Distributed Name spacE phase I

High Performance Data

Di Wang

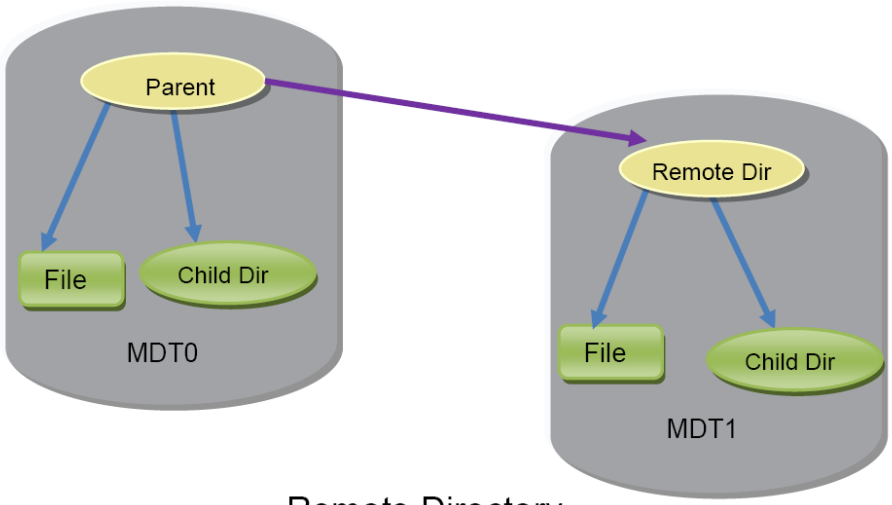
04/16/2013

Agenda

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

Introduction

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



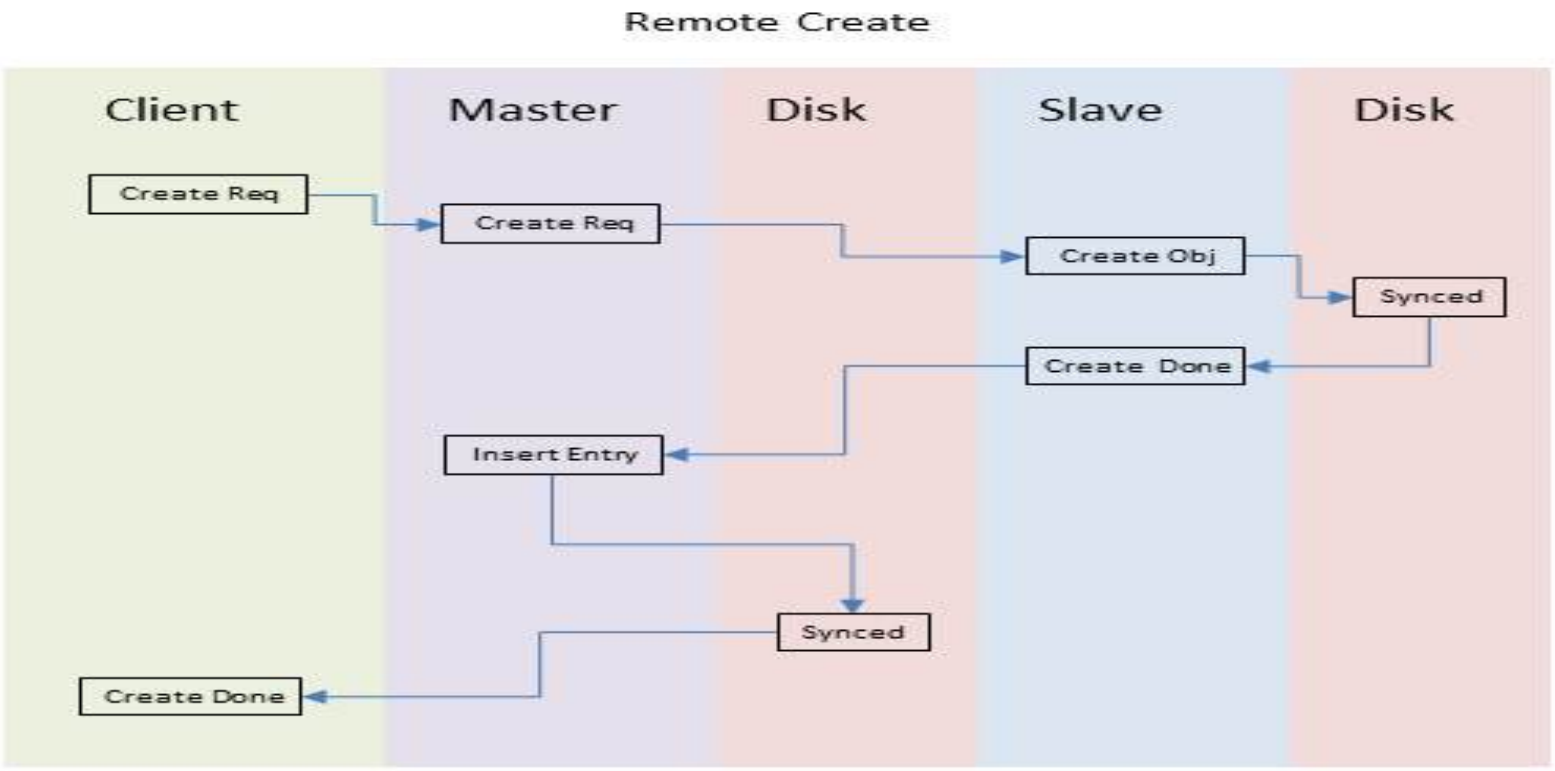
Remote Directory

Remote directory

- Create child on the remote MDT by special lfs commond
  - Only the administrator can create the remote directory on MDT0
    - lfs mkdir –i n remote_dir # create remote directory on MDT n
    - rmdir remote_dir # remove remote directory
