Final Report for the *Remote Directories* of the Distributed Namespace Project of the SFS-DEV-001 Contract

Revision History

Date	Revision	Author
06/10/13	Original	R. Henwood
06/20/13	Links to OpenSFS site.	R. Henwood
06/26/13	Disclaimer included.	R. Henwood
07/17/13	Updated disclaimer	R. Henwood

1. Contents

Notices

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

2. Executive Summary

This document finalizes the activities undertaken during the Distributed Namespace project, Sub Project 2.1: Remote Directories project within the OpenSFS Lustre* software development contract SFS-DEV-001 signed July 30th 2011.

Notable highlights of this project include:

- Demonstrated linear scaling of metadata requests and showed a peak performance on four metadata servers of over 150K operations per second.
- FID-in-Dirent and LinkEA was implemented with around 10K lines of code on top of the Orion code and was completed and landed on Lustre Master for inclusion in release version 2.4 on May 31st 2013.
- All relevant assets from the project are attached to the public ticket LU-1187.

3. Statement of Work

Remote Directories distributes the Lustre file system namespace over multiple metadata targets (MDTs) under administrative control using a Lustre specific mkdir command. Whereas normal users are only able to create child directories and files on the same MDT as the parent directory, administrators can use this command to create a directory on a different MDT. The contents of any directory remain limited to a single MDT. Rename and hardlink operations between files and directories on different MDTs return EXDEV, forcing applications and utilities to treat them as if they are on different file systems. This limits the complexity of the implementation of this sub project while delivering capacity and performance scaling benefits for the entire namespace in aggregate.

Metadata update operations that span multiple MDTs are sequenced and synchronized to create and/or increment the link count on an MDT object before it is referenced by the remote directory entry and to update the remote directory entry before decrementing the link count and/or destroying the MDT object it referenced. Although this may result in an orphan MDT object under some failure conditions, it ensures that the Lustre file system namespace remains intact under any and all failure scenarios. All other metadata operations avoid synchronous I/O and execute with full performance.

4. Summary of Scope

1. In Scope

*Other names and brands may be claimed as the property of others.

- DNE code development takes place against the Lustre file system version 2.x Sequoia development branch.
- Interoperability of 1.8/2.1 clients with with Lustre file system version 2.x FID enabled OSTs.
- Multiple MDTs running on the same MDS node.
- Failover/failback of MDT to active backup MDS node.
- Administrative documentation in the form of man page for new user tools and update to Lustre file system version 2.x manual.
- Upgrade of existing 2.1 single-MDT file system to DNE-capable Lustre file system, with shutdown/restart.
- Addition of MDTs to existing DNE-capable Lustre file system, with shutdown/restart (at least initially).
- Accessing a DNE files system with more than 1 MDT from pre-DNE clients should fail with an error message.
- Remote directories with their parent directory on MDT0. Remote directories with parent directories on other MDTs will be possible with administrator override.

2. Out of Scope

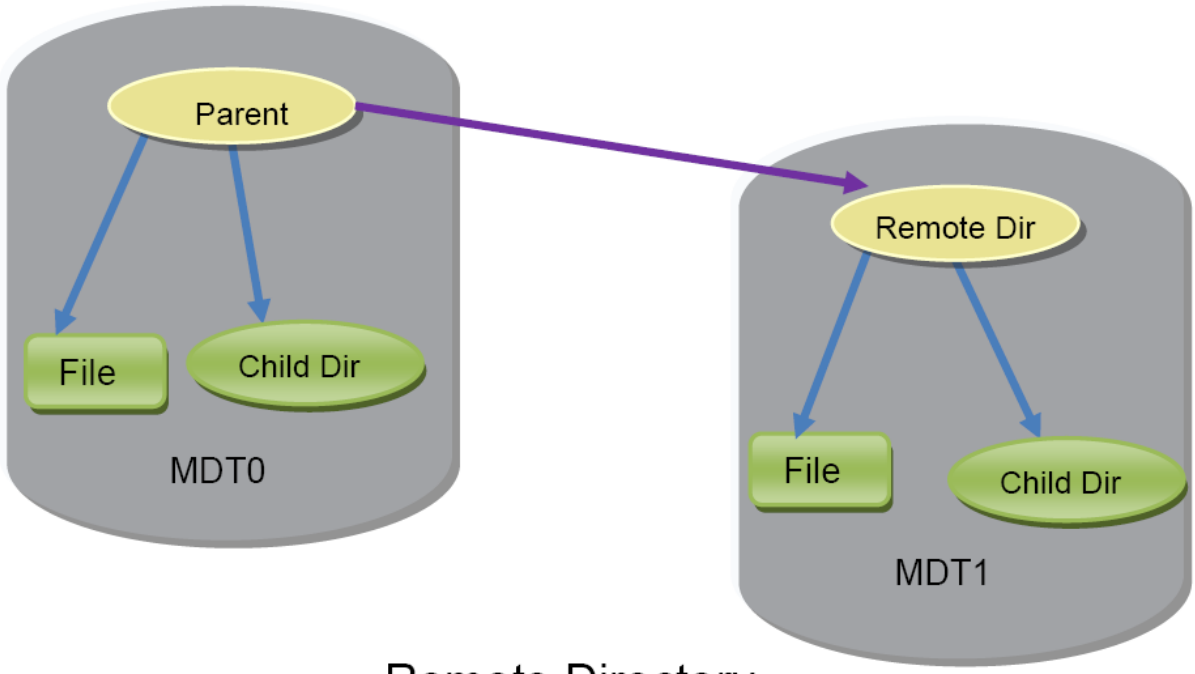
- Accessibility of remote DNE directories with 1.8 or 2.1 clients.
- Interoperability of DNE-enabled MDTs and non-FID-based 1.8/2.1 OSTs.
- Rename and hard-link operations will not work across directories (returning -EXDEV).
- Avoiding unreferenced remote directories (orphans) in case of interrupted mkdir/rmdir operations (depends on distributed transaction mechanism).
- Distributed consistency check of DNE-enabled filesystem (implemented via LFSCK phase III project).
- Recovering from a permanent failure when a remote directory has a parent directory on any MDT other than MDT0.

