will try file locations in your region first before trying to go overseas.

# XRootD data transfers

❖ To perform a transfer:

```
xrdcp root://<redirector_or_server>//store/path/to/file  
/some/local/path
```

For example:

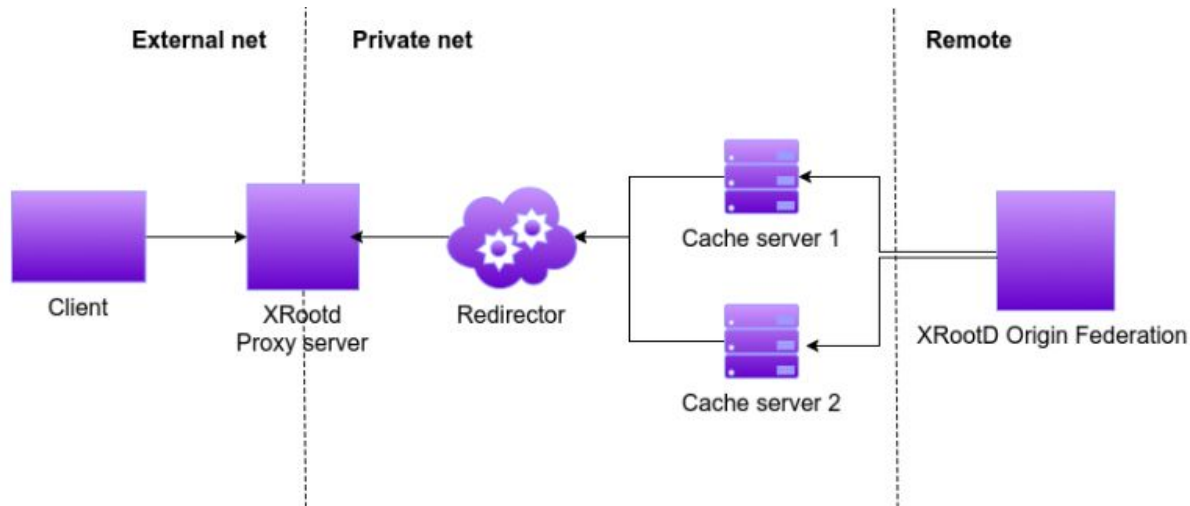
```
xrdcp \  
root://xrootd-cms-02.cr.cnaf.infn.it:1094///store/test/xrootd/T1  
_IT_CNAF/store/mc/SAM/GenericTTbar/AODSIM/CMSSW_9_2_6_91X_mcRun1  
_realistic_v2-v1/00000/A64CCCF2-5C76-E711-B359-0CC47A78A3F8.root  
local_copy
```

# On-Demand XCache cluster

- ❖ XCache is a plugin built upon XRootd
- ❖ It implements **Caching** as well as **Proxy** functions
- ❖ **On-demand-Caching** is a deployment framework for XCache implemented and developed within the CMS collaboration, efforts also from **XDC Project** → **INFN Perugia Group**
  - **Daniele Spiga, Diego Ciangottini**

# XCache components

- ❖ The **clients** are configured to request files to a proxy/cache server on their same network
- ❖ The redirector federates the caches (any new cache is configured to register to the redirector)
- ❖ The Caching servers contacts the remote federation
- ❖ Tested on several cloud providers as well as CNAF



**<https://cloud-pg.github.io/CachingOnDemand/>**

# XCACHE Demo

```
root@vnode-0:/home/cloudadm# helm install cache/cachingondemand -f config.yaml
NAME: veering-sloth
LAST DEPLOYED: Mon Feb 18 09:16:40 2019
NAMESPACE: default
STATUS: DEPLOYED

RESOURCES:
==> v1/Service
NAME          AGE
xcache-service 16s
xcache-proxy  16s

==> v1/Deployment
xcache-pod  16s
xredis-pod  16s
proxy-pod   16s

==> v1/Pod(related)

NAME                                READY  STATUS             RESTARTS  AGE
xcache-pod-6958d966f4-xpdkz        0/1    Pending            0          16s
xredis-pod-68b7bd4c4d-tjg99        0/1    ContainerCreating  0          16s
proxy-pod-7fc99f6644-m55w6         0/1    ContainerCreating  0          16s
```

