Distributed Namespace Environment Phase I

High Performance Data Division

Di Wang

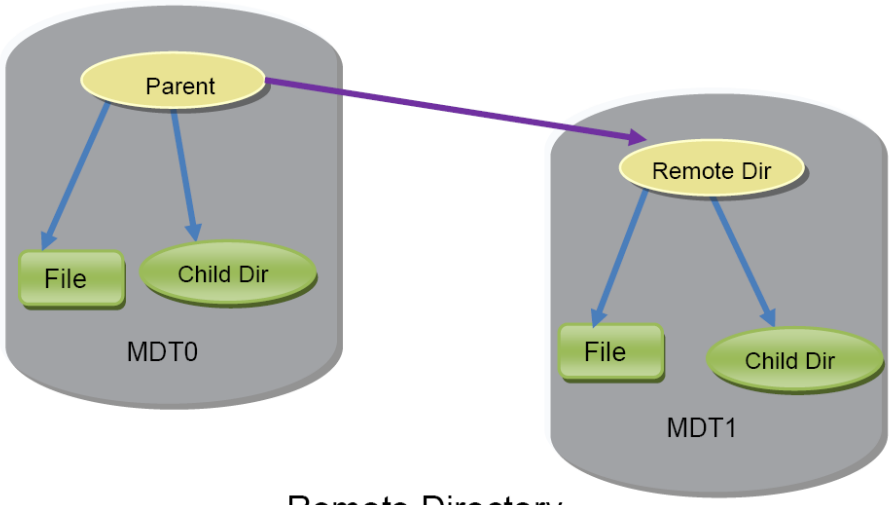
04/16/2013

Agenda

- Introduction
- Phase I
  - Remote directory
  - Failover
  - Disk layout
  - Performance
  - Limitation
- Phase II

Introduction

- DNE is sponsored by OpenSFS, and Phase I will be released in Lustre 2.4
- DNE Phase I distributes Namespace by remote directory