http://wiki.opensfs.org/images/d/df/DNE_RemoteDirectories_ScopeStatement.pdf

5. Summary of Solution Architecture

The goal of the Distributed Namespace (DNE) project is to a deliver a documented and tested implementation of Lustre file system that addresses this scaling limit by distributing the file system metadata over multiple metadata servers. This project is planned and executed in two phases. This document is concerned with the first phase: 'Remote Directories'.

With Remote Directories, Lustre file system sub-directories are distributed over multiple metadata targets (MDTs). Sub-directory distribution is defined by an administrator using a Lustre file system specific mkdir command. This is illustrated in the figure below:



Remote Directory

Remote Directories in itself is an ambitious engineering project that will take place over a period of many months. It requires considerable engineering and testing resources to cover the following aspects of a successful solution:

- Performance:
  - Scalability.
  - Isolation.
- Cross-MDT Operations:
  - Non-local rename and hard-link.
  - Non-local mkdir and rmdir.
- Compatibility:
  - Upgrading to DNE.
  - Adding an MDT.
  - Disabling an MDT.
  - Removing an MDT.
- Resilience:
  - Operations affecting multiple MDTs.
  - Remote directories.
  - MDT failover.

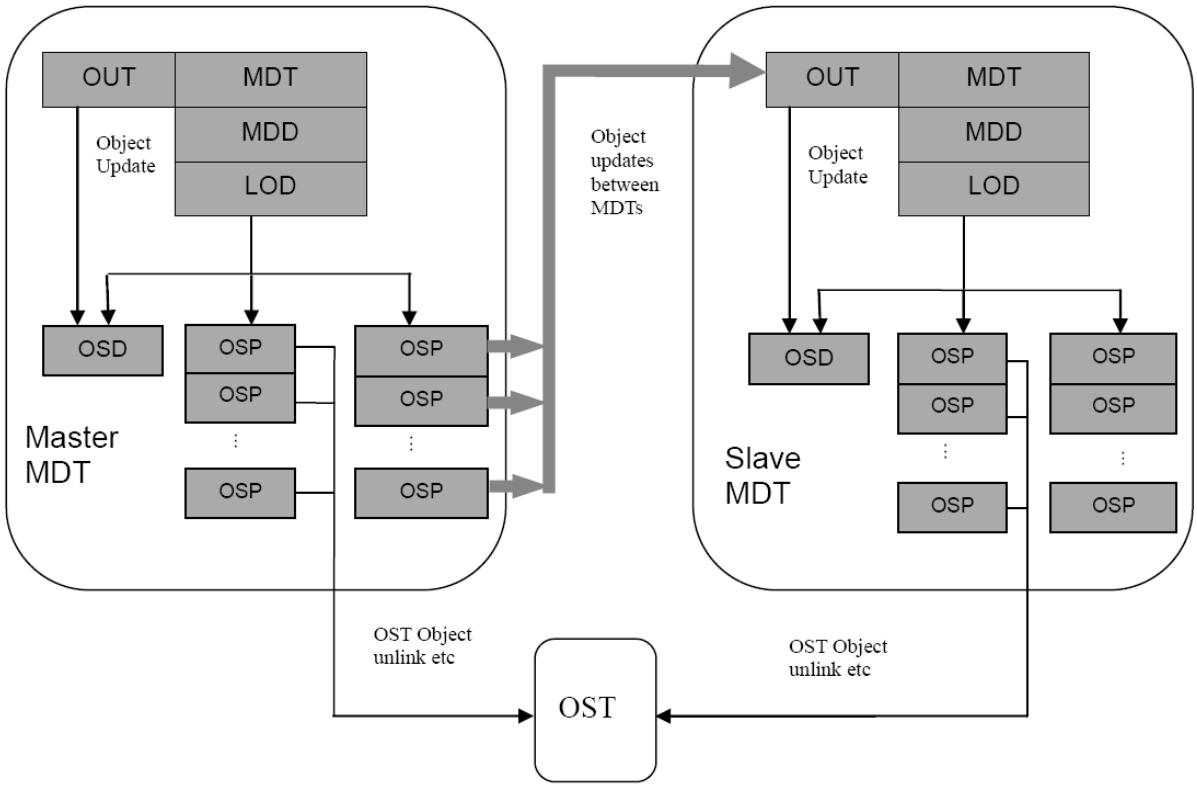
- Solution Proposal
    - Distributed metadata operations including
        - Remote directory creation
        - Remote create recovery
        - Resend between the Master MDT and client.
        - Resend between the Master MDT and the slave MDT.
        - Remote rmdir
        - Remote unlink recovery.
    - MDT Failover
        - Active-active failover.
- Integration test plan.
- Recovery test.
- Compatibility test.
- Performance test.
- Acceptance criteria.