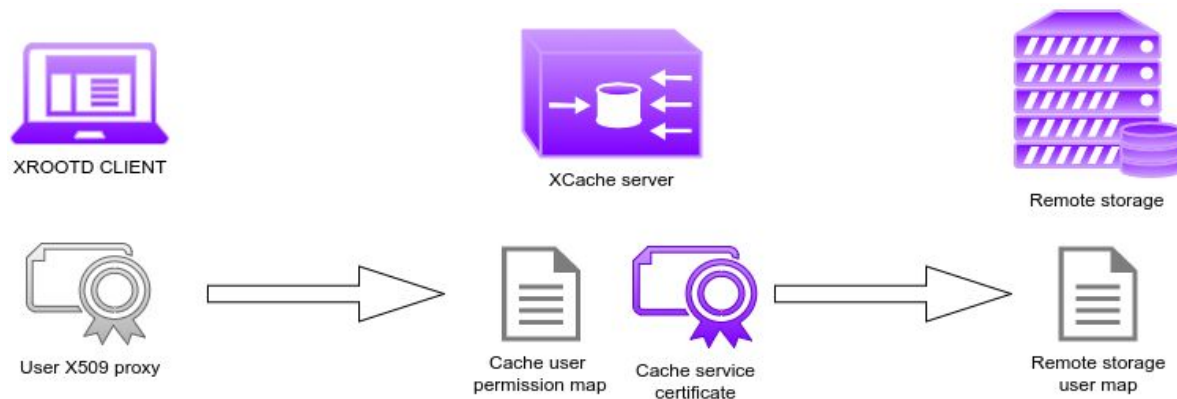
K8s cluster deployed  
using helm






# AuthN/Z mode in XCache

- ❖ The **clients** authenticate only against the Cache service
- ❖ If authorized the file is served from the cache (if present)
- ❖ If not present the cache downloads the file with its own credentials
- ❖ Available also OIDC authentication



# XCache Demo

```
(11:43 dciangot@lxplus101 ~) > FILE=/store/data/Run2017B/ZeroBias/AOD/17Nov2017-v1/60000/E8E47D88-BDDA-E711-8090-549F35AD8BD6.root
(11:44 dciangot@lxplus101 ~) > CMS_DEFAULT_REDIRECTOR=xrootd-cms.infn.it
(11:44 dciangot@lxplus101 ~) > CMS_CACHE_REDIRECTOR=cloud-vm90.cloud.cnaf.infn.it:31194
(11:44 dciangot@lxplus101 ~) > xrdcp -f root://$CMS_DEFAULT_REDIRECTOR/$FILE /dev/null
[1.67GB/1.67GB][100%][=====][7.635MB/s]
(11:48 dciangot@lxplus101 ~) >
(11:48 dciangot@lxplus101 ~) >
(11:48 dciangot@lxplus101 ~) >
(11:48 dciangot@lxplus101 ~) >
(11:48 dciangot@lxplus101 ~) > xrdcp -f root://$CMS_CACHE_REDIRECTOR/$FILE /dev/null
[1.578GB/1.67GB][ 94%][=====> ][11.97MB/s]
```



