



Chimera

Dmitry Litvintsev

SCF leaders meeting
11/16/11



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)



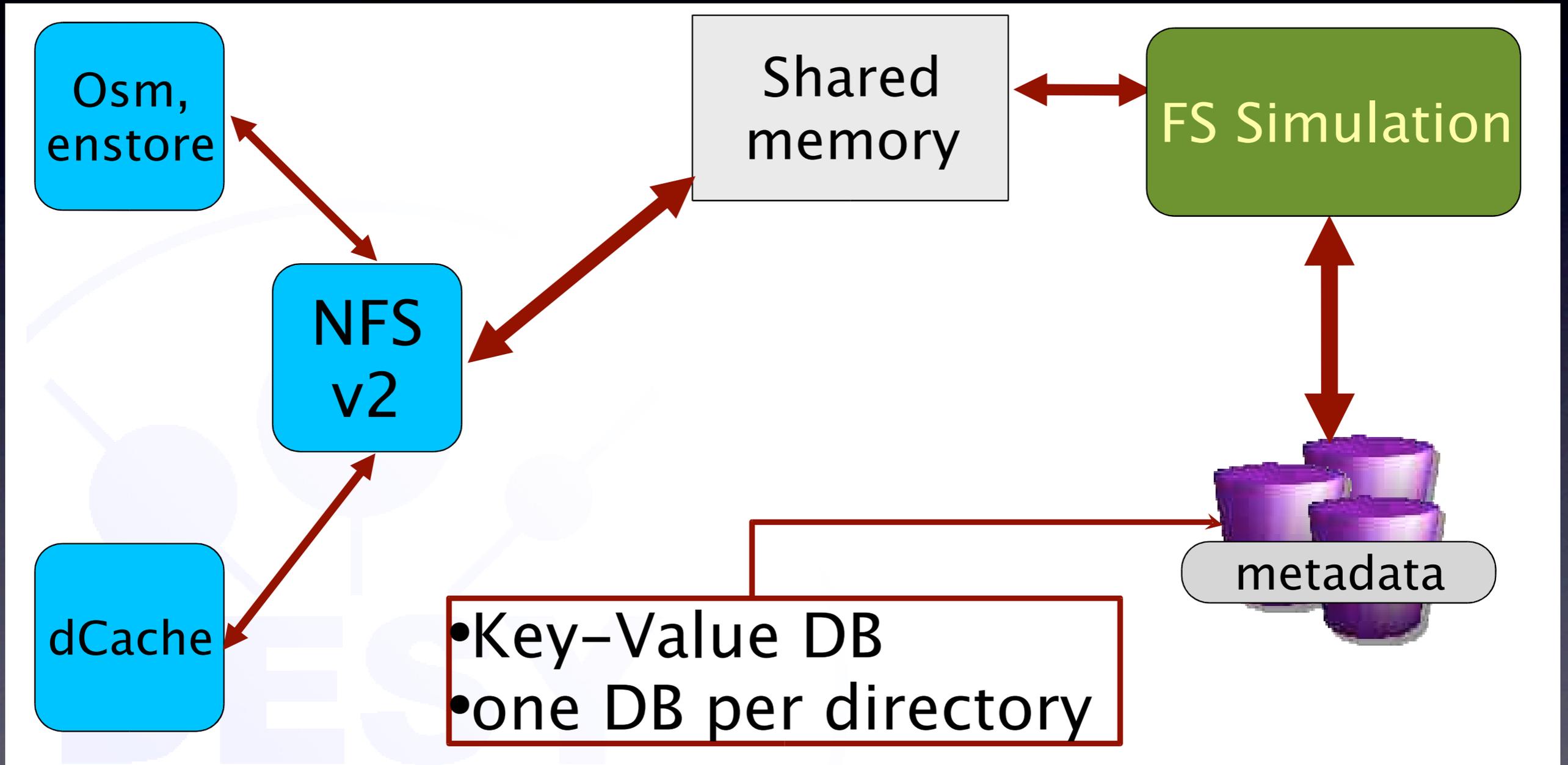
PNFS

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 architecture



stolen from Tigran Mkrtchyan's talk



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 fcdftcache91 kernel: NFS: server cdfensrv1.fnal.gov error: fileid changed
```

