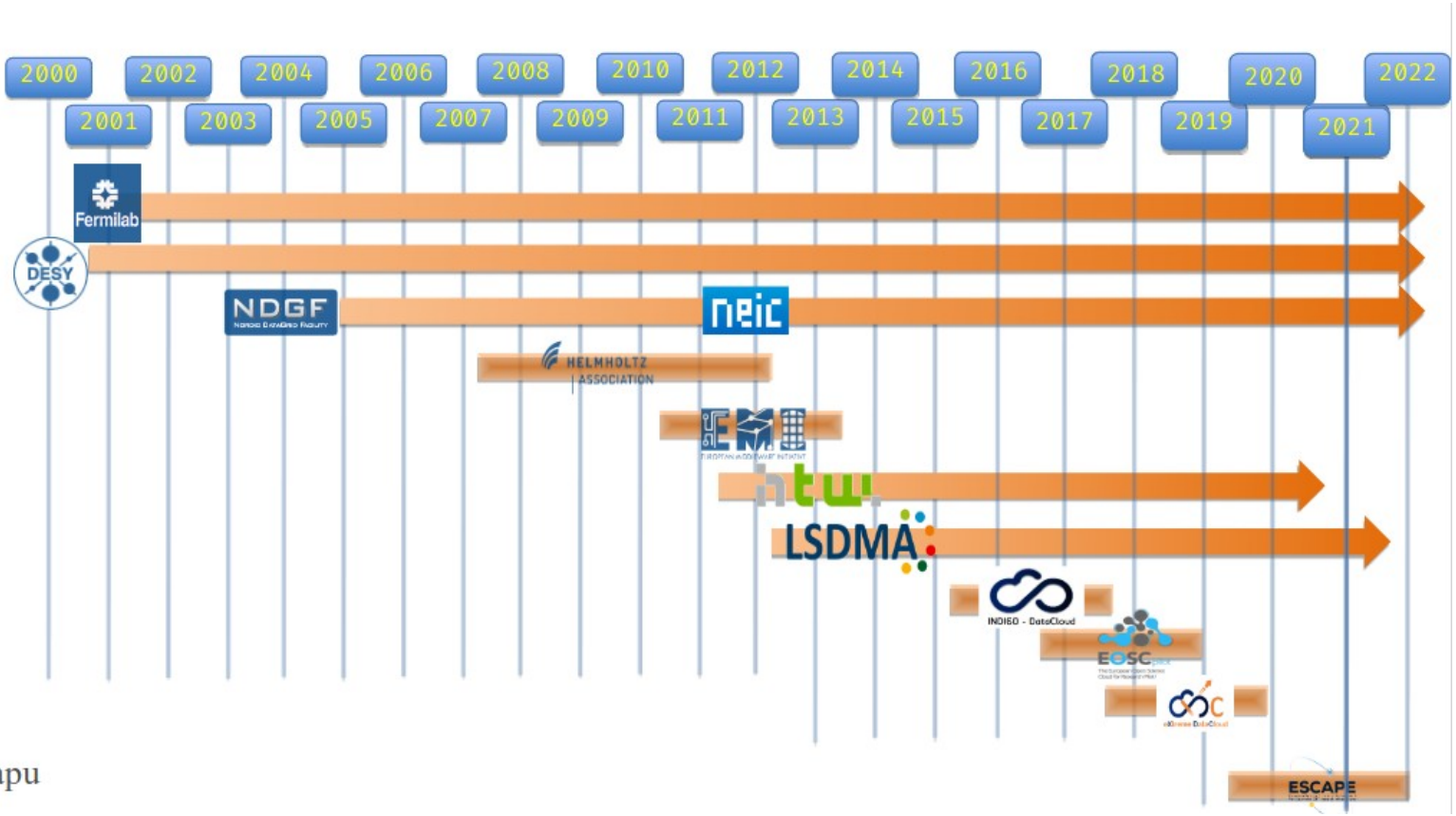
dCache – Inter-disciplinary storage system

Webinaire RI3 – Adrien Georget

dCache history

Born in September 2000 at DESY, designed at first for HEP community
Open source project developed in Java

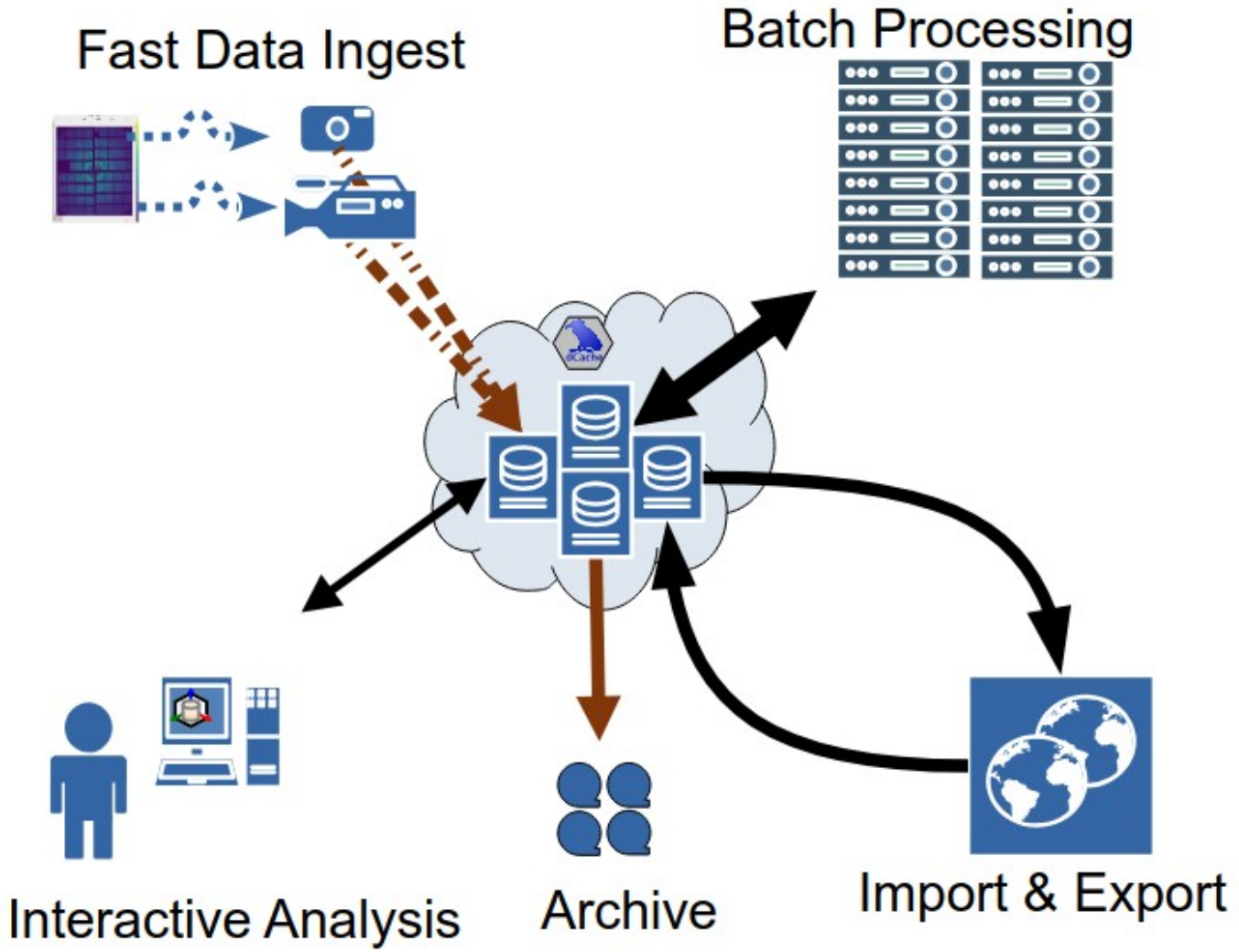
- **DESY**
  - Leo Eckert
  - Svenja Meyer
  - *Paul Millar**
  - Tigran Mkrtchyan
  - Lea Morschel
  - Marina Sahakyan
- **FermiLab**
  - Dmitry Litvintsev
  - Albert Rossi
- **NeIC**
  - Krishnaveni Chitrapu



dCache deployments around the world

dCache goals

- Ingest
  - Multiple parallel streams.
  - High data rate.
  - Large number of files.
- Analysis
  - High CPU efficiency.
  - Chaotic user access.
  - Standard access protocols.
- Sharing&Exchange
  - Effective WAN access.
  - In-flight data protection.
  - Federated Identity handling.
- Long-Term Archival
  - High reliability.
  - Automatic technology migration.