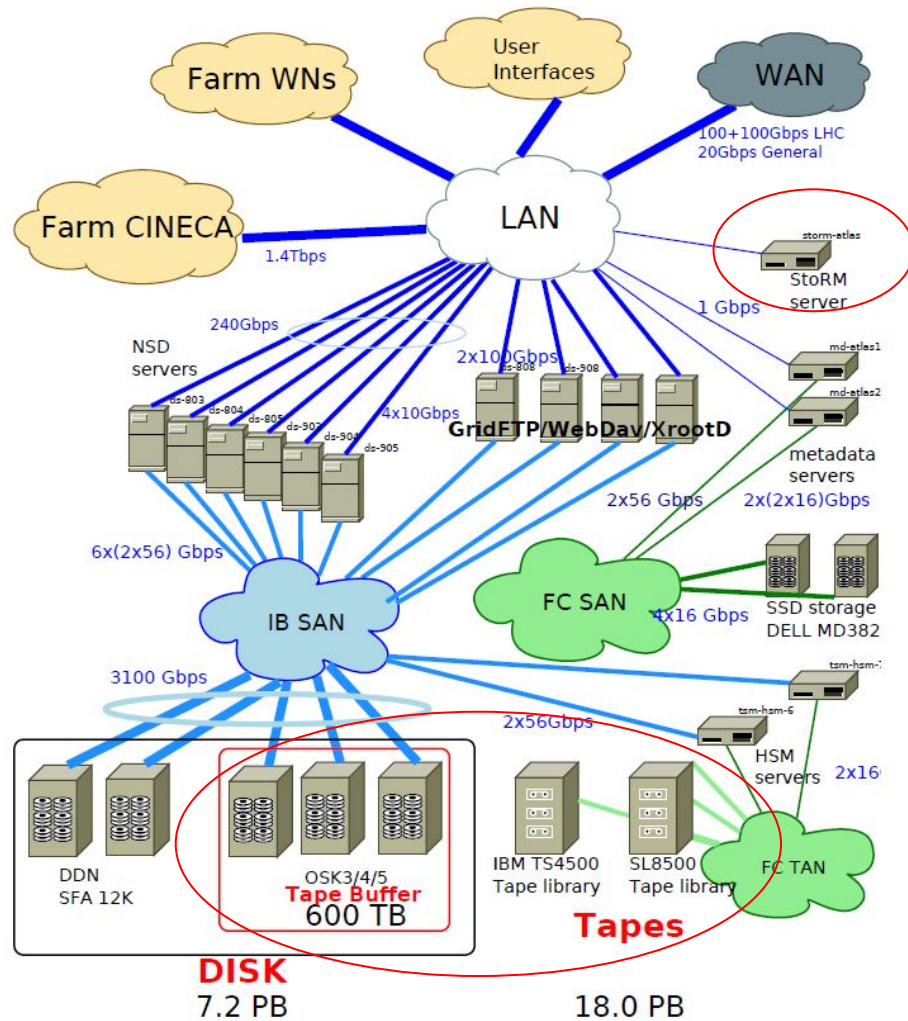
```
(12:30 dciangot@lxplus096 ~) >
(12:31 dciangot@lxplus096 ~) > xrdcp -f root://$CMS_CACHE_REDIRECTOR/$FILE /dev/null
[1.67GB/1.67GB][100%][=====][95.02MB/s]
(12:31 dciangot@lxplus096 ~) >
```



I

# Tape

- Buffer
- Where is my file?
- Recall



# Tape Area Network (TAN)

- ❖ Is 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 every server on the network



# CNAF mass Storage System

- ❖ 1 Oracle StorageTek SL8500 tape library
- ❖ 10000 slots, 85PB capacity with present technology
- ❖ 17 T10000D tape drives
- ❖ 250 MB/s throughput
- ❖ Disk buffer to perform writing and reading operations with tapes
- ❖ 80 PB of data
- ❖ Especially scientific RAW data
- ❖ Backup of CNAF service configurations, logs, repositories, etc.



# The new tape library @CNAF

- ❖ IBM TS4500
- ❖ 6200 slots, total capacity 120 PB
- ❖ Will be added to the old Oracle Library
- ❖ 19 tape drives TS1160 - data rate 400MB/s each
- ❖ In production from Jan 2020

# QoS, Migration and Recall

- ❖ Migration ⇒ moving a file from disk to tape
- ❖ Recall ⇒ moving a file from tape to disk
- ❖ QoS ⇒ Ensures that a particular application or workload always gets a certain performance level
  - Typically expressed as IOPS
  - An increasing number of storage systems now claim to offer some form of QoS



