DE LA RECHERCHE À L'INDUSTRIE

cea

www.cea.fr

Lecture on Parallel Filesystems

Lustre (1/4)

Jacques-Charles Lafoucriere

ENSIIE| 2018

Lustre Introduction

What is Lustre

- A storage architecture for clusters
  - Power the largest HPC systems in the world
  - Can manage multiple File System in a single Storage Cluster
- A POSIX standard compliant UNIX file system interface
- A kernel Open Source File System
- Can achieve performance of TB/s and manage Petabytes of storage
- Designed to integrate multi-cluster data centers
  - LNet Routers
- A scalable design
  - Client #
  - Server #
  - Storage system #

What is not Lustre

- A peer-to-peer cluster
  - Optimized for client attached storage
  - Designed to support multiple copies
- A non Linux FS
- A user space file-system
- A work-station file system for Campus Area Network
- A distributed file-system for Wide Area Network

Lustre components

- MGS: Management Server
  - Only one in a Lustre Cluster
- MDS: MetaData Server
  - A machine
  - Host MetaData Targets (MDT)
- OSS: Object Storage Server
  - A machine
  - Host Object Storage Targets (OST)
- Client
  - A machine
  - Run POSIX compliant interface in Linux Kernel

Lustre Scalability and Performances

Feature	
Client Scalability	100-50 000-131 072
Client Performance	Single client = 90% Network BW, 1 000 MD op/s
OSS Scalability	32 OST/OSS, 1 000 OSS (with up to 4 000 OSTs) Single OST = 300 M obj, 128 TB (ldiskfs), 500 M obj, 256 TB (ZFS)
OSS Performance	Single OSS = 15 GB/s
MDS Scalability	4 MDT/MDS, 256 MDS (with up to 256 MDTs) Single MDT = 4 G files (ldiskfs), 64 G file (ZFS)
MDS Performance	50 000 create/s, 200 000 stat/s
FS Scalability	max file size: 32 PB (ldiskfs), 2^63 (ZFS) Max FS size: 512 PB, 1 trillion file

Lustre Features

- ldiskfs backend
  - An enhanced version of ext4
- ZFS backend
  - Provide ZFS data integrity
- POSIX standard compliance
  - All POSIX (even mmap())
- High performance heterogeneous networking
  - RDMA based and tcp/ip
  - LNet Routing support
- High-availability
  - Support many HA managers
  - Add Multi Mount Protection (MMP) to avoid data corruption
  - Active/active on MDT
- Small file optimizations

Lustre Features (cont.)

- Security
  - TCP privileged port
  - Kerberos support
  - Network data integrity with checksum
- Access Control List, Extended attributes
- Interoperability: mixed endian clusters
- Byte granular file and fine grained MD locking
  - Many clients car modify/read same file/directory concurrently
- Quotas
- Capacity growth/reduction
- Controlled file layout: per file, per directory, per FS
  - Progressive File Layout
- MPI-I/O optimized (ADIO driver)
- NFS/CIFS exports
- Disaster recovery tool: online LFSCK

Lustre Server Components

MGS

- Store configuration information for all the Lustre file systems
  - One per cluster / per file system
- Provide configuration to all other components
- Each target register to MGS
- Each client contact MGS to retrieve information

MDS

- Maintains MD which includes information seen via stat()
- Make FS MD stored in MDT's available to clients
- MDT host the name space and file MD in a Linux local FS (ldiskfs or ZFS)

OSS

- Make FS data stored in OST's available to clients
- OST hosts data file in a Linux local FS (ldiskfs or ZFS)