Key features

- Capable of managing petabytes of data
- Fault tolerance against server failures (HA)
- Support for commodity servers
- Easy scalability by adding new pool nodes
- Transparent data distribution and replication across multiple nodes
- Fine-grained authorization with POSIX file permissions and NFS-style ACLs
- Quality of Service (QoS) management
- HSM integration
- dualstack IPv4/v6

Core components

Doors

User Protocol-specific entry point

Pools

Handles data storage

PoolManager

Manages and configures pools, handling data flow and pool selection

PnfsManager

Interface to the namespace

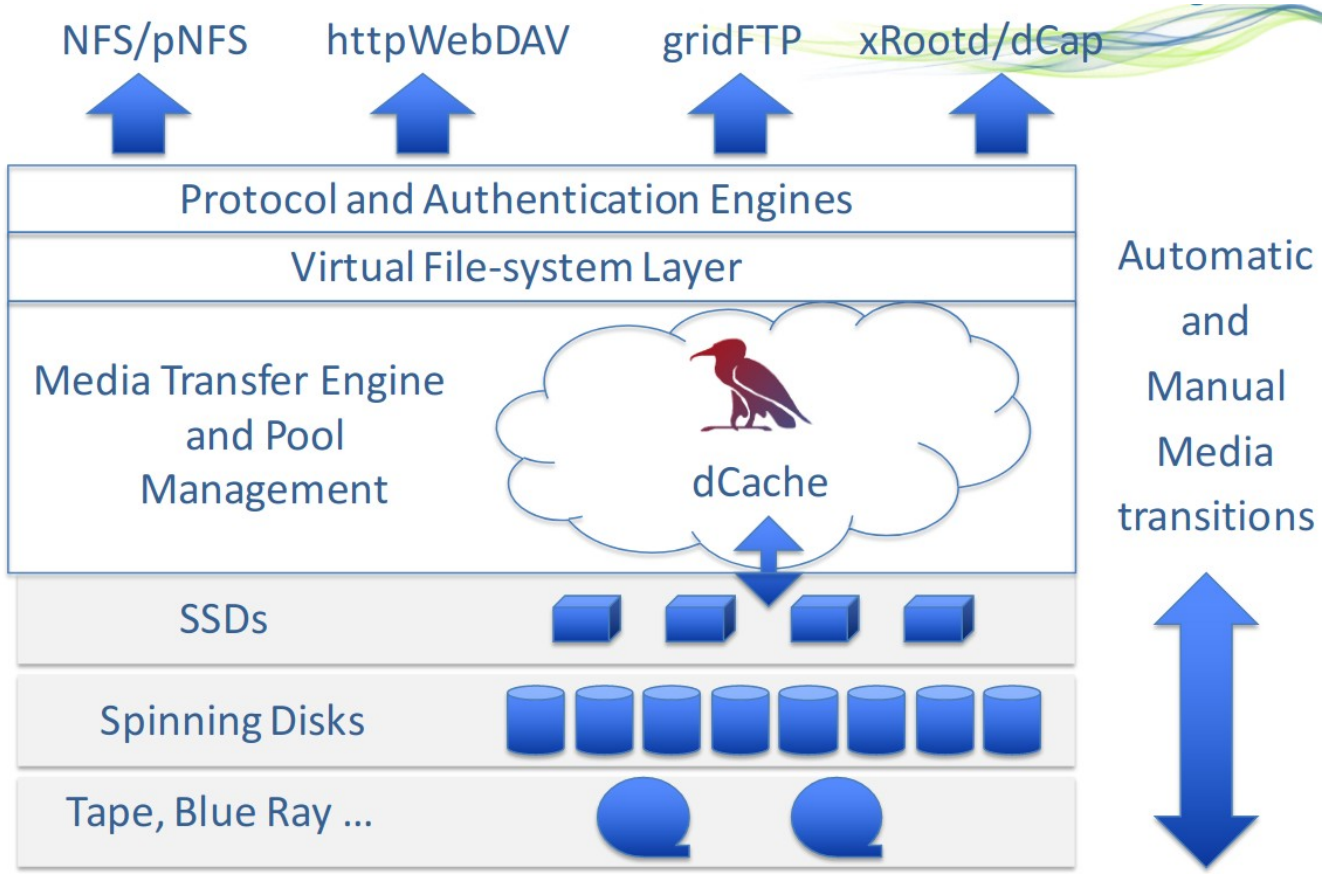
Admindoor

Administration interface

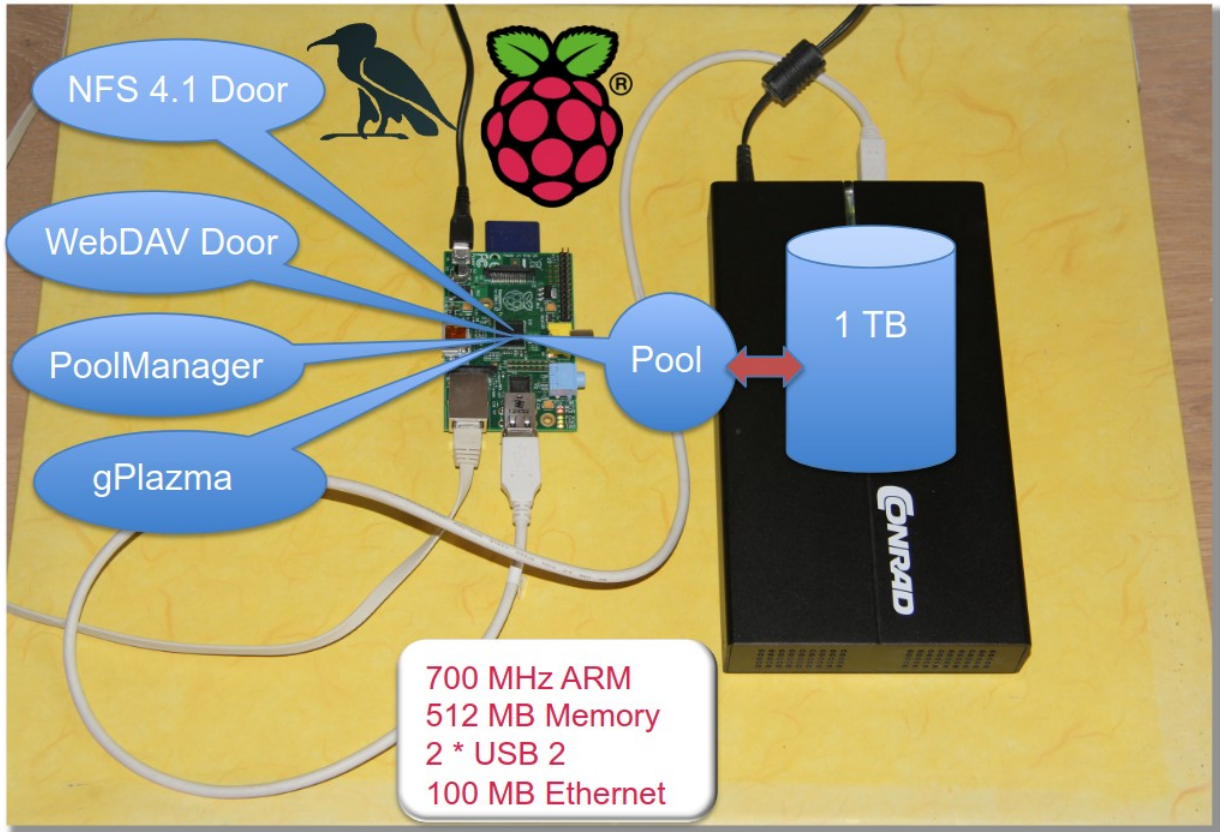
Databases

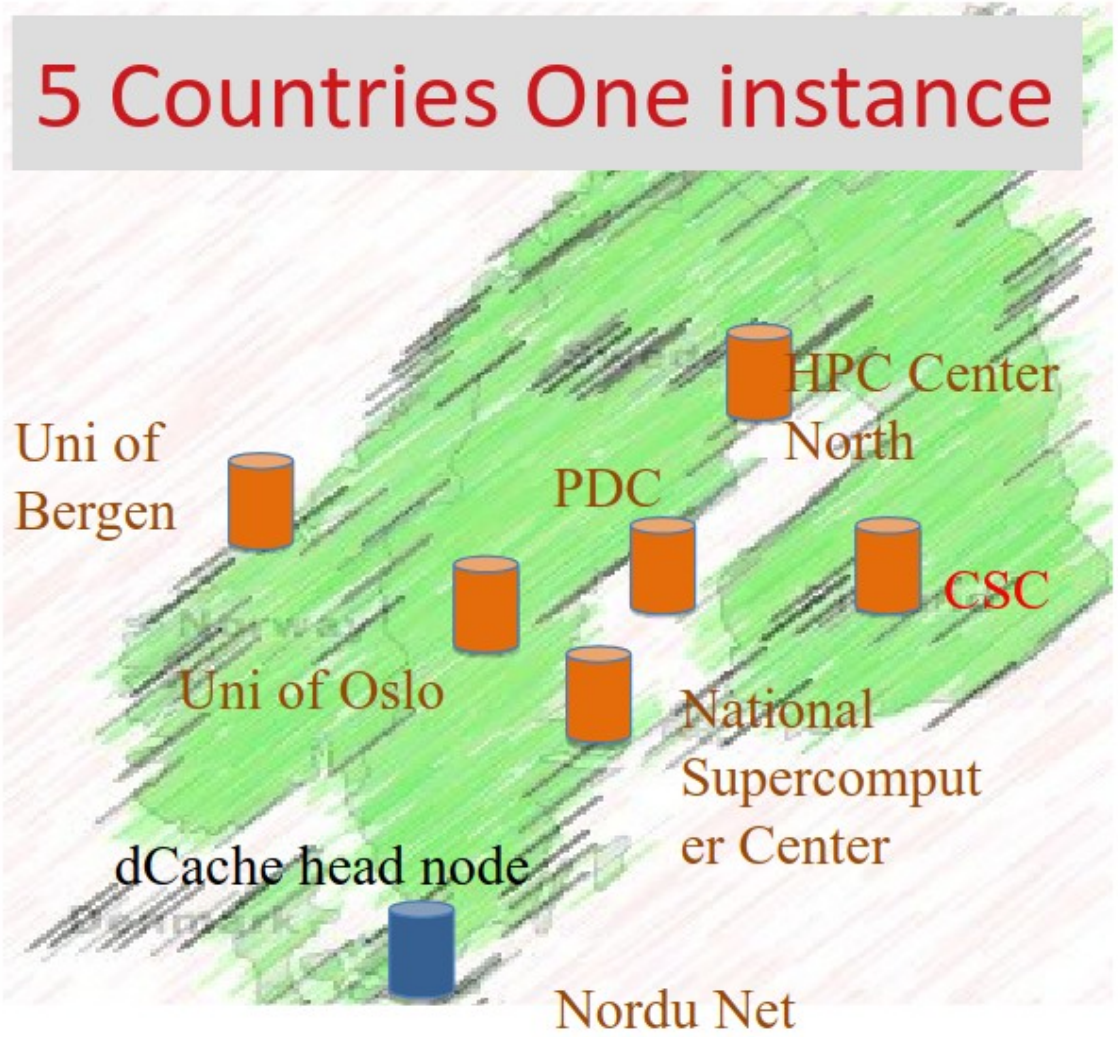
metadatas

space management, quotas, pins, billing, bulk



From tiny to huge





Requirements

For headnodes

PostgreSQL (metadatas, spacemanager DBs)

Apache Zookeeper

For poolnodes

mounted FS (xfs, ...)

Minimum System Requirements

- Hardware:
  - Contemporary CPU
  - At least 1 GiB of RAM
  - At least 500 MiB free disk space
- Software:
  - OpenJDK 11
  - Postgres SQL Server 13.0 or later
  - ZooKeeper version 3.5 (embedded)

- Minimal dCache installation guide :

  https://github.com/dCache/dcache/blob/master/docs/TheBook/src/main/markdown/dcache-minimal-installation.md

In production since 2005 @CC-IN2P3 from 35TB to 65 PB

IN2P3dCachesetup

for the Tier II dCache workshop, June 2006
by Lionel Schwarz, IN2P3

1. Head node setup

Right now all head node services are located on a single machine which is a (V20Z bi-opteron 2GHz, 2GB RAM). There are plans to separate the pnfs server and its DB to another host, same hardware. The backup is done once a day with pg_dump and saved to our TSM backup system.

