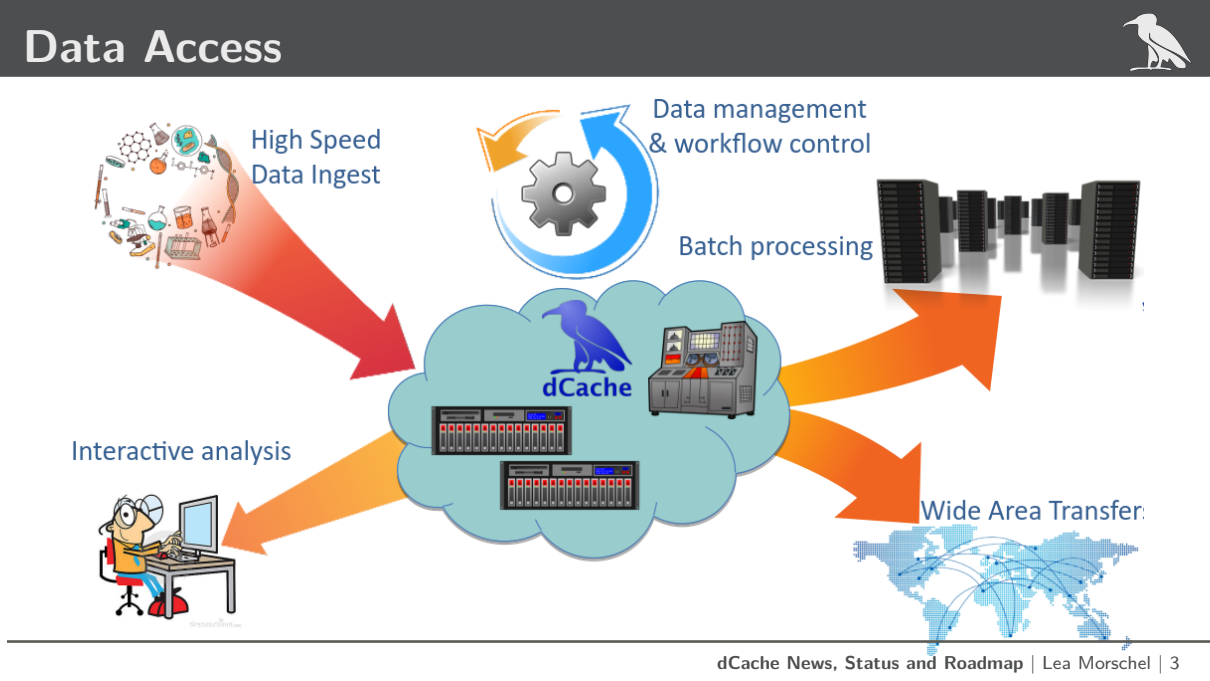
dCache News, Status and Roadmap

18^{th} International dCache Workshop

6 June 2024

The Big Picture.

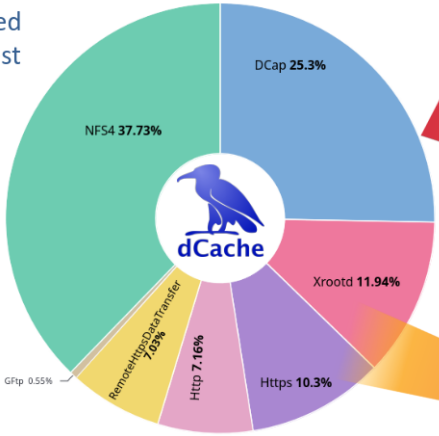
Data Access


High Speed
Data Ingest
Data management
& workflow control
Batch processing
dCache
Interactive analysis
Wide Area Transfer