    - Parameters 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=xx

Remote directory

- Remote operations are synchronous to avoid recovery problems

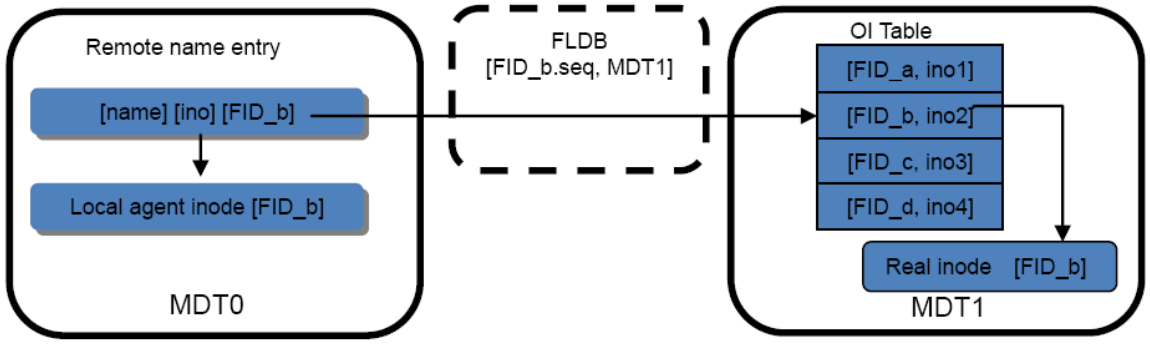


Failover

- Active-Active failover
  - Allows multiple MDTs to be exported from one MDS
  - Supports active-active failover for metadata as it already does for data
- Permanent MDT failure
  - Failure of MDT0 can make the whole file system inaccessible
  - Failure of other MDTs will isolate any of its subsidiary directory trees

