

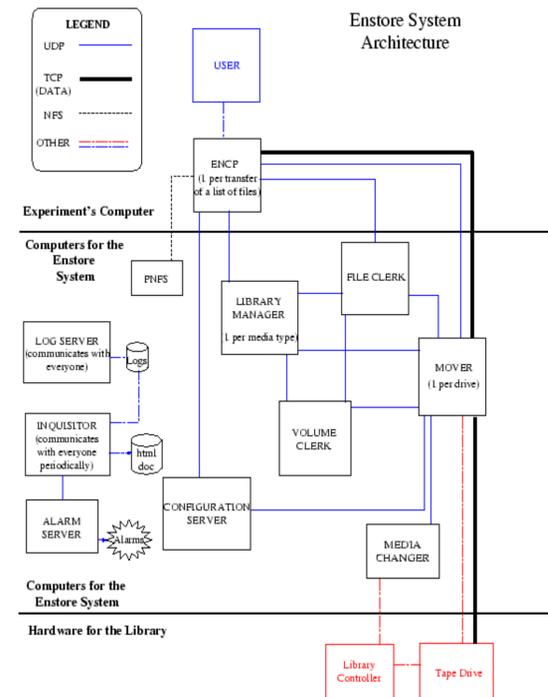
Enstore with Chimera namespace provider

D.Litvintsev, A.Moibenko, G.Oleynik, M.Zalokar
(Fermi National Accelerator Laboratory)

Enstore

Enstore is a Mass Storage Management system developed and operated by Fermilab. It provides seamless access to the data stored on permanent media by client applications distributed across IP network.

Enstore is client-server application



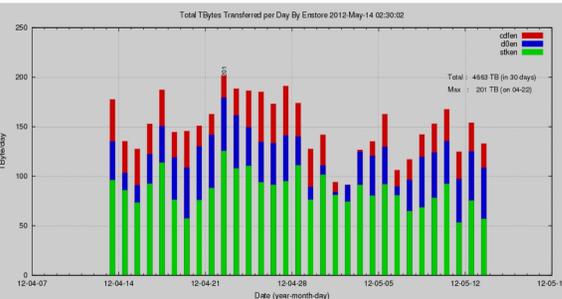
Client side:

- Enstore client, encp, provides “cp” like functionality to retrieve/store files from/to tape.
- End-to-end data integrity verification

Server side is a multicomponent ensemble of distributed servers that provide:

- Hierarchical view of files stored on tapes as if they were located on a Unix file system. PNFS namespace provider developed by DESY is used.
- Configuration service.
- Management of user files.
- Distributed access to tape drives.
- Interface to Robotic Tape Libraries
- Resource management (tapes, drives)
- Tape allocation accounting per storage group, media type.
- Self-monitoring, error-reporting and alarm services.
- Periodic data integrity checks.
- Web-interface.
- File migration from older media type to new media type (e.g. LTO3 → T10KC).
- Small File Aggregation.

The system is highly scalable. Fermilab currently operates three Enstore instances with total data stored approaching 40 TB, 50M files.



Transfer rates on all 3 production Enstore instances in Fermilab in the last 30 days

dCache and Enstore

Enstore can operate stand-alone or can use dCache, a disk caching system as a front-end. dCache system is developed jointly by DESY, Fermilab and NDGF.

References

- <http://www-ccf.fnal.gov/enstore/>
- <http://www.dcache.org/>

Acknowledgement

We would like to thank CMS T1 management and CMS Data & Debugging Services group for providing client nodes for Chimera acceptance testing.

Namespace provider

Namespace provides association between file names and data distributed across multiple locations and on variety of media.

Namespace supplies:

- Unique file ID independent from file name.
- Path to ID mapping.
- File metadata storage.
- Directory tags inherited by sub-directories to store extra metadata associated with storage system.
- Callbacks on file system events.

Both dCache and Enstore use(d) PNFS implementation of Namespace provider developed by DESY in late 90-s.

PNFS is NFSv2 filesystem on top of the database:

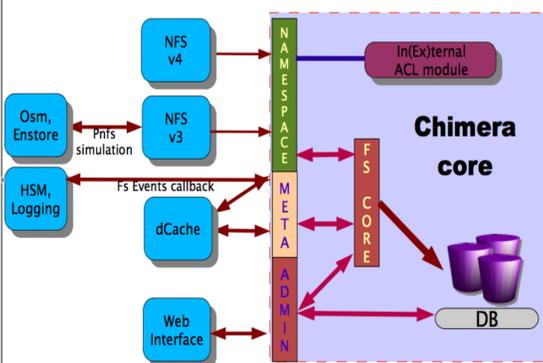
- Supports all NFSv2 namespace operations.
- Actual I/O is performed by HSM tools
- Allows storage of user defined metadata associated with files and directories.

PNFS limitations:

- Maximum file size is 20 GB.
- Metadata access only through NFS server.
- No direct path for Storage System clients.
- Storage System performance degrades when users perform heavy metadata lookups and vice-versa, heavy Storage System usage impacts NFS operations.
- Metadata stored as (key, value) pair. The value stored as BLOB. No efficient metadata query functionality. No selective metadata querying.
- No support for ACLs
- No security.
- Not easily portable.