The complete Solution Architecture is available at:

http://wiki.opensfs.org/images/4/43/DNE_RemoteDirectories_SolutionArchitecture.pdf

6. Summary of High Level Design

Remote Directories work is based on the Orion MDS stack. The Orion MDS stack is illustrated below:



New MDS layer

The high level design covers the following areas in detail:

- Transaction design.
- FID design.
- Create remote directory.
- Remove remote directory.
- Mode/attribute update of remote directory.
- FID on OST.
- MDT failover and recovery.
- Permanent MDT failure.
- Upgrade Lustre file system 2.1 to Lustre file system with DNE.
- Use cases.
- Implementation Milestones.

Remote Directories project was implemented with three milestones:

1. Demonstrate working DNE code. The sanity.sh and mdsrate-create tests will pass in a DNE environment. Suitable new regression tests for

the remote directory functionality will be added and passed, including functional Use Cases for upgrade and downgrade.

2. Demonstrate DNE recovery and failover. Suitable DNE-specific recovery and failover tests will be added and passed.
3. Performance and scaling testing will be run on available testing resources. The Lustre software Manual will be updated to include DNE Documentation.

The complete High Level Design is available at:

http://wiki.opensfs.org/images/e/ec/DNE_RemoteDirectories_HighLevelDesign.pdf

5. Summary of Implementation

Remote Directories is implemented in the following patches.

Commit	Patch Subject	Review
edafc19	LU-1445 ost: enable fid on OST support on OST	4791
a4a66a1	LU-1445 mdt: allow lightweight connection even if no OST	4920
ec654e2	LU-1445 fld: allow fld lookup during recovery	4919
9f82e92	LU-1445 osd: Add multiple sequence support for osd-zfs	4837
b36763d	LU-1445 osd: add fld look up in osd.	4330
57070c9	LU-1445 osp: rollover to the new seq synchronously	4790
6c4c51e	LU-1445 osp: Use FID to track precreate cache.	4789
6239198	LU-1445 lod: Add FLD lookup to LOD	5047
a6db240	LU-1445 fid: start ptlrpc service for OST FID	4328
8b5cfb9	LU-1445 fid: Add DATA fid type in fid_request.	4787
00e814f	LU-1445 ofd: Add fid support on OFD	4326
3269ac0	LU-1445 ofd: set index during server_data_init	4918
9b92218	LU-1445 fld: change md_site to seq_server_site	4805
c942006	LU-1445 ofd: remove ofd_seq_count and add ocd_seq	4325
33c7936	LU-1445 osd: replace OFD_GROUP0_LAST_OID with [seq, 0]	4324
7e5be18	LU-1445 fld: Checking lsr_flage after gotten from the cache.	3164
069deb0	LU-1187 osd: add remote entry insert for ZFS DNE.	4933
aa36847	LU-1187 out: add resend check for update.	4343
75ae281	LU-1187 mdt: Add MDS_INODELOCK_PERM lock for remote dir	4346
5fd303d	LU-1187 mdt: directory remote open fix	4934
11b08d4	LU-1187 mdd: a few missing stuff in MD stack for DNE.	4930
39a9eeb	LU-1187 mdt: add sanity check for rename and link	4348
4112a29	LU-1187 mdt: unlink remote directory	4339
2ad263c	LU-1187 utils: add lfs setdirstripe/getdirstripe	4341

