Lustre Metadata Performance and Solutions

February, 2015

Bill Loewe

Agenda

File System Metadata, a growing issue

Parallel File System - Lustre Overview

Metadata and Distributed Namespace

Test setup and implementation for metadata testing

- Scaling Metadata Servers
- High Availability

Metadata Performance

File System Performance typically viewed in Bandwidth

Bandwidth problem largely addressed, but metadata is a growing issue.

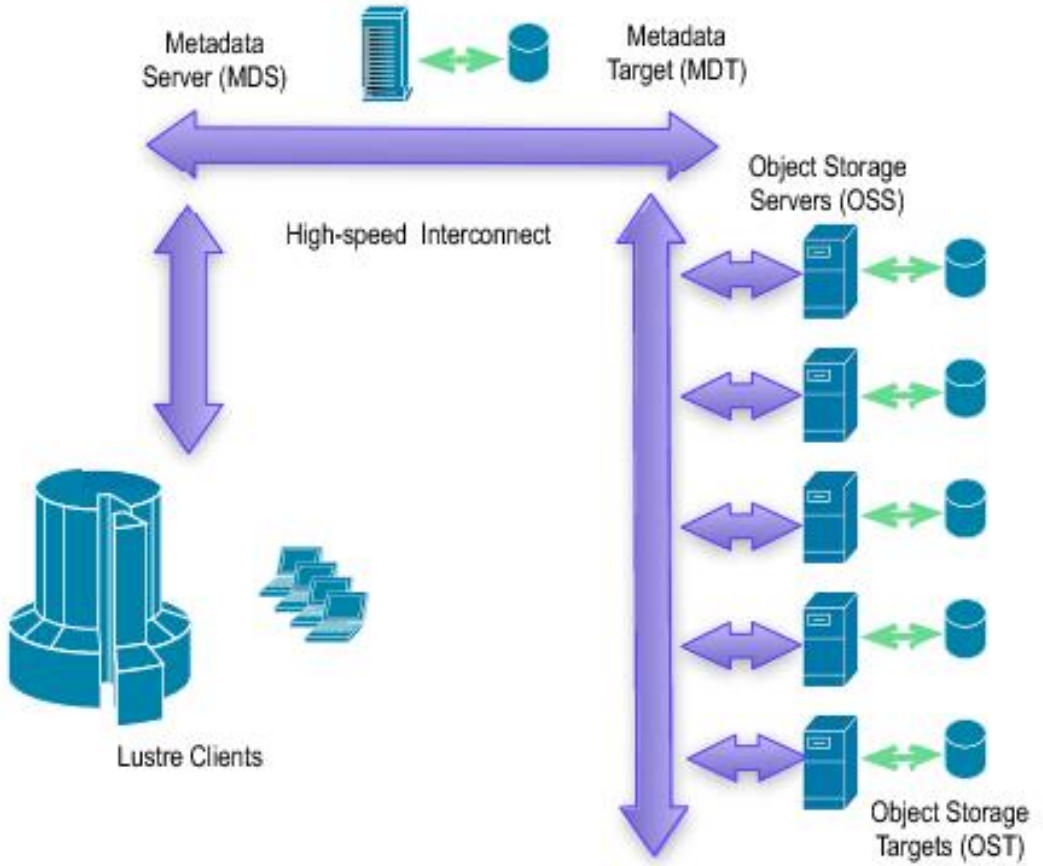
We see this in workloads with high numbers of files to access and process.

- **Genome processing**
- **CPU Chip manufacturing**
- **Video compositing/rendering**

Lustre Parallel File System

Lustre is an open source, distributed parallel file system

- Object-based design provides extreme scalability
- Compute clients interact directly with storage servers
- Comprised of:
  - Clients
  - Metadata Servers and Targets
  - Storage Servers and Targets



Lustre Distributed NamespacE (DNE)

Distributed NamespacE (DNE) is a new feature available in Lustre 2.5 that allows multiple MDS / MDT components to participate in a single file system.

DNE allows the namespace to be divided across multiple metadata servers.