Lustre Server Sizing

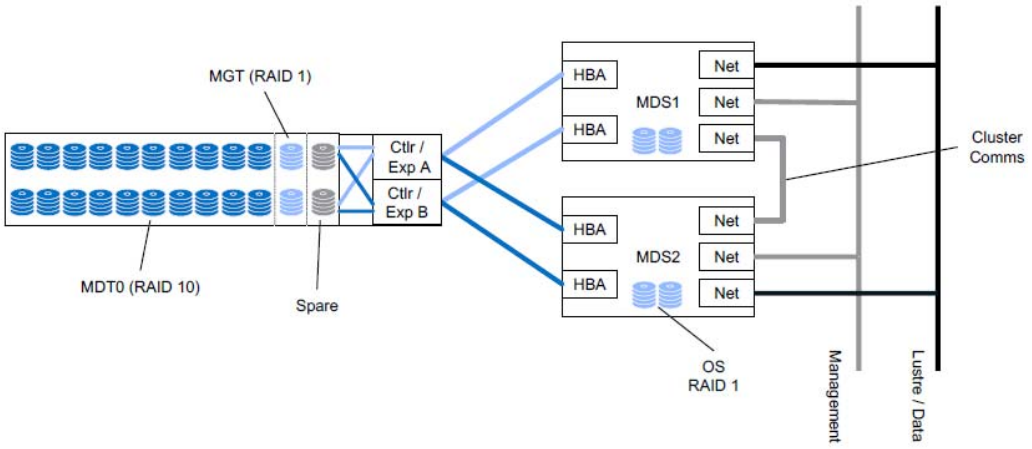
MDS

- 1-2% of FS capacity
- A lot of memory for caching
  - Formula available for sizing

OSS

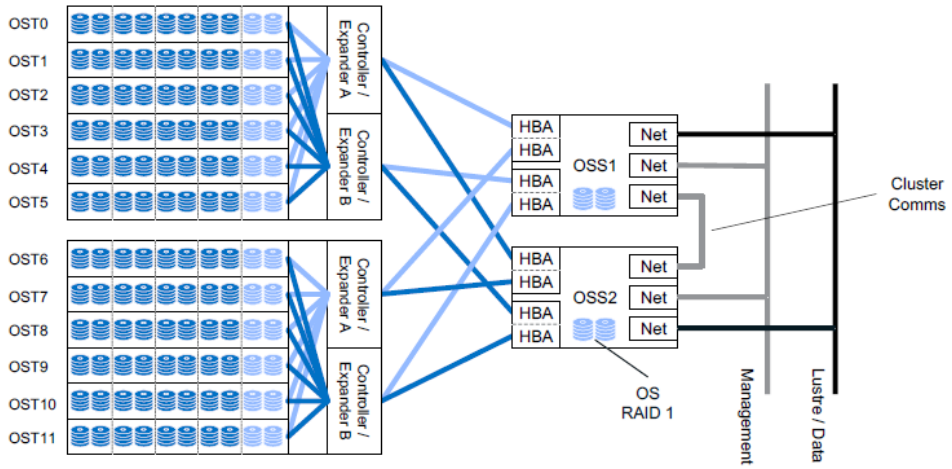
- 1-128 TB per OST, 1-8 OSTs per OSS
- Some memory for caching
  - Formula available for sizing

MGS/MDS Reference Design



SFP 2018 Parallel Filesystems / Lustre | PAGE 11

OSS Reference Design



SFP 2018 Parallel Filesystems / Lustre | PAGE 12

Lustre Client Components

Management Client (MGC)

- Process handles RPCs with the MGS
- All servers (even the MGS) run one MGC
- Every Lustre client runs one MGC for every MGS

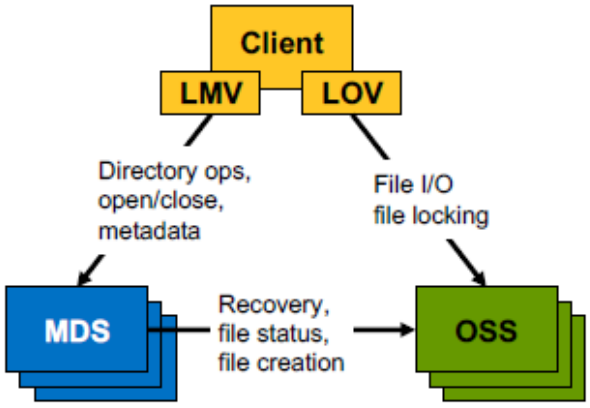
Metadata Client (MDC)