PNFS has been de-supported by DESY in favor of the next generation product – Chimera.

Chimera Namespace provider



Chimera is high performance replacement for PNFS:

- Built on top of relational DB allowing efficient metadata queries.
- Clean API for namespace operations, metadata manipulations and administration.
- dCache accesses metadata directly, bypassing NFS server achieving higher throughput.
- Relies on stored procedures for transactional security.
- Platform independent:
 - Pure Java implementation
 - JDBC driver for DB access eliminating DB flavor specific bindings.
- Pluggable permission handler.
- NFS version supported:
 - V2 (legacy)
 - V3 (legacy) overcomes 2GB file size limit.
 - V4 with GSS authentication
 - V4.1 allows parallel POSIX I/O on distributed data. dCache + Chimera becomes a real filesystem.

Chimera and Enstore

Adaptation of Chimera in Enstore entails modifying encp only. Filesystem specific implementation of generic interface is instantiated at run time based on directory tags.

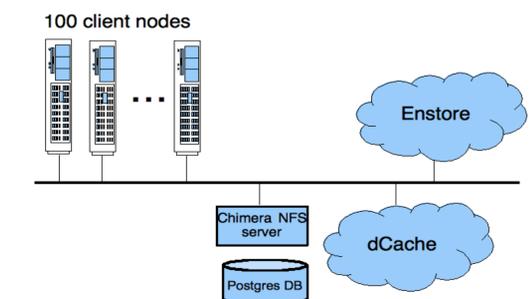
Chimera support is available in encp v3_10e

Acceptance test

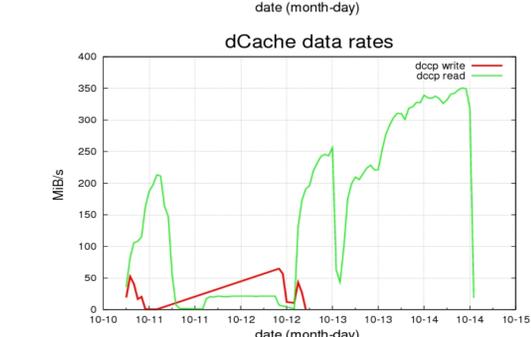
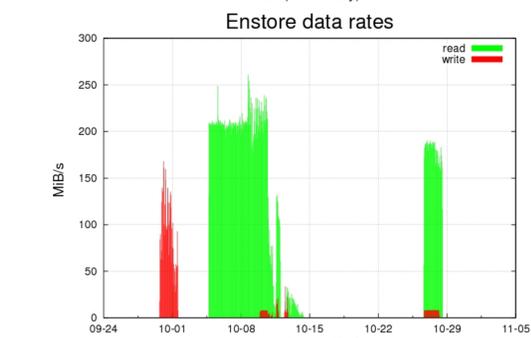
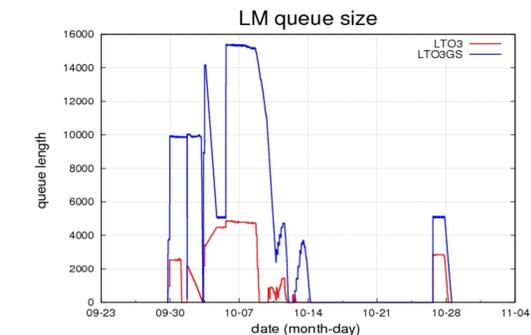
The goal of the acceptance test was to determine that Enstore with Chimera namespace provider is stable under heavy load in close to production environment. Tests involved performing direct encp transfers as well as dccp transfers using dCache.

Test setup:

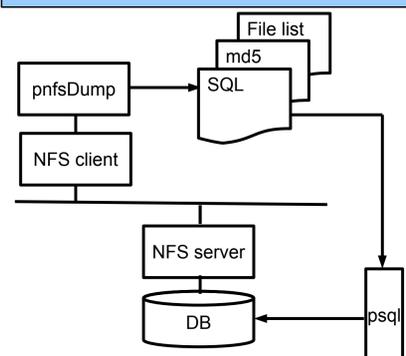
- 100 client nodes
- Fully functional Enstore system attached to SL8500 robotic library:
 - Separate Namespace node
 - Separate node for Enstore databases and database servers
 - Two LTO3 Library Managers (LMs)
 - 6 LTO3 tape drives
- dCache system:
 - 10 pool nodes, 1 dcap door per node
 - Separate head node
 - dCache version 1.9.5-28 (support for Enstore and Chimera is available in dCache starting 1.9.5-24)



- Spawned on average 130 simultaneous encp/dccp per client node
- Created load similar or exceeding the typical load on production Enstore systems in terms of number of elements in LM queues:
 - Up to 15K request in LM queue
- No errors or dropped requests observed



PNFS->Chimera migration



action	stken	CDF	D0
pnfsDump	6h52m	5h32m	12h53m
SQL Import	9h 47m	5h38m	8h28m

Timing of pnfs → Chimera migration on 3 production Enstore instances @ Fermilab

