Overview of dCache Systems at BNL

Iris Wu
Hiro Ito
Jane Liu

dCache workshop 2018
DESY, Hamburg Germany
Outline

- Overview
- Future operation
- Needs help from developers
Puppet Management
• Efficient
• Automation
**dCache**

- Disk pool nodes
  - Primary
  - Secondary
- Resilience Management

**Disk cache**

**US ATLAS Tier1**

dCache @BNL

**Admin nodes**

**Data transfer nodes**

**Tape pool nodes**

- **NFS4.1/pNFS**
- **DAV**
- **HTTP/WebDav**
- **Globus Online**
- **GFTP**
- **XRootD**

**US DEPARTMENT OF ENERGY**

**Scientific Data and Computing Center**
Protocols

- NFS 4.1
  - Mounted in the Linux farm
  - US ATLAS Tier3 facility

- XRootD
  - Read is in production
  - Write is tested through
  - GSI authentication

Support of XRootD third party transfer?
Protocols

**dCache Native Gridftp Server**

- **Pros**
  - It is a native dCache!!!
  - Performance and stability is great
- **Cons**
  - Can dCache developer keep supporting?
  - Can it be officially supported by Globus

**Globus Gridftp Server + dCache NFS4.1**

- **Pros**
  - It is a native Globus. The server is always supported by Globus.
  - NFS4.1 is a standard.
- **Cons**
  - The performance is not as good as the dCache gridftp option.
  - The stability issue has been reported.
Pros
- Easiest to transfer data in/out dCache
- Automatic sync

Cons
- SFTP can be a problematic.
- Can we natively support?
  - Like CERN box with EOS?

BNLBox (Cloud storage/Owncloud)
Customization - Deletion

**namespace**
Files were removed from namespace by clients

**pool**
Tune Cleaner to efficiently remove files physically from pool
favor pools with most deleted files
query t_location_trash

**database**
Remove records from srm spacefile in srm database on time

- Dark Data
## Customization— schema change

The customization involves changing the schema of the `srmSpacefile` table in the DCache system.

### Current Schema
```
id  vogroup  vorole  spacerservation  sizeinbytes  creationtime  pnfsid  state
```

### Proposed Schema
```
id  vogroup  vorole  spacerservation  sizeinbytes  creationtime  pnfsid  state  chimerapath
```
Future operation - Splitting read pools for Tape area

- Disk space crunch!
  - Various experiments and users have more data than the size of spinning disks available to them. And, it is expected to grow further.
- The tape storage is still cheaper than the spinning disks (or SSDs).
- Would like to use the archive storage (aka TAPE) as effective as reasonably possible.
  - TAPE requires a particular access to get the optimum performance.

The callback option for the future?
Future operation — Ceph Pools

• Pros:
  • Separate pool service hosts from actual storages.
  • remove the association of files in Ceph pools with t_location
  • Erasure codes allows the resiliency without duplicate copy

• Questions
  • Performance of partial read???
  • Scalability of Erasure code???
Needs help from developers

- continue the support for Ceph pools
- continue the support for Globus online
- support XRootD 3rd party transfer
- the issue with Resilience Management
Thank You!