2. Pool Nodes

We have 13 disk servers in dCache serving about 35TB. We use various disk configuration (direct attached disk/disk array) on various hardware (Transtec bi-Xeon 4GB RAM/V40Z quad-pro 8GB RAM...). All nodes are installed under SL3. We plan to install nodes under SL4 and Solaris10 in the future. All nodes have 2 1Gb interface, 1 on the outside and 1 on the inside (workers and HPSS connection), so that GridFTP traffic does not mix with migration/stage.

3. Installation

All installations/upgrades are done manually. We plan to use some automatic tools like yaim in the future.

IN2P3 experience

Conclusions

- We are confident in using dCache in production but aware of the amount of work in order to reach a good level of administration
- We feel that we would benefit from having the sources...
- We are ready to participate to any effort (development, tools, documentation...)
- Suggestions
  - Have a dCache admins mailing list
  - Have a regular (once a year?) meeting with dCache developers and admins (and users?)

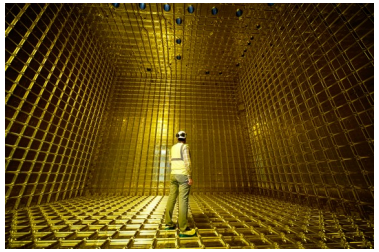
dCache workshop – 30 Aug – 1 Sept 2005 11

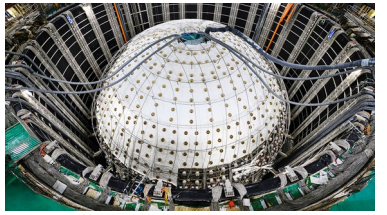
3 instances

- **LCG (Atlas / CMS / LCHb)**
  - 43PB / 123M files
  - 165 servers (Dell R740XD2, HPE Apollo 4200)
  - weakly : 3PB imported, 5PB exported, 4PB read analysis
  - up to 300TB staged from tapes per day

- **Rubin Observatory (LSST)**
  - 18PB / 258M files
  - 65 servers
  - 2500 images per night (20TB), +5PB per year

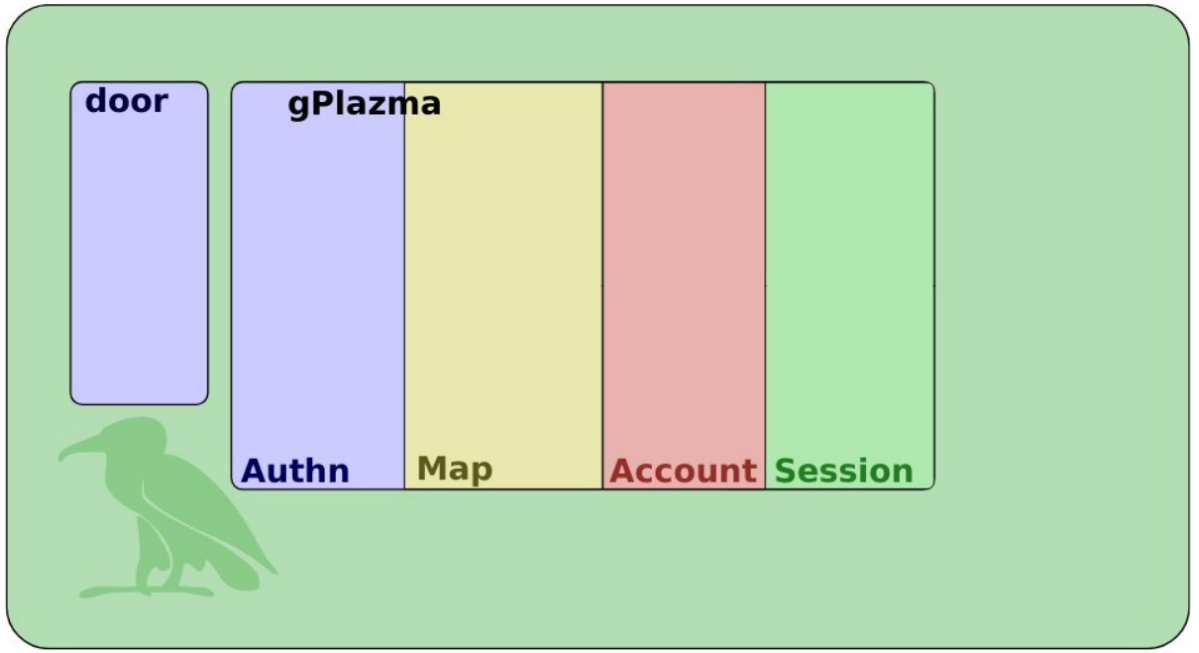
- **EGEE (Dune, Belle2, Juno, Xenon, ...)**
  - 2.5PB / 36M files
  - 13 servers