- Handles RPCs with the MDS
- Only Lustre clients initiate RPCs with the MDS
- Each client runs an MDC process for each MDT

Object Storage Client (OSC)

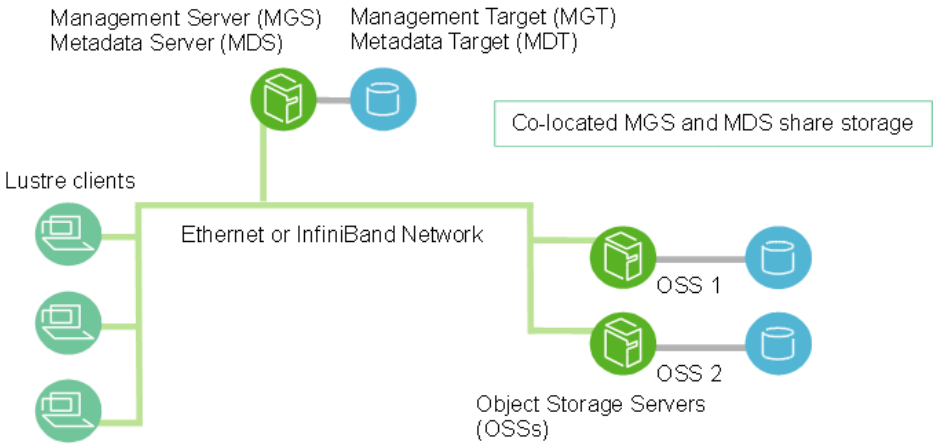
- Manages RPCs with a single OST
- Both MDS and Lustre clients initiate RPCs to OSTs
- Each of these machines runs one OSC per OST

Lustre I/O Operations

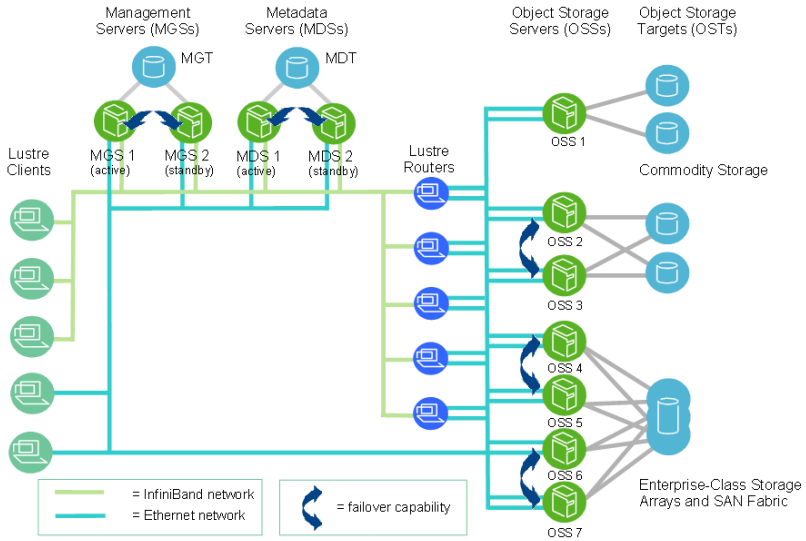


- Logical Metadata Volume (LMV)
  - aggregate the MDCs
  - present a single logical metadata namespace to clients
  - provide transparent access across all the MDTs
- Logical Object Volume (LOV)
  - aggregate the OSCs
  - provide transparent access across all the OSTs

