Installing & Configuring Lustre on KVMs

Naomi Cheeves, Gabe Maxfield

August 2, 2022

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Team Members

- Naomi
  - Senior at University of California, Davis
  - Computer Science
  - Graduating December 2022

- Gabe Maxfield
  - Sophomore at Brigham Young University
  - Computer Science
  - Graduating 2025

Goals

Our mission was to find a scalable filesystem for our HPC environment.

We decided to create a small, proof of concept cluster on some kernel-based virtual machines.

Our goals were as follows:

- Investigate Lustre
- Create a miniature lustre cluster utilizing three kvms
- Investigate and install a backend filesystem for lustre (zfs vs ldiskfs).
- Run some benchmark tests

Lustre

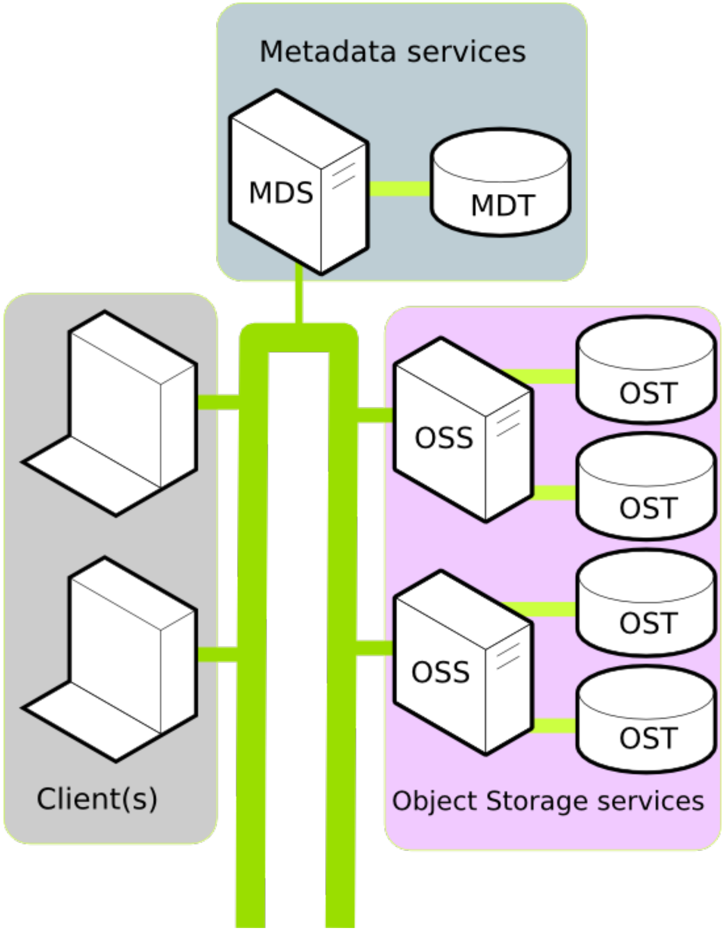
- Why use Lustre in HPC?

  Lustre has…

  - Exascale Capacities
    - Lustre uses distributed, object-based storage managed by servers.
      - Very large files can be distributed amongst different data objects and amongst several servers.
  - Data Integrity (only ZFS)
    - Data is written frequently in the event of a node crashing
  - Massive Scalability
    - Object Storage is low cost and efficient
    - I/O thoroughput and capacity are easily scaled by dynamically adding servers
  - POSIX-compliant on a Linux-based OS

The Lustre Architecture

- Components of a Cluster
  - MGS: Management Server, stores configuration information.
  - MDS: Metadata Server, manages MDTs.
  - MDT: Metadata target that stores file information/ location.
  - OSS: Object Storage Server, manages OSTs.
  - OST: Object Storage Target, storage device, hosts files.
  - Client(s): Access and use the data.



Why ZFS over LDISKFS?

- Larger Capacities
- Integrated Data Integrity

Feature	LDISKFS	ZFS
Max Volume Size	32PB	512PB
Maximum Lustre File Size	512PB	8EB
Native Data Protection	None	Mirror, RAIDZ{1,2,3}, DRAID (Future)
Detect/ repair silent data corruption	None	Yes
File system repair	Offline: FSCK	Online: ZFS Scrub

What we accomplished

In the end, our setup consisted of:

- 3 Centos 7 Kernel-based virtual machines
- A single MGS/MDS/MDT and OSS/OST setup with one client
- A small Lustre cluster built from source utilizing a ZFS backend filesystem

- What was not accomplished:

- Benchmark tests
  - Lustre utilizing local storage
  - Lustre utilizing JBODs

Challenges

- Outdated resources
  - Tutorials written for an older RHEL version
  - More niche technology with few public discussions
  - Poor documentation (Except Lustre manual)
    - Documentation used old methods and syntax
- Installing Lustre
  - Lustre packages hard to find/ Don’t support ZFS
  - Setting up the development environment
    - LNET
    - Mounting Lustre Devices

Possible Future Goals

- High Availability
  - Utilize pacemaker to maintain high availability in the event of server failure
- Explore Hierarchical Storage Management (HSM)
  - Integrate Cheap Long-term Storage Solutions
  - Automatically move old data to/ from a cheaper/ slower storage medium
  - Invisible to end client
- Explore Clustered Trivial Database (CTDB)
  - Interface for the Server Message Block protocol (Windows protocol)
  - Extend filesystem support for different clients and operating systems

Sources

- https://wiki.lustre.org/
- https://wiki.whamcloud.com/
- https://en.wikipedia.org/wiki/Lustre_(file_system)
- https://wiki.lustre.org/images/6/64/LustreArchitecture-v4.pdf
- https://www.netapp.com/data-storage/storagegrid/what-is-object-storage/

Lawrence Livermore
National Laboratory

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.