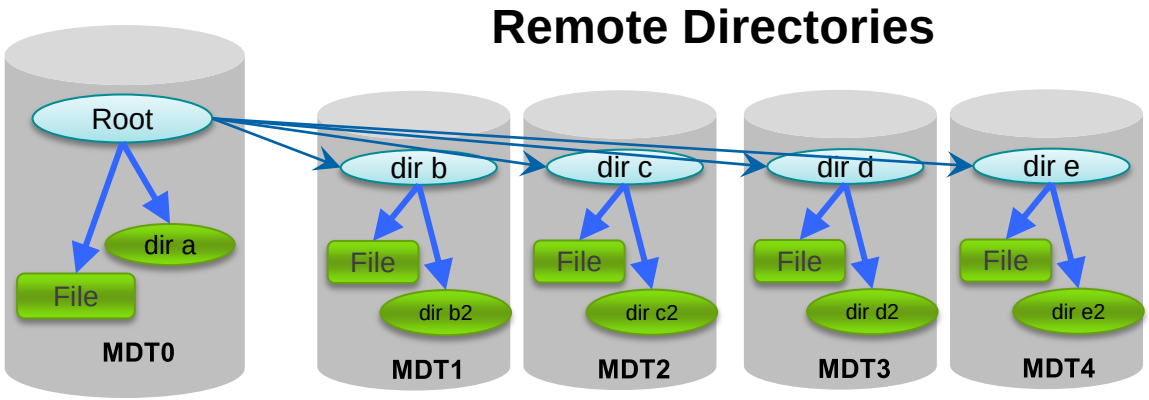
Enables the size of the namespace and metadata throughput to be scaled with the number of servers.

The Lustre DNE project is comprised of 2 phases.

Phase 1, Lustre 2.5 Release

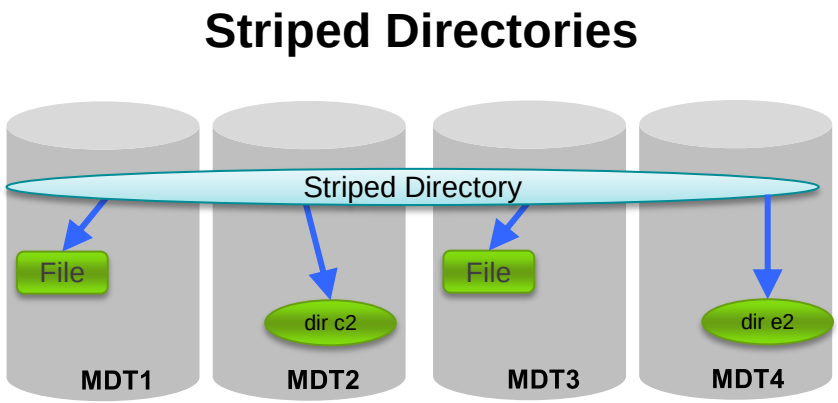
Remote Directories -- Lustre sub-directories are distributed over multiple metadata targets (MDTs).

Sub-directory distribution is defined by an administrator.



Phase 2, Lustre 2.7

Striped Directories -- The contents of a given directory are distributed over multiple MDTs.

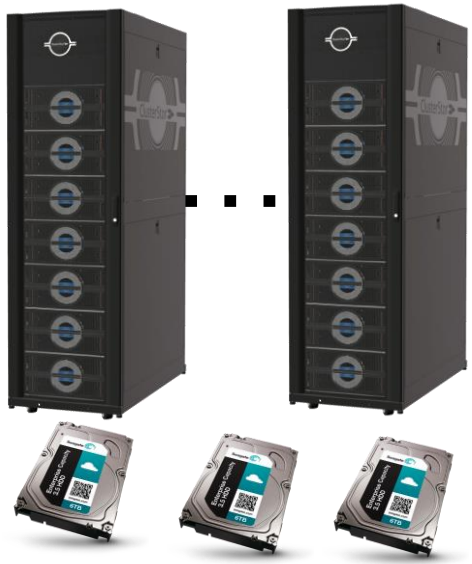


Engineered Storage Solutions for HPC, Big Data & Cloud

ClusterStor

- High speed networking (IB/40GB/e)
- Parallel file system/Object
- Data protection
- High availability
- Flash optimization
- File system (Ext4)
- Linux OS
- BIOS/IPMI
- GEM diagnostics
- Custom X86 embedded server
- Seagate storage platforms
- Seagate Storage Devices