```
Nov 15 15:40:41 fcdftcache91 kernel: fsid 0:16: expected fileid 0x1d7ce677, got 0x1d7ce670
```

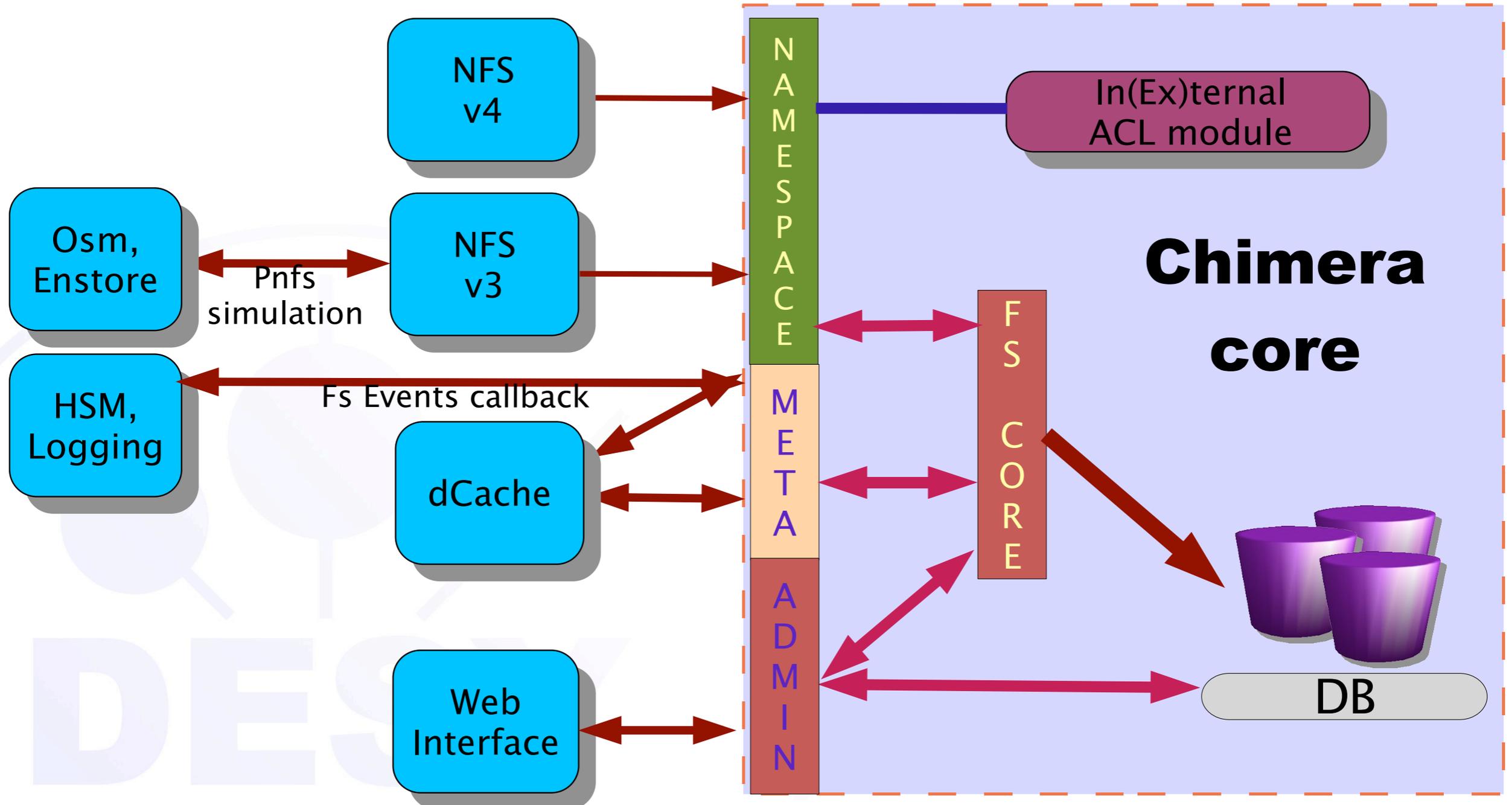
- **No expertise in PNFS code base**
- **In our tactical plan we identified reliance on PNFS as a significant risk factor**



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.

Chimera Architecture



stolen from Martin Radicke's talk



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.



Enstore/Chimera testing

- setup “real” (attached to tape library) Enstore instance:
 - two LTO3 library managers
 - 5 tape movers. 60 LTO3 tapes.
