Lustre* – Manual Installation

Legal Disclaimer

- THIS DOCUMENT AND RELATED MATERIALS AND INFORMATION ARE PROVIDED "AS IS" WITH NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION, OR SAMPLE. INTEL ASSUMES NO RESPONSIBILITY FOR ANY ERRORS CONTAINED IN THIS DOCUMENT AND HAS NO LIABILITIES OR OBLIGATIONS FOR ANY DAMAGES ARISING FROM OR IN CONNECTION WITH THE USE OF THIS DOCUMENT.
- All products, product descriptions, plans, dates, and figures are preliminary based on current expectations and subject to change without notice. Availability in different channels may vary.
- Intel and the Intel logo are trademarks or registered trademarks of Intel Corporation in the United States and other countries.
- *Other names and brands may be claimed as the property of others.
-

Module Overview

- Types of Lustre* installations
- Manual Lustre* installation – ldiskfs
- Manual Lustre* installation – ZFS*
- Starting and stopping the Lustre* file system
- Summary

Overview – Types of Lustre* Installation

Types of Lustre* Installations

1. Manual Installations
   - OpenSFS Lustre*
   - Intel® Foundation Edition for Lustre* Software
   - Intel® Enterprise Edition for Lustre* Software, monitored by Intel® Manager for Lustre* software (IML)
2. Using Intel® Manager for Lustre* Software to deploy Intel® EE for Lustre* Software
3. Using Amazon* & Microsoft* to deploy Intel® Cloud Edition for Lustre* Software

Manual Installations – Similarities & Differences

- Each follows the same process, mostly
- Different origin of the packages
  - Naming scheme and content will differ
- Different hardware
  - So, different drivers and settings
- Different post-install configuration
  - Intel® Manager for Lustre* software vs. Other monitoring software
- Many more similarities than differences

Storage Targets - Differences using ZFS*

- Traditional Lustre* storage targets have been ldiskfs
- Both ldiskfs and ZFS* installations can be performed manually
  - Processes for each are similar, but they have significant differences
    - The code from the OpenZFS releases is used
      - Intel is unable to release Lustre* compiled with OpenZFS
      - Therefore, Lustre* with OpenZFS requires on-site compilation
      - Lustre* is compiled using Dynamic Kernel Module Support (DKMS)
    - Configuration of storage targets differs
      - ZFS* vs ldiskfs

Additional Software and Configuration

Vendor-required drivers

- Disk controllers, network adapters
- May be necessary to use the vendor release of a driver
- Integrator can assist in compiling custom packages to suit your hardware

High Availability (HA)

- Configuration is not trivial
- Pacemaker, Corosync, etc. – Complimentary Packages

Monitoring and Reporting

- Intel® Manager for Lustre* Software, etc.

Manual Lustre* Installation – ldiskfs

Installing Lustre* Manually - Overview

1. Verify the prerequisites have been met
2. Ensure that each component performs as expected
3. Download and install Intel® Solutions for Lustre* Software packages
4. Reboot servers to load new kernel
5. Configure LNET on all nodes
6. Format storage targets
7. Mount storage targets
8. Mount Lustre* file system on clients

Installation Prerequisites

Ensure Lustre* and the OS version(s) are compatible

- Commonly used distributions include:
  - Red Hat (RHEL*, CentOS*, Scientific Linux*) and SLES*
- Linux distributions may be possible - talk to Intel for details

A Proven Infrastructure

- Verify network and storage performance before running Lustre*
- Lustre* has "survey" tools to do bottom-up testing – Learn and use them!
  - sgpdd_survey, obdfilter_survey, ost_survey, mds_survey, LNET Self Test
- After the Lustre* filesystem is up, benchmark the I/O
  - IOR will serve well for large, sequential I/O
  - Ensure the benchmark matches the expected I/O pattern(s)

Installation Prerequisites (cont.)