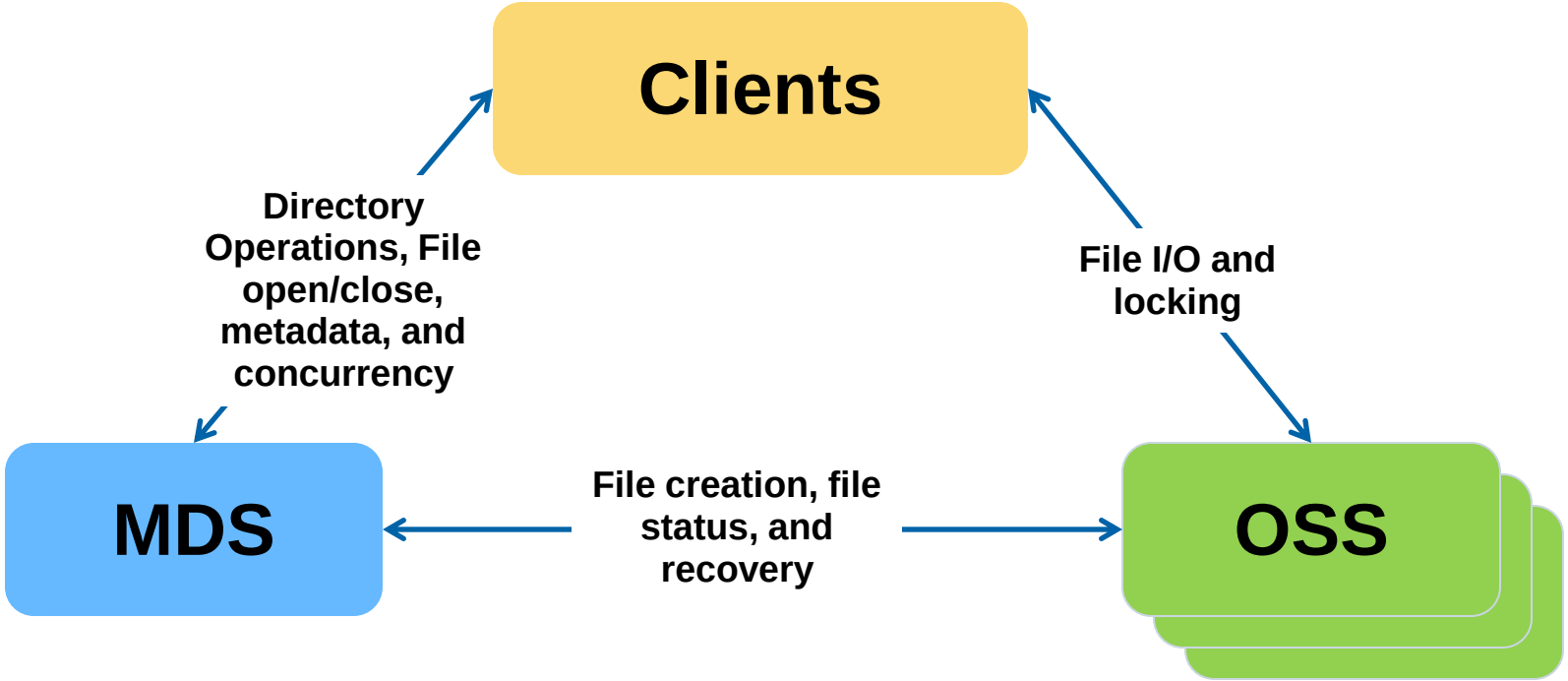
lustre®



ClusterStor™ M·A·N·A·G·E·R

- ✓ ***Architected***
- ✓ ***Integrated***
- ✓ ***Optimized***
- ✓ ***Qualified***
- ✓ ***Supported***

Lustre Components



ClusterStor Management Unit (CMU): Management and Metadata (MDS/MDT)

CSM Manager and MDS/MGS Nodes

- 2RU 4-node Sandy Bridge Servers
  - Server 1: CSM Mgmt
  - Server 2: Boot
  - Server 3: MGS
  - Server 4: MDS

