dCache, agile adoption of storage technology

Paul Millar
CHEP-2012 New York, 2012-05-24
Overview

- Some news
- Flexibility
- Future directions
- Summary
Funding

• dCache is our **contribution to WLCG:**
  from Germany, the Nordic countries and USA/Fermilab,

• has been funded (independently from WLCG) for **over 10 years**

• Funding for dCache **is secure** for after EMI:
  Without EMI, funding only drops by ~20–25%
Community

- **3rd International workshop:**
  - 57 participants, from 13 countries
  - New user-communities presented how they wish to use dCache
- **Forging links with industry:**
  DESY and IBM form “large data” strategic partnership based on dCache storage competence (CeBIT)
- **Establishing a Stack Exchange site**
  http://area51.stackexchange.com/proposals/40050/dcache
Evolution

- Within WLCG:
  - Strong involvement with TEG groups
  - Working in collaboration on federated storage
    (both xrootd and HTTP)
- Outside WLCG:
  - OGF standardisation
  - Engaging new communities
- Improve dCache modularity:
  Allow dCache to be easily adapted to novel environments
  *Agility is a process, not a target*
News: under the hood

• Splitting the code into smaller, **reusable** pieces:
  
  • **Chimera**: enstore
    
    See *Enstore with Chimera namespace provider* by D. Litvintsev
  
  • **jrpc**: BACNET, a Swiss Bank, ...
    
    See *dCache: Implementing a high-end NFSv4.1 server using a Java NIO framework* by T. Mkrtchyan.
  
  • **xrootd4j**: (ALICE?)

• dCache is adopting Free/Open-source license
  
  • Mostly **AGPLv3**, the rest is LGPL or BSD
  
  • Needed to get dCache into distributions
News: NFSv4.1 / pNFS

- Industry standard protocol
- Client availability:
  - RHEL/SL 6.x,
  - RHEL/SL 5.x (with Oracle kernel + nfs-utils upgrade),
  - Fedora 15,
  - Debian 7.0 (“Wheezy”),
  - Windows 7 (with driver from CITI),
  - Windows 8,
  - Solaris “Oracle (..) will deliver implementations of (a client and server) in future releases of Solaris” (1)
- Hardware vendor support:
  - NetApp OnTap 8.1
  - Panasas “in 2012” (1)
  - BlueArc,
  - IBM “key part of SONAS Active Cloud Engine” (1)

(1) Source is “FAST 2012 pNFS BoF” 2012-02-15
News: dCache & pNFS

- NFS v4.1 / pNFS has been supported since 2009.
- Deployed in production (at DESY) for over a year.
- Fermilab's REX dept. evaluated dCache NFSv4.1 for their Intensity Frontier experiments:
  
  "Results look promising, throughput scales well with number of pool nodes"

- Supports:
  - authn: trusted-host and Kerberos
  - all three GSS security modes.
Flexibility

(plugins and extension points)
Plugins: who should be interested & why

- **Core developers:**
  - New functionality can be added as a plugin
  - Backwards compatibility by keeping old plugins
  - Can test-deploy new features at friendly sites

- **dCache sites:**
  - Integrating with local, site-specific services

- **User-communities:**
  - Add some experiment-specific behaviour

- **External developers / trail-blazer sites:**
  - Experiment with exciting new features
What can I enhance?

- xrootd door
- Namespace
- PoolManager
- gPlazma
- Pool
- A dCache service

Note that some details have been glossed over.
What can I enhance?

- xrootd door
- Namespace
- PoolManager
- gPlazma
- Pool
- Billing
What can I enhance?

- xrootd door
- Namespace
- PoolManager
  - PoolSelection
    - Links & Units
- gPlazma
- Pool
- Billing
What can I enhance?

- xrootd door
- Namespace
- PoolManager
  - PoolSelection
    - Links & Units
- gPlazma
- Pool
- Billing

Naming Convention
What can I enhance?

For further details see *A strategy for load balancing in distributed storage systems* by M. Wadenstein and G. Behrmann.
What can I enhance?

Network topology, bandwidth and latency aware
What can I enhance?

- xrootd door
- gPlazma
- Billing

Namespace

PoolManager
  - PoolSelection
  - Partition
What can I enhance?
What can I enhance?

Overlay namespace
What can I enhance?

- xrootd door
- gPlazma
- Billing

Namespace

NamespaceProvider

Chimera

PoolManager

PoolSelection

Partition
What can I enhance?

Diagram showing the structure of namespace and pool manager with layers such as xrootd door, namespace provider, filesystem provider, and Chimera.
What can I enhance?

- xrootd door
- gPlazma
- Billing

Namespace

NamespaceProvider

Chimera

FilesystemProvider

JdbcFs

HDFS' namespace

cassandra
What can I enhance?

Namespace

NamespaceProvider

Chimera

FilesystemProvider

JdbcFs

PoolManager

PoolSelection

Partition

gxrootd door

gPlazma

Billing
What can I enhance?
What can I enhance?
What can I enhance?