Authentication

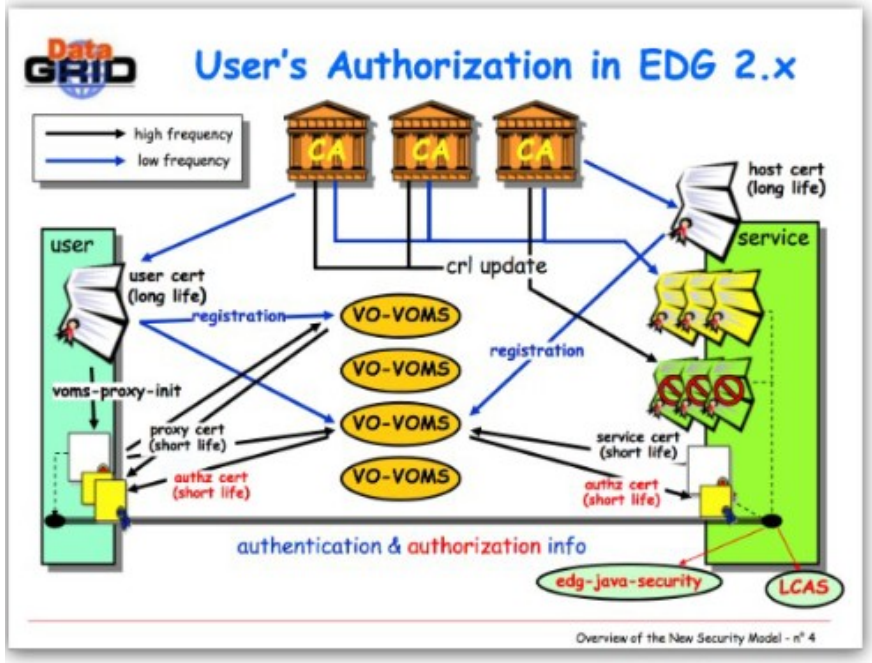
- dCache auth plugins :
  - X509
  - voms
  - kpwd
  - htpasswd
  - jaas
  - oidc

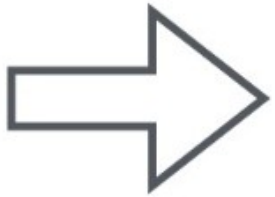


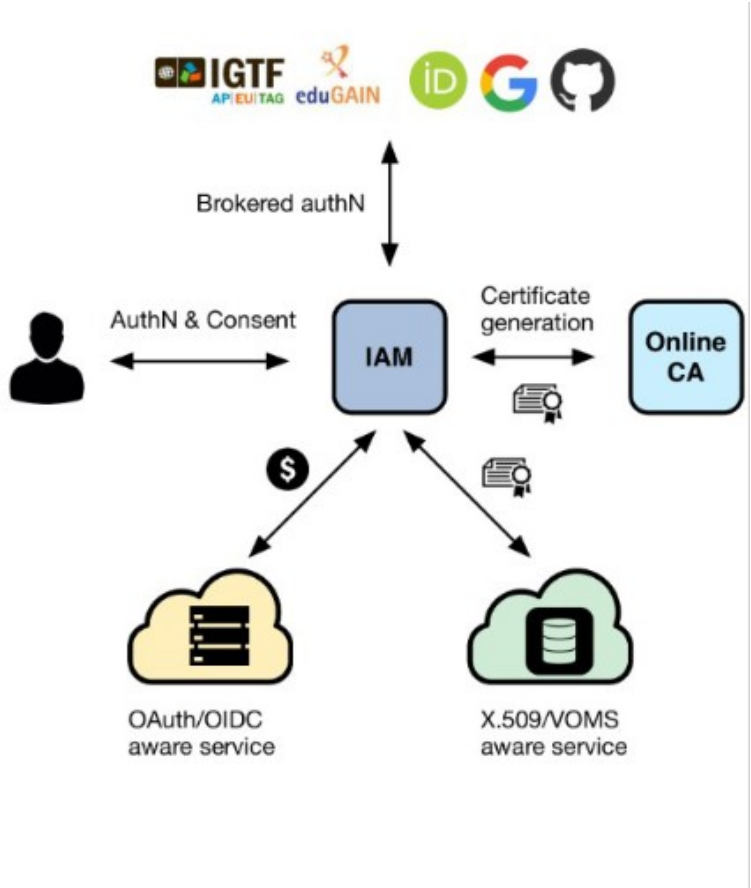

```
auth     optional     x509
auth     optional     voms
auth     optional     oidc
map      optional     vorolemap
map      sufficient   multimap gplazma.multimap.file=/etc/dcache/multi-mapfile.wlcg_jwt
map      optional     authzdb
session  requisite    authzdb
```

Authentication

- **VOMS -> OIDC**







Multi protocols

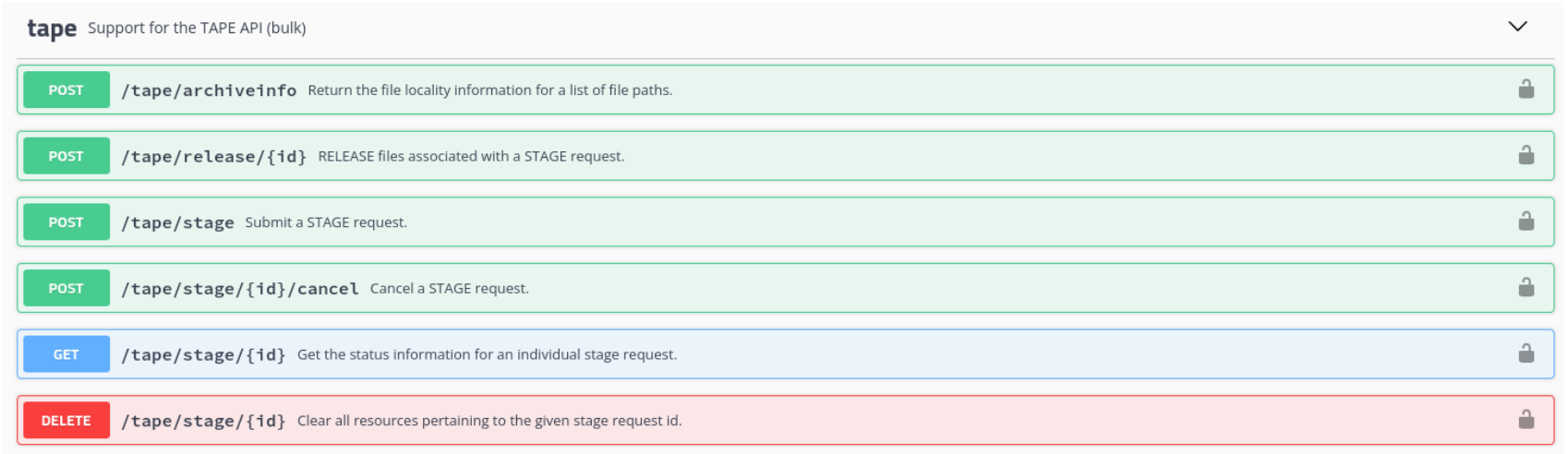
dCache access protocols

- ~~GsiFTP~~
- HTTP/WebDAV
- XRootD
- NFSv4.1

Tape interface

Native to dCache

- dCache = disk cache on front of tape
- The essential part of the dCache design
- Transparent for the users
- Stage Protection
- Supports multiple HSM on a single instance
- Tape REST API



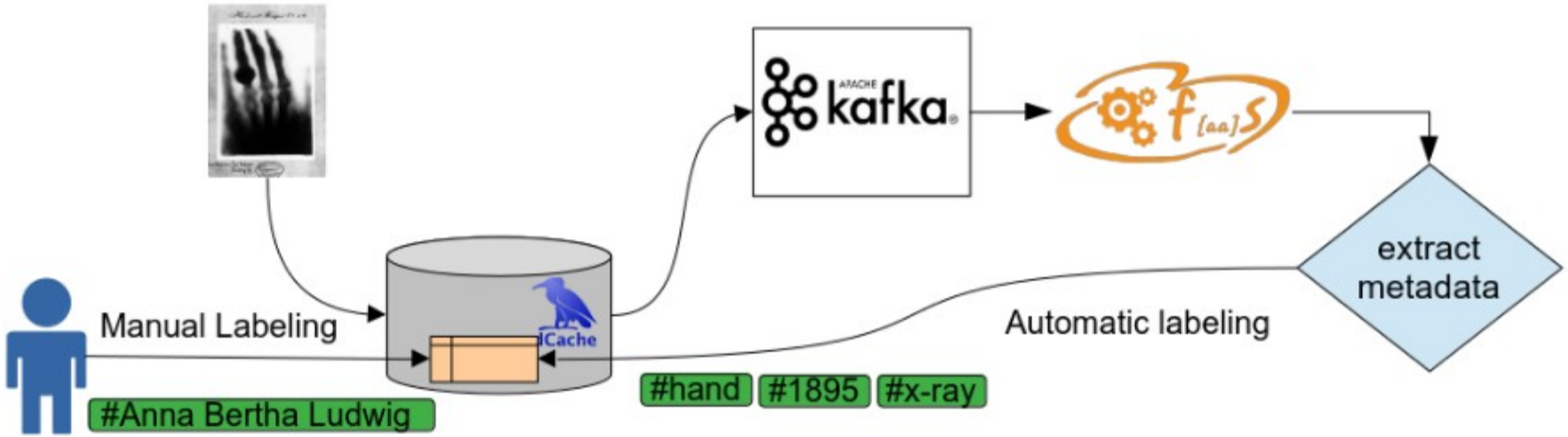
QoS

- **HEP : Single copy (tape or disk)**
- **Cloud : 2 disk copy + 1 tape**
- **Double copy : 2 tapes on different media types**
- **...**

```
" name ": "my - policy ",
" states ": [
 {
   " duration ": " P10D ",
   " media ": 2x DISK
 },
 {
   " duration ": " P1M ",
   " media ": 1x DISK , 1x HSM
 },
 {
   " media ": 2x HSM
 }
]
```