Fault Tolerance (active/passive)

Serviceability

2U24 JBOD – MDT

- SAS JBOD for MDS/MGS/Management
- Disk Configuration
  - Qty 4 Lustre Management (MGS)
  - Qty 4 ClusterStor Management and NFS
  - Qty 2 Global Hot spares
  - Qty 14 Drives for MDT

Scalable Storage Unit (SSU)

SSU

- 5U84 Enclosure
- 2 Object Storage Servers's per SSU
- Two (2) trays of 42 HDD's each for Object Storage Targets
- H/A on each SSU
- Infiniband QDR/FDR and 40Gb Ethernet data network connectivity

ClusterStor & Lustre 2.5 DNE Hardware

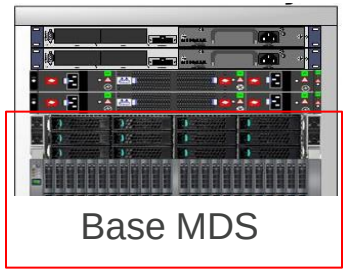
DNE is available in ClusterStor v2.0

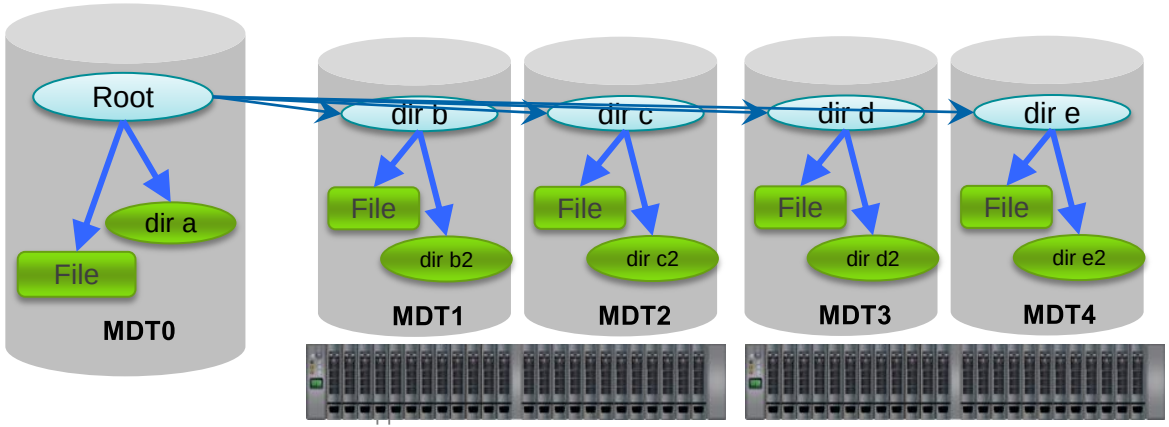
- **MDT0 is master and default in DNE environment**

