What’s new in dCache-9.2

Tigran Mkrtchyan for dCache team
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

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.
Prominent Changes

- QoS & BULK Service
- TPC improvements
- NFSv4.1/pNFS improvements
- XROOT evolution (TLS, tokens, TPC, proxy-IO)
- Namespace performance improvements
- HSM connectivity
What's new in 9.2

- High Speed Data Ingest
- Interactive analysis
- Data management & workflow control
- Batch processing
- 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.
What's new in 9.2

### Tunable Consistency

<table>
<thead>
<tr>
<th>Consistency</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong</td>
<td>A creation of a filesystem object will right away update parent directory's mtime, ctime, nlink and generation attributes (POSIX).</td>
</tr>
<tr>
<td>weak</td>
<td>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).</td>
</tr>
<tr>
<td>soft</td>
<td>Same as the weak, however, reading of directory attributes will take into account pending attribute updates (POSIX).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>(wcc)</th>
<th>Score</th>
<th>Error</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>createFile</td>
<td>weak</td>
<td>14791.269 ± 1287.317</td>
<td>ops/s</td>
<td></td>
</tr>
<tr>
<td>createFile</td>
<td>strong</td>
<td>203.099 ± 17.556</td>
<td>ops/s</td>
<td></td>
</tr>
<tr>
<td>createFile</td>
<td>soft</td>
<td>1955.169 ± 908.004</td>
<td>ops/s</td>
<td></td>
</tr>
</tbody>
</table>
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
What's new in 9.2

Java Flight Recorder (9.1+)

```
[dcache-lab] admin > jfr start
enabled with config: default
[dcache-lab] admin > jfr stop
recorded into /tmp/core_xyz.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
A single door can now be configured to support all authentication protocols as an ordered chain:

```bash
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

```bash
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
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 defined 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
    }
]
```
Integration with CTA

What's new in 9.2

- cta-admin cli
- Integration with CTA
- EOS frontend
- dCache frontend
- Tape server
- Scheduler
- Catalog
- dCache
- cta-admin cli
- 2023-10-11
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 $\infty$
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
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

```bash
$ 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

- **dCache frontend**
  - Tape server
  - Catalog
  - Object Store (NFS)

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