dCache, the Peta-Scale storage element

or: About Managed Storage

Patrick Fuhrmann
for the
dCache Team

Presented at the Linux Cluster for Super Computing workshop in Linköping, Sweden
Preliminaries

The Team

Responsibility, dCache
Patrick Fuhrmann  Rob Kennedy

Core Team (Desy and Fermi)
Jon Bakken
Ted Hesselroth
Alex Kulyavtsev
Birgit Lewendel
Dmitri Litvintsev
Tigran Mrktchyan
Martin Radicke
Owen Synge
Elena Tews
Vladimir Podstavkov

Responsibility, SRM
Timur Perelmutov

External

Development
Nicolo Fioretti, BARI, Italy
Abhishek Singh Rana, SDSC, US

Support and Help
Maarten Lithmaath, CERN
N.N, CERN

Patrick Fuhrmann  LCSC 2006, Linköping  Oct 18, 2006
Road-map for today

- Why do we need managed storage (WLCG)
- Requirements to managed storage systems
- dCache specification
- Selected Topics
- Performance Considerations and scaling
- Who is using dCache
Why do we need Managed Storage?
Do you want to do this manually;

- Honor MoU's on storage quality
- Data transport has to be organized
- Global available space needs to be managed
- Tape vs. Disk needs to be honored (including transitions)
Zoom into a atomic storage entity

(View before there was norduGrid)
Prepares for data transfer (not transfer itself) by storage URL
Negotiates data transfer protocol (theoretically).
May initiate restore of data from back-end storage systems.
Delivers 'transfer URL' (TURL) for subsequent transfer (gsiFtp, httpg).
Supports directory functions including file listings.
Supports space reservation functionality (implicit and explicit via space tokens)
Supports 'property spaces':

File Properties resp. Property classes:

<table>
<thead>
<tr>
<th>Media Quality</th>
<th>Persistence</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tape</td>
<td>permanent</td>
<td>how long does it take to get this file ready for I/O</td>
</tr>
<tr>
<td>Output</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Replica</td>
<td>volatile</td>
<td></td>
</tr>
</tbody>
</table>
Some Details

**Requirements to managed storage**

### Basics

- Stores data in the order of Petabytes
- Total-throughput scales with the size of the installation
- Supports several hundreds to thousands of clients
- Adding / removing storage nodes w/o system interruption
- Supports posix-like access protocols (dCap/rfio/xroot)
- Supports wide area data transfer protocols (gsiFtp/https)

### Advanced

- Drives back-end tape systems (generates tape copies, retrieves non cached files)
- Selects storage areas based on rules (client IP, file type, directory location) -> Storage Ownership by experiments
- System improves access speed by replicating 'hot spot' datasets
- Supports being 'managed' -> SRM
Now ... dCache
dCache manages storage and does exactly what is demanded on the previous transparency.

and more ...
dCache manages peta bytes of storage, distributed among thousands of storage nodes.

dCache manages multiple internal or external copies of a dataset associated to a single file-system entry.

dCache autonomously manages the number and location of the internal copies to optimize overall data throughput.

For data transport, dCache supports a variety of posix-like and wide area protocols. (gsiFtp, dCap, xRoot)

dCache name space is managed by NFS2/3/(4) and ftp.

dCache supports the SRM storage management protocol.

dCache can drive a tertiary (e.g. tape) storage back-end.
**Insight dCache**

**File multiplicity and pool selection**

### File system view
- `/pnfs/desy.de`
- `/pnfs/fs`
- `usr`
- `cms`
- `atlas`
- `public`

### File content view
- **CMS Pools**
- **Atlas Pools**
- **Public Pools**

Patrick Fuhrmann  
LCSC 2006, Linköping  
Oct 18, 2006
**Pool Selection is a two phase process**

(I) Select a **set of pools** which matches the following attributes

- Protocol
- Data flow direction (put, get, pool to pool, retrieve from tape)
- Directory subtree
- Client IP address

(II) Out of those pools, select the one, with the best value concerning the number of already running movers and the available (removable) space.

