What's new in dCache-9.2

Tigran Mkrtchyan for dCache team

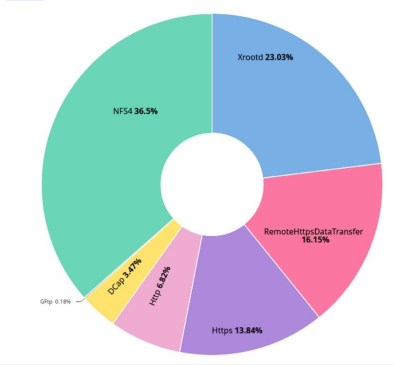
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

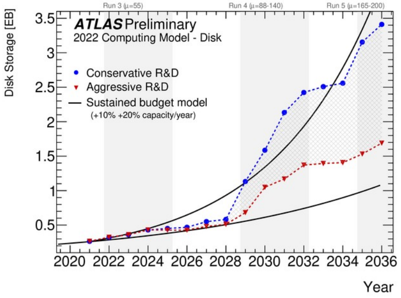
Golden Release (or LTS)

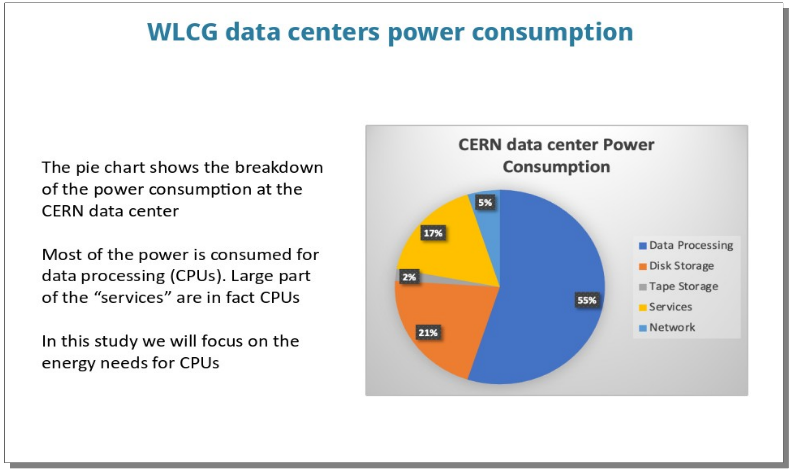
- 2 years support
  - Bug fixes and important fixes
  - All
- Compatible with previous two LTS versions
  - 7.2 pool can work with 9.2 core services
  - (sometime we break it, sorry) 😳

The Challenges

- Data is going to grow… A lot…
  - High ingest data rates
  - More movements between sites
- Shared Computing Resources
  - Analysis Facilities
  - Grid Farms
  - HPC
  - Cloud resources (CPU&Storage)
- Standard analysis tools
  - ROOT
  - Jupyter Notebooks, non-ROOT analysis
- Competing Tape Operations







WLCG data centers power consumption

The pie chart shows the breakdown of the power consumption at the CERN data center

