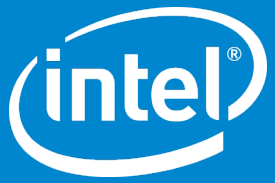
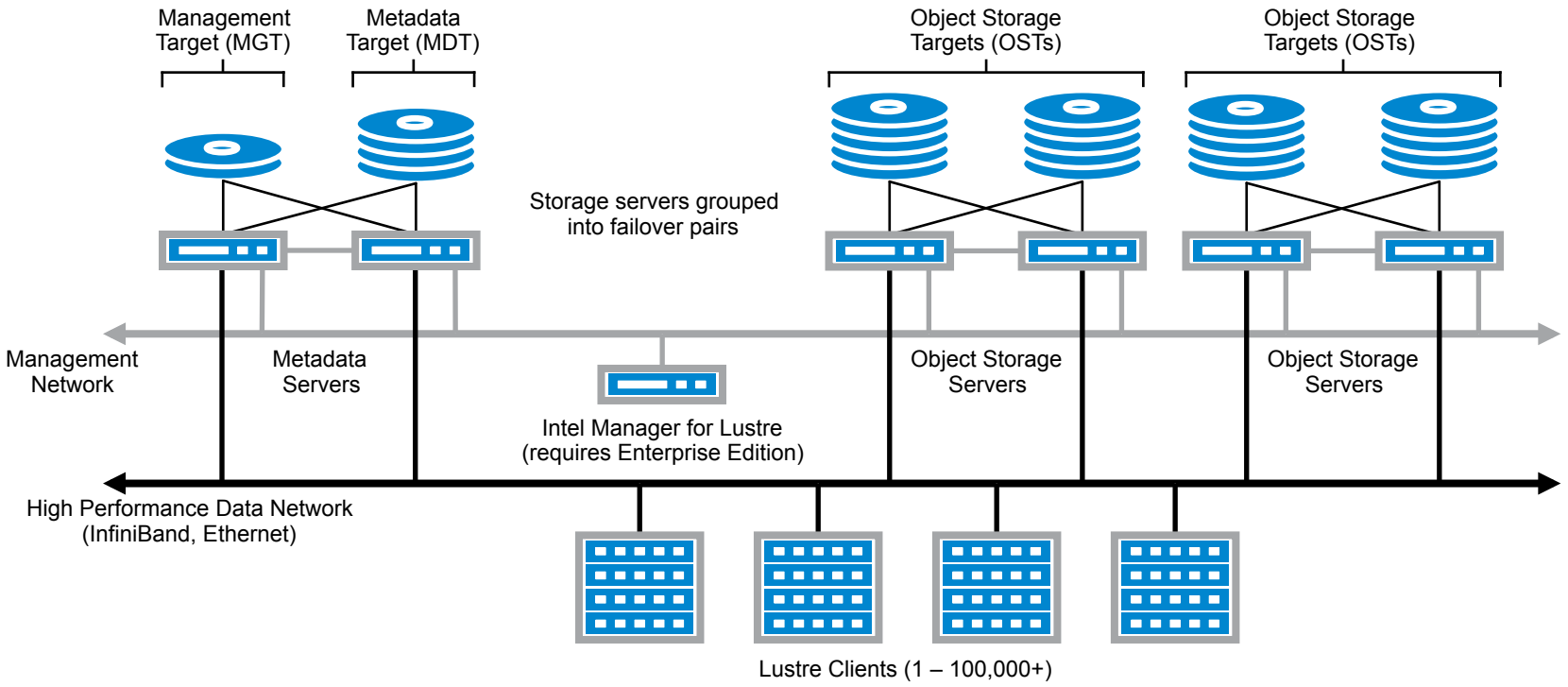


Disruptive Storage Workshop

Lustre* ZFS and HSM Overview

2015

Lustre* HW Overview



Storage Servers and Storage Targets

Servers

- Management Server (MGS)
- Metadata Server (MDS)
- Object Storage Server (OSS)

