Lustre Directory Migration and Restripe

Lai Siyao 2018

Whamcloud

Directory Migration

Directory Migration Overview

- Migrate directory to specified MDTs to improve scalability
- Can be used to empty certain MDT for maintainence
- lfs migrate –m <start_mdt_index> -c <stripe_count> -H <hash_type> <directory>

Directory Migration Internals

Whamcloud

- Client send migrate RPC with new layout to MDS
- MDT lock all objects involved, including all link parents of source
- MDD create target with new layout, append source stripes to new layout, destroy source and update namespace with target FID
- Client iterate sub files under source, and migrate them to target
- Client send SETXATTR RPC to remove source stripes from target

Whamcloud

Directory Restripe

Directory Restripe Overview

- Balance MDT usage
- Improve directory scalability when it grows
- Move as little data as necessary to minimize system impact
- Transparent to user

Directory Restripe Functionality Spec

Whamcloud

- Basically it's directory migration originated by MDT, one thread per MDT
- Restripe plain directory only
- Start directory restripe when its size exceeds limit, e.g. 100k bytes
- Restripe to a certain number of MDTs, e.g. 8, which can be set by user
- FID mapping to support transparent restripe and NFS reexport
  - FID mapping cache is stored in OI, and it's an LRU cache
  - if current RPC should be sent to the target MDT where FID is located, but the mapped FID is on another MDT, return –EREMOTE, and client will retry with mapped FID
  - else MDT just uses the mapped FID to locate object
- Recovery
  - Stateless
  - Current migration replay as before
  - MDT will continue directory restripe if directory is marked RESTRIPE, but not in current restripe list

Whamcloud

DNE Status Update

Major Issues

- LU-6848: Support multiple modify RPCs in flight for MDT-MDT connection
- LU-7426: Current llog format for remote update llog
- LU-7427: multiple entries for BATCHID
- LU-10784: mkdir() automatically create remote directory on MDS which has more space
- LU-11213: remote mkdir() in ROOT/ by default
- LU-10329: REMOTE_PARENT_DIR scalability
- LU-10888: lctl abort_recovery' allow aborting recovery between MDTs

Thanks!