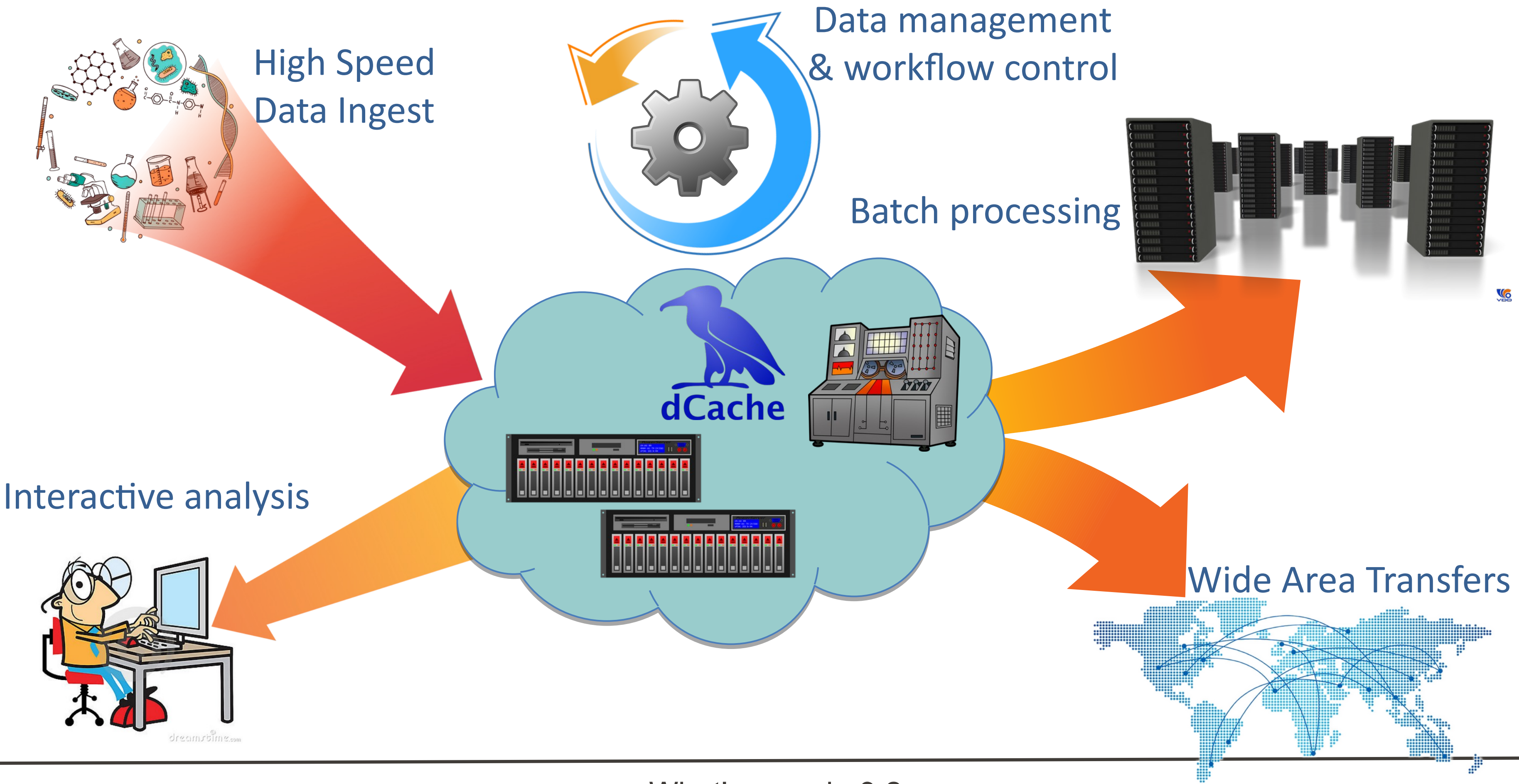
Most of the power is consumed for data processing (CPUs). Large part of the "services" are in fact CPUs

In this study we will focus on the energy needs for CPUs

CERN data center Power Consumption
5%
17%
2%
21%
55%
Data Processing
Disk Storage
Tape Storage
Services
Network

Prominent Changes

- QoS & BULK Service
- TPC improvements
- NFSv4.1/pNFS improvements
- XROOT evolution (TLS, tokens, TPC, proxy-IO)
- Namespace performance improvements
- HSM connectivity



High Speed Data Ingest

Data management & workflow control

Batch processing

dCache

Interactive analysis

Wide Area Transfers

POSIX Constraints

- According to POSIX standard, on new file system object creation the parent directories *modification time* should be updated.
- To track the directory changes that happen at a higher rate than the precision of mtime attribute Linux kernel has an additional attribute *iversion* that is incremented whenever the inode's data is changed.
- To reduce unnecessary directory listing requests to the servers, the NFSv4 clients utilize the *iversion* attribute to identify the directory content changes and use the locally cached copy of the directory entry list as long as last known *iversion* attribute value matches the remote one.

