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

Revision History

Date	Revision	Author
03/16/16	Original	R. Henwood

Executive Summary

This document finalizes the activities undertaken during the Distributed Namespace project, Sub Project 2.2: Striped Directories project within the OpenSFS Lustre[1] Development contract SFS-DEV-001 signed July 30th 2011.

Notable highlights of this project include:

- Demonstrated good scaling of metadata requests with the addition of MDTs.
- Successfully completed acceptance testing at scale on Hyperion.
- Delivered code and documentation into the 2.8 community release.

Statement of Work

Striped directories allow single directories to be distributed over multiple MDTs under administrative control to scale both the capacity and throughput of those directories. Distributed operations needed to operate on these directories are sequenced and synchronized as described above; however, file creation within such directories remains local to their directory stripe and, therefore, avoids synchronous I/O and executed with full performance. Distributed rename and hardlink operations will be supported so that they work as expected within a single striped directory.

Striped Directories remove the limit on the maximum number of entries in a single Lustre directory (currently 10M for ldiskfs) and increases both the maximum single client open files limit and the peak rate at which multiple threads in a single client can create files in a single directory.

Summary of Scope

In Scope

- Migration tool will move individual inodes from one MDT to another MDT.
- LFSCK OI Scrub, LinkEA, FID-in-Dirent agent and proxy parent interaction with DNE 1 verified.

1Other names and brands may be claimed as the property of others.

- Asynchronous updates between MDTs.
- commit-on-share
- striped directory
- man page and manual documentation.
- test plan and captured regression tests.
- code landed to Lustre master branch.

The complete scope statement was agreed on 2013-05-13 and is available at:

http://wiki.opensfs.org/images/3/3d/DNE_StripedDirectories_ScopeStatement.pdf

Summary of Solution Architecture

DNE Phase 1: Remote Directories made multiple metadata servers a reality on a Lustre* file system. This first phase included some limitations, namely:

1. The namespace can only be distributed to other MDTs by creating sub directory, i.e. a directory on a different MDT (aka remote directory). Typically an administrator may create remote directories for individual users, then those users will do operations in their directory serviced by a given MDT. A typical user will not benefit from simultaneously using multiple MDTs on their own jobs.

2. Except create/unlink remote directory, other cross-MDT operations return -EXDEV. In addition, cross-MDT operation are synchronous to simplify recovery.

3. All name entries in one directory can only exist on a single MDT. Single directory performance for operations like open/create files under one shared directory is the same as single MDT file system.

4. Moving a file onto a remote directory currently requires the file to be copied within the namespace. This is currently inefficient as redundant traffic between OSTs is generated by the copy operation.

DNE Phase 2: Striped Directories address these limitations and enables multiple MDTs on multiple MDS nodes serving a single directory.

The complete Solution Architecture was agreed on 2013-06-13 and is available at:

http://wiki.opensfs.org/images/0/07/DNE_StripedDirectories_SolutionArchitecture.pdf