Targets

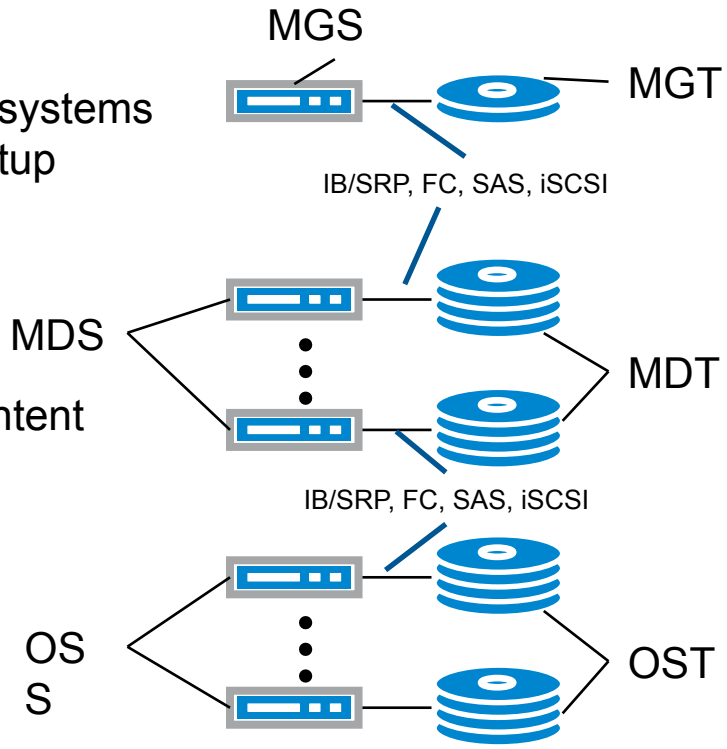
- Management Target (MGT)
- Metadata Target (MDT)
- Object Storage Target (OST)

Storage Servers and Storage Targets

Management Service and Target
- Provides configuration information for Lustre file systems
- All Lustre components register with MGS on startup
- Clients retrieve information on mount

Metadata Service and Targets
- File system name space
- File Layout (location of data objects), no data content
- Scalable (DNE)

Object Storage Service and Targets
- File content, stored as objects
- Files may be striped across multiple targets
- Massively Scalable



Intel Lustre* storage servers

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

Management Server (MGS)

One per site / per file system

- Each Lustre file system needs an MGS
- A MGS can serve one or more Lustre file systems
- Manages filesystem configuration/tunable changes

Minimal storage requirements

- Small (~100 MB) data footprint
- Recommend running on a separate node (e.g. backup MDS)
- Historically, has been combined with a Metadata Server

Registration point

- New components ***register*** with MGS during configuration
- Clients obtain Lustre configuration from MGS during mount
- Clients obtain configuration updates from MGS after mount

Metadata Server (MDS)

One or more per file system

Significant multi-threaded CPU use

- Simultaneous access from many clients

Maintains the metadata, which includes

- Information seen via stat()
  - Owner, group, filename, links, ctime, mtime, etc
- Extended attributes
  - File Identifier (FID)
  - Mapping of FID, OSTs and Object ID's
  - Pool Membership, ACLs, etc.

Object Storage Server (OSS)

Bulk data movers

- Use large network RPCs for high bandwidth

Moves data between block storage and network

Provide clients access to OSTs

- Clients do not communicate with MDS for read/write requests

Typically there are many of these

- Clients communicate directly with OSSes in parallel

Intel Lustre* storage targets

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

Backend file system supported for Lustre

- Lustre* targets run on a local file system on Lustre* servers. Object Storage Device (OSD) layer supported are:
  - ldiskfs (modified version of EXT4) is the commonly used driver
  - ZFS is the 2nd use of the OSD layer based on OpenZFS implementation
- Lustre* targets can be different types (hybrid ldiskfs/ZFS is possible)
- Lustre* Clients are unaffected by the choice of OSD file system
- Lustre* ZFS is functional since Lustre* 2.4

Storage targets – MGT / MDT / OST

A storage volume formatted as type ldiskfs or ZFS

- Type ldiskfs is a modified version of ext4 (has additional code)

Mostly, it's just a regular file system:

- Contains few special directories for config files, etc.