 - separate node hosted chimera
 - database servers on separate node
- kind folks @ CMS T1 allowed us to use about 100 of their worker nodes to run encp clients for about 4 weeks.
- 10 of these nodes were also used to setup dCache (1.9.5-28) with Enstore backend.

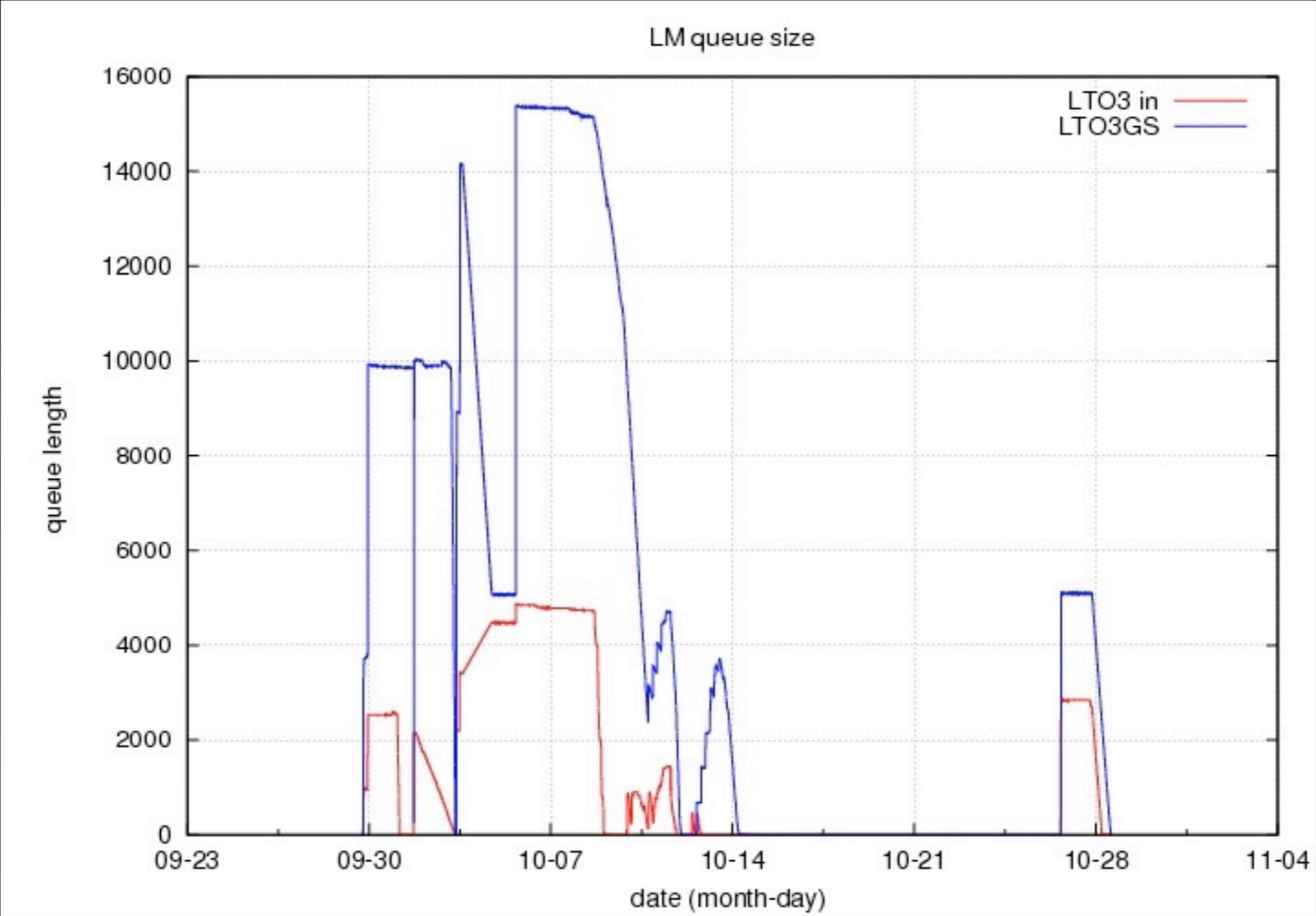


Enstore/Chimera testing

- deployed encp v3_10zzzzb (pre-production release of v3_10d) on 100 CMS worker nodes
- created load similar or exceeding typical production Enstore loads : up to 18K request in LM queue:
 - submitted about 130 simultaneous encp to write data to Enstore