Tunable Consistency

Consistency	Behavior
strong	A creation of a filesystem object will right away update parent directory's mtime, ctime, nlink and generation attributes (POSIX).
weak	A creation of a filesystem object will eventually update (after 30 seconds) parent directory's mtime, ctime, nlink and generation attributes. Multiple concurrent modifications to a directory are aggregated into a single attribute update (near-POSIX).
soft	Same as the *weak*, however, reading of directory attributes will take into account pending attribute updates (POSIX).

```
Benchmark     (wcc)        Score      Error  Units
createFile     weak   14791.269 ± 1287.317  ops/s
createFile   strong     203.099 ±   17.556  ops/s
createFile     soft    1955.169 ±  908.004  ops/s
```

Java Flight Recorder

- A profiler built into JVM
- Starting dcache 7.2 attach listener is enabled by default
- Low overhead – can be enabled on production systems
- Starting dcache 9.1 added admin commands to start/stop recording

Java Flight Recorder (9.1+)

```
[dcache-lab] admin > jfr start
enabled with config: default
[dcache-lab] admin > jfr stop
recorded into /tmp/core_xxx.jfr
[dcache-lab] admin >
```

Xroot Improvements

- Proxying through the xroot door
- Relative paths in the xroot URLs
- Resolution of symlinks in paths
- `ls -l` efficiency

Xroot Multi AuthN Support

A single door can now be configured to support all authentication protocols as an ordered chain:

```
     xrootd.plugins=gplazma:gsi,gplazma:ztn,gplazma:none,authz:scitokens
```

This means the door will first tell the client to try *gsi*; if that fails, it will ask for *ztn*; failing that, it will allow anonymous access. *gsi* is tried first so that TLS is not turned on if not requested by the client (whereas it is enforced for *ztn*).

Thus all protocols are supported out of the box, but this configuration can be modified if desired using the property as before.

NOTE: for scitokens authorization, the default

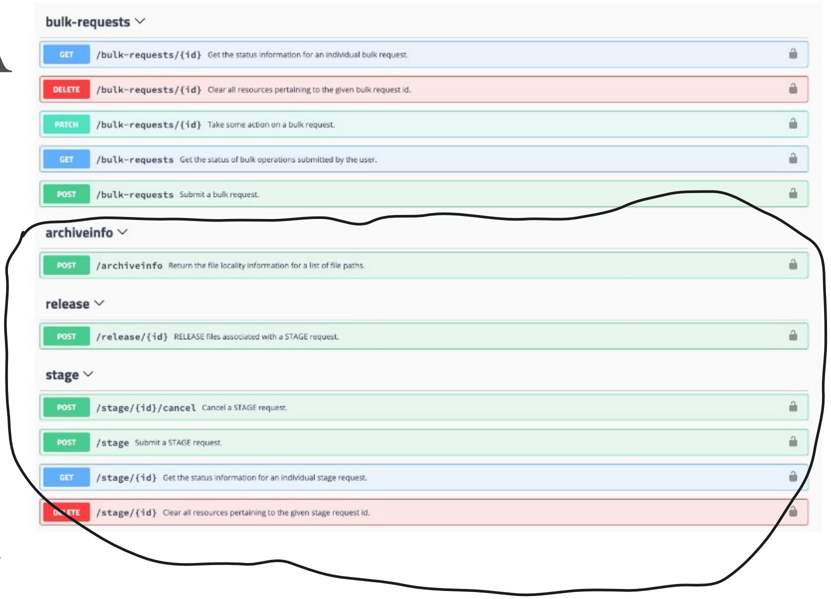
```
   xrootd.plugin!scitokens.strict=false
```

should be used with doors that allow non-token authentication and token-based TPC.

Bulk Service (the backend of tape API)

- Throughput improvements, HA
- Archiving/removing complete requests
- Request statistics
- More options to control default behavior
  - Various request lifetimes

https://example.org:3880/api/v1



bulk-requests

GET `/bulk-requests/{id}` Get the status information for an individual bulk request.

DELETE `/bulk-requests/{id}` Clear all resources pertaining to the given bulk request id.

PATCH `/bulk-requests/{id}` Take some action on a bulk request.

GET `/bulk-requests` Get the status of bulk operations submitted by the user.

POST `/bulk-requests` Submit a bulk request.

archiveinfo

POST `/archiveinfo` Return the file locality information for a list of file paths.

release

POST `/release/{id}` RELEASE files associated with a STAGE request.

stage

POST `/stage/{id}/cancel` Cancel a STAGE request.