# Tiered Storage

- ❖ Tiered storage is a data storage environment consisting of two or more kinds of storage delineated by differences in at least one of these four attributes:
  - Price
  - Performance
  - Capacity
  - Function
- ❖ Any significant difference in one or more of the four defining attributes can be sufficient to justify a separate storage tier
  - Examples:
    - Disk and tape: two separate storage tiers identified by differences in all four defining attributes.
    - Old technology disk and new technology disk: two separate storage tiers identified by differences in one or more of the attributes
    - High performing disk storage and less expensive, slower disk of the same capacity and function: two separate tiers



**Different QoS**

# Tape Buffer

- ❖ 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 - garbage collection
- ❖ To put files into buffer, you just copy the file to `/storage/gpfs_archive/<exp>`

# Tape - Where is my file?

- ❖ To know if a file is on the disk, it is sufficient to check file dimension with the command “`ls -ls`”.  
If the file has null dimension, it is not physically present on the disk (it is on tape).

```
-bash-4.2$ ls -ls
/storage/gpfs_tsm/cms/cms/store/test/rucio/cms//store/mc/RunIIFall18wmLHEGS/SUSYG
luGluToBBHToBB_M-600_TuneCP5_13TeV-amcatnlo-pythia8/GEN-SIM/102X_upgrade2018_real
istic_v11-v1/280000/A61D92B2-C74A-6045-8325-869194181F9E.root

0 -rw-rwxr--+ 1 storm storm 1790274828 Jul 10 18:33
/storage/gpfs_tsm/cms/cms/store/test/rucio/cms//store/mc/RunIIFall18wmLHEGS/SUSYG
luGluToBBHToBB_M-600_TuneCP5_13TeV-amcatnlo-pythia8/GEN-SIM/102X_upgrade2018_real
istic_v11-v1/280000/A61D92B2-C74A-6045-8325-869194181F9E.root ## ON TAPE
```

# Tape - Where is my file?

- ❖ To check if a file is on the disk using Grid tools, you can use the “lcg-ls” command with the “-l” option:

```
lcg-ls -c 100 -v -l srm://storm-fe-archive.cr.cnaf.infn.it:8444/pamela/data/file  
SE type: SRMv2  
-rw-rw-rw- 1 2 2 681491712 ONLINE_AND_NEARLINE /pamela/data/file  
[...]
```

- ❖ In output of the command, next to the file, there will be its status.  
ONLINE\_AND\_NEARLINE means the file is present both on disk and tape,  
while NEARLINE means it is only on tape.  
NB: for SL7 “lcg-utils” and so “lcg-ls” are deprecated

## Tape - Where is my file?

- ❖ Another way to check where is a file is to use the following command (used with VOMS Proxy):

```
clientSRM ls -l -v NIG -e <endpoint> -s <file-SURL>
```

Based on the information shown in the output, we can locate the file:

- “retentionPolicyInfo=(2,0)” : on tape
- “retentionPolicyInfo=(0,0)” : only on disk

# Tape - Where is my file?

```
# file on 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/icarus/test-srm
```

```
[...]  
[0] retentionPolicyInfo=(2,0)  
[...]
```

```
#file only on DISK:  
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/icarusdata/std.err
```

```
[...]  
[0] retentionPolicyInfo=(0,0)  
[...]
```

# gfal-xattr

```
-bash-4.2$ gfal-xattr srm://storm-fe-archive.cr.cnaf.infn.it:8444/icarus/test-srm
```

```
user.replicas = gsiftp://gridftp-storm-archive.cr.cnaf.infn.it:2811//storage/gpfs_archive/icarus/raw/test-srm
```

**user.status = ONLINE\_AND\_NEARLINE**

srm.type = StoRM

**NOTE: gfal-xattr ALWAYS recall a file! - this will be fixed:**

<https://cern.service-now.com/service-portal/view-request.do?n=RQF1463791>

# Tape - Recall

```
clientSRM bol -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444 \  
-s srm://storm-fe-archive.cr.cnaf.infn.it:8444/srm/managerv2?SFN=/pamela/data/file
```

Where:

“-e” option provides the end-point to contact,

“-s” option provides the SURL of the files.