Simple Configuration



Larger Configuration



Lustre FS Storage and I/O

Lustre Object structure

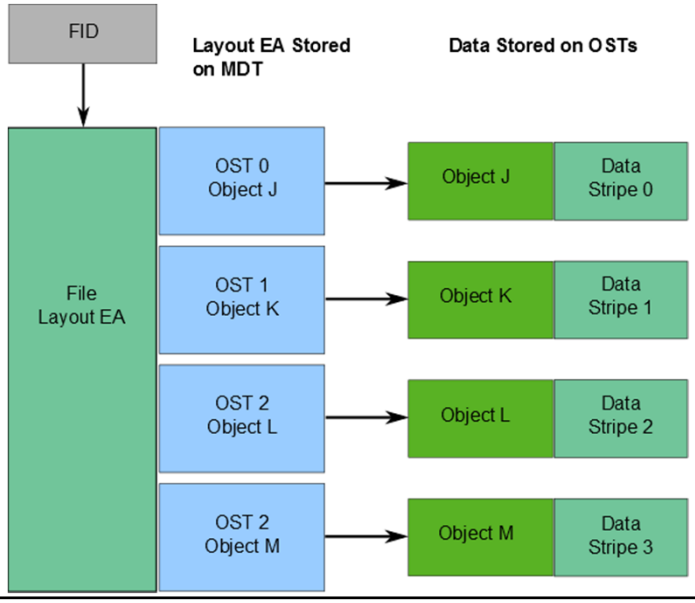
File Identification

- Each Lustre file is associated to a FID
  - Equivalent of Linux Inode in local FS
  - 128 bit length
    - Sequence = 64 bits, locate storage target, unique across all MDTs, OSTs
    - Object ID (OID) = 32 bits, reference to the object within the sequence
    - Version number = 32 bits
  - Sequences are granted to clients by servers at client connection to the FS
- FID is stored as part of the name in the file the parent directory

Object MD

- Located on MDT
- Stored in Lustre layout: an extended attribute of MDT object = a ldiskfs file
- MDT object is identified by FID

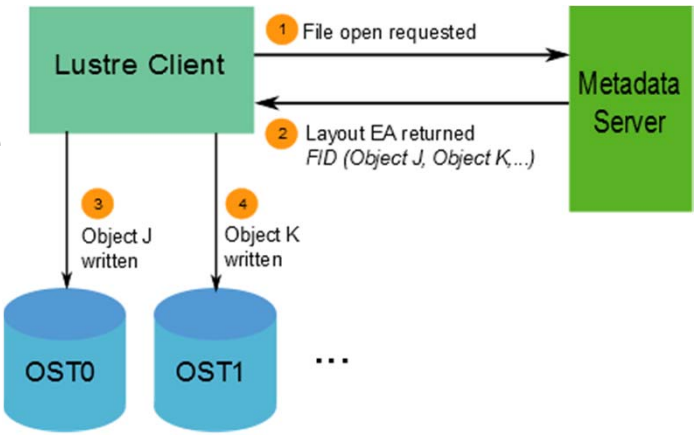
Lustre Regular File



- MDT object points to file data on OSTs

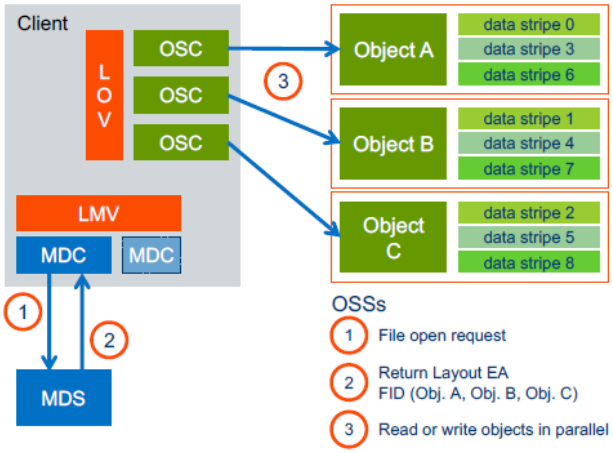
Lustre Regular File Access

- File access
  - Client lookup file to get FID
  - Client read layout on MDT
  - Client access data on OSTs
- Small file can use File On MDT feature
  - Client-MDT only dialog
  - Reduce network traffic
  - Reduce file access latency