---

Tuning:

- Equally distributed movers on all pools
- Fill up pools equally
File hopping

- Automatic data set replication on hot spot detection.
- File replication on client read request (pools may be disallowed for reading)
- Dataset replication on arriving of datasets. (configured)
File hopping (cont.)

Pool is configured for write only. So a read will copy the file to a read pool prior to file delivery.

From Client

Write Only Cache

To Client

Read Only Cache

Replicate on high load

To Client

Patrick Fuhrmann

LCSC 2006, Linköping

Oct 18, 2006
- Datasets collected in write pools and flushed according to rules.
- Centrally controlled (Smart) flushing -> (Alternated Flushing)
- Datasets restored if requested but no longer in cache.
- Intermediate restore pool for HSM optimization.
**HSM interactions**
overcoming hardware deficiencies

Intermediate Restore Pools

Read Only Cache

Ping Pong flush pools

From Client

Insight dCache
Resilient dCache (pools on worker nodes)

- Controls number of copies for each dataset in dCache
- Makes sure \( n < \text{copies} < m \)
- Adjusts replica count on pool failures
- Adjusts replica count on scheduled pool maintenance

Attractive because:

- \( N \)-pool nodes may be in maintenance mode without affecting the overall availability of datasets in the dCache system.
- Improves overall performance by read striping
- Makes use of unused space on worker/farm nodes.
Flexible (growing) setup
or: The dCache toolkit

Insight dCache

Patrick Fuhrmann
LCSC 2006, Linköping
Oct 18, 2006
• Destination pool selection by IP, directory, protocol, I/O direction.
• Final pool selection by space cost and pool node load.
• dCache instance partitioning.
• Extended proxy (certificate) support (OSG and LCG)
• Draining of pools for maintenance.
• Rich command line interface (via ssh).
• First version of GUI for admin and cpu/space cost analysis.
• Highly improved file system emulation (chimera) in evaluation phase.
• See 'dCache, the Book' for details.
Performance characteristics

Speed

- Number of 'opens' per second
  - 500 Opens/sec

- Number of pools
  - 5000 pools

Storage Space

- Amount of disk space
  - 2 Peta Bytes

- Number of bytes per second
  - 1 – 2 Peta Bytes/day

- Number of pools
  - 1000 Pools

- Amount of disk space
  - 300 TB/Day - 3 GB/sec

Insight dCache

- 50 Opens/sec

- 300 Tbytes
  (FermiLab)

- 1000 Pools
  (Manchester)

Patrick Fuhrmann
LCSC 2006, Linköping
Oct 18, 2006
Future dCache

Coming sooner (or later)

- SRM 2.2 interface
  - Space Tokens
  - Storage Classes

- Chimera (Improved file system engine)

- Acl's

- Quotas

- nfs4.1 (including data transport)

- Improved Hsm connectivity
In Use at

Tier I centers:
- FNAL (enstore)
- BNL (HPSS)
- Triumph (TSM)
- SARA (DFS)
- RAL (Home Grown)
- IN2P3 (HPSS)
- GridKa (TSM)
- Nordu (Home Grown)

Tier II centers:

Germany
- LCG: Aachen, DESY, Freiburg, Dortmund, Darmstadt (GSI)
- d-Grid: Juelich (ZAM), Berlin (ZIB)

UK
- 30% of gridPP, UK

US
- CMS: 7 sites
- ATLAS 7 sites in preparation

Italy
- INFN: Bari, Torino

Poland, Bulgaria, Spain

Canada

Patrick Fuhrmann
LCSC 2006, Linköping
Oct 18, 2006
The End

References

dCache, the Book

www.dCache.ORG

need specific help for you installation or help in designing your dCache instance.

support@dCache.ORG

dCache user forum

user-forum@dCache.ORG