d6ab8d4	LU-1187 osd: add agent/local inode for remote directory	4931
bc962bd	LU-1187 mdt: Only create remote dir on MDT0	4336
e39a2a9	LU-1187 lod: prepare default stripe EA for remote directory	4335
1012193	LU-1187 mdt: Add pre_cleanup phase in MDT stack cleanup	4929
a7241f1	LU-1187 mdt: add out handler for object update	4928
303ea89	LU-1187 osp: add osp_md_object for remote directory.	4927
85a5da0	LU-1187 lod: move seq client init from MDT to LOD	4337
d7cf1f8	LU-1187 lod: Add remote object for DNE	4924
82643b2	LU-1187 tests: chmod +x for dir_remote.sh	4997
aac646e	LU-1187 tests: add cleanup for sanity 24q.	4521
d7be37f	LU-1187 mdt: return lock to client for remote dir.	5026
b3a5326	LU-1187 lmv: allocate lmv tgts array by index	4936
74ec683	LU-1187 lod: Fix config log and setup process for DNE	4922
7ff7b6e	LU-1187 lod: reorganize lod_ost	4921
f31c60c	LU-1187 tests: Add mntfmt/mntcount/mdtcount to mdsrate	4614
96a0f6e	LU-1187 dne: add remote dir check in replay-vbr.	4321
7d1927e	LU-1187 lmv: Locate right MDT in lmv.	4356
3d0b6e4	LU-1187 mdt: enqueue rename lock locally for phase I.	4355
5e91e5b	LU-1187 lmv: remove obsolete lmv object.	5011
82bea2e	LU-1187 tests: Add parallel sanity tests to dne	4318
0a2a9b7	LU-1187 tests: Add dne specific tests to sanityN	4366
467521a	LU-1187 osd: allocate osd_compat_objid_seq dynamically	4323
50f2b9d	LU-1187 ofd: Allocate ofd group dynamically.	4322
5fd6406	LU-1187 test: add dne test to insanity.sh	4319
3a84f1d	LU-1187 tests: add create remote directory to racer	4365
f394dce	LU-1187 fld: fix fldb proc after moving range lookup to fld.	4350
13b269a	LU-1187 tests: Add test_mkdir in sanity for DNE.	4359
3ac7c32	LU-1187 tests: Add DNE test cases in sanity.	4358
9ed6c21	LU-1187 tests: sanity fixes for multiple MDT	4523
5a9f589	LU-1187 tests: Add DNE tests into recovery-small.	4360
71bdcf9	LU-1187 tests: Add DNE tests to conf sanity.	4367
8e2f5a3	LU-1187 tests: add DNE test cases in replay-single	4362
caf5bdf	LU-1187 tests: Add DNE tests cases in replay-dual	4361
edefa75	LU-1187 tests: define MGSDEV in right way.	4774
682f3e1	LU-1187 tests: Fixes in test-framework for DNE	4520

