dCache introduction

Paul Millar

on behalf of the dCache team.

EU dCache developer meeting
High-level Overview
dCache is...

**software** for providing scalable, managed storage for huge amounts of data.

**deployed** at research institutes throughout the world and used by a diverse collection of user-communities.

**supported** through the dCache.org collaboration, which provides:

- regular feature releases that are maintained with subsequent bug-fix releases.
- Support and advice through a variety of channels.
# dCache history

<table>
<thead>
<tr>
<th>Era</th>
<th>Disk cache</th>
<th>Grid Storage</th>
<th>Generic Storage</th>
<th>Cloud Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1" alt="hermes" /> <img src="image2" alt="ZEUS" /></td>
<td><img src="image3" alt="ALICE" /> <img src="image4" alt="ATLAS" /></td>
<td><img src="image5" alt="Fermilab" /> <img src="image6" alt="Intensity Frontier" /></td>
<td><img src="image7" alt="egi" /> <img src="image8" alt="LSDMA" /></td>
</tr>
<tr>
<td>Additional Communities</td>
<td><img src="image9" alt="CDF" /> <img src="image10" alt="MM" /> <img src="image11" alt="IceCube" /></td>
<td><img src="image12" alt="CMS" /> <img src="image13" alt="LHCb" /></td>
<td><img src="image14" alt="European XFEL" /> <img src="image15" alt="Belle II" /></td>
<td><img src="image16" alt="Industry" /> <img src="image17" alt="SNIC" /> <img src="image18" alt="CFEL" /></td>
</tr>
<tr>
<td>Additional Authentication</td>
<td>Trusted host</td>
<td>X.509, Kerberos</td>
<td>Username+PW</td>
<td>SAML, OpenID, OAuth, Token, ...</td>
</tr>
</tbody>
</table>

- **dCache** is a distributed file system for high-energy physics experiments.
- **Grid Storage** includes projects like ALICE, ATLAS, and CDF.
- **Generic Storage** includes collaborations such as Fermilab and SNIC.
- **Cloud Storage** includes cloud providers like egi and LSDMA.
What is dCache today?

LHC data stored on each storage system:
- dCache (96 PB)
- DPM (34 PB)
- EOS (0 PB)
- StoRM (20 PB)
- CASTOR (14 PB)
- BeStMan (7.6 PB)
- Globus FTP (6.1 PB)
- ARC (0.01 PB)
- xrootd (22 PB)

Source: BDII (2014-11-14)

Collaborations:
- EGI
- Globus Online
- LSDMA
- TERA
- Open Grid Forum
- Open Science Grid

Student mentor programme:
- Hochschule für Technik und Wirtschaft Berlin
- 3 students

Core team:
- DESY 8 FTEs
- Fermilab 2 FTEs
- NEIC 1 FTEs
Current and future project funding

Standardization
- 2010: NFS 4.1 / pNFS
- 2013: HTTP / WebDAV
- Contributing to the Dynamic Federation

Deploying new technologies into Production and exploring new communities
- 2015: AAI

Data Life Cycle
- 2018: INDIGO DataCloud
  - Multi Tier Storage
  - Quality of Service
  - Migration Archiving
dCache key features include...

- Users see a single POSIX filesystem (hard- & soft-links, etc),
- Transparent support for tertiary (tape) storage,
- Scalable bandwidth,
- Steerable target when reading and writing,
- Space management,
- Resilience to storage node failure,
- Supports transparent storage device life-cycle,
- Hot-spot detection and mitigation,
- Differentiable quality of service,
- Pluggable authentication,

...
The scientific cloud vision

**HPC & Grid Clusters**
- Low latency access

**Cloud storage**
- Standard back-end for clusters and portals

**Fast data ingest**
- Standard devices at high data rates

**Bulk WAN transfer**
- Moving huge datasets

**DropBox-like storage**
- Devices synchronise with storage

**Remote access**
- Rich access via web-browser

**dCache**
Some details* on how dCache operates...

* Some details are deliberately omitted to keep slides manageable.
dCache – under the hood

- Message passing layer
  - JVM
  - JVM
  - JVM

- Door(s)
  - (clients entry point)

- Pool Manager
  - (requests scheduler)

- Name Space
  - (MetaData Server)

- Pools
  - (Data Server)

- DBMS

Protocols:
- dcap
- http
- nfs
- ftp
Core components when transferring

Diagram showing the core components of the dCache system, including:

1. Client
2. Door
3. Namespace
4. PoolManager
5. Pool
6. gPizzma

The diagram illustrates the flow and interaction between these components during a data transfer process.
Importance of redirection

Pools

dCache

Door

Data

Metadata & Control

Client nodes

Data

Metadata & Control

Client nodes
Operational experience
Storage at DESY

- 6 dCache instances: Hera, CMS, ATLAS, Photon, “DESY” and Cloud:
  - **Hera** is officially switched off,
  - **CMS & ATLAS** for WLCG: CERN LHC experiments,
  - **Photon** is for various photon user-communities,
  - **Cloud** is for sync-and-share service,
  - **DESY** is for the rest.
## Comparative numbers

<table>
<thead>
<tr>
<th>CMS</th>
<th>ATLAS</th>
<th>Photon</th>
<th>DESY</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>~5x10^6 files</td>
<td>~1x10^7 files</td>
<td>~8x10^7 files *</td>
<td>~1x10^7 files</td>
<td>~2x10^6 files</td>
</tr>
<tr>
<td>~3 PiB</td>
<td>~3 PiB</td>
<td>~2.5 PiB *</td>
<td>~3 PiB</td>
<td>~10 TiB</td>
</tr>
<tr>
<td>~300 pool-nodes</td>
<td>~300 pool-nodes</td>
<td>~30 pool-nodes</td>
<td>~30 pool-nodes</td>
<td>~6 pool-nodes</td>
</tr>
<tr>
<td>~580 GiB/s ‡</td>
<td>~200 GiB/s ‡</td>
<td>~12 GiB/s ‡</td>
<td>~3 GiB/s ‡</td>
<td></td>
</tr>
<tr>
<td>~400 Hz (read)†</td>
<td>~180 Hz (write)†</td>
<td>~200 Hz (read)†</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Photon instance accepts ~1 TiB per month as ~1x10^7 files.
‡ Value is peak observed bandwidth aggregate over all clients within last 7 days.
† Value is peak observed open rate (either read rate or write rate) observed within last 7 days.
### Other dCache instances

<table>
<thead>
<tr>
<th></th>
<th>NT1</th>
<th>US-CMS T1</th>
<th>BNL</th>
<th>SARA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Files</td>
<td>~5x10^7 files</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td>~6.3 PiB (2.1 PiB tape; 4.2 PiB disk)</td>
<td>~20 PiB (disk)</td>
<td>~15 PiB (disk)</td>
<td>~6.2 PiB (disk)</td>
</tr>
</tbody>
</table>
Backup slides
Guaranteeing QoS for write
Guaranteeing QoS for tape activity