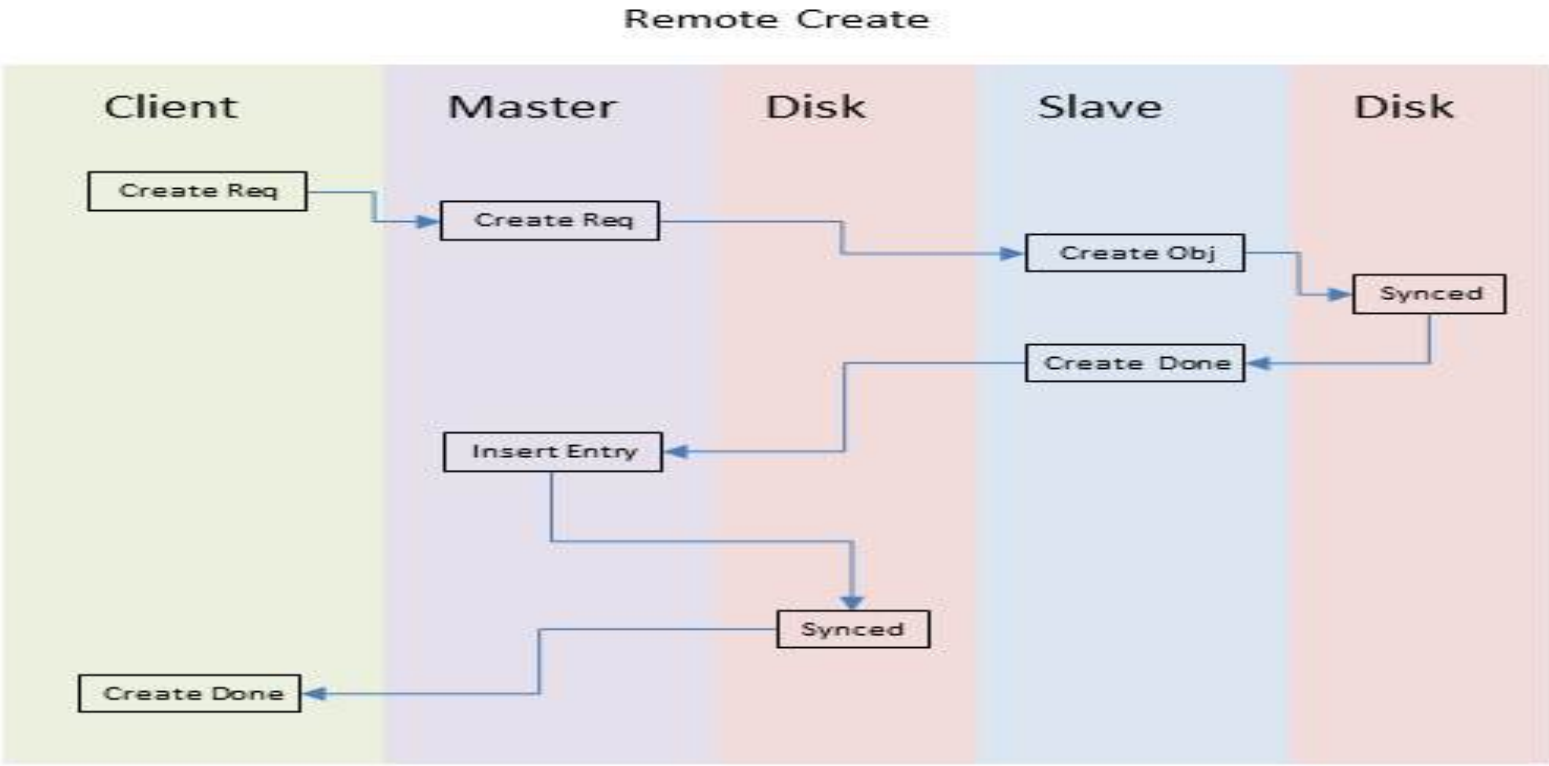
Remote Directory

Remote directory

- Create child on remote MDT by special lfs command
- Only admin can create remote directory on MDT0
  - *lfs mkdir –i n remote_dir* #create remote directory on the nth MDT
  - *rmdir remote_dir* #remove remote directory
- Tunable to allow normal users to create remote directory on other MDT
  - *lctl set_param mdt.fsname-MDT0000.enable_remote_dir=1*
  - *lctl set_param.mdt.fsname-MDT0000.enable_remote_dir_gid=allowed_gid*
- DNE will work for both ldiskfs and ZFS

Remote directory

The remote operations are synchronized to avoid recovery problems

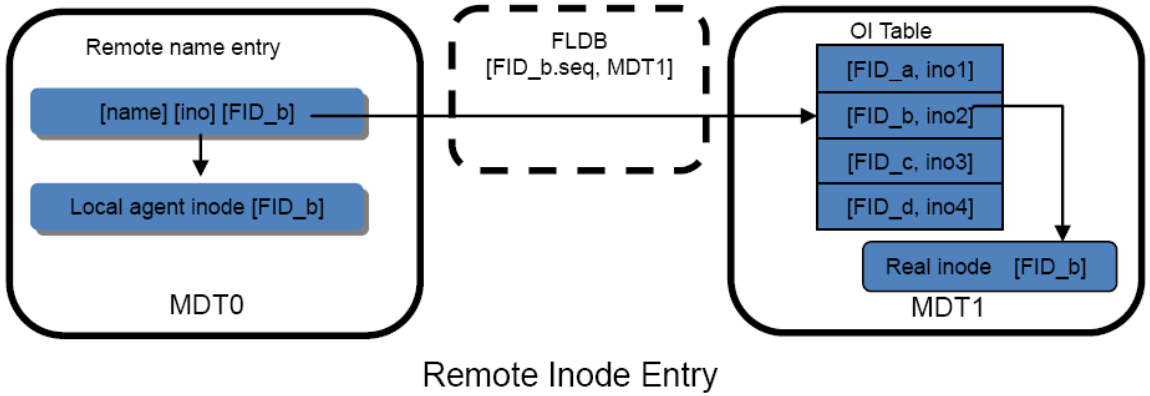


Failover

- Active-Active failover
  - Allowing multiple MDTs to be exported from one MDS, Lustre* can support active-active failover for metadata as it already does for data
- Permanent MDT failure
  - The failure of MDT0 is an extreme case which can make the whole file system inaccessible
  - The failure of other MDTs will isolate any of its subsidiary directory trees