POST `/stage` Submit a STAGE request.

GET `/stage/{id}` Get the status information for an individual stage request.

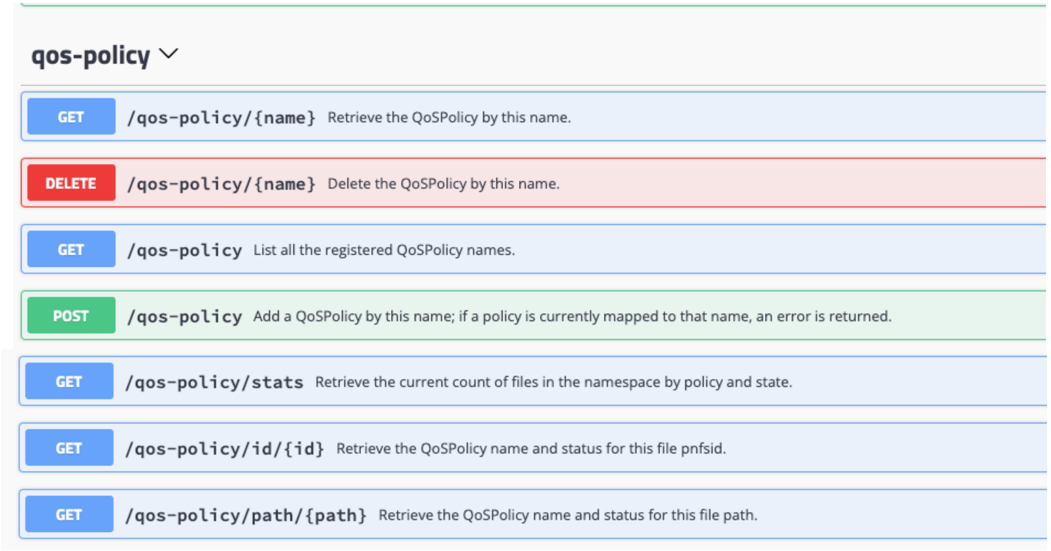
DELETE `/stage/{id}` Clear all resources pertaining to the given stage request id.

QoS “Rule Engine”

- The policy contains a ordered list of QoS transitions (or media changes)
- Admins can associate a qos-policy with a file
  - New policy can be assigned to files on create
  - New “QosPolicy” directory tag
- The policy uploaded through front-end REST-API
- The policy is defied as a JSON document

QoS Policy (pseudo) Example:

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



qos-policy

GET /qos-policy/{name} Retrieve the QoSPolicy by this name.

DELETE /qos-policy/{name} Delete the QoSPolicy by this name.

GET /qos-policy List all the registered QoSPolicy names.