Can be almost ***any*** Linux block device

- Whole LUN / block device preferred, but can run on a partition
- Will run on LVM
  - LVM useful for taking snapshots / making backups
- ZFS has included RAID protection, volume management capabilities, snapshots, etc

MGT – Management Target

A storage volume formatted as type ldiskfs or ZFS

Can support one (1) or more Lustre file systems

Think of it as a Registration Point

- Knows about clients, servers and targets
- Provides information upon request

Should be not co-located with the metadata target (MDT).
Some critical recovery capability of Lustre are disabled if MDT and MGT are co-located.

MDT – Metadata Target

A storage volume formatted as type ldiskfs or ZFS

- Could run on LVM – useful for snapshot backups

File system is unavailable if MDT is unavailable

Holds all the metadata for one (1) Lustre file system

Size is based on the amount of files in file system expected

Contains no Lustre block data – only “Lustre inodes”

Lustre Inodes

Lustre inodes are MDT inodes

MDT is formatted with many large inodes

- Default inode size is 2K
- The maximum number of inodes in LDISKFS is 4 billions no practical limits for ZFS
- Maximum of 4096 MDTs with DNE

Lustre inodes hold all the metadata for Lustre files

Lustre inodes contain:

- Typical metadata from stat() (UID, GID, permissions, etc.)
- Extended Attributes (EA)

Extended Attributes contain:

- References to Lustre files (OSTs, Object ID, etc.)
- OST Pool membership, POSIX ACLs, etc.

OST – Object Storage Target

A storage volume formatted as type ldiskfs or ZFS

- Whole LUN or disk device preferred

Contains:

- A few special directories for config files, etc.
- File system data accessed via object IDs

Each object is either: (based on stripe_count)

- A complete file (stripe_count == 1)
- Part of a file (stripe_count > 1)

Limits each OST:

- 128TB with LDISKFS
- 256TB with ZFS

Intel Lustre* clients

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

Other types of Lustre nodes

Client compute nodes

- Despite what people think, these are the brains of the operation
- Present the distributed Lustre file system as a single namespace
- Typically compute, visualization, or large SMP nodes

Lustre software clients *(details on next slide)*

- Software clients that communicate with storage services
- Client/Server design, so each service has a client counterpart
- Each Lustre node runs multiple software clients

Lustre router nodes bridge network types

- Efficiently connect different network types
- Only run limited Lustre Networking software stack
- Use RDMA network transfers for efficiency

NFS / SMB server clients

- Lustre clients that re-export Lustre file system to non-Linux clients

Lustre Software Clients

Management Client (MGC)

- MGC handles RPCs with the MGS
- All servers (even the MGS) run one MGC
- Every Lustre client runs one MGC per MGS
- Basically, "everybody" runs one

Metadata Client (MDC)

- MDC handles RPCs with the MDS
- Only Lustre clients do RPCs with the MDS
- So, all Lustre Clients run one MDC per MDT

Object Storage Client (OSC)