 - submitted about 130 simultaneous encp clients to read data from Enstore
- performed mixed write/read tests with and without dCache as end-to-end exercise.

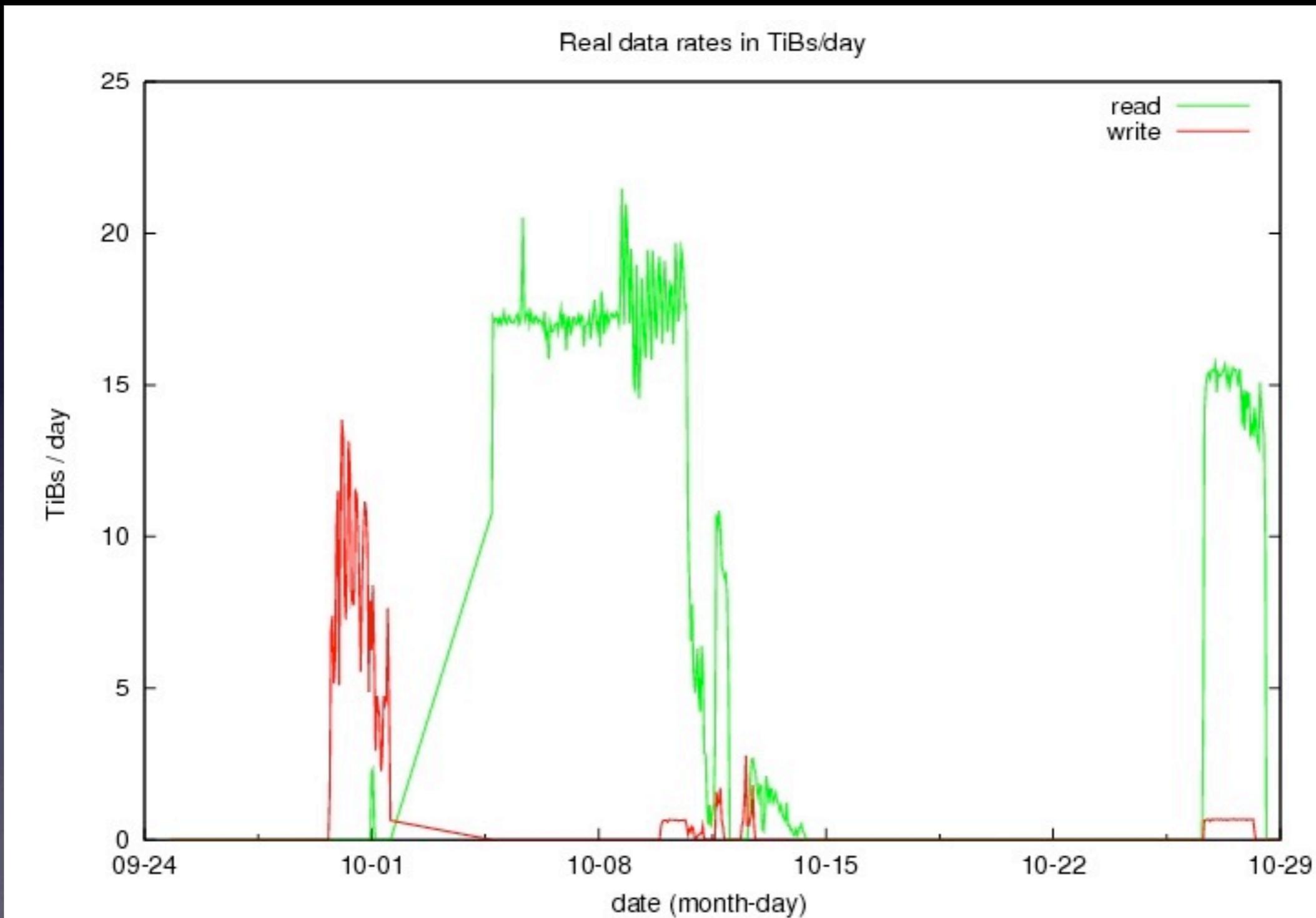


LM queues





Enstore/Chimera testing



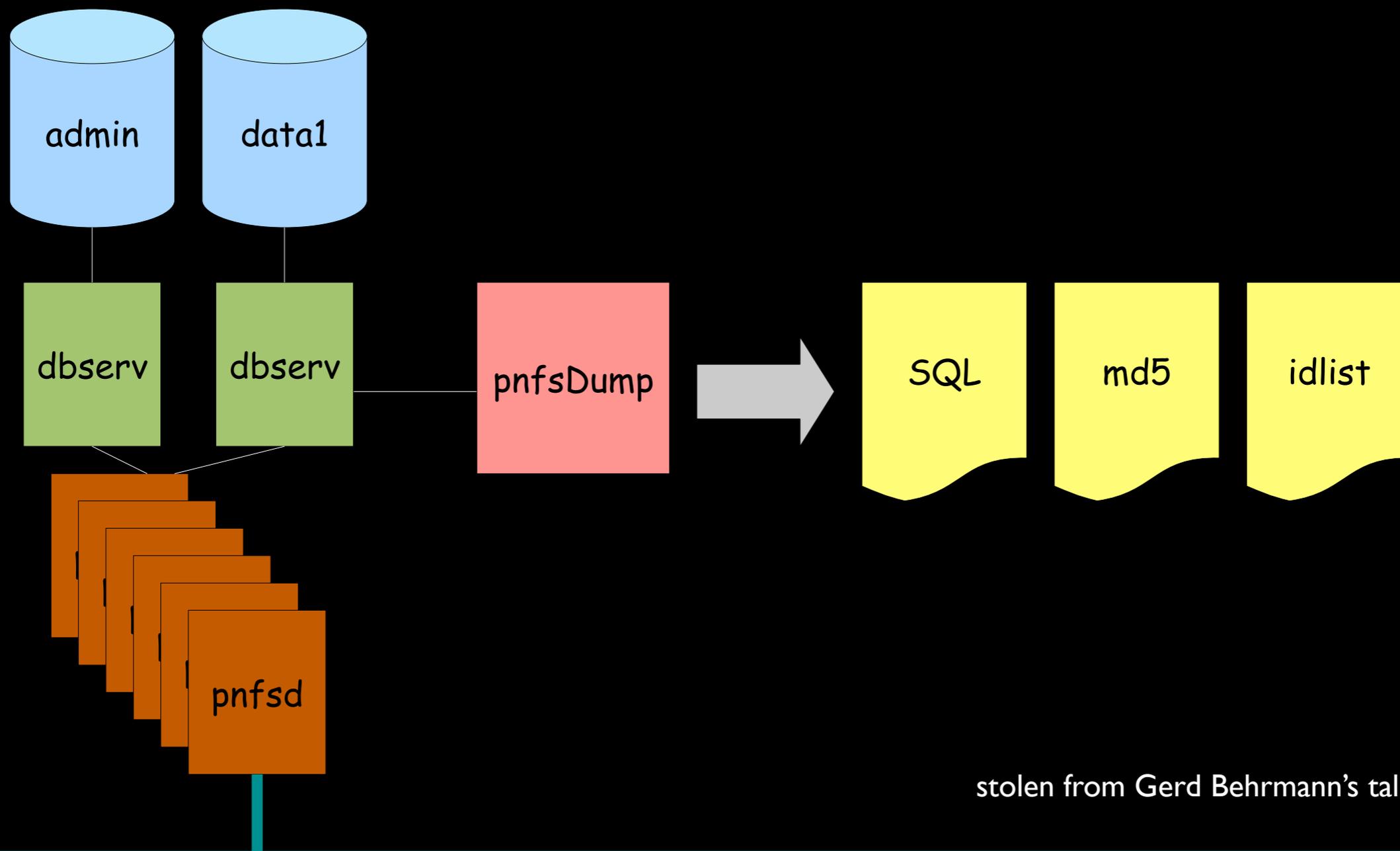


Enstore/Chimera testing

#transfers	TiBs	Enstore/dCache	r/w
83198	161.5	Enstore	read
23749	25.4	Enstore	write
1215	2.3	dCache	read
9869	1.4	dCache	write

No errors observed

PNFS -> Chimera



stolen from Gerd Behrmann's talk



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



Chimera

all files on tape (in GiBs) :

```
select sum(ti.isize)/1024./1024./1024. as total from t_inodes ti,  
t_locationinfo tl where ti.ipnfsid=tl.ipnfsid and tl.itype=0;
```

total

2427106.978934593486

(1 row)



Chimera

all files in dCache pools (in GiBs):

```
select sum(ti.isize)/1024./1024./1024. as total from t_inodes ti,  
t_locationinfo tl where ti.ipnfsid=tl.ipnfsid and tl.itype=1;
```

total

107004.133253824893

(1 row)



Chimera: todo

- Go thru existing SSA procedures that use PNFS and checking if these procedures work with Chimera
- Provide detailed migration plan