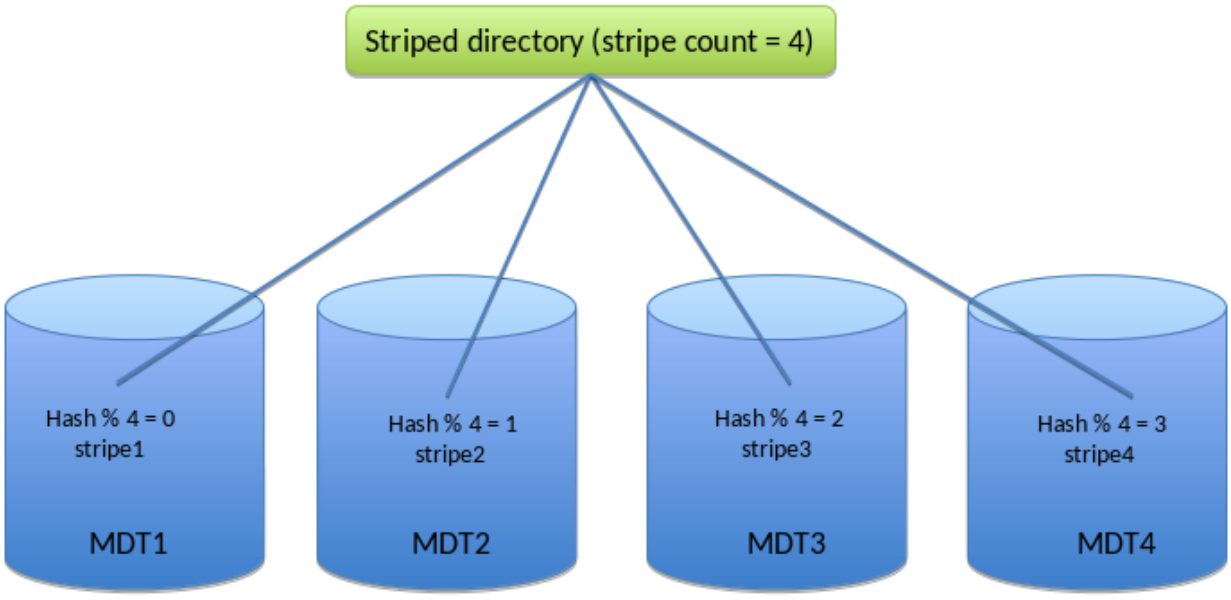
Summary of High Level Design

Introduction

In DNE Phase I all of name entries of one directory will be only in a single MDT. As a result, single directory performance is expected to be the same as single MDT file system. In DNE Phase II a striped directory will be introduced to improve the single directory performance. This document will discuss how striped directory will be implemented. It assumes the knowledge of DNE phase II async cross-MDT operation High Level Design and DNE phase I Remote Directory High Level Design.

Functional Statement

Similar to file striping, a striped directory will split the name entries across multiple MDTs. Each MDT keeps directory entries for certain range of hash space. For example, there are N MDTs and hash range is 0 to `MAX_HASH`, first MDT will keep records with hashes `[0, MAX_HASH/N - 1]`, second one with hashes `[MAX_HASH / N, 2 * MAX_HASH / N]` and so on. During file creation, LMV will calculate the hash value by the name, then create the file in the corresponding stripe on one MDT. It will also allow the user to choose different hash function to stripe the directory. The directory can only be striped during creation and can not be re-striped after creation in DNE phase II.



The complete High Level Design was agreed on 2013-07-24 and is available at:

http://wiki.opensfs.org/images/f/ff/DNE_StripedDirectories_HighLevelDesign.pdf

Summary of Implementation

Remote Directories code was developed and landed over three release cycles. The patches that implemented this feature include (but not limited to):

Commit	Ticket
e88992a	LU-2430 mdt: Add global rename lock.
0209add	LU-2430 mdd: add lfs mv to migrate inode.
370de92	LU-3531 mdt: delete striped directory<
3c216b9	LU-3531 llite: fix "lfs getdirstripe" to show stripe info
7117ff4	LU-3531 mdc: release dir page cache after accessing
4e0c8ae	LU-3531 llite: move dir cache to MDC layer
7f6f701	LU-5420 mgc: MGC should retry for invalid import
3e28034	LU-3536 lod: Separate thandle to different layers.
07c9244	LU-3534 osp: move RPC pack from declare to execution phase
67fe9ef	LU-3534 lod: record update for cross-MDT operation
31bb2c2	LU-3564 mdt: move last_rcvd obj update to LOD
9ec47fe	LU-3564 LOD: add distribution id to identify updates
a603212	LU-3534 lod: write updates to update log
548a70e	LU-3546 lod: cancel update log after all committed
9f71978	LU-3564 lod: update recovery thread
2e6dbe1	LU-3537 mdt: allow cross-MDT rename and link
59aa0b7	LU-3536 osp: send updates by separate thread
0fe99fb	LU-3536 update: change sync updates to async update

The complete record of Implementation was agreed on 2014-12-23 and is available at:

http://wiki.opensfs.org/images/e/e2/DNE_StripedDirectories_Implementation.pdf

Summary of Demonstration

Remote Directories successfully completed the Demonstration milestone on August 21st 2015. The purpose of this milestone was to show the code performs acceptably in a production-like environment. To achieve this, tests were executed on a variety of hardware platforms including the OpenSFS Functional Test Cluster and the public cloud.

Scaling tests showed that striping a directory over multiple MDS improves file creation rate approximately linearly with each new MDS. Tests with multiple MDTs per MDS were completed. These showed that apparently optimal performance scaling with DNE striped directories was with two MDTs per MDS.

The complete Demonstration milestone was agreed on 2015-08-21 and is available at:

http://wiki.opensfs.org/images/6/6c/DNE_StripedDirectories_Demonstration.pdf

Delivery

A complete list of code reviews and landings includes (but is not limited to) the patches listed in the implementation summary:

http://wiki.opensfs.org/images/e/e2/DNE_StripedDirectories_Implementation.pdf

DNE2 feature became available with the release of Lustre software version 2.8. Details of the 2.8 release can be found here:

http://wiki.lustre.org/Release_2.8.0

Documentation

The Lustre manual update completed review at LUDOC-306 dne2: update with new command for DNE2:

http://review.whamcloud.com/4773

The update includes the following topics:

- Create a striped directory.
- Migrate a striped directory.
- Administrate a striped directory.