DNE Servers are configured in active / active pairs

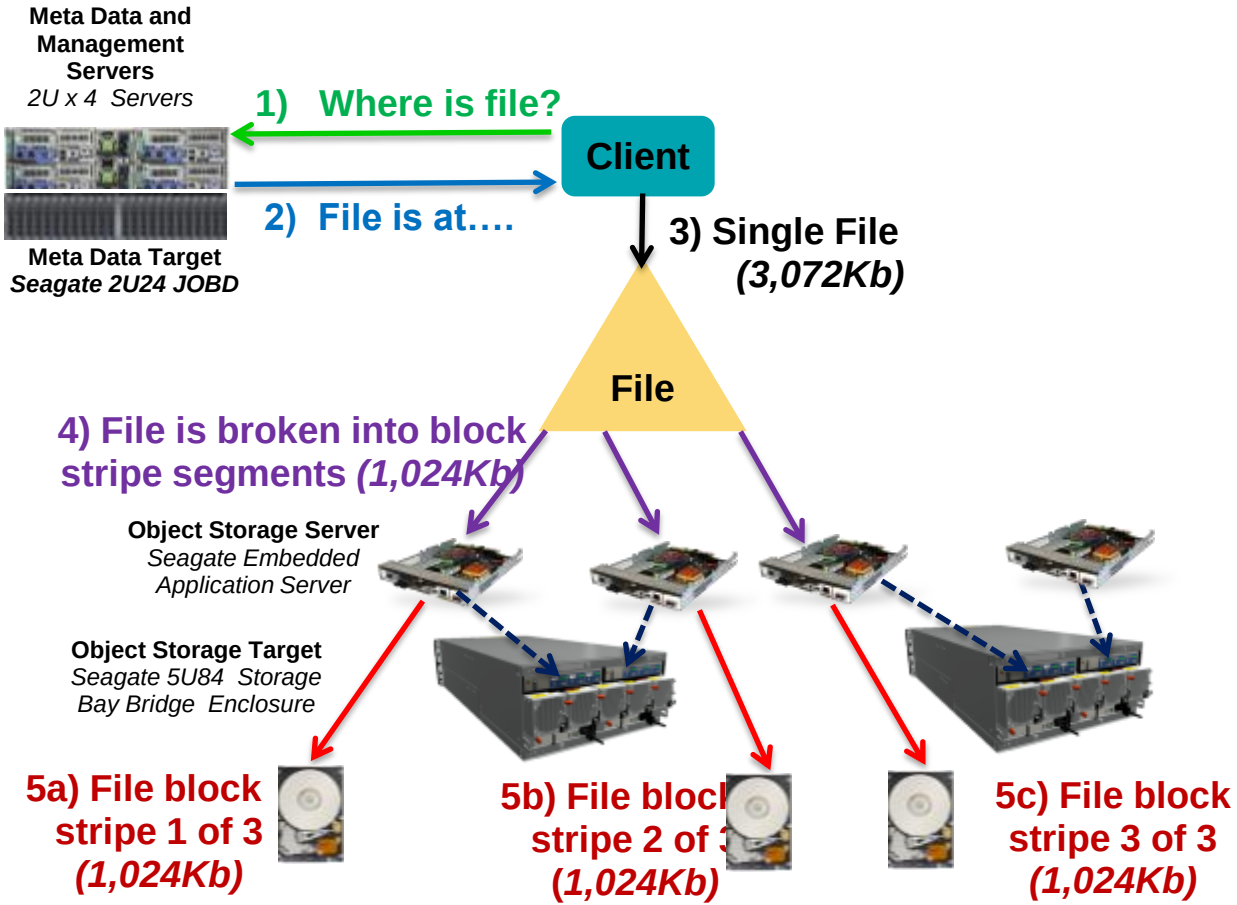
- **Seagate 2U24 with 2 MDS embedded server modules**