Data Access Protocols

Speed Ingest

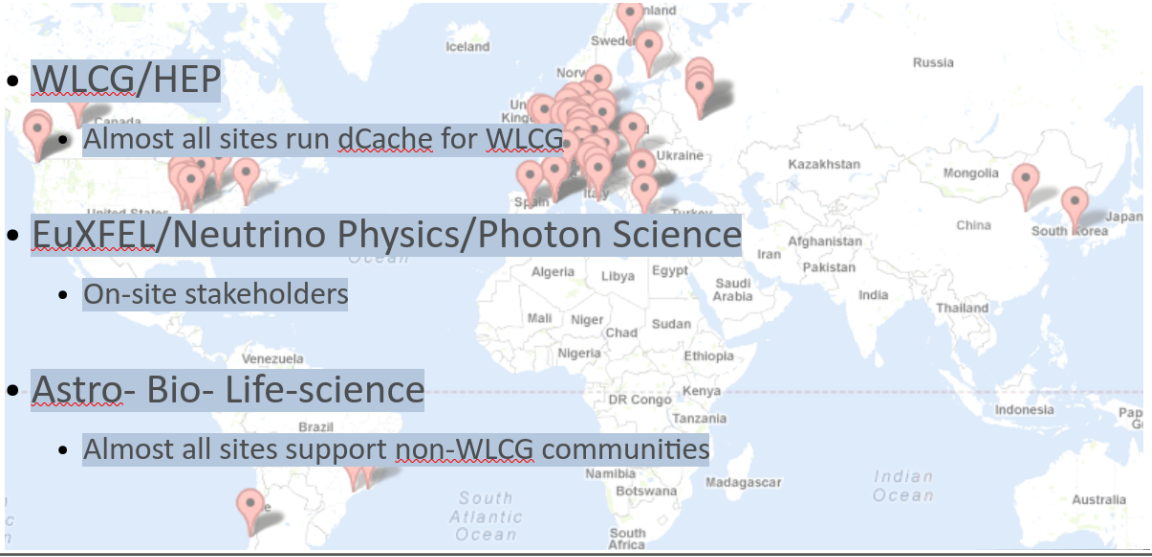
Batch processing



Interactive analysis

Wide Area Transfers

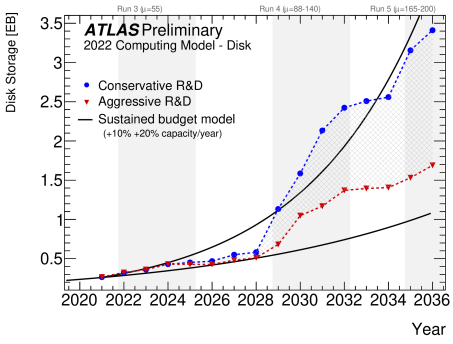
Strategic Communities



- WLCG/HEP
  - Almost all sites run dCache for WLCG
- EuXFEL/Neutrino Physics/Photon Science
  - On-site stakeholders
- Astro- Bio- Life-science
  - Almost all sites support non-WLCG communities

The Callenges

- 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

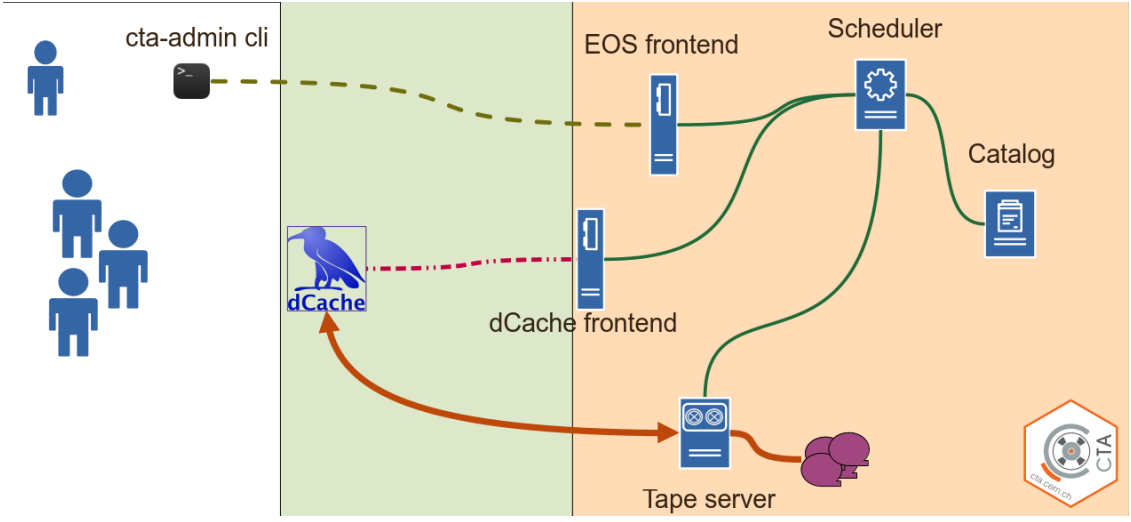


So – What Happened since the last Workshop?

