Chimera is upon us

Dmitry Litvintsev

Nucomp meeting 02/15/12

Namespace function

- Unique file ID independent from file name
- Path to ID mapping
- Mechanism to store file metadata
- Directory tags inherited by subdirectories
- Callbacks on FS events (at least rm)

PNFS

PNFS (perfectly normal file system) developed in 1997 by DESY. NFSv2 filesystem on top of database

- supports all NFSv2 namespace operations
- actual I/O performed by HSM utilities
- allows for storage of user defined metadata associated with files and directories

Adopted as namespace provider by Enstore HSM and dCache systems.

PNFS limitations

- max file size is 2GB
- metadata access only thru NFS
  - no direct path for storage system client
  - heavy metadata access by storage system impact regular NFS operations
- metadata stored as BLOBs (no efficient metadata query functionality)
- no ACLs
- no security

PNFS Status

- De-supported by DESY in favor of new product Chimera
- Only 2 sites remain that use PNFS:
  - CMS T1 @ Fermilab, CDF, D0 and public Enstore systems
  - Spain T1 site @ PIC
- Issues with PNFS and newer kernels:
  - encp is known to hang client nodes mounting pnfs (hopefully fixed only recently in encp) PBI000000000147, PBI000000000184
  - issues with fileid mismatch:

```
Nov 15 15:40:41 fcdfcache91 kernel: NFS: server cdfensrv1.fnal.gov error: fileid changed
Nov 15 15:40:41 fcdfcache91 kernel: fsid 0:16: expected fileid 0x1d7ce677, got 0x1d7ce670
```

- No expertise in PNFS code base
- In our tactical plan we identified reliance on PNFS as a significant risk factor

Chimera

- High performance replacement for PNFS
- Build on top of relational DB allowing efficient metadata queries. Isolation of queries for different metadata types for better throughput
- Well defined API for namespace operations, metadata manipulations and admin interface
- dCache accesses metadata directly, bypassing NFS for higher throughput
- Platform independent:
  - pure java implementation
  - JDBC without DB specific binding.

dCache

Chimera

- plugin interface for permission handler
- NFS versions supported:
  - v2 (legacy)
  - v3 (legacy) overcomes 2GB file limit
  - v4 GSS authentication
  - v4.1 one protocol for namespace and data file access. Allows parallel POSIX I/O on distributed data. A real filesystem.

Chimera @ Fermilab

- 2007 : early evaluations. Determine performance and stability levels.
- 2010 : encp has been modified to work with both Chimera and PNFS namespace
- 2010 : functionality testing directly with Enstore and dCache/Enstore.
  - added Enstore specific triggers (on write and update of layer 4 files)
  - dCache modified to extract Enstore specific data from layer 4
  - file deletion (marking files deleted on tape) adopted to use Chimera DB directly
- fall 2011: production level acceptance tests.

PNFS->Chimera

- Copy of production pnfs from stken:
  - 62 databases, 14718667 files
- on quad core Xeon CPU @ 2.33GHz, 8GB RAM

pnfsDump	7h41m
SQL import	8h27m
enstore2chimera	1h9m
import of companion	20m
md5sum verification	30h39m

dCache

What is it for Users

We will have 20 hour downtime of public dCache and STKEN Enstore (excluding CMS) 02/22/12 6:00PM - 02/23/12 2:00PM

- direct encp users:
  - No write access during downtime
  - Read access is OK
- public dCache users
  - public dCache is unavailable for 20 hours for read and write

After Upgrade

- public dCache users :
  - No Difference
- direct encp users:
  - must upgrade to encp v3_10e to be able to work with chimera
  - v3_10e production version will be available tomorrow 02/16/12 from KITS. A test version is available now.
  - Please start upgrading your encp tomorrow. We will send announcement

List of direct encp users (storage groups)

B DMS cdf cms coupp ccstlogs
d0lib-archive des fermigrid
minerva miniboone minos sdss

We will be contacting reps. of these storage groups individually with instructions/help on how to update encp to v3_10e