Universal UID / GID

- Required for user access enforcement and by quota services

Consistent clocks / Network Time Protocol

- Lustre* doesn't itself require this
- HA packages DO require consistent clocks
- Unified time simplifies debugging

Tools to distribute commands and read output

- Example: pdsh and dshbak

Lustre* Packages

Intel releases of Lustre* are RPM packages

Releases are **packaged into Server and Client tarballs**

Lustre* Server packages include:

- Patched Linux kernel and kernel modules
- Command line utilities, Lustre* version of Linux tools (e.g. e2fsprogs)
- Intel® EE for Lustre* Software server bundle contains ~11 RPMs

Lustre* Client packages include:

- Kernel modules, command line utilities, LNET, etc.
- Intel® EE for Lustre* Software client bundle contains ~3 RPMs

Lustre* Server Packages

```
$ tar ztf lustre-2.5.37.7-bundle.tar.gz|grep rpm
./kernel-headers-2.6.32-504.30.3.el6_lustre.x86_64.rpm
./kernel-devel-2.6.32-504.30.3.el6_lustre.x86_64.rpm
./lustre-dkms-2.5.37.7-1.el6.noarch.rpm
./lustre-2.5.37.7-2.6.32_504.30.3.el6_lustre.x86_64.x86_64.rpm
./lustre-osd-zfs-mount-2.5.37.7-2.6.32_504.30.3.el6_lustre.x86_64.x86_64.rpm
./lustre-osd-ldiskfs-mount-2.5.37.7-2.6.32_504.30.3.el6_lustre.x86_64.x86_64.rpm
./lustre-osd-zfs-2.5.37.7-2.6.32_504.30.3.el6_lustre.x86_64.x86_64.rpm
./lustre-modules-2.5.37.7-2.6.32_504.30.3.el6_lustre.x86_64.x86_64.rpm
./kernel-2.6.32-504.30.3.el6_lustre.x86_64.rpm
./kernel-firmware-2.6.32-504.30.3.el6_lustre.x86_64.rpm
./lustre-osd-ldiskfs-2.5.37.7-2.6.32_504.30.3.el6_lustre.x86_64.x86_64.rpm
```

Lustre* Client Packages

```
$ tar ztf lustre-client-2.5.37.7-bundle.tar.gz|grep rpm
./lustre-client-source-2.5.37.7-2.6.32_504.30.3.el6.x86_64.x86_64.rpm
./lustre-client-2.5.37.7-2.6.32_504.30.3.el6.x86_64.x86_64.rpm
./lustre-client-modules-2.5.37.7-2.6.32_504.30.3.el6.x86_64.x86_64.rpm
```

Installing Lustre* packages

On all the servers:

- Install a new Linux kernel that has been patched with support for Lustre* (ldiskfs as backend filesystem)
  - Note that Lustre* servers with ZFS as backend filesystem, do not need Lustre*-patched kernels
  - Note that clients do not need Lustre*-patched kernels
- Install the Lustre* server packages
- Reboot the nodes to load the Lustre*-patched kernel

On all the clients:

- Install a version of the Linux kernel that has Lustre* client packages available
- Install the Lustre* client packages
- Reboot client nodes if a new kernel was installed

LNET Configuration

LNET is configured between clients and servers

If node only has a single interface:

- No manual configuration is needed
- Lustre* auto-magically configures LNET on the first interface

If a node has multiple interfaces, networks should be specified using either:

- 'networks'
- 'ip2nets'

LNET Configuration Example

Using LNET `networks` parameter to map a Linux interface to a Lustre* network:

- General syntax is:

```
options lnet networks="<net type><#>(<if name>)"
```

For example:

```
# cat /etc/modprobe.d/lustre.conf
options lnet networks="tcp1(eth1)"
```

For more information, see the Lustre* Operations Manual:

High Performance Data Division (HPDD) wiki: https://wiki.hpdd.intel.com

or

Lustre* Documentation - http://lustre.org

Storage Targets - Initialization

Initialization includes three steps:

1. Formatting of the storage target - “mkfs.lustre” command
2. Assignment of a unique index number
3. Integration into the Lustre* file system
   - Accomplished via initial mount of the target on a storage server
     - Formatting sets a on-disk flag - indicates configure the target when mounted
     - Target mounted on server - server contacts MGS - target becomes registered
     - Initial mount pending disk flag is cleared - target is available for use

Storage Target Formatting - Simple Example

MGS

```
mkfs.lustre --mgs /dev/sda
```

MDS

```
mkfs.lustre --mdt --fsname=fs1 --index=0
\
```

OSS

```
mkfs.lustre --ost --fsname=fs1 --index=0
\
  --mgsnode=192.168.3.1@o2ib0  /dev/sdc
mkfs.lustre --ost --fsname=fs1 --index=1
```

Storage Target Formatting - HA Example

MGS

```
mkfs.lustre --mgs --reformat --servicenode=192.168.3.1@o2ib0 \
  --servicenode=192.168.3.2@o2ib0 /dev/mapper/mgt
```