Disk Layout

- Remote directory



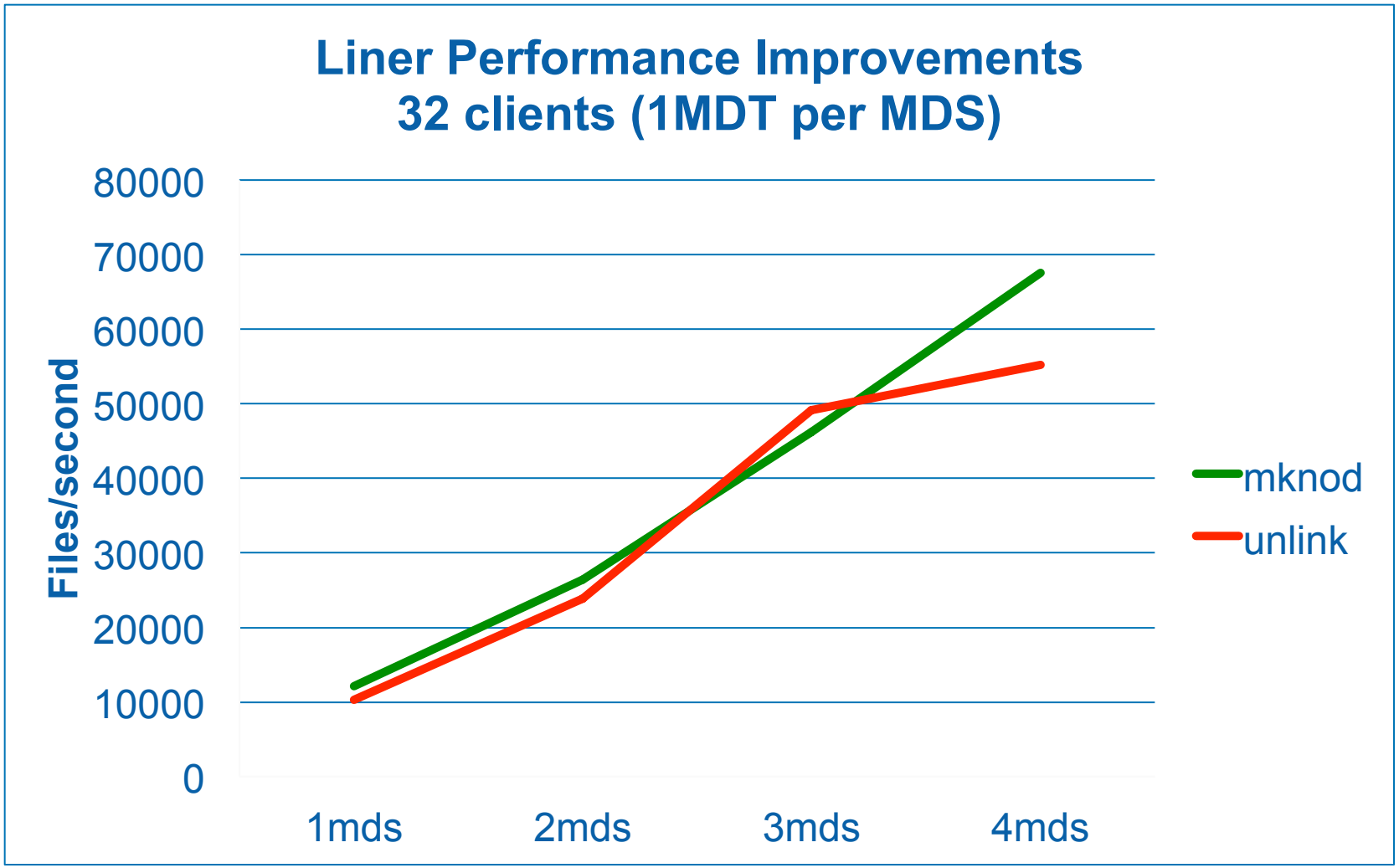
Remote Inode Entry

  - FID will be stored both in directory entry and EA(LMA) of the directory
  - LFSCK phase III will check and fix remote directories online
    - Off-line check is not supported for DNE

Upgrade to DNE

- All Lustre servers and clients are either 1.8/2.x.
- Shutdown MDT and all OSTs, then 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
  - Non-DNE clients can still access the DNE servers, but only files on MDT0

DNE performance

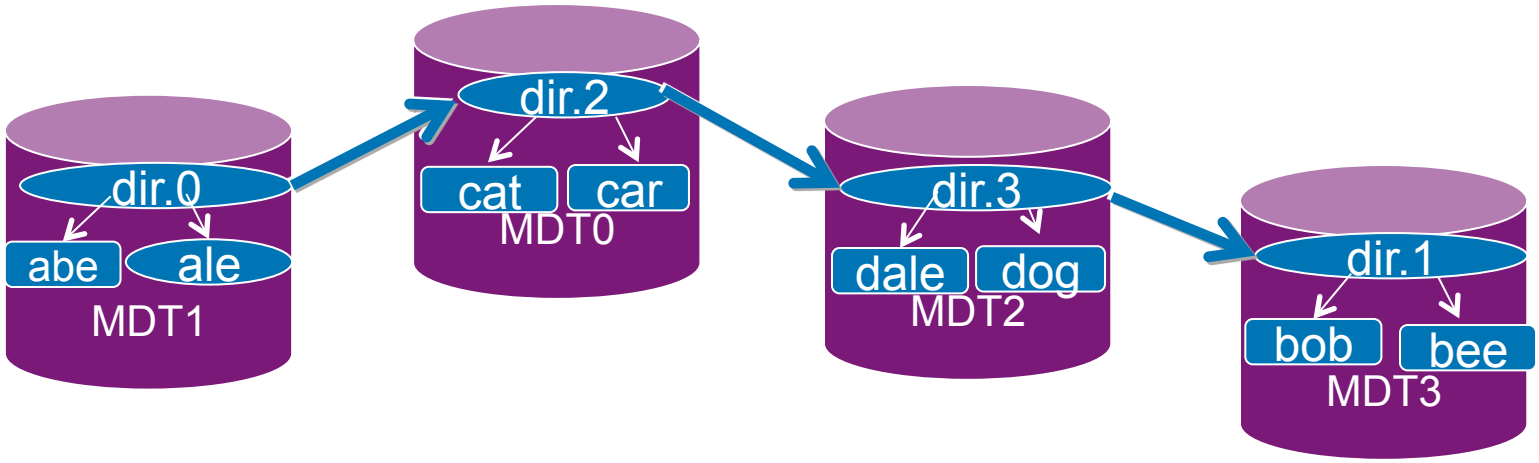


Limitation

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

DNE phase II

- Fully functional DNE
  - Migration tool
  - Any metadata operations can be cross-MDT
  - Normal users can do remote operation
  - No synchronization for cross-MDT operation
  - Shard directory



DNE phase III

- MDT pools
- Space balancing between MDT and QOS