Now, your recall request is queued.

```
clientSRM sbol -e httpg://storm-fe-archive.cr.cnaf.infn.it:8444 \  
-s srm://storm-fe-archive.cr.cnaf.infn.it:8444/srm/managerv2?SFN=/pamela/data/file -t  
<token>
```



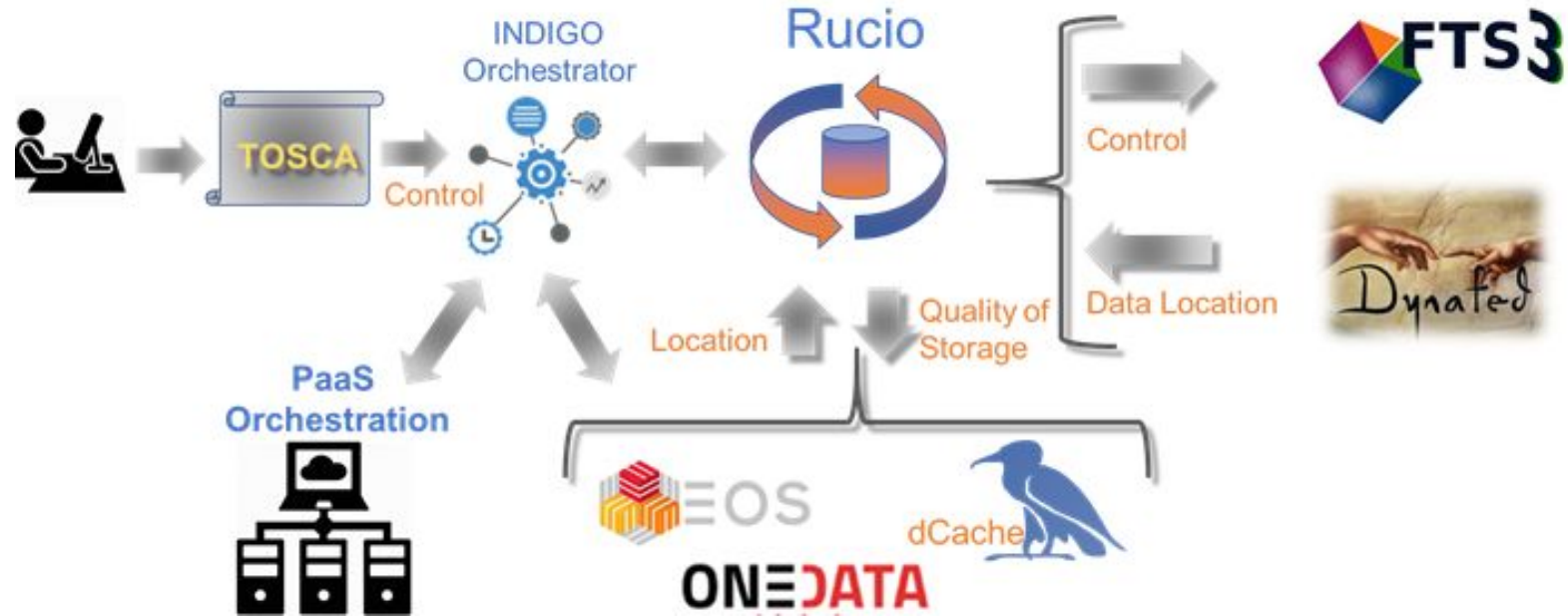
# Tape - Recall

- ❖ To recall files from tape using VO, you can use “clientSRM” command with the “bo1” option (*Bring On Line*)
  - BoL is performed also when running a PtG
  - BoL operations are authenticated via VOMS proxies
- ❖ To recall files from tape without VO, it is necessary to provide the list of the files to be recalled to CNAF that will recall them