User Metadata/Labeling

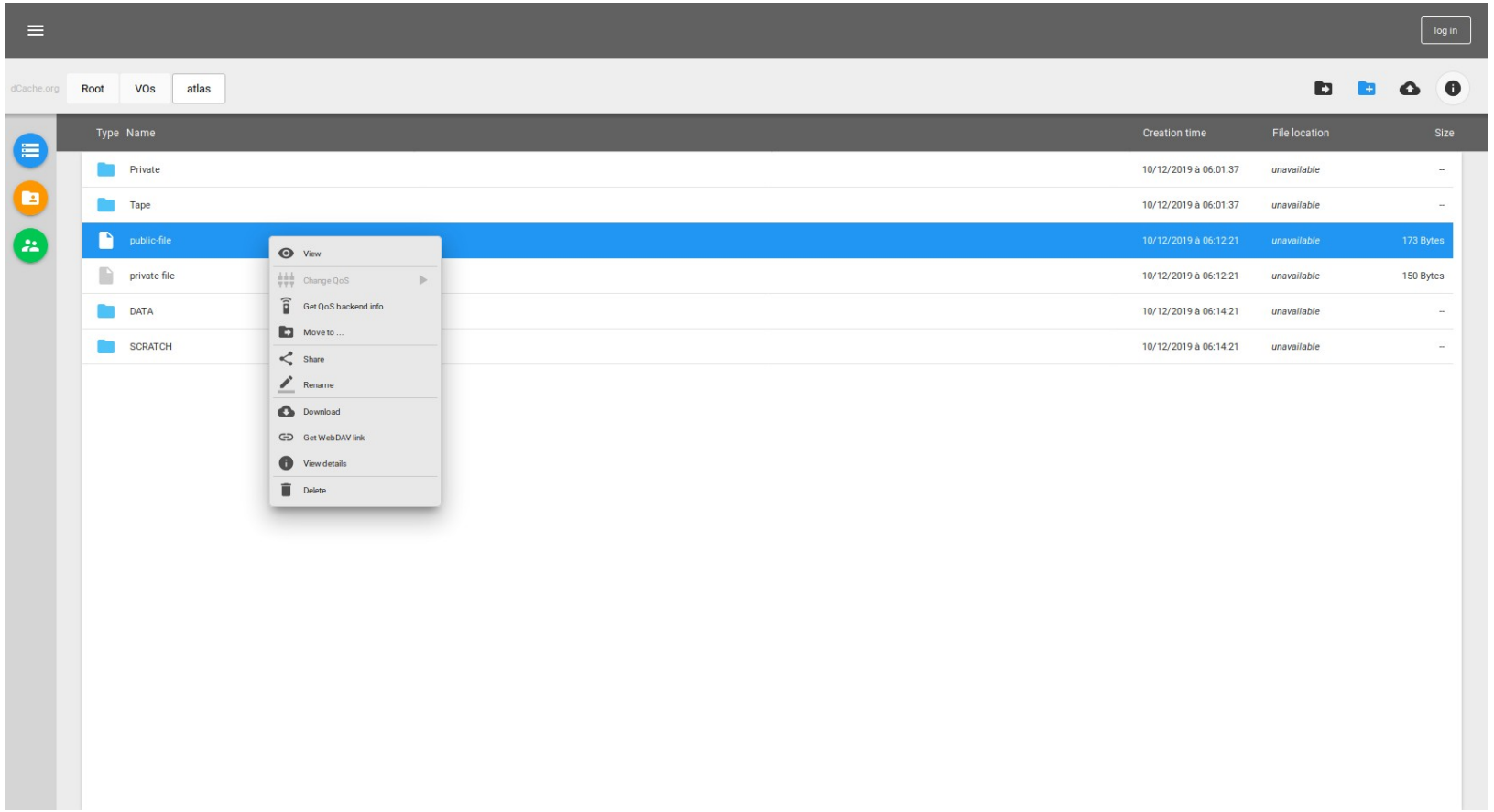
- **Extended attributes for files/directories**
  - Exposed via NFS, WebDAV, REST
  - Labels can be attached to a file
  - Query files by label using the REST API



Monitoring

- **dCache Web interface**
  - **frontend**
  - **alarms**
  - **RESTful API**

- **Billing log files**
- **Billing database**
- **Kafka events**



Monitoring

namespace Files, directories and similar objects

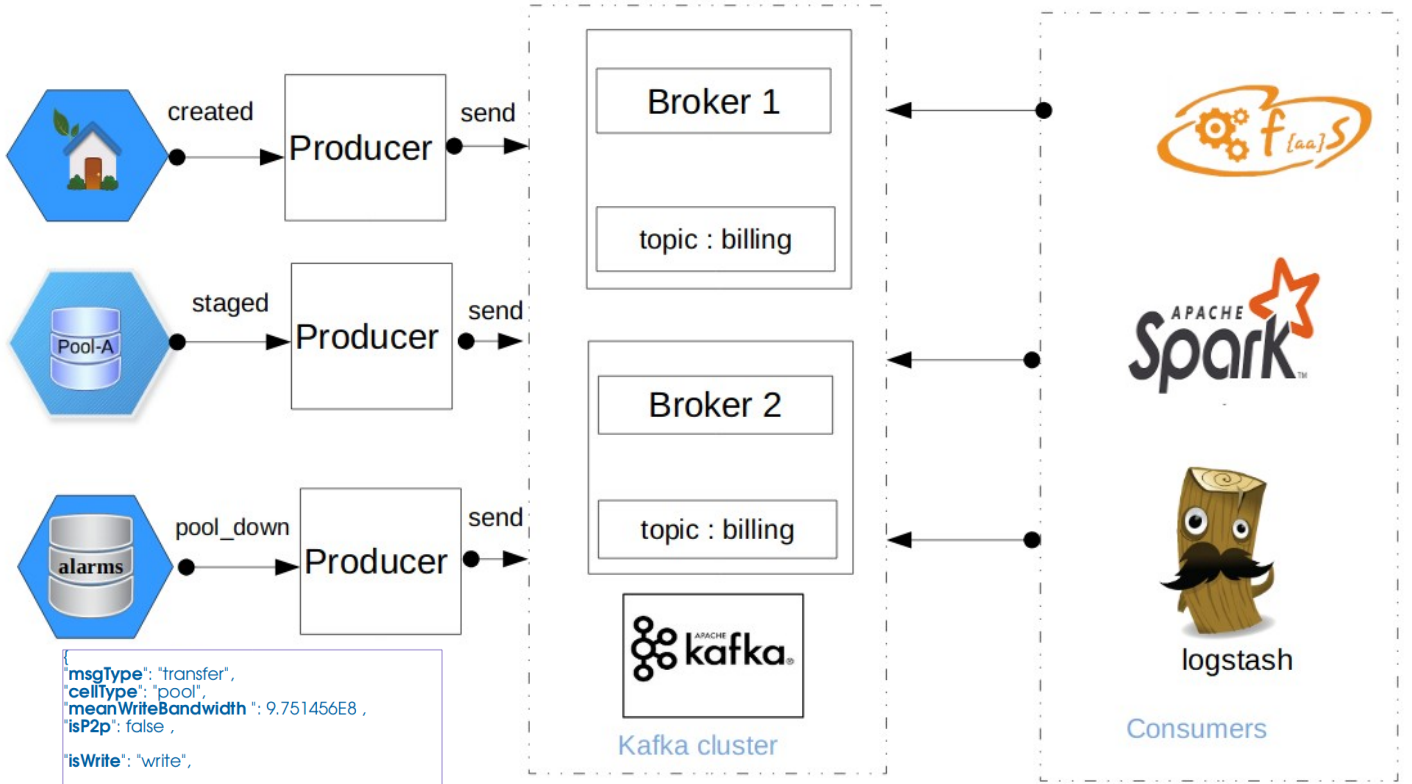
GET	`/namespace/{path}`	Find metadata and optionally directory contents.
POST	`/namespace/{path}`	Modify a file or directory.
DELETE	`/namespace/{path}`	delete a file or directory
GET	`/id/{pnfsid}`	Discover information about a file from the PNFS-ID.
GET	`/qos-policy/stats`	Retrieve the current count of files in the namespace by policy and state.
GET	`/qos-policy/id/{id}`	Retrieve the QoSPolicy name and status for this file pnfsid.
GET	`/qos-policy/path/{path}`	Retrieve the QoSPolicy name and status for this file path.

poolmanager Data placement and selection decisions

GET	`/poolgroups/{group}`	Get information about a poolgroup.
GET	`/poolgroups`	Get a list of poolgroups. Results sorted lexicographically by group name.
GET	`/poolgroups/{group}/pools`	Get a list of pools that are a member of a poolgroup. If no poolgroup is specified then all pools are listed. Results sorted lexicographically by pool name.
GET	`/poolgroups/{group}/usage`	Get usage metadata about a specific poolgroup.
GET	`/poolgroups/{group}/queues`	Get pool activity information about pools of a specific poolgroup.
GET	`/poolgroups/{group}/space`	Get space information about pools of a specific poolgroup.

