

Project Analysis

INTRODUCTION

The Lustre® file system is an open-source, parallel file system that supports many requirements of leadership class HPC simulation environments.It is generally used for large-scale [cluster computing](https://en.wikipedia.org/wiki/Cluster_computing).It is highly scalable and can support many thousands of client nodes, petabytes of storage, and hundreds of gigabytes per second of I/O throughput.

Lustre is purpose-built to provide a coherent, global POSIX-compliant namespace for very large scale computer infrastructure, including the world's largest supercomputer platforms. It can support hundreds of petabytes of data storage and hundreds of gigabytes per second in simultaneous, aggregate throughput. Some of the largest current installations have individual file systems in excess of fifty petabytes of usable capacity, and have reported throughput speeds exceeding one terabyte/sec.

Project Summary

| Website | <https://www.lustre.org/> |
| --- | --- |
| Lustre Repository | <https://git.whamcloud.com/> |
| Organization | [OpenSFS](http://www.opensfs.org/) and [EOFS](http://www.eofs.org/) |
| License | [GPL v2](https://en.wikipedia.org/wiki/GNU_General_Public_License) , [LGPL](https://en.wikipedia.org/wiki/GNU_Lesser_General_Public_License) |
| Open/Proprietary | Open |
| Source code | <https://git.whamcloud.com/?p=fs/lustre-release.git;a=summary>  <https://github.com/lustre/lustre-release> (Github mirror) |
| Documentation | <https://doc.lustre.org/lustre_manual.xhtml#part.intro> |

PROJECT DETAILS

Lustre Features

• POSIX Compliance: With few exceptions, Lustre passes the full POSIX test suite. Most operations are atomic to ensure that clients do not see stale data or metadata. Lustre also supports mmap() file IO.

• Online file system checking: Lustre provides a file system checker (LFSCK) to detect and correct file system inconsistencies. LFSCK can be run while the file system in online and in production, minimizing potential downtime.

• Controlled file layouts: The file layouts that determine how data is placed across the Lustre servers can be customized on a per-file basis. This allows users to optimize the layout to best fit their specific use case.

• Support for multiple backend file systems: When formatting a Lustre file system, the underlying storage can be formatted as either ldiskfs (a performance-enhanced version of ext4) or ZFS.

• Support for high-performance and heterogeneous networks: Lustre can utilize RDMA over low latency networks such as Infiniband or Intel OmniPath in addition to supporting TCP over commodity networks. The Lustre networking layer provides the ability to route traffic between multiple networks making it feasible to run a single site-wide Lustre file system.

• High-availability: Lustre supports active/active failover of storage resources and multiple mount protection (MMP) to guard against errors that may result from mounting the storage simultaneously on multiple servers. High availability software such as Pacemaker/Corosync can be used to provide automatic failover capabilities.

• Security features: Lustre follows the normal UNIX file system security model enhanced with POSIX ACLs. The root squash feature limits the ability of Lustre clients to perform privileged operations. Lustre also supports the configuration of Shared-Secret Key (SSK) security.

• Capacity growth: File system capacity can be increased by adding additional storage for data and metadata while the file system is online.

| **Dates recorded** | modification (mtime), attribute modification (ctime), access (atime), delete (dtime), create (crtime) |
| --- | --- |
| **Date range** | 2^34 bits (ext4), 2^64 bits (ZFS) |
| **Date resolution** | 1 s |
| **Forks** | No |
| **Attributes** | 32bitapi, acl, checksum, flock, lazystatfs, localflock, lruresize, noacl, nochecksum, noflock, nolazystatfs, nolruresize, nouser\_fid2path, nouser\_xattr, user\_fid2path, user\_xattr |
| [**File system permissions**](https://en.wikipedia.org/wiki/File_system_permissions) | [POSIX](https://en.wikipedia.org/wiki/POSIX), [POSIX.1e ACL](https://en.wikipedia.org/wiki/Access_Control_List), [SELinux](https://en.wikipedia.org/wiki/Security_Enhanced_Linux) |
| **Transparent compression** | Yes (ZFS only) |
| [**Transparent encryption**](https://en.wikipedia.org/wiki/Filesystem-level_encryption) | Yes (network, storage with ZFS 0.8+, fscrypt with Lustre 2.14.0+) |
| [**Data deduplication**](https://en.wikipedia.org/wiki/Data_deduplication) | Yes (ZFS only) |
| **Copy-on-write** | Yes (ZFS only) |