- **Al** left ☹, **Chris** and **Karen** joined ☺
- We released versions **9.1**, **9.2** (golden), **10.0**
- Some interesting changes on master
- Adventurous site upgrades to 9.2

Tape and Related.

dCache + CTA

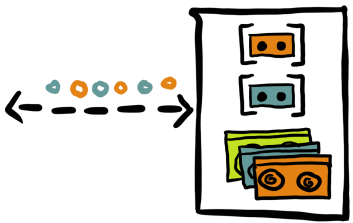

cta-admin cli
EOS frontend
Scheduler
Catalog
dCache
dCache frontend
Tape server
CTA
cta.cern.ch

dCache + CTA Status

- Integration with dCache **merged into upstream CTA code** at CERN
  - Starting CTA release {4,5}.7.12
- **Existing ENSTORE/OSM tape format supported for READ**
  - Tape catalog conversion successfully tested at DESY, Fermilab, PIC
- dCache + CTA **deployed at DESY for all experiments**
  - 2PB/week (3.4 GB/s, 9 drives)
- dCache + CTA **deployment replicated by other HEP sites**
  - Fermilab & PIC Barcelona: successful setup replication (currently dCache + ENSTORE)
  - RAL (UK) plans ORACLE $\rightarrow$ PostgreSQL based on our experience

General Changes To and From Tape

- Added pool option to **dynamically add/remove nearline storage providers** [9.1]
- **Restriction checking for staging operations** via `storage.stage` OIDC claims [master]



Bulk and Tape REST API

dCache bulk API

WLCG Tape API

Tape REST API V1 (like SRM, but different)

- **STAGE** – request to stage many files at once
- **CANCEL** – cancel bulk request
- **DELETE** – cancel bulk request + clear history/status
- **EVICT** – unpin cached copy
- **PIN** – pin cached copies with a lifetime
- **FILEINFO** – request status of many files at once (locality, checksum)

StoRM

Bulk Changes [9.2]

- **Major redesign in 9.2**, requires DB schema changes, properties deprecated
- Major performance improvements
- Periodic archiving of requests (configurable)
- Path resolution (for relative paths)
- Added HA mode (**currently broken!**)

QoS.

QoS Requirements

- HEP
  - Single copy (tape or disk)
- Photon Science
  - 2 tape copies, different media types (Jag+LTO)
- XFEL
  - 2 media copies (disk+tape => tape+tape)
- NextCloud
  - 2 disk copies + tape

Changes to QoS [9.2]

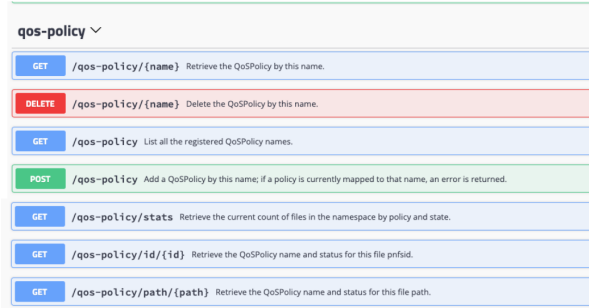
- Consider **Resilience as superseded by QoS** (but still there)
- Support for migration using new pool mode "*DRAINING*"
- DB namespace endpoint configurable: scanner may use replica
- **Role-based authorization** of QoS transitions (can be disabled)
- **Added QoS Rule Engine**: define QoS policy
  - Through directory tag or via a requested transition
  - Engine tracks state transitions over time

QoS Rule Engine [9.2]

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

QoS Rule Engine [9.2]

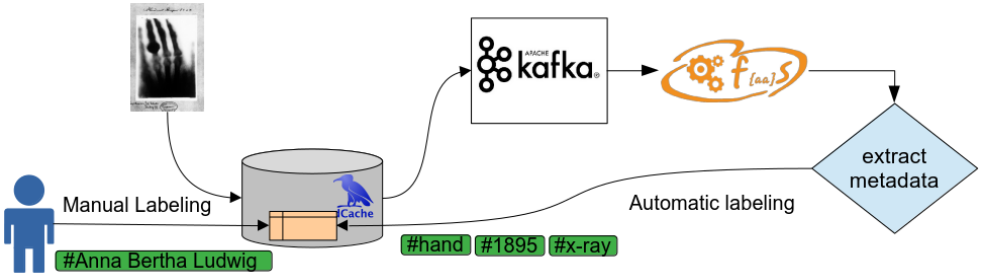
```
"name": "my-policy",
"states": [
  {
    "duration": "P10D",
    "media": 2x DISK
  },
  {
    "duration": "P1M",
    "media": 1x DISK, 1x HSM
  },
  {
    "media": 2x HSM
  }
]
```



Metadata/Labeling.