Kafka

- **A message broker system allowing services to send and receive messages**
- **load balancing and resilience**
- **Widely used in the IT industry and scales very well**



dCache

Transfers monitoring Kafka+Opensearch

LCG write

Events	Write	rate /s
268,111	293.16TB	3.45GB

LCG read

Events	Read	rate /s
128,732	498.47TB	5.87GB

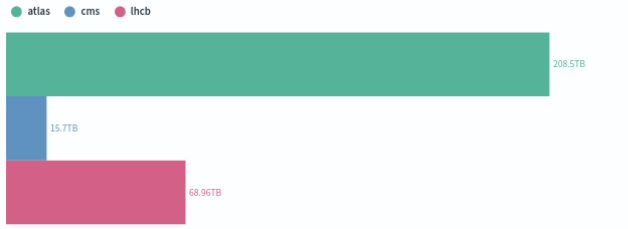
LCG XRootD analysis

Events	XRootD read size	rate /s
974,367	386.4TB	4.5GB

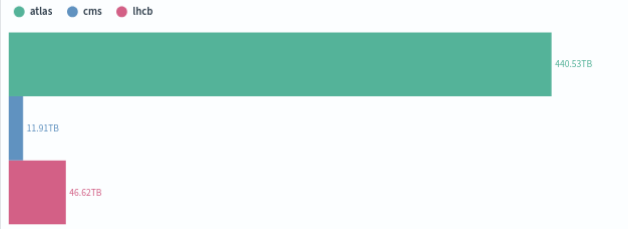
LCG Pool2Pool

Events	P2P volume
220,033	512.72TB

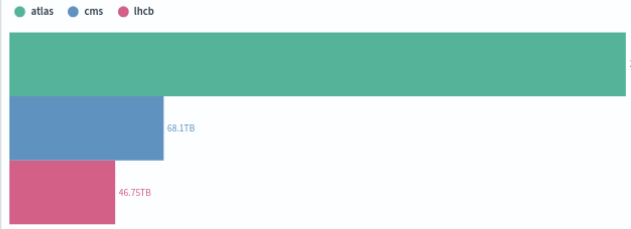
LCG write volume / VO



LCG read volume / VO



LCG XRootD analysis



LCG deletion

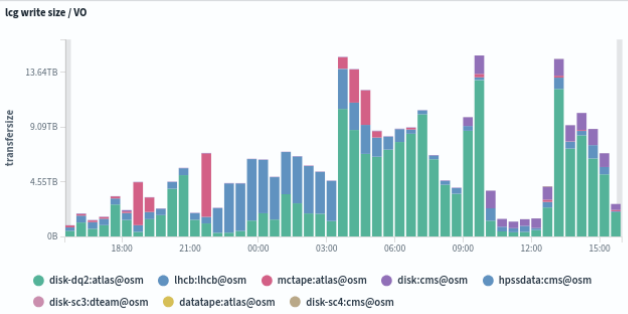
Events	Deletion Size
330,086	318.7TB

LCG hsm migration details

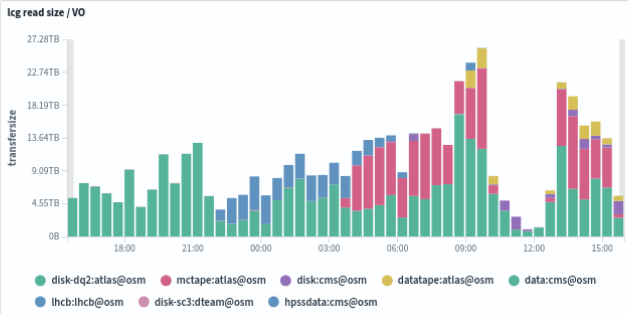
storageInfo: Desc...	Count	migration volume
mctape:atlas@osm	23,759	18.9TB
lhcb:lhcb@osm	2,244	5.6TB
datatape:atlas@osm	29	31.5GB
	26,032	**24.6TB**

1

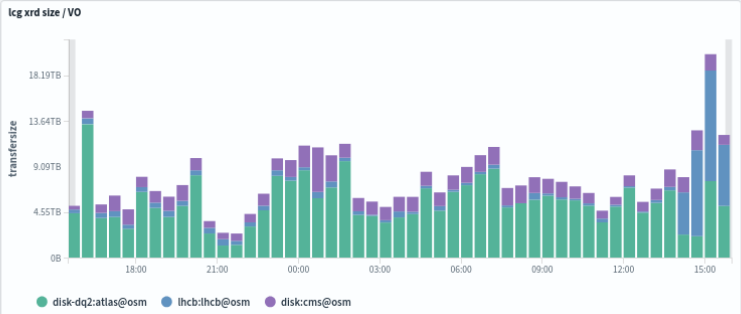
lcg write size / VO



lcg read size / VO



lcg xrd size / VO

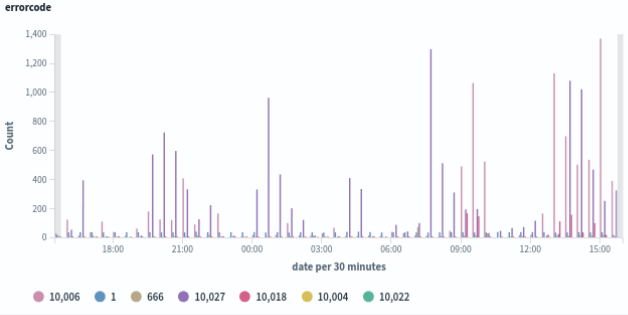


LCG hsm staging details

storageInfo: Desc...	Count	staging volume
mctape:atlas@osm	12,905	93.4TB
lhcb:lhcb@osm	10,225	46.5TB
datatape:atlas@osm	677	3.7TB
	23,848	**143.8TB**

1

errorcode



error table

error code	error msg	Count
10,006	No connection from client after 300 s	6,476
10,027	rejected PUT: 507 UNKNOWN_RESPO	3,345
10,006	No connection from client after 5 sec	2,012
10,027	rejected PUT: 501 NOT_IMPLEMENTE	1,715
10,018	File not online. Staging not allowed.	696
10,006	Request to [>PoolManager@dCacheD	411
10,027	rejected GET: 500 Server Error	333

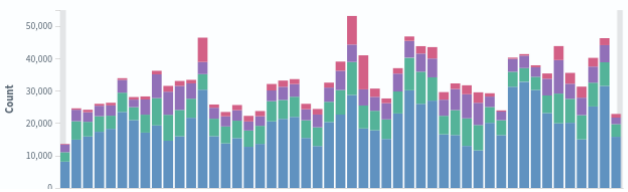
Top 10 DN

DN	Count
atlagrid	235,086
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo2/CN=663	182,536
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614	175,402
cmsgrid	173,970
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=chiw/CN=87859	124,940
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atulupov/CN=84	119,162
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=amaltaro/CN=71	106,699

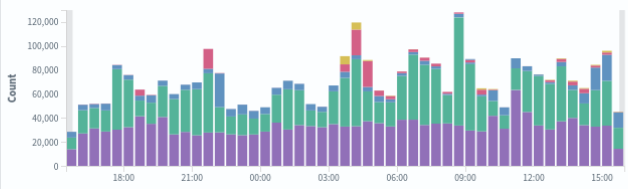
LCG hsm errors count

store - errors	restore - errors
1,775	1

lcg event by protocol



LCG event / VO



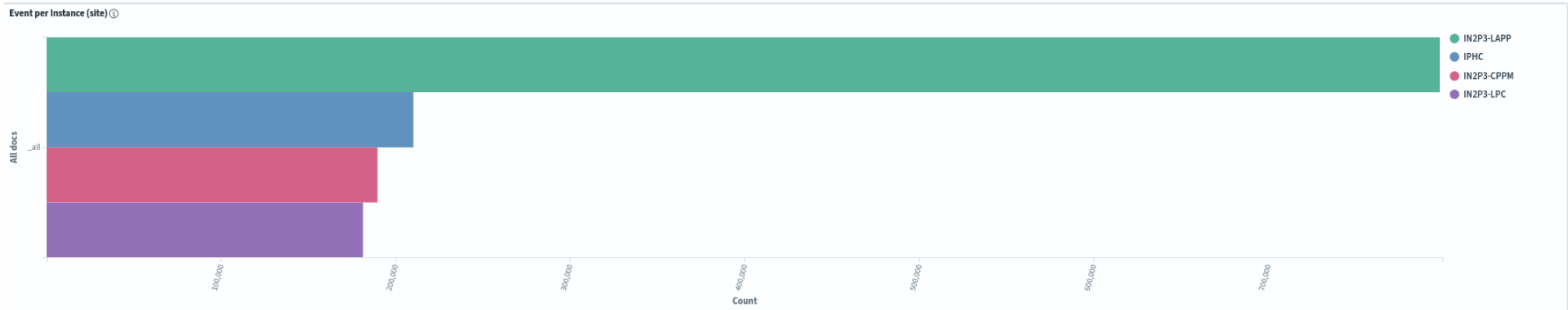
LCG Top 10 clients

Host	Count
2620:6a:0:8420:0:0:0:132	11,193
2001:1458:d00:24:0:0:100:1b	11,031
2001:638:50a:125:1:7cd8:0:1	10,222
2001:660:5009:9:193:48:99:6	10,175