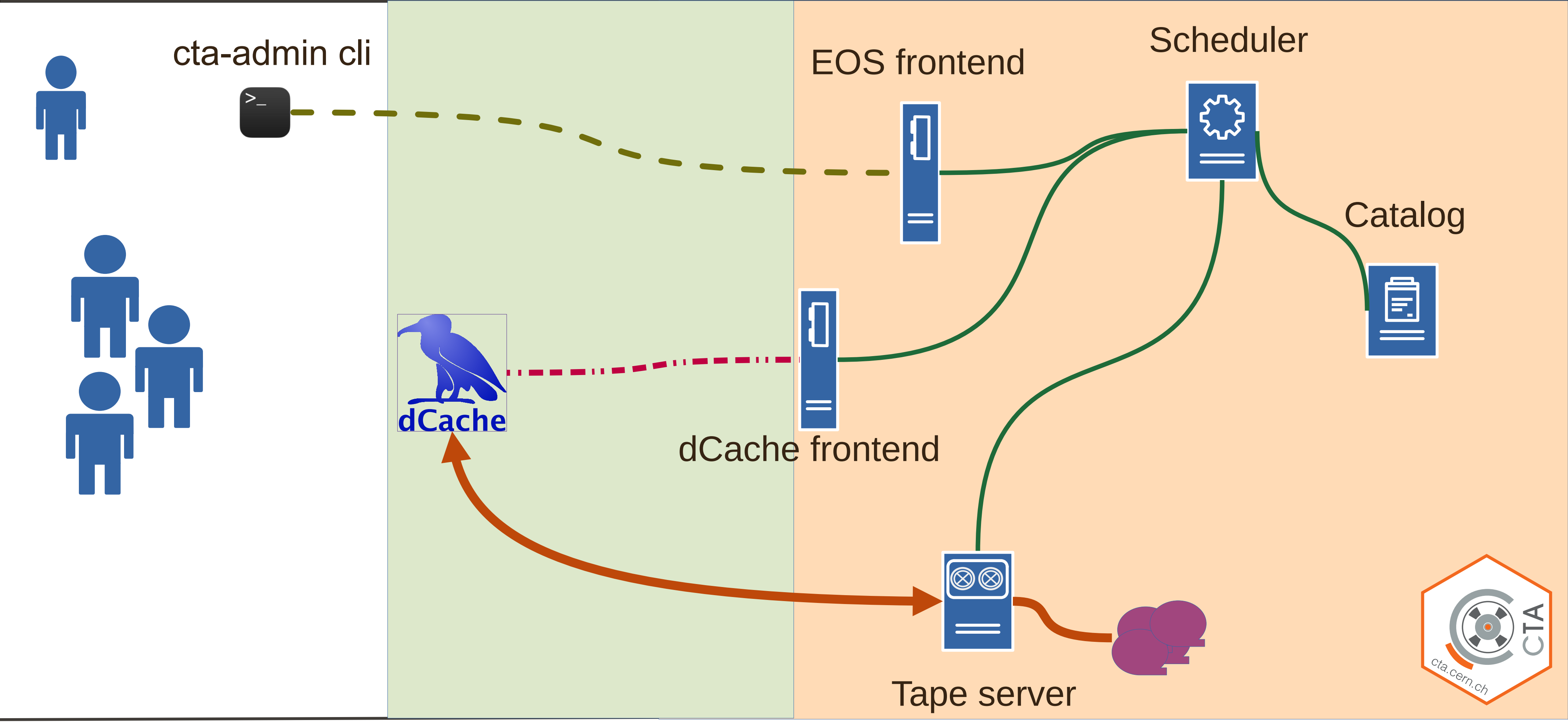
POST /qos-policy Add a QoSPolicy by this name; if a policy is currently mapped to that name, an error is returned.

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.

Integration with CTA



dCache+CTA Status

- Seamless integration with dCache is merged into upstream CTA code at CERN
  - Starting CTA release {4,5}.7.12
- The existing ENSTORE/OSM tape format is supported for READ
  - The ENSTORE/OSM tape catalog conversion procedures are successfully tested at DESY and Fermilab.
- dCache+CTA is deployed at DESY for BELLE-II, EuXFEL
  - ~2PB/week (3.4 GB/s, 9 drives)
- dCache+CTA deployment replicate to by other HEP sites
  - Fermilab and PIC Barcelona have successfully replicated our setup (currently dCache + ENSTORE).
  - RAL in UK plans to migrate to PostgreSQL from ORACLE based on our experience

Bits and Pieces…

- Native SSL for better performance
- Locality, ID and the checksum exposed as xattrs
- Nested Pool groups
  - Pool groups can be built from other pool groups
- Local endpoint in billing information
  - Make happy *WLCG ops* and *Packet Marking* teams
- No default HSM operation timeout
  - Practically there was only two values used: N or ∞