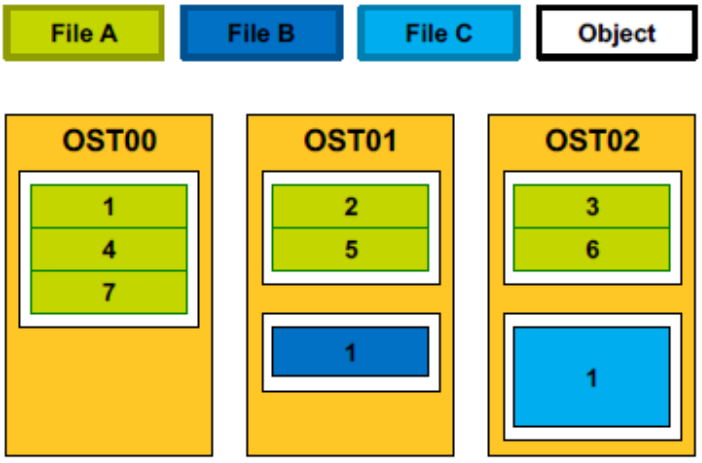


Lustre File Striping

- Striping allows parallelism between OSTs
- Striping is done in a round-robin fashion over allocated object
  - Not always the same order
- File size limit (2000) comes from the max size of the layout EA
- Software layering



Lustre File Striping: Global View



LNet

LNet

LNet is Lustre Network layer

- Originally derived from a project called Portals
- Designed to be lightweight and efficient
- Message passing for RPC request processing
- Bulk transfers for data movement
- Lightweight and versatile, capable of operating over different network fabrics
- Implemented as a Linux kernel module (Lustre Network Driver = LND)
  - Pluggable driver modules
- All participants in a Lustre file system must have a valid LNet configuration
- The LND provides an interface abstraction between the upper level LNet protocol and the kernel device driver for the network interface
- Multiple LNDs can be active on a host simultaneously

LNet (Cont.)

RPC model + Bulk transfers

- Intense use of call backs
  - Client send a request and register a call back function
  - Server do the action and send the reply
  - Client receives the reply and calls the CB functions
  - Allow the client to work during server work
    - Parallelism

LNet (Cont.)

Supported Networks through Lustre Network Driver (LND)

- Intel Omni-path(OPA) (o2iblnd)
- Mellanox InfiniBand (IB) (o2iblnd)
- Atos BXI (Bull eXtreme Interconnect) (ptlflnd)
- Cray Aries (gnilnd)
- RDMA over Converged Ethernet (RoCE) (o2iblnd)
- TCP/IP (socklnd)

- o2iblnd is OpenFabrics Enterprise Distribution (OFED) Lnet Driver
  - A “standard” RDMA interface

LNet (Cont.)

LNet Address

- NID or lustre Network IDentifier
- <IPv4 address>@<LND protocol><lnd#>
  - Address in network
  - Protocol
  - Interface number
- example: 192.68.1.10@tcp0

LNet Routers

Goals

- Used to direct Lustre I/O between different networks
- Can be used to bridge different network technologies
- Can be used as routing between independent subnets

Description

- Dedicated servers that do not participate as clients of a Lustre file system but provide a way to efficiently connect different networks

Use

- Enable centralization of Lustre server resources
- File systems can be made available to multiple administrative domains within a data centre
- Connect Lustre storage to multiple HPC clusters

Thank you for your attention

ENSIIE | 2018

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM-Ile de France | 91297 Bruyères-le-Châtel Cedex
T. +33 (0)1 69 26 40 00 | F. +33 (0)1 69 26 70 86

Direction des applications militaires
Département sciences de la simulation et de l'information
Service informatique scientifique et réseaux

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019