

Federated ATLAS XRootd: Implementation and discussion

Charles G Waldman
University of Chicago/USATLAS

Constraints

- T1 and T2 sites do not store files according to “global namespace” (PDP)
 - Prefixes for different storage areas
 - `_DQ-xxx` suffixes left behind by storage agents
 - Differing LFN->PFN conventions over time
- Currently, LFC lookup cannot be avoided
- But, we want to eliminate (or at least consolidate) LFC!

Implementation

- Must be fast – xrd redirector expects prompt replies (~200ms)
- Must not cause undue load on LFC or storage system
- In-memory caching of LFC and pnfs queries
 - Configurable TTL and max cache size (2 hours / 500k entries)
 - Cache positive results only (files may show up)
 - Idea - cache negative results with short TTL?

Name2Name

- Input: global namespace path
/atlas/xxx/yyy/DSN/file
- /atlas → /grid/atlas can be handled by symlink in LFC
- Symlinks do not help in all cases, e.g.
/grid/atlas/dq2 vs /grid/atlas/pathena,
/grid/atlas/users, etc (where should symlink point?)
- Various heuristics applied to rewrite input path to cope with varying conventions (can we renormalize LFC paths?)

Name2Name, 2

- More expensive search: looking for `_dis` and `_sub` datasets

- Random example:

```
grid/atlas/dq2/mc09_10TeV/HITS/mc09_1TeV  
.105802.JF17_pythia_jet_filter.simul.H  
ITS.e469_s595_tid09528755_sub04316149/  
HITS.095287._552779.pool.root.1
```

- Need to go up 1 level, list LFC subdirs, look for matches
- Search currently disabled, pending review

dCache bypass mode

- xroot is more lightweight than dCache, so read files direct from pool rather than via Java layer
- Pools run both dCache and Xrd software
 - Xrd has small VM footprint, uses 'sendfile', etc
- Lookup & cache PNFS id, search in pools
 - `/dcache/pool3/51213512` (e.g.)
- Probably want a local xrd redirector, if doing this

xrd-lfc

- “Name2Name” plugin (.so)
- Config. Params:
 - lfc_host (LFC_HOST env. var.)
 - lfc_cache_ttl, lfc_cache_maxsize
 - root (start of filesystem path in SFN, e.g. /pnfs/)
 - match (for shared LFC, etc)
 - nomatch (avoid e.g. tape files)
 - dcache_pool[s]
 - force_direct

Example config (UC)

```
xrd.port 1094
all.role server
all.manager atl-grdr.slac.stanford.edu:1213
all.export /atlas r/o
oss.namelib /etc/xrootd/XrdOucName2NameLFC.so \
root=/pnfs match=uchicago.edu \
lfc_host=uct2-grid5.uchicago.edu \
dcache_pools=/dcache/pool*/data force_direct
xrootd.fslib /usr/lib64/libXrdOfs.so
all.adminpath /var/run/xrootd
all.pidpath /var/run/xrootd
xrootd.trace emsg login stall redirect
ofs.trace none
xrd.trace conn
cms.trace all
```


Status

- Sites are interested in participating in tests
 - Currently: MWT2, SWT2, SLAC
 - Next up: SMU
 - Some interest also from European sites (Wuppertal)
- Performance tests: starting next week
- Code (comments welcome):

<http://repo.mwt2.org/viewvc/xrd-lfc>

cgw@hep.uchicago.edu