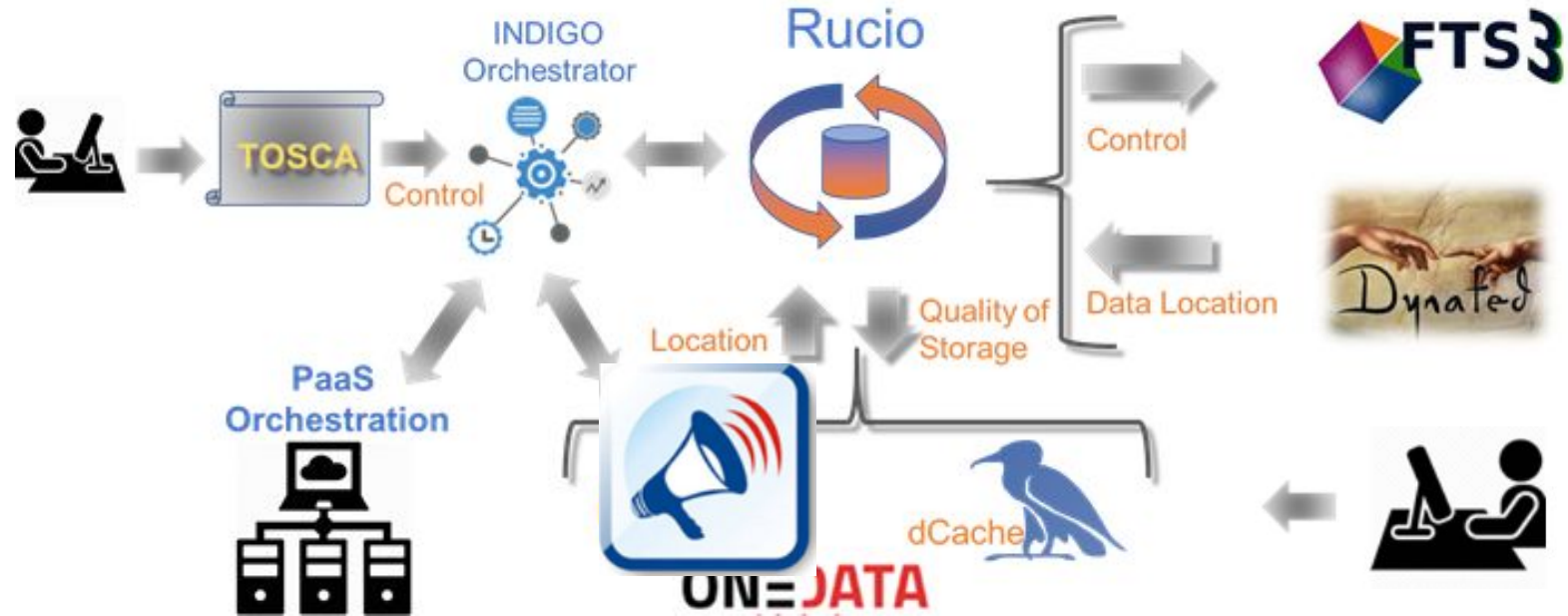
# Data Management Orchestration

- FTS
- Rucio
- INDIGO Orchestrator
- Storage Notifications

# Services for Orchestration



# Services for Orchestration



# FTS

- ❖ FTS, File Transfer Service
  - WLCG data transfer workhorse.
  - Transfers around 1 Exabytes of WLCG data per year between hundreds of storage sites around the world.
  - Performs request queueing and network shaping.
  - Can be used as “micro service” or with GUI (WebFTS).
  - Support X509 and token based authentication for endpoints.

# RUCIO

- ❖ Rucio
  - Originally LHC ALTAS data management tools
  - Recently adopted by a growing number of other communities
  - Already provides interfaces to most Storage systems and FTS components

# Indigo-PaaS Orchestrator

- ❖ INDIGO PaaS Orchestrator
  - Based on INDIGO-DataCloud developments
  - Allows to coordinate complex deployments on hybrid clouds featuring advanced scheduling and federation capabilities
  - Orchestrates compute resources and provides data-aware scheduling of jobs through data placement plugins
  - Integrates with Rucio for data location and transfer orchestration
  - Operates with an professional BPM system. (Flowable)

# Handling Software

- /opt/exp\_software
- CVMFS
- Containerized CVMFS



# Data vs Software

Software	Data
POSIX interface	put, get, seek, streaming
File dependencies	Independent files
$O(\text{kB})$ per file	$O(\text{GB})$ per file
Whole files	File chunks
Absolute paths	Relocatable
WORM (“write-once-read-many”)	
Billions of files	
Versioned	

