



lustre®

Lustre Interoperability

Huang Hua

Lustre Group

Sun Microsystems, Inc.

Interoperability Overview

Bug #:

Bug 11930: for wire protocol (done)
Fid alloc, fld

Bug 11824 for 1.8 recovery, 17911 for 2.0

Bug 11826: 2.0 changes tracking
lgif

Mount (config log and disk files)

Fid in EA

Orphan

Readdir

Object in group zero

Interoperability Overview (cont.)

Motivation

- Make main releases interoperable

Objectives:

- Full & seamless interoperability between 1.6 & 1.8: same code base

 - ptlrpc_body

 - AT enabled in 1.8

- Simplified interoperation between 1.8 & 2.0

 - 1.8 clients + 2.0 servers

 - Upgrade order: mds, oss, clients

 - Downgrade order: clients, oss, mds

Interoperability Overview (cont.)

Design Highlights

- Default, 2.0 uses the same disk format as 1.8
 - For 2.0 CMD, IAM disk, not compatible.

- 1.8 client speaks both 1.8 protocol and 2.0 protocol

- The same code base in HEAD to handle IOP and CMD

1.8 client changes

OBD_CONNECT_FID flag

In mdc/osc, packing different request according talking to 1.8 servers or 2.0 servers, parsing different reply.

mdt_body instead of mds_body

mdt_rec_* have the same size

In mdc, allocate FID for new file if needed while talking to 2.0

Use 2.0 recovery algorithm while talking to 2.0

2.0 changes

IOP mode (default, disk compatible)

Standard ext3/4 directory format

Store FID in EA: trusted.lma

The same orphan list format as 1.8

Objects are created in group 0

Generic local objects handling

CMD mode

IAM disk format:

directory format: {name, fid}

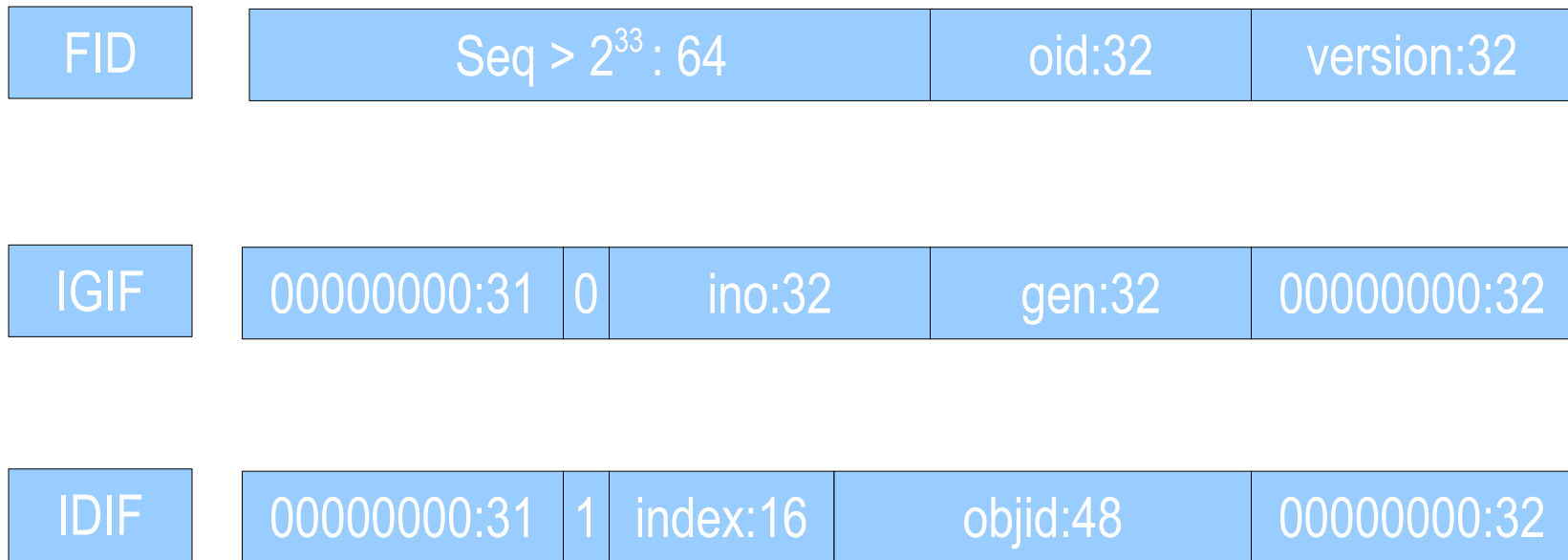
Objects are in group 0, 3, 4, 5, ...

IAMDIR=yes, mkfs.lustre --iam-dir

The same wire protocol and code base

FID, IGIF, IDIF

Some special FID are reserved.



Use case: create

Create request: parent FID, name, child FID

Lookup and permission checking

Create ost objects

Create file/dir/... on mds:

- store fid in EA if needed (IOP)

- Insert object index

Add dir entry

Use case: readdir

MDS_READPAGE from client to mdt

Iterate over directory from specified position:

IOP: {name, ino}, retrieve FID from EA, IGIF if needed.

CMD: {name, FID}

Pages of {name, FID} are returned

2.0 changes for lprocfs

/proc/fs/lustre/mds -> mdt

/proc/fs/lustre/mds -> osd

 blocksize filesfree filestotal fstype kbytesavail kbytesfree
 kbytestotal

 mntdev

/proc/fs/lustre/osc/lustre-OST0000-osc-MDT0000/

/proc/fs/lustre/lov/lustre-MDT0000-mdtlov/

Current Status

Interop for 1.6 & 1.8: full

Interop for 1.8.0 and 2.0 (HEAD)

- 1.8.0 client talks well to 2.0 servers

- 2.0 is full compatible with 1.8 disk

- upgrade/downgrade: cold, or no replay/resend

Simplified interoperation enables live upgrade

- Targeted to land to 1.8.1 and 2.0

- Some changes are expected.

Use case: Upgrade

Upgrade all clients & servers to 1.8.x

Upgrade mds to 2.0: a normal failover

Upgrade oss to 2.0: a normal failover

Upgrade clients to 2.0: mixed versions are allowed

No error, no client/server eviction

Use case: Downgrade

All clients/servers are 2.0

Downgrade ALL clients to 1.8.x

Downgrade OSS to 1.8.x: normal failover

Downgrade MDS to 1.8.x: normal failover, abort
recovery

Clients are evicted, reconnection

Resources

Arch page:

[http://arch.lustre.org/index.php?
title=Interoperability_fids_zfs](http://arch.lustre.org/index.php?title=Interoperability_fids_zfs)

[http://arch.lustre.org/index.php?
title=Simplified_Interoperation](http://arch.lustre.org/index.php?title=Simplified_Interoperation)

HLD and DLD

Team:

Huang Hua

Rahul Deshmukh

Pravin Shelar

Nikita Danilov (desinger)



THANK YOU