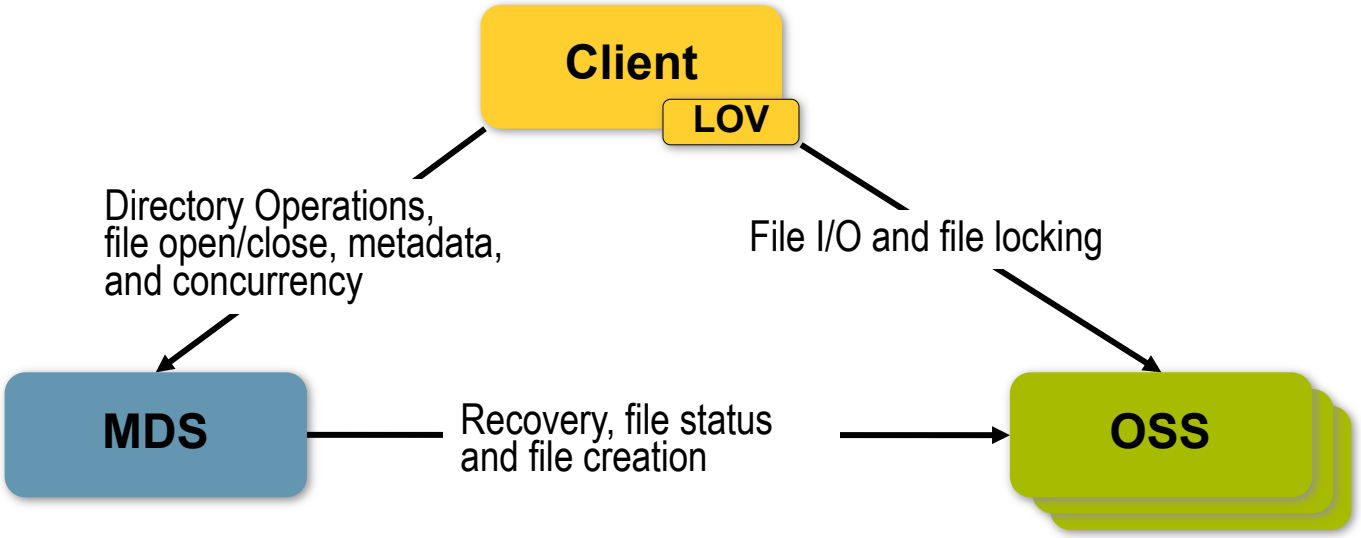
- OSC handles RPCs with a single OST
- Both the MDS and Lustre clients initiate RPCs to OSTs
- So, all MDS's and Lustre Clients each run one per OST on the OSS