Software is massive not in volume but in number of objects and meta-data query rates

# The Old Style

- ❖ /opt/exp\_software storage area made available in each site
- ❖ Available ReadOnly via POSIX in all WNs and UIs
- ❖ Special roles (sgm) can write on it
- ❖ Need manual deployment in all the sites of newer files/releases
- ❖ In general it is a file systems optimized for small files

# CernVM-Fs

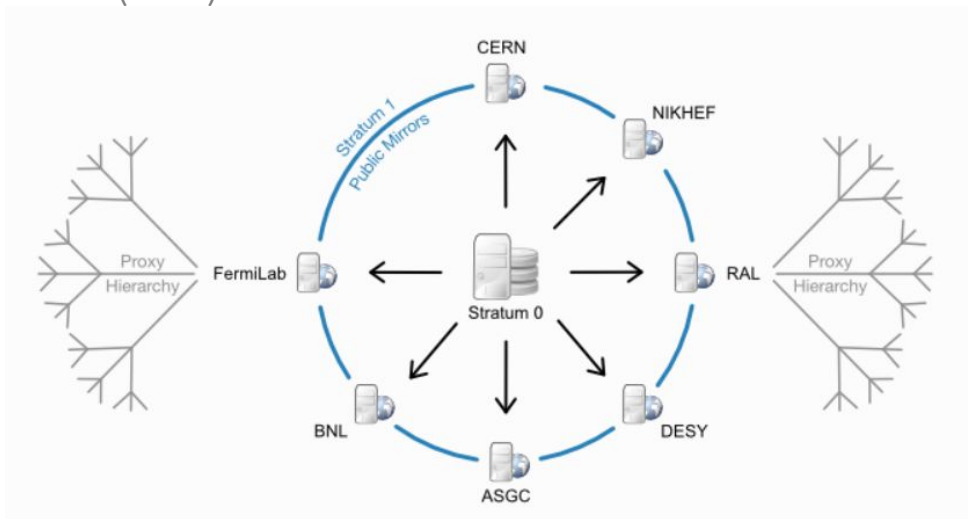
- ❖ CernVM-FS is a special-purpose virtual file system that provides a global shared software area for many scientific collaborations
- ❖ Asynchronous writing (publishing) key to meta-data scalability
- ❖ Provides a scalable, reliable and low- maintenance software distribution service.
- ❖ Implemented as a POSIX read-only file system in user space (a FUSE module).
- ❖ Files and directories are hosted on standard web servers and mounted in the universal namespace /cvmfs.
- ❖ Internally, CernVM-FS uses content-addressable storage and Merkle trees in order to maintain file data and meta-data
- ❖ Uses outgoing HTTP connections only
  - Avoids most of the firewall issues of other network file systems.
  - Transfers data and meta-data on demand and verifies data integrity by cryptographic hashes.
- ❖ Aggressive caching and reduction of latency, focuses specifically on the software use case

# CernVM-Fs -Stratum 0

- ❖ CernVM-FS is a file system with a single source of (new) data
- ❖ This single source, the repository Stratum 0, is maintained by a dedicated release manager machine or publisher.
- ❖ A read-writable copy of the repository is accessible on the publisher.
- ❖ The CernVM-FS server tool kit is used to publish the current state of the repository on the release manager machine.
- ❖ Publishing is an atomic operation
- ❖ All data stored in CernVM-FS have to be converted into a CernVM-FS repository during the process of publishing.

# CernVM-Fs -Stratum 1

- ❖ While a CernVM-FS Stratum 0 repository server is able to serve clients directly, a large number of clients is better served by a set of Stratum 1 replica servers.
- ❖ Multiple Stratum 1 servers improve the reliability, reduce the load, and protect the Stratum 0 master copy of the repository from direct accesses.
- ❖ Stratum 0 server, Stratum 1 servers and the site-local proxy servers can be seen as a Content Distribution Network (CDN).



# Scale of Deployment