Lustre Components

• Management Server (MGS) - Provides configuration information for the file system. When mounting the file system, the Lustre clients will contact the MGS to retrieve details on how the file system is configured (what servers are part of the file system, failover information, etc.). The MGS can also proactively notify clients about changes in the file system configuration and plays a role in the Lustre recovery process.

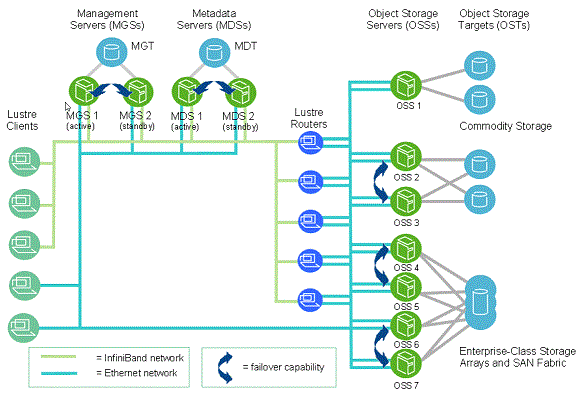
• Management Target (MGT) - Block device used by the MGS to persistently store Lustre file system configuration information. It typically only requires a relatively small amount of space (on the order to 100 MB).

• Metadata Server (MDS) - Manages the file system namespace and provides metadata services to clients such as filename lookup, directory information, file layouts, and access permissions. The file system will contain at least one MDS but may contain more.

• Metadata Target (MDT) - Block device used by an MDS to store metadata information. A Lustre file system will contain at least one MDT which holds the root of the file system, but it may contain multiple MDTs. Common configurations will use one MDT per MDS server, but it is possible for an MDS to host multiple MDTs. MDTs can be shared among multiple MDSs to support failover, but each MDT can only be mounted by one MDS at any given time.

• Object Storage Server (OSS) - Stores file data objects and makes the file contents available to Lustre clients. A file system will typically have many OSS nodes to provide a higher aggregate capacity and network bandwidth.

Architecture



Lustre File System Users

Due to the size of their machines and I/O requirements, early adopters of Lustre were the Department of Energy National Laboratories including Lawrence Livermore, Sandia, Oak Ridge and, more recently, Los Alamos’ Cielo supercomputer is supported by the Lustre file system.

# Lustre-Current Projects

* [Directory bulk operations](https://wiki.lustre.org/index.php?title=Directory_bulk_operations&action=edit&redlink=1)
* [File Level Redundancy (FLR) - Phase 3 Erasure Coding](https://wiki.lustre.org/File_Level_Redundancy_Solution_Architecture#Phase_4:_Erasure_Coded_Striped_Files)
* [Metadata Writeback Cache](https://wiki.lustre.org/Metadata_Writeback_Cache)
* [Enhanced Adaptive Compression in Lustre](https://wiki.lustre.org/Enhanced_Adaptive_Compression_in_Lustre)

# Lustre-Future Projects

* [RTDS(Real-Time Dynamic Striping)](https://wiki.lustre.org/index.php?title=RTDS(Real-Time_Dynamic_Striping)&action=edit&redlink=1)
* [File Level Redundancy (FLR) - Phase 2 Immediate Write Mirroring](https://wiki.lustre.org/File_Level_Redundancy_Solution_Architecture#Phase_2:_Immediate_Redundancy_and_Optimized_Resync)
* [External HSM Coordinator](https://wiki.lustre.org/index.php?title=External_HSM_Coordinator&action=edit&redlink=1)
* [HSM control of data migration](https://wiki.lustre.org/index.php?title=HSM_control_of_data_migration&action=edit&redlink=1)