Remote Directories Scope Statement

This document has been submitted for review on 2011-10-05. The document was accepted on: 2011-10-19.

Revisions since 2011-10-19

Date	Revision	Author
12/21/2011	Wording for non-MDT0 parent directory and ldiskfs clarification.	R Henwood

Introduction

The following scope statement applies to the Distributed Namespace project within the SFS-DEV-001 contract/SOW dates 08/01/2011.

Problem Statement

Today, Lustre filesystem routinely have thousands to tens of thousands of clients. As client numbers continue to increase, a single Metadata Server (MDS) for a single Lustre filesystem becomes a performance and scalability constraint. Distributed NamespacE (DNE) is a project to spread the Lustre filesystem metadata namespace horizontally over multiple MDS nodes and Metadata Targets (MDTs), in a similar manner that Lustre spreads filesystem data over multiple Object Storage Server (OSS) nodes and Object Storage Targets (OSTs). This enables namespace size and metadata throughput to be increased by the addition of MDTs. In addition, DNE enables an administrator to allocate metadata resources to specific directories within the filesystem.

Project Goals

DNE Phase 1 enables deployment of multiple MDTs on multiple MDS nodes. Functionality is restricted in this phase by specially-created directories to remote MDTs.

1. Modified Lustre protocol to access OST objects using Lustre File IDentifiers (FIDs) so that multiple MDTs can concurrently create objects on the same OST.
2. User tool to create directories on a specific remote MDT.
3. User tool to display on which MDT a directory or file is located.
4. Benchmark metadata throughput with one, two, and four MDTs (WC-Lustre including DNE code).
5. Show a 1.5x increase in metadata throughput with doubling of MDT count when serving multiple clients from separate directories.
6. Code landed on WC-Lustre master branch.
7. Demonstrate that DNE reacts well and remains manageable in various error conditions, including failover of any MDT, recovery
   from complete server cluster failure and damage limitation after permanent failure of any MDT, including the root MDT.
8. Demonstrate metadata performance isolation between subtrees located on separate MDSs.

In-Scope

- DNE code development will take place against WC-Lustre 2.x Sequoia development branch.
- Interoperability of 1.8/2.1 clients with with WC-Lustre 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 WC-Lustre 2.x manual.
- Upgrade of existing 2.1 single-MDT filesystem to DNE-capable Lustre filesystem, with shutdown/restart.
- Addition of MDTs to existing DNE-capable Lustre filesystem, with shutdown/restart (at least initially).
- Accessing a DNE filesystem 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.

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.

Project Constraints

- Wang Di is the only engineer with the correct expertise available for this work.

Project Assumptions

- DNE Phase 1 code is contingent upon restructured MDS and OSS code being developed in the Sequoia project.
- Landing of DNE Phase 1 code is contingent on landing of required Sequoia functionality to Master WC-Lustre 2.x.
- Current schedule assumes hardware will be available from completion of the design document.
- Test hardware (128 clients, 8 servers) configuration at client site will be available for access by any Whamcloud engineer regardless of national citizenship at least 20 hours/week during the assessment and implementation phases of the project.

Key Deliverables

- Signed Milestone documents for project phases:
    - Solution Architecture.
    - High-Level Design.
    - Implementation & Test.
    - Acceptance Testing (OpenSFS executed).
- Test Plan.
- Source code that meets feature requirements and runs with WC-Lustre 2.x on customer’s site.
- Source code for new test cases.
- DNE Phase 1 code landed in the Master WC-Lustre 2.x.

Key Milestones

Scope Statement delivery 2011-10-05
Solution Architecture delivery 2011-11-02

High-Level Design delivery 2011-11-23
Implementation delivery 2012-07-04
Acceptance Testing delivery 2012-09-26
Demonstration and close 2012-10-10

Glossary:

directory – a filesystem sub-tree.
metadata target (MDT) – a filesystem component that services and persists metadata operations.
namespace – the abstract tree structure of directories and files on the filesystem.
Object Storage Target (OST) – a filesystem component that services storage requests from clients.
File IDentifier (FID) – a unique 128bit identifier for objects across all OSTs and MDTs.
WC-Lustre – Whamcloud community release of Lustre.