Disk Layout

- Remote directory



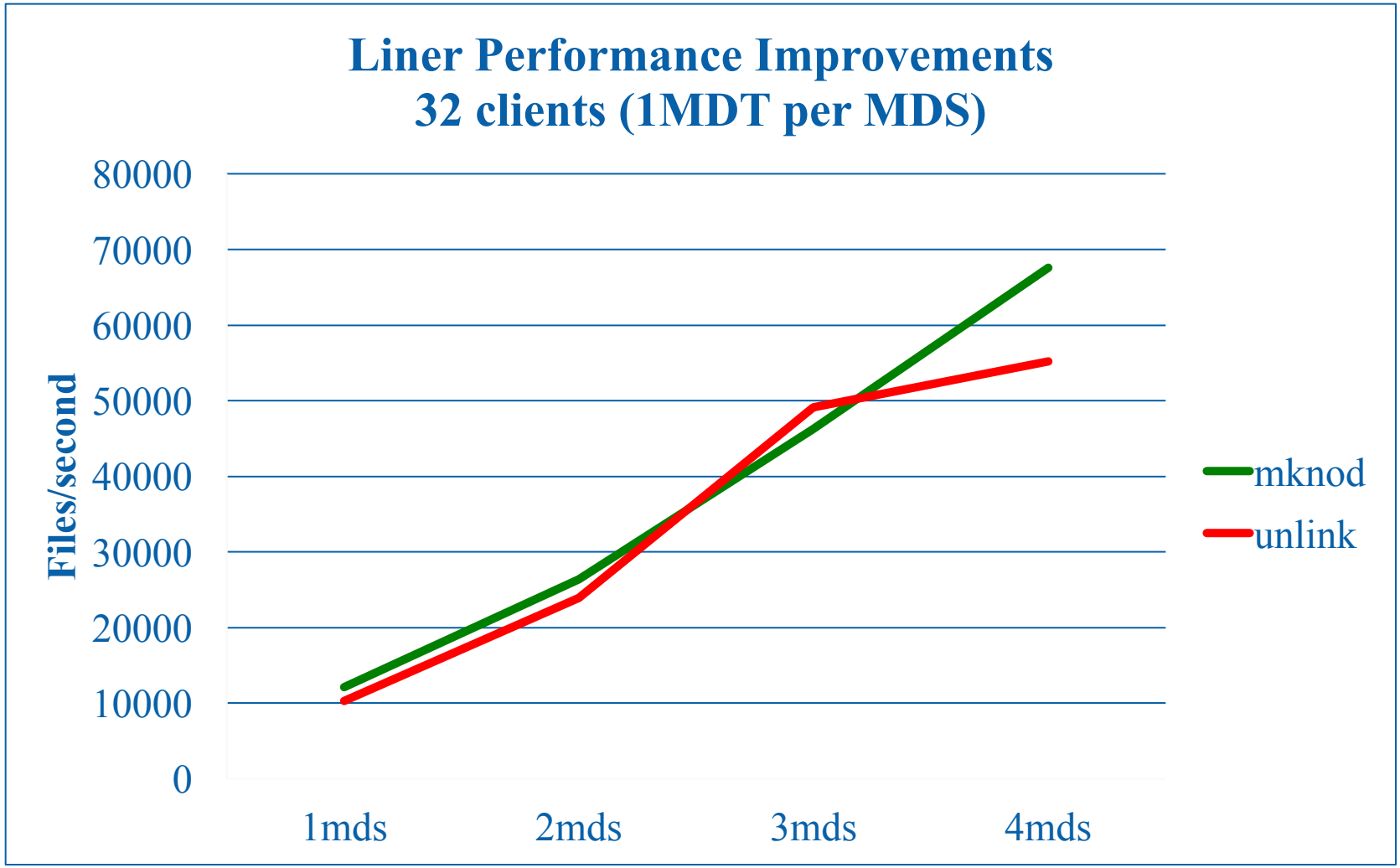
Remote Inode Entry

  - FID will only be put in xattr (LMA), 2.4 will also store it in directory entry
  - LFSCK phase III will check and fix remote directories online.
    - Only off-line check is currently supported
    - But the consistency of remote directories is not being checked/fixed right now

Upgrade to DNE

- All Lustre servers and clients are either 1.8/2.x
- Shutdown MDT and all OSTs
- Upgrade MDT and all OSTs to Lustre 2.4
- Remount MDT and OSTs
  - Erase the config log with tunefs.lustre, if upgrading from 1.8 to DNE
- Adding new MDT by
  - *mkfs.lustre --reformat –mgsnode=xxx –mdt --index=1 /dev/{mdtn_devn} mount -t lustre -o xxxx /dev/{mdtn_devn} /mnt/mdtn*
- Upgrade clients to Lustre version with DNE
  - Old clients (pre 2.4) can still access the filesystem, but only MDT0

DNE performance

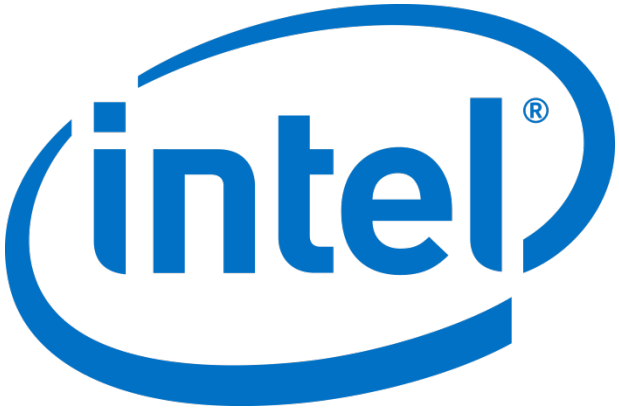


Limitations

- Only remote directory creation/unlink are allowed, and other remote operations will return –EXDEV
- No FS checking tool for DNE
- Might leave some orphans
- Only using copy/remove to migrate directories/files to the new MDTs
- Cross MDT operations are synchronized

DNE phase II

- Fully functional DNE
  - Directory migration tool to move inodes to new MDT
  - Any metadata operations can be cross-MDT (rename, link)
  - Normal users can do remote operation
  - No synchronous cross-MDT operations


intel