Even More Bits and Pieces…

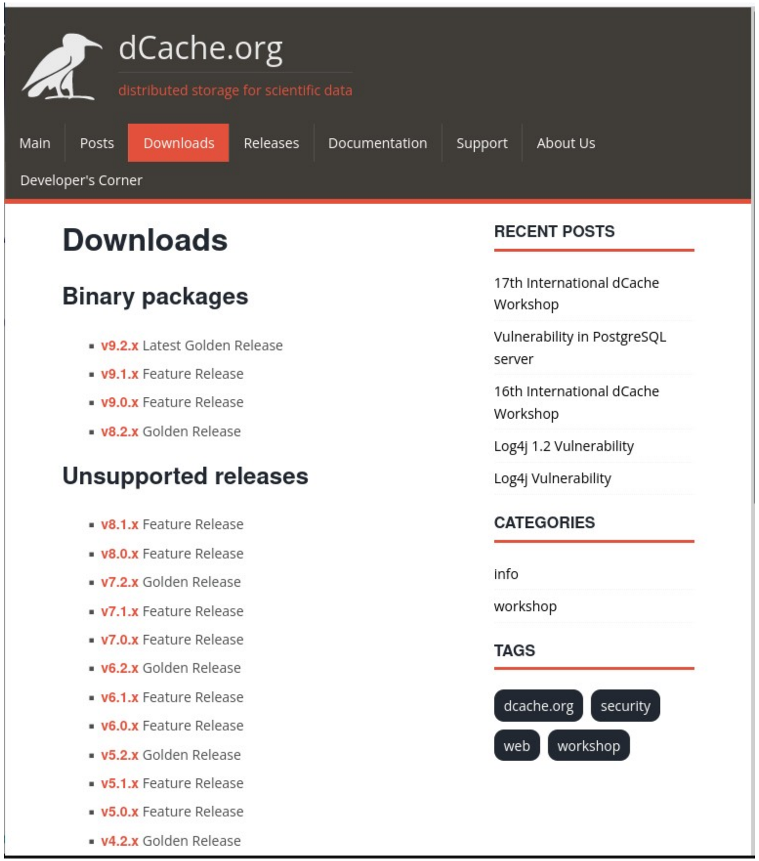
- Split disk and tape cleaners
- Dynamic reload of HSM drivers (ENDIT, CTA)
- Bulk cancellation of HSM requests
- User root for xroot door
- and many, many more…

Breaking Changes

- 9.0
  - `cleaner` service evolution ⇒ cleaner-disk, cleaner-tape
  - IPv6 link local addresses not published by SRM/SRR/…
  - DCAP door always in passive mode (client connects to a pool)
  - No default HSM ops timeout
  - Dropped experimental message serialization format
- 9.1
  - The link on directories counts only sub-directories
  - Dropped XACML gplazma plugin
- 9.2
  - Default configuration of NFS door incompatible with RHEL 6

Supported OS platforms

- 6.2 - 8.2
  - RHEL 7, 8, 9
  - JVM 11
- 9.0 – 9.2
  - RHEL 7, 8, 9
  - JVM 11, 17
- 10.0 (~ 1Q 2024)
  - RHEL 8, 9
  - JVM 17



Downloads

Binary packages

- v9.2.x Latest Golden Release
- v9.1.x Feature Release
- v9.0.x Feature Release
- v8.2.x Golden Release

Unsupported releases

- v8.1.x Feature Release
- v8.0.x Feature Release
- v7.2.x Golden Release
- v7.1.x Feature Release
- v7.0.x Feature Release
- v6.2.x Golden Release
- v6.1.x Feature Release
- v6.0.x Feature Release
- v5.2.x Golden Release
- v5.1.x Feature Release
- v5.0.x Feature Release
- v4.2.x Golden Release

RECENT POSTS

17th International dCache Workshop

Vulnerability in PostgreSQL server

16th International dCache Workshop

Log4j 1.2 Vulnerability

Log4j Vulnerability

CATEGORIES

info

workshop

TAGS

dcache.org security web workshop

Build Infrastructure: GitLab + k8s

- Documented release/test process
- Shareable build pipelines
- Can be replicated at sites
- Transparent release process
- Code will stay on Github

K8S Based Testing

- Sites can reproduce our release process
- dCache containers available at docker hub
- Helm carts to deploy dCache with three commands

```
$ helm install dcache-db bitnami/postgresql
$ helm install cells bitnami/zookeeper
$ helm --set image.tag=9.2.0 my-tier-2 dcache/dcache
```

Thank You!