User Metadata/Labeling

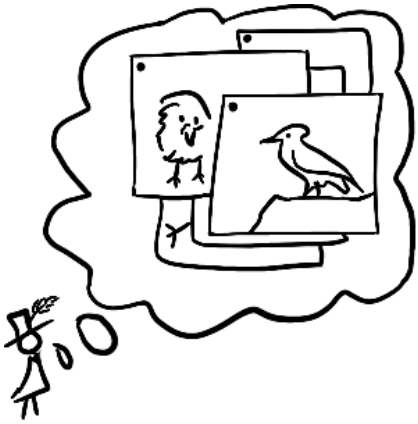
- Extended attributes for files/directories
  - Exposed via NFS, WebDAV, REST
- Attributes via HTTP(S) available to the flush process



File Labels (WIP) [master]

- Labels can be attached to a file
- **Query files by label** using the REST API
- **Label-based virtual read-only directories** via NFS:

```
.(collection)(bird)
```



Specific Services.

gPlazma

- Empty and non-existent **banfiles** treated the same [9.1.16+]
- Dropped gplazma support for **XACML** [9.1]
- **Dicontinued gPlazma 1** (no more Revocation entries) [9.1]
- **OIDC plugin**: allow suppression of offline verification & audience claim verification [9.1]
- Added **RolePrincipal** to support use of role definitions in multimap file [9.2]
  - Supported: `admin`, `qos-user` and `qos-group`

Namespace / PnfsManager

- The nlink count for directories shows only number of sub-directories [9.1]
- Dropped reference count tracking for directory tags, **improves concurrent dir creation/deletion rate** [9.2]
- New property `chimera.attr-consistency` (deprecates Java properties `chimera_soft_update` and `chimera_lazy_wcc`) [9.2]
  - Controls **behavior of parent dir attribute update policy**
  - Values: `strong, weak, soft`

PinManager [9.1]

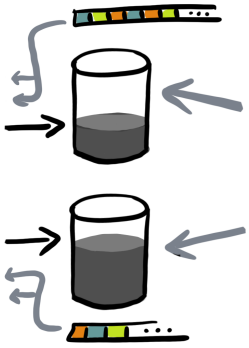
- New admin command to **remove *FAILED_TO_UNPIN* pins for pool**:

  ```
  clear unavailable mypool
  ```

- **On pin expiry, directly remove pin** from pins DB
  → Pool will delete sticky pin itself
  → Speeds up unpinning

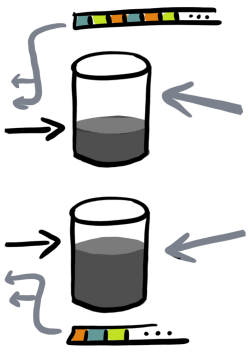
Pool

- The `st/rh/rm kill` admin commands accept `*` to **cancel all running requests** [9.1]
- Re-introduce ***wrandom* partition type** (*WASS* partition that ignores LRU metric and number of movers) [9.2]
- Added `pool.mover.https.port.min` and `pool.mover.https.port.max` to **control TPC port number used by HTTPS mover** [9.2]



Pool [10.0]

- `csm info` – show enabled digests for checksum calculation
- `pool.path.meta` – **set location of pools metadata dir**
- `csm use directio on|off` – **bypass filesystem cache during checksum calc.**
- Dropped idle timeout handler in netty-based movers (xroot, http) → just jtm
- `pp set max active` de-deprecated. Otherwise, **many-to-one p2p can overwhelm pool** (IO starvation)



Frontend

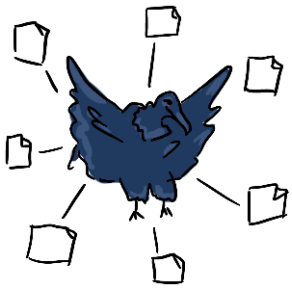
- **Query info on pool migration jobs** on new REST API `/api/v1//migrations` [9.1]
- Support for `.well-known/security.txt` added to frontend, WebDav ports [9.2]
- `frontend.wellknown` paths have been deprecated in favour of `dcache.wellknown` [9.2]
- Authz checks have been removed in Quota GET methods [9.2]

Xroot(d)

- **Proxying through the xroot door** [9.2]
- **Relative paths** are supported in the xroot URL [9.2]
- Resolution of **symlinks in path** prefixes and paths is supported [9.2]
- **Improved efficiency** of stat list (`ls -l`) [9.2]
- Add ability to **automatically reload TLS certificate** [9.2]