Top 10 doors

Door	Count
xrootd-ccdcacli432Domain	252,296
xrootd-ccdcacli433Domain	220,865
xrootd-ccdcacli423Domain	219,495
webdav-ccdcatli367Domain	206,349
xrootd-ccdcacli521Domain	118,937
webdav-ccdcatli406Domain	114,136
webdav-ccdcacli537Domain	113,447

Monitoring Kafka+Opensearch multi sites



LCG deletion

230,827	48.8TB
Events	Deletion Size

Top 10 doors

Door	Count
doorsDomain	400,049
lapp-dccentral01_doorsDo	196,089
lapp-dp3s03_doorsDomai	7
lapp-dp10_doorsDomain	6
lapp-dp16_doorsDomain	6

LCG write

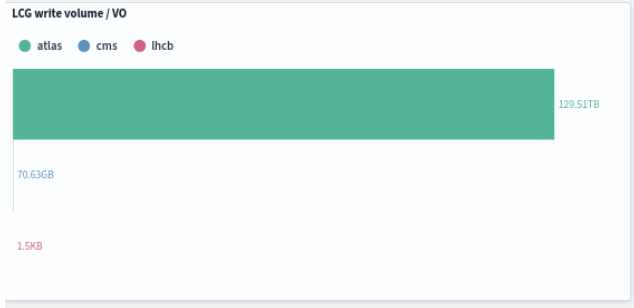
96,531	129.58TB	1.53GB
Events	Write	rate /s

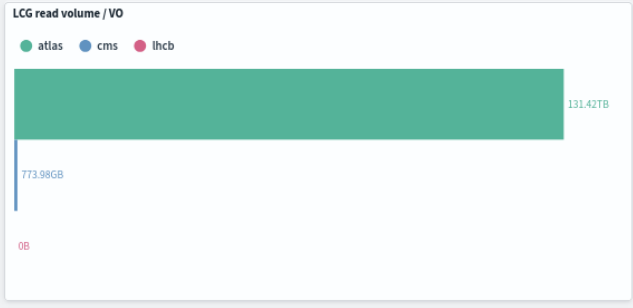
LCG read

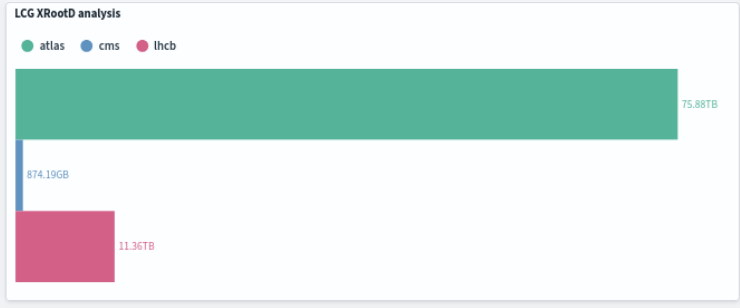
72,509	131.85TB	1.55GB
Events	Read	rate /s

LCG XRootD analysis

337,558	125.2TB	1.5GB
Events	XRootD read size	rate /s

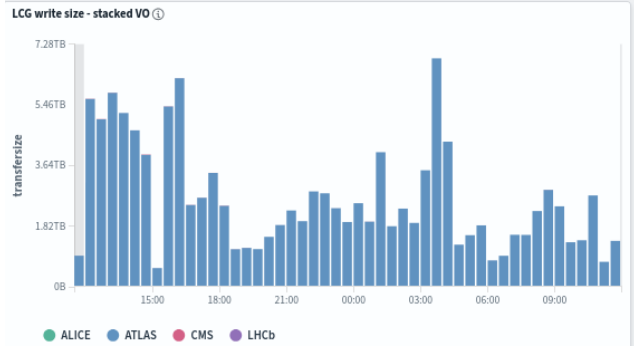


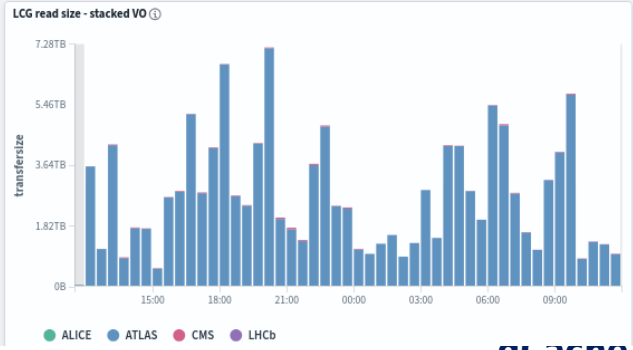


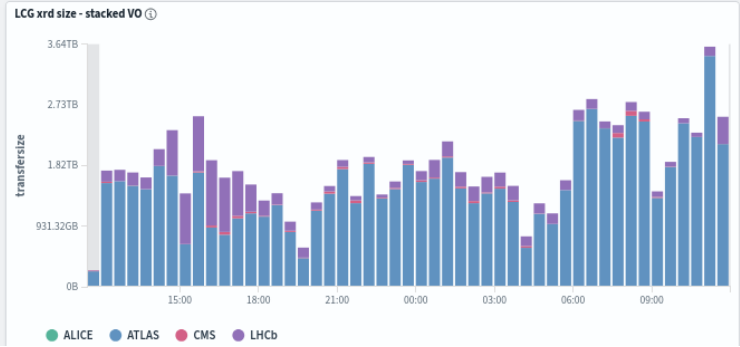


LCG Top 10 pools

Pool	Count
DP_06-ATLAS-1	305,908
DP_06-ATLAS-2	64,837
DP_37-ATLAS-2	6,530
Sbgpool1_201	6,509
DP_302-OTHER-2	6,083
DP_09-ATLAS-1	6,048
DP_36-ATLAS-2	6,020
DP_37-ATLAS-1	5,879
DP_12-ATLAS-2	5,835
DP_12-ATLAS-1	5,759