More info:

https://dcache.org

To steal and contribute:

https://github.com/dCache/dcache

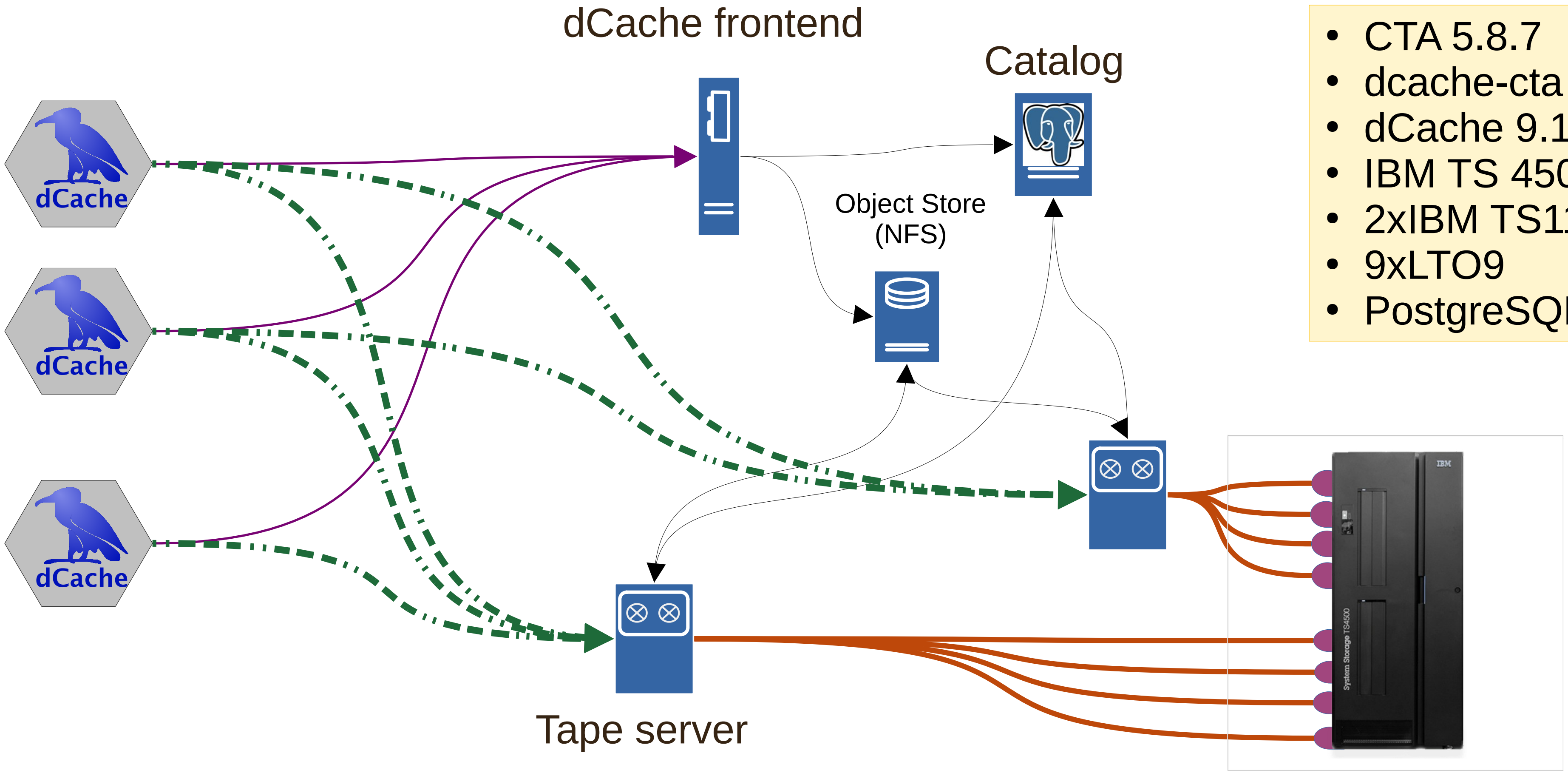
Help and support:

support@dcache.org, user-forum@dcache.org

Developers:

dev@dcache.org

Production Deployment at DESY



- CTA 5.8.7
- dcache-cta 0.10.0
- dCache 9.1.6
- IBM TS 4500
- 2xIBM TS1160
- 9xLTO9
- PostgreSQL 14.1