FIDs on OST
brainstorming

Mikhail Pershin
Lustre Group
2009-11-16

Summary

We introduce new FIDs on OST to unique identify all objects in Lustre.

Why?

- Unification of code, remove old code
- OFD works with FIDs already though they are IDIFs still
- LOV-OSC new stack will need that
- Anything else?

lu_fid

```
/**
 * File IDentifier.
 *
 * FID is a cluster-wide unique identifier of a file or an object (stripe).
 * FIDs are never reused.
 */
struct lu_fid {
        /**
         * FID sequence. Sequence is a unit of migration: all files (objects)
         * with FIDs from a given sequence are stored on the same server.
         * Lustre should support 2^64 objects, so even if each sequence
         * has only a single object we can still enumerate 2^64 objects.
         */
        __u64 f_seq;
        /** FID number within sequence. */
        __u32 f_oid;
        /**
         * FID version, used to distinguish different versions (in the sense
         * of snapshots, etc.) of the same file system object. Not currently
         * used.
         */
        __u32 f_ver;
};
```

What is done already

- IDIFs are defined and introduced

```
f_seq = IDIF_SEQ_START | (ost_idx << 16) | ((objid >> 32) & 0xffff)
f_oid = (objid & 0xffffffff)
f_ver = 0
```

See http://arch.lustre.org/index.php?title=Interoperability_fids_zfs

- OFD is big step forward, uses lu_fid internally
- FID/SEQ service is working part of MDS and it is easy to run it on OST as well

What should be done

- FLDB, SEQ
  - services on OST
  - Client part rework
  - FLDB rework to refer OSTs too
- Precreation/orphans to work with FIDs
- Interoperability

Where allocate FIDs

- Client
  - Precreation becomes very difficult as we will need (clients * OSTs) precreation pools
  - No orphans recovery
- MDS
  - Precreation will work like now if FIDs are allocated by MDS.
  - Old objids model can be still used for interoperability
  - We could use llog to cleanup orphans

FLDB, SEQ services

- FID/SEQ client works on top of MDC now, need to be re-done
- FLDB uses just MDS index to identify MDS.
  - Uses plain index to identify MDS
- Configure FLDB through MGC?

Precreation

- Can be still used if OST FIDs allocated by MDS
- Should be replaced with CROW eventually (anyway?)
- Objids represent the f_oid in newly allocated FIDs

Orphans

- Objids model can be used still, good for interoperability but has flaws
  - Id's sequence may have holes
- We could use llog
  - Place all precreated object ids in unlink llog immediately after precreation done
  - Cancel them in llog upon using for real create
  - After recovery all unused objects will be deleted during ordinary llog processing from OST

Interoperability

- What are interop cases?
- Old objects are mapped to IDIF
- Old objids files on MDS
  - Cleanup orphans and switch to new format?
  - Backward: re-create objids files
- Wire
  - OBDO has o_id/o_gr