Scale Metadata Capacity / Performance with DNE Server pairs



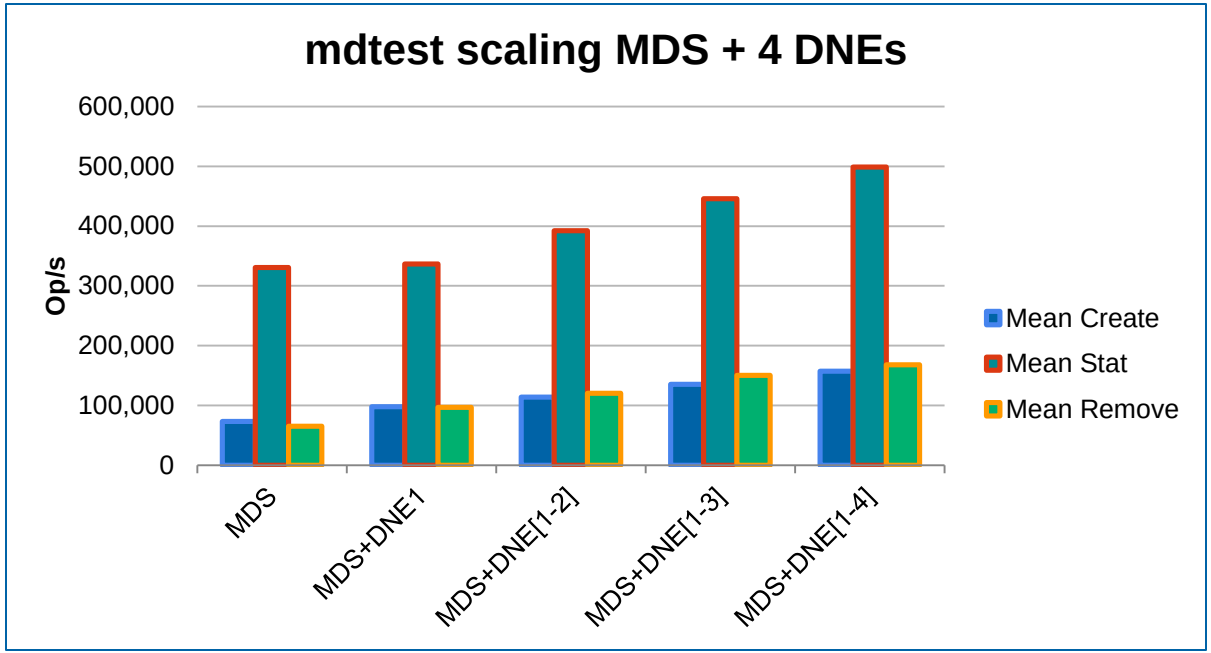


ClusterStor Hardware and the Lustre File System



Scaling MDS and DNEs

- **MDS + 4 DNE Servers (2 ADUs)**
- **mdtest create/stat/del**
- **Mean of 5 iterations**



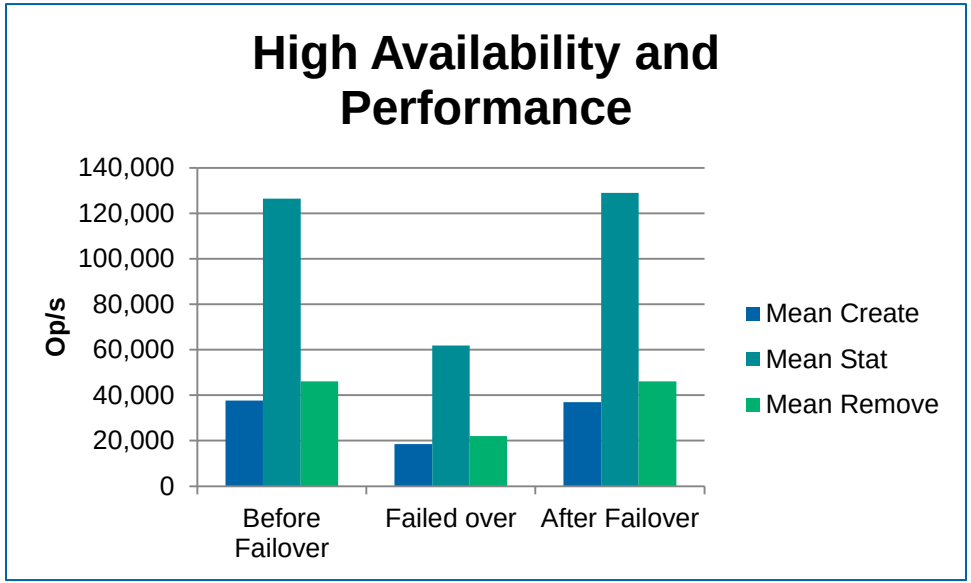
Metadata High Availability

MDT failover will ensure that the Lustre filesystem remains available in the face of MDS node failure

Based on existing OSS pair failover model

Failover is graceful, quick, and non-disruptive

Failback is automatic and non-disruptive



Green Machine: Environmentally-Aware Cold Storage Solution

Space

Light weight
Small foot print
Cold storage optimized design

Cooling

Zero heat emission
Ambient cooling/No fans
High operating temp. tolerant HDDs

Power

Dynamic power management
Low power servers
Aggressive TCO goals

Green

Recyclable chassis
Reduced metal
Responsible disposal of old chassis

Lowest Operating Cost
Reduced Carbon footprint

"Best for the Planet"

Seagate

Typical Use cases

- Retrieve content, photographs etc. from deep archive while maintaining consistent user experience
  - Online pictures/Social media store use cases
    - Pictures >45 days in cold storage
  - Retrieve MRIs/X-rays of a patient
  - Use cases leveraging Tape-based solutions




Seagate

Seagate

Thank you !