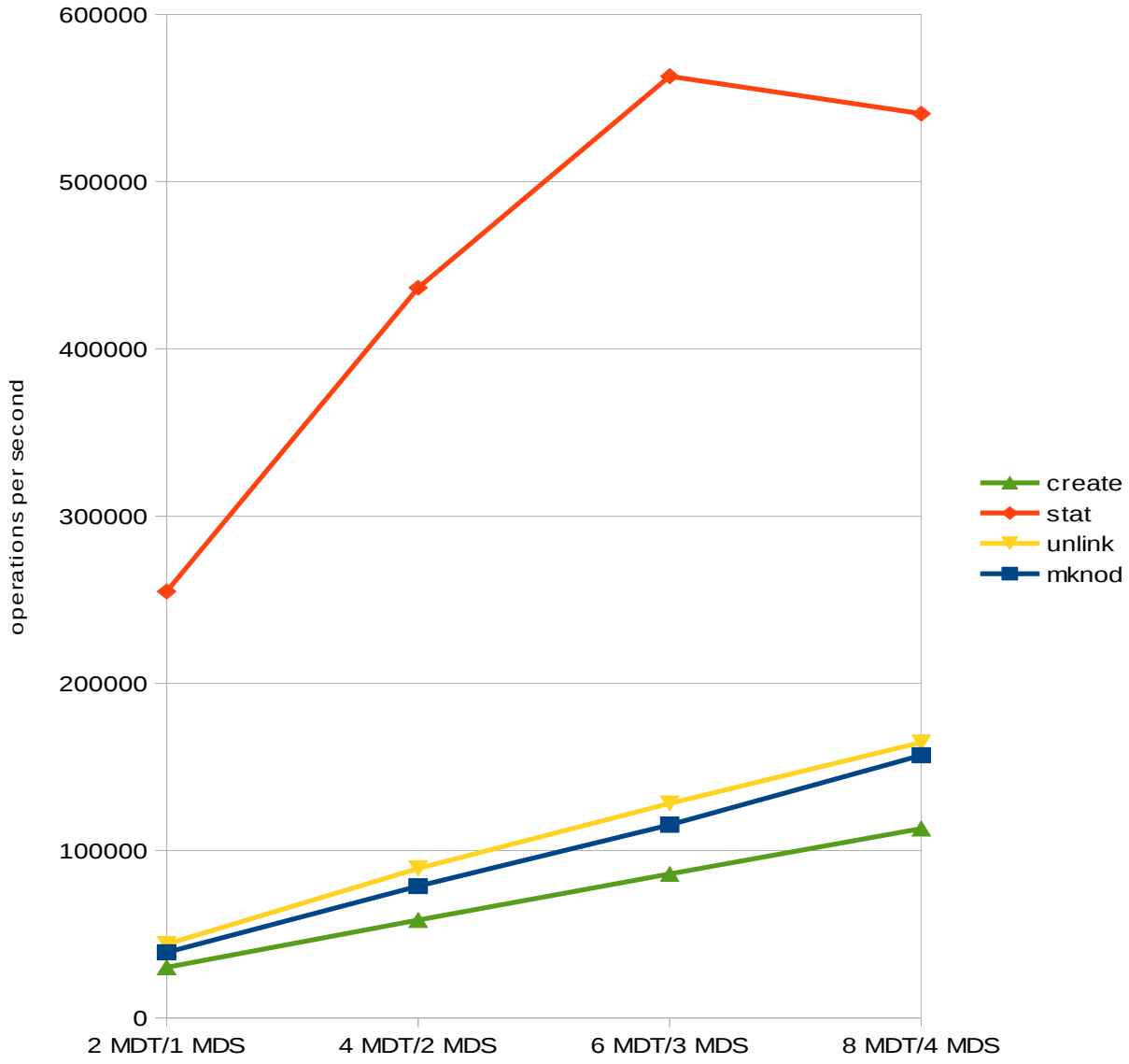
Implementation was completed as three Milestones. They are available at:

http://wiki.opensfs.org/images/8/84/DNE_RemoteDirectories_Implementation1.pdf
http://wiki.opensfs.org/images/f/f3/DNE_RemoteDirectories_Implementation2.pdf

http://wiki.opensfs.org/images/9/95/DNE_RemoteDirectories_Implementation3.pdf

6. Summary of Demonstration

Remote Directories successfully completed the Demonstration milestone on May 15th 2013. The important result showing approximately linear metadata performance scaling is:



In this test, each MDS is configured with two MDTs. The performance of create, unlink and mknod all show linear scaling with the addition of MDTs.

The stat performance increases with additional MDTs until six MDTs are present at which point it flattens out.

Additional demonstration highlights include:

1. The create performance measured with mds-survey of approximately 45K IOPS is close to the performance measured by mdsrate of approximately 55K IOPS. This result increases confidence in the value of mds-survey results, which can be run without the need for a large number of clients to generate testing load.
2. The absolute performance of a single metadata server is satisfactory.
3. Two MDTs attached to a single MDT performs measurably better than a single MDT attached to a MDS – excluding the case of a single MDS. This effect may be even more noticeable if a large number of disk operations are required (e.g. stat from disk).

The complete milestone report is available here:

http://wiki.opensfs.org/images/3/39/DNE_RemoteDirectories_DemonstrationMilestone.pdf

7. Delivery

A complete list of code reviews and landings is provided in Section 5, Summary of Implementation. The work, landing and designs are recorded on the tickets LU-1445 and LU-1187

https://jira.hpdd.intel.com/browse/LU-1445

https://jira.hpdd.intel.com/browse/LU-1187

1 Documentation

The Lustre* software manual update completed review at:

http://review.whamcloud.com/4773

The update includes the following topics:

- add an MDT.
- remove an MDT.
- upgrade to multiple MDT configurations.
- designing active-active MDS configurations.
- warns against having chained remote directories.

*Other names and brands may be claimed as the property of others.