- xrootd door
- Namespace
  - NSP
- PoolManager
  - PoolSelection
  - Partition
- gPlazma
- Billing

pool
What can I enhance?

- xrootd door
- Namespace
  - NSP
- PoolManager
  - PoolSelection
  - Partition
- Pool
- File store
  - Flat directory
- gPlazma
- Billing
What can I enhance?

Hierarchical Storage

Pool

File store

Flat directory

xrootd door
Namespace Manager

gPlazma

Billing

Hadoop

HDFS

lustre

IBM GPFS

amazon web services
What can I enhance?

- xrootd door
- Namespace
  - NSP
- PoolManager
  - PoolSelection
  - Partition
- Pool
- File metadata
  - Files
  - Berkeley DB
- gPlazma
- Billing

---
What can I enhance?

- xrootd door
- Namespace
- PoolManager

Pool

File metadata

- Files
- Berkeley DB

File store

gPlazma

Billing

PostgreSQL
cassandra
What can I enhance?

- xrootd door
- Namespace
  - NSP
- PoolManager
  - PoolSelection
  - Partition
- gPlazma
  - File metadata
- Billing
What can I enhance?

xrootd door

Namespace
  NSP

PoolManager
  PoolSelection
  Partition

gPlazma
  AuthN
  Map
  Account
  Session
  Identity

Billing

File metadata
What can I enhance?

- xrootd door
- Namespace
  - NSP
- PoolManager
  - PoolSelection
  - Partition
- gPlazma
  - AuthN
  - Map
  - Account
  - Grid
  - Identity
- Pool
  - File store
  - File metadata
- Billing
What can I enhance?

xrootd door
Namespace
PoolManager

gPlazma
Pool

Billing
Storage

Files
Database
What can I enhance?
What can I enhance?
What can I enhance?

xrootd door

AuthN
- none
- GSI

Billing
- Storage

PoolManager
- PoolSelection
- Partition

File store
- File metadata
What can I enhance?
What can I enhance?

- **xrootd door**
  - AuthZ + namespace
  - AuthN
  - Session
  - Identity

- **PoolManager**
  - PoolSelection
  - Partition

- **Billing**
  - Storage
What can I enhance?

- Replace prefix (CMS)
- Call out to external service (ATLAS)
What can I enhance?

- xrootd door
  - AuthN
  - AuthZ + namespace

- Namespace
  - NSP

- PoolManager
  - PoolSelection
  - Partition

- gPlazma
  - AuthN
  - Map
  - Session
  - Identity
  - File store
  - File view

- Pool
  - movers
    - dcap
    - FTP
    - HTTP
    - NFS v4.1
    - xrootd
What can I enhance?

- xrootd door
  - AuthN
  - AuthZ + namespace
- Namespace
  - NSP
- PoolManager
  - PoolSelection
  - Partition

**Pool**
- Movers
  - dcap
  - FTP
  - HTTP
  - NFS v4.1
  - xrootd

**new-protocol door**

- SFTP (SSH File Transfer Protocol)

CDMI
gPlazma: logging in
gPlazma: identities

The login process

Authn

Map

Account

Session

Identity process

Plugin-1

Plugin-2

Plugin-3

UID: 200
GID: 310

UID: 300 ?

fred@example.org

NFS door

Is command
Future directions
HTTP and WebDAV

- How do we support **non-HEP users**?
- Dcap, SRM, rfio, xrootd
  
  Nobody outside of HEP has heard of these (HEP is 1% of scientists)

**HTTP & WebDAV**

- Everyone has a web-browser
- WebDAV is commonly available on platforms
- Used by some cloud storage providers (Microsoft SkyDrive, Deutscher Telekom, ..)

- Deployed **in production**: DESY, PIC, BNL, ...
Federating storage

“Collection of disparate storage resources managed by co-operating but independent administrative domains transparently accessible via a common name space.”

Hey, we can do this with a **standard protocol**: HTTP!

**Benefits:**

- Get **high-performance client** for free,
- **Loads of free software** (Apache, Squid, Varnish, …)

**Two stage approach:**

- **Web front-end** to existing catalogues (LFC, …)
- **Dynamically** discovering available data using WebDAV
  - All replicas of a file are discoverable (c.f. **dark data** problem)

For further details, see *Dynamic federations: storage aggregation using open tools and protocols* by F. Furano
Missing files

A user may ask for a file that doesn't exist

A user may ask for a file that should exist, but the pool is broken
Missing files

Maybe dCache should do “something” in these cases. That “something” should be highly configurable; i.e., a plugin.

For further details, see **SYNCAT – Storage Catalogue Consistency** by F. Furano
Faster storage
3 Tier Model

For further details see *Evaluation of benefits of a three tier data model for WLCG analysis* by D. Ozerov and P. Fuhrmann
Summary

The dCache project is independent of WLCG and EMI funding.

dCache has the flexibility to adapt to new deployments, scenarios and technology.

The dCache community is growing.
Thanks for listening