LOV – Logical Object Volume

A software layer in the client stack

Aggregates multiple OSC together

Presents a single logical volume to the client

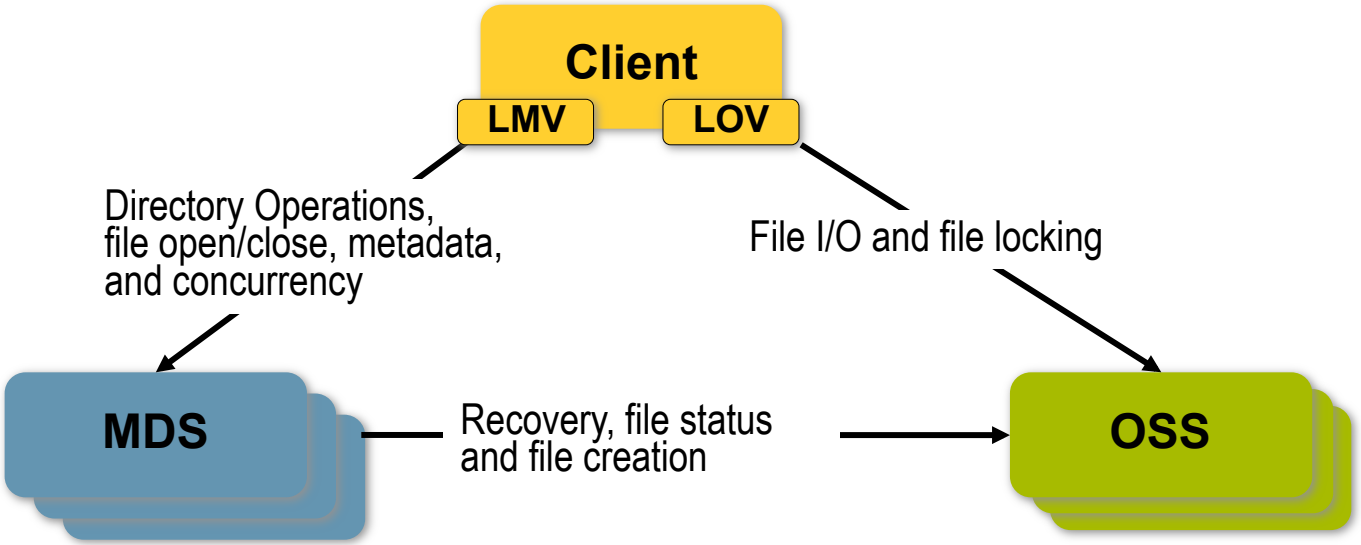


LMV – Logical Metadata Volume

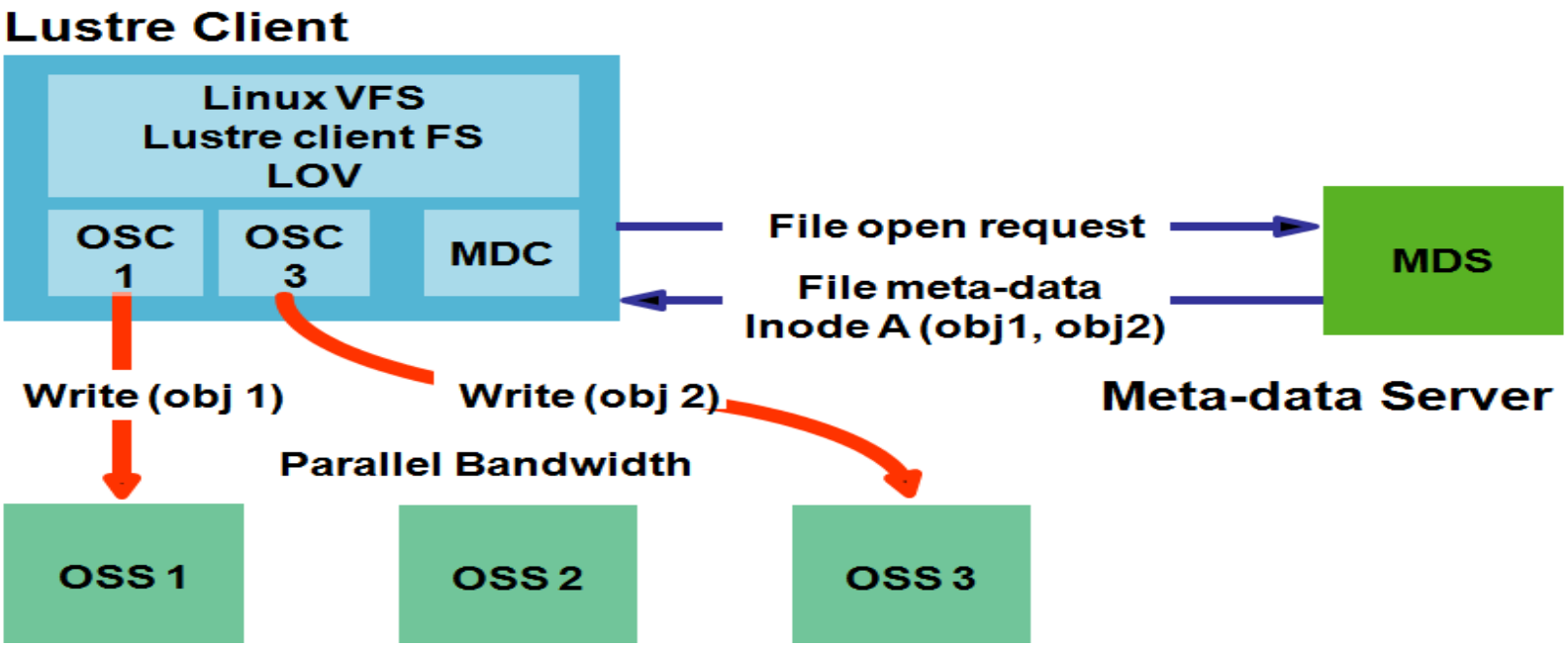
A new software layer in the client stack since lustre 2.4

Aggregates multiple MDCs together

Presents a single logical metadata space to client



Basic Lustre I/O Operation



Just so there's no confusion… ☺

MGS – Management Server
MGT – Management Target
MGC – Management Client

MDS – Metadata Server
MDT – Metadata Target
MDC – Metadata Client

OSS - Object Storage Server
OST - Object Storage Target
OSC - Object Storage Client
OSD – Object Storage Device

LOV - Logical Object Volume
LMV - Logical Metadata Volume
LWP - Light Weight Proxy