MDS

```
mkfs.lustre --mdt --fsname=fs2  --reformat --index=0 \
  --servicenode=192.168.3.1@o2ib0 --servicenode=192.168.3.2@o2ib0 \
  --mgsnode=192.168.3.1@o2ib0 --mgsnode=192.168.3.2@o2ib0 \
  --mkfsoptions="-J size=4096" /dev/mapper/mdt
```

OSS

```
mkfs.lustre --ost --fsname=fs2 --reformat --index=0 \
  --servicenode=192.168.3.3@o2ib0 --servicenode=192.168.3.4@o2ib0 \
  --mgsnode=192.168.3.1@o2ib0 --mgsnode=192.168.3.2@o2ib0 \
  --mkfsoptions="-J size=4096" /dev/mapper/vol1
mkfs.lustre --ost --fsname=fs2 --reformat --index=1 \
  --servicenode=192.168.3.3@o2ib0 --servicenode=192.168.3.4@o2ib0 \
  --mgsnode=192.168.3.1@o2ib0 --mgsnode=192.168.3.2@o2ib0 \
  --mkfsoptions="-J size=4096" /dev/mapper/vol2
```

Manual Lustre* Installation – ZFS*

ZFS* - What Is It?

ldiskfs is a fast, reliable file system - but it is not ZFS*

ZFS* is a file system - and more!

ZFS* provides features beyond the typical file system, to include:

- RAID - mirroring, striping, striping with parity
- Storage pools
- Compression
- Snapshots
- SSD Caching, and more

Installing Lustre* Manually – ZFS*

OpenZFS Configuration Guide - Provided within Intel® EE for Lustre* Software
Configuration guide covers:

- Prerequisites
- Supported configurations
- Lustre* with ZFS* installation
- ZFS* configuration
  - Create a MDT on a Mirrored VDEV
  - Create OSTs on RAIDZ2
  - Create DKMS packages
  - Enable ZFS* compression
- Enable monitoring using Intel® Manager for Lustre* software

Installation – Process Overview

Documented installation process is as follows:

- Download and extract the Intel EE tarball
- Run the “create_installer” script
- Copy the resulting tarball to each of the storage servers
- Extract the tarball and run the “install” script
- Create the storage targets
- Mount the storage targets
- Install Lustre* client software to the clients
- Mount Lustre* on the clients

Starting and Stopping the Lustre* File System

Starting a Lustre* File System

Lustre* servers run its "services" as kernel threads

- MGT/MDT/OST server threads are started when a target is mounted
- MGC/MDC/OSC clients also start when the targets mount

To mount a target:

```
# mkdir -p /mnt/ost0

# mount -t lustre /dev/sdb /mnt/ost0
```

Starting Lustre* on a Client

From a client, mount the Lustre* file system via the MGS

```
# mkdir /scratch

# mount -t lustre mgs@o2ib0:/fs1 /scratch
```

Syntax to mount Lustre* on boot - /etc/fstab:

```
<1st MGS NID>[:<2nd MGS NID>]:/<fname> <mount pt> lustre defaults,_netdev 0 0
```

Example:

```
mds1@o2ib0:mds2@o2ib0:/fs1 /scratch lustre defaults,_netdev 0 0
```

Starting and Stopping Lustre*

Preferred order of starting:

MGS → OSTs → MDT → Clients

- Starting MDT after OSTs prevents new I/O until ALL OSTs are up

Initial start of file system, or after an upgrade:

MGS → MDT → OSTs → Clients

Preferred order of stopping:

Clients → MDT → OSTs → MGS

- Stopping MDT before OSTs prevents new I/O

Installing Lustre* – Process Overview

1. Ensure Lustre* and OS versions are compatible
2. Download and install Lustre* packages
3. Reboot servers to load new [patched] Lustre* kernel
4. Ensure network drivers and storage drivers are installed and configured
5. Install HA framework packages
6. Configure the HA framework and create resource groups to manage Lustre*
7. Add STONITH capability to the HA framework and configure fencing devices
8. Configure LNET on all nodes
9. Format storage targets
10. Mount storage targets using the HA framework command line tools
11. Mount Lustre* file system on clients

Summary – Manual Lustre* Installations

Summary

- Different types of installations
- When manual Lustre* installations are used
  - Intel® FE for Lustre* Software and Intel® EE for Lustre* Software (with Intel® Manager for Lustre*software monitoring)
- How to perform a manual installation of Lustre* for:
  - Traditional Lustre* installations using ldiskfs
  - ZFS* backend storage targets instead of ldiskfs
- How to start and stop a Lustre* file system

Congratulations! You have completed:
Lustre* - Manual Installation