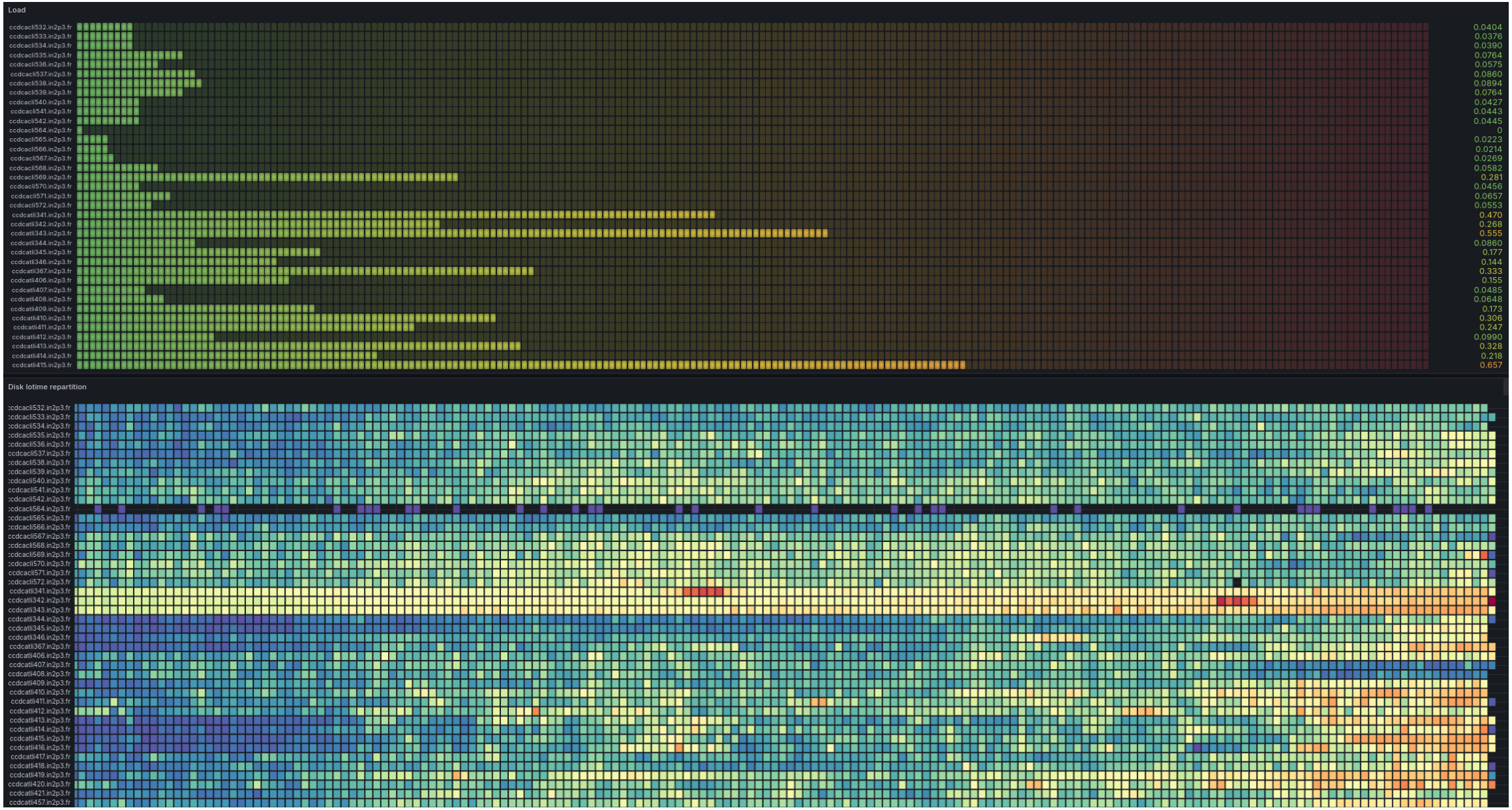




dCache

Global monitoring

Load monitoring



Deployment

- dCache soft/conf with Ansible playbook
- ecosystem with Puppet (sys conf, grid, probes)

Download	Rel. Date	md5 hash	Release Notes
dCache 10.2.10 (Debian package)	25.02.2025	29d0269388fc04845b9c9bc1381f8d8c	10.2.10
dCache 10.2.10 (rpm)	25.02.2025	deefbefaeb6c64bc40f0521e73e57aea	
dCache 10.2.10 (tgz)	25.02.2025	e2a338f70d7f9fa75021bfc2e456ffff	

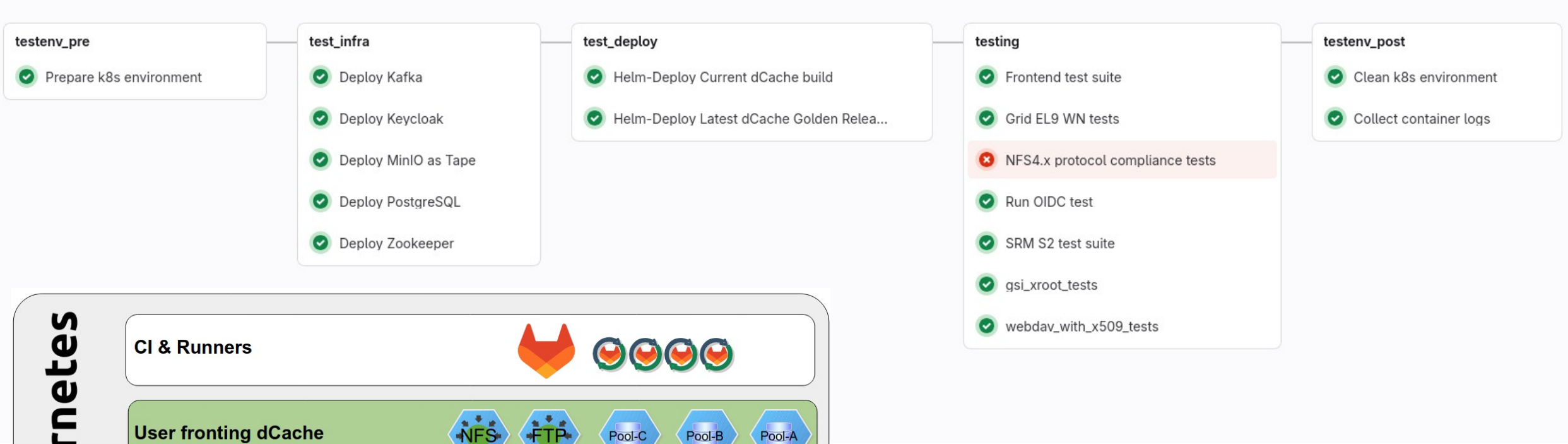
```
# Host specific settings for ccdcatli537
# Pools infos
dcache_poolinfo:
  pool-lhcb-dst:
    poolgroup: "pgroup-lhcb-dst"
    poolname: "pool-lhcb-dst-li537a"
    poolsize: "162000G"
  pool-atlas-dq2:
    poolgroup: "pgroup-atlas-import-disk"
    poolname: "pool-atlas-dq2-li537a"
    poolsize: "6000G"
  pool-cms-hpssdata:
    poolgroup: "pgroup-cms-hpssdata"
    poolname: "pool-cms-hpssdata-li537a"
    poolsize: "8000G"

# Doors infos
dcache_doorinfo:
  webdav:
    doorname: "webdav-ccdcacli537"
    root: "/pnfs/in2p3.fr/data/cms/"
    tag: "glue storage-descriptor"
```

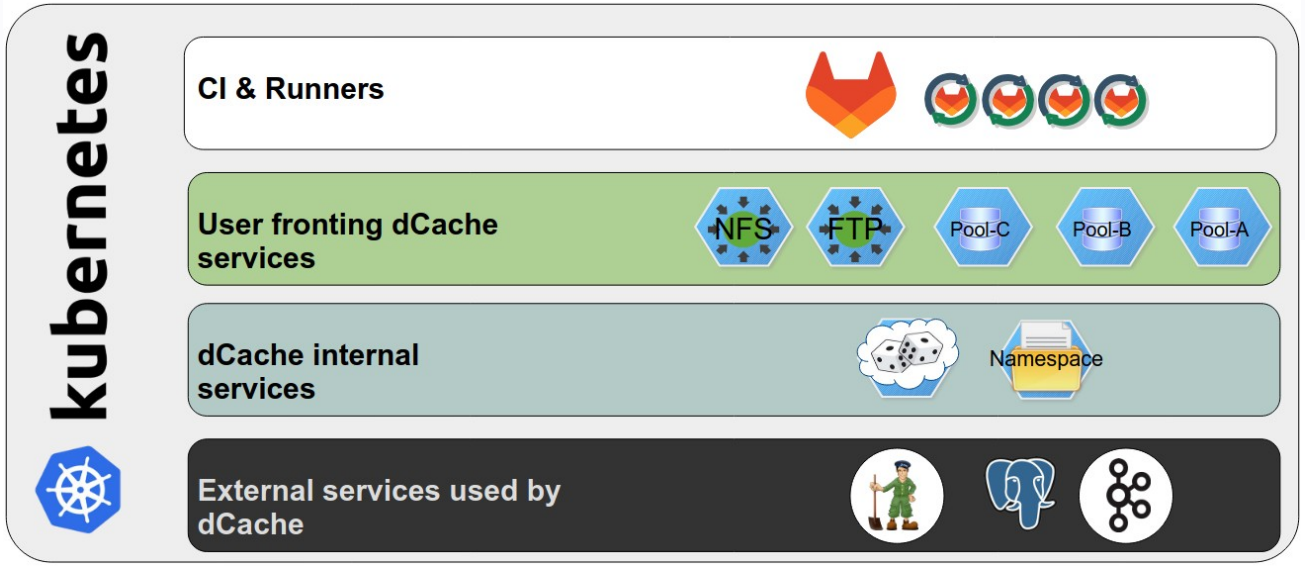
- dCache on Kubernetes (Helm Chart)
  https://github.com/dCache/dcache-helm

```
$ helm install chimera bitnami/postgresql
$ helm install cells bitnami/zookeeper
$ helm install billing bitnami/kafka
$ helm install –set image.tag=9.2.33 my-store dcache/dcache
```

Testing

- **dCache CI**
  https://github.com/dCache/dcache/blob/master/.gitlab-ci.yml





More ?

- **dCache international workshop host by CC-IN2P3 in May 20-21**

  https://indico.desy.de/event/48191/

19th International dCache Workshop

May 20–21, 2025
IN2P3 computing center
Europe/Berlin timezone

Enter your search term

Overview
Accomodation
Transportation

Info & Support
workshop@dcache.org

The 19th International dCache workshop 2025 will be held in person from 2025-05-20 to 2025-05-21 and hosted by IN2P3 in Lyon. As with earlier workshops, the dCache team is eager to maintain and strengthen its relationship with dCache system administrators, experienced or novice. Contributions to the workshop will focus on presenting mechanisms helping sysadmins run secure and fault-tolerant dCache systems.


THANK YOU
DCACHE
DCACHE
DCACHE
DCACHE
DCACHE

- More info: https://dcache.org
- Steal and contribute: https://github.com/dCache/dcache
- Help and support: supportϕdcache.org, user-forumϕdcache.org
- Developers: devϕdcache.org