dCache introduction

Paul Millar
On behalf of the dCache team.

EGI-EISCAT-3D ad-hoc 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.
Association and funding

- **2000**: Fermilab
- **2001**: DESY
- **2002**: NDGF
- **2003**: neic
- **2004**: HELMHOLTZ ASSOCIATION
- **2005**: SNIC
- **2006**: EMI
- **2007**: LSDMA
- **2008**: Open Source
- **2009**: Open Source
- **2010**: Open Source
- **2011**: Open Source
- **2012**: Open Source
- **2013**: Open Source
- **2014**: Open Source
- **2015**: Open Source
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>hermes, ZEUS</td>
<td>ALICE, ATLAS</td>
<td>Fermilab, Intensity Frontier</td>
<td>egi</td>
</tr>
<tr>
<td></td>
<td>CDF, HII, IceCube</td>
<td></td>
<td>European XFEL, Belle II</td>
<td>LSDMA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Industry, SNIC, CFEL</td>
<td>PETRA III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional Communities</td>
<td>Trusted host</td>
<td>X.509, Kerberos</td>
<td>Username+PW</td>
<td>SAML, OpenID, OAuth, Token, ...</td>
</tr>
<tr>
<td>Additional Authentication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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)

Core team

- DESY
  - 8 FTEs
- Fermilab
  - 2 FTEs
- NeIC
  - 1.5 FTEs

Collaborations

- EGI
- globus online
- LSDMA
- EMI
- SNIC
- Open Grid Forum
- Open Science Grid
- Physics at the Tera Scale
- Helmholtz Alliance

Student mentor programme

- HTW
  - Hochschule für Technik und Wirtschaft Berlin
  - 3 students
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**
- Multi Tier Storage
- Quality of Service
- Migration Archiving

INDIGO DataCloud
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

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

**Bulk WAN transfer**
- Moving huge datasets

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

**dCache.org**
- NFS
- CDMI
- FTP
- HTTP
- WebDAV
Some details* on how dCache operates...

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

Message passing layer

- Door(s) (clients entry point)
- Pool Manager (requests scheduler)
- Name Space (MetaData Server)
- Pools (Data Server)

Protocols:
- dcap
- ftp
- http
- nfs
Core components when transferring
Importance of redirection

Client nodes

metadata & control

Data

Pools

dCache

Door

Data

Client nodes
Pools

dCache

Door

Metadata & Control

Client node
Client node

Pools

dCache

Door

Data

Metadata & Control
Guaranteeing QoS for write

Write

Read

WRITE

READ

READ

File on write pool (sticky) on read pool (cached; locked)

File on write pool (sticky) on read pool (cached)

File on write pool (sticky) on read pool 1 (cached)

File on write pool (sticky) on read pool 2 (cached)
Guaranteeing QoS for tape activity
Operational experience
Storage at DESY

- 6 dCache instances: Hera, CMS, ATLAS, Photon, “DESY” and Cloud:
  - **Hera** is officially switched off,
  - **CMS & ATLAS** for WLCG 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