NFS

- New default: **door publishes all available layout types**, not just `nfs4_1_files` [9.2]
  - Old behavior can be enforced by `lt=nfsv4_1_files` export option.
- NFS door supports **RPC-over-TLS protocol extension** for data-in-transit protection (linux kernel 6.5+) [master]



Admin

- **Admin interface-based start of JFR** [9.1]
  - Capture JVM performance stats to investigate high CPU load, memory consumption, file descriptor leaks, etc.
  - Call via `System@thecell jfr start` and `System@thecell jfr stop`
- Output **paging added to admin interface** [9.2]

Other

- WebDAV door: Added limited support for **metalink format** (XML that describes how to download multiple files) [9.2]
- Add support for **injects of environment variables into configuration** (common in container world) [10.0]

  e.g. `dcache.net.wan.port.min=$env.DCACHE_WAN_PORT_MIN`
- `billing.format.json` – **write logs in JSON format** [10.0]

Non-Functional Developments.

Release- and Test-Process

- Documented test/release process
- Shareable build pipelines
  - Can be replicated at sites
- K8S based deployment
- **Code will stay on GitHub**

Kubernetes-Based Testing

- **dCache containers available at docker hub**
  → Sites can reproduce our release process
- Helm charts 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
```

Incompatibilities and Breaking Changes.

Incompatibilities

- **Compatibility between 8.2 and 9.2 is broken!**
  - Upgrade entire instance
  - Drop SrmManager `*requests` and `*filerequests` tables
- Dropped **gplazma-xacml plugin** [9.2]
- **Dropped support for native CEPH** → mounted [10.0]
- Starting with dCache 10.1, **Java 17 is required** at runtime

Other Important Bug Fixes

- HA RTM: **simultaneously started transfers get same ID** → one orphaned after other completes [9.2+]
- Accidentally **create loop on directory move** [8.2+]
- Fixed **bulk truncating path to 256 characters** [9+]
- On RHEL9, **NFS read returned incorrect last block** [9.2+]

9.2 Post Mortem.

9.2 Post Mortem – Problems and Fixes

- **Broken 8.2 – 9.2 compatibility** → global upgrade
- **No perf markers, orphaned/failed transfers** → HA RTM fix
- On **RHEL9** or clones → **enable SHA1** (for certain grid certs):

  ```
  update-crypto-policies --set DEFAULT:SHA1
  ```

- PoolManager not loading part of its config → fixed

9.2 Post Mortem – Problems and Fixes

- Bulk **staging error due to truncated paths** → fixed
- HA Bulk error → **run single bulk** (now)
- **Higher pool memory requirements**
  - Grizzly issue, fixed as a workaround
  - More predictable buffer initialization in pool instead of first NFS read

Outlook.

Summary of Prominent Changes Since 9.1

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

Photo by Dr. Raju Kasambe

Technical Directions

- Scaleout
  - Namespace
  - Number of pools (cells)
- Token-based authentication
- Better analysis facility support
  - POSIX access and compliance
  - HPC workload support (DDoS protection)
- QoS
- Tape integration

- **You can contribute** with ...
  - Code
  - Configuration
  - *Tests*
  - HW setup
  - Knowledge
- **You can make dCache visible** with ...
  - Sharing your use case
  - Demonstrate dCache use in various projects

- support φ dcache.org
  - User request tracking system
  - Place to ask for a help from developers
  - Accessible by all team members
- security φ dcache.org
  - Request tracking system
  - To report security issues or incidents
  - Accessible by selected people
- user-forum φ dcache.org
  - Mailing list for sysadmins/self-help group
  - To ask for advice or share experiences
  - Used by (almost) all sysadmins and developers
- dev φ dcache.org
  - Shared mailbox
  - Address to contact developers. Not for support
  - Developers can send e-mail from this address
- srm-deployment φ dcache.org
  - Tier-1 coordination mailing list
- workshop φ dcache.org
  - Shared mailbox
  - An e-mail used to organize workshops
- GitHub issues
  - Request tracking system
  - To report software defects and feature requests
  - Public
- GitHub pull-requests
  - Request tracking system
  - To provide code changes
  - Public

Thank You!

- More info: `https://dcache.org`
- Steal and contribute: `https://github.com/dCache/dcache`
- Help and support: support φ dcache.org, user-forum φ dcache.org
- Developers: dev φ dcache.org