**Albano GRP 1.0.0-2015**

(Revision 2015-03-26)

**Research Findings**

**for Lustre Distributed File System**

**Graduate Research Project**

Sponsor

**The Department of Electrical, Computer, Software & Systems Engineering at**

**Embry-Riddle Aeronautical University**

Released March 26, 2015

**Justin Albano**

**Abstract:** Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sit amet dolor turpis. Nam posuere lorem nibh, nec posuere lorem ultrices et. Proin est diam, volutpat nec leo ac, congue ultricies odio. Fusce turpis sapien, porta sed nunc eget, interdum dictum odio. Vestibulum id est id lacus feugiat dictum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec viverra augue orci, ac posuere elit interdum id. Vestibulum justo orci, suscipit non elit sed, placerat consectetur mi. Vivamus et odio ullamcorper, semper lorem nec, auctor urna. Proin risus nisi, ullamcorper a varius eget, elementum vel lorem. Maecenas justo ligula, dignissim et diam et, rhoncus lobortis erat. Nam molestie lorem ac mauris blandit eleifend. Nunc gravida sodales nisl, sed eleifend ante dapibus vitae. Duis convallis quam sit amet rutrum placerat. Aenean blandit in elit eu luctus. Maecenas nec tortor vitae leo rhoncus placerat.

Department of Electrical, Computer, Software & Systems Engineering

Senior Design 2013-2014

Embry-Riddle Aeronautical University

600 S. Clyde Morris Blvd.

Daytona Beach, FL 32114

This document is property of the Justin Albano, 2015, at Embry-Riddle Aeronautical University and may not be reproduced in any form or altered without the permission of the aforementioned author.

# Revision History

|  |  |  |
| --- | --- | --- |
| **Date** | **Reason for Change** | **Version** |
| Mar. 19, 2015 | Initial draft of document | 1.0.0 |

**Contents**

[Revision History ii](#_Toc414985584)

[1. Introduction 1](#_Toc414985585)

[2. Background 2](#_Toc414985586)

[3. Installation Procedures 2](#_Toc414985587)

[3.1 Lustre Server Installation & Configuration Procedures 8](#_Toc414985588)

[3.2 Configuring Server Nodes 14](#_Toc414985589)

[4. Solution 15](#_Toc414985590)

[Glossary 19](#_Toc414985591)

[Acronyms & Abbreviations 25](#_Toc414985592)

[References 26](#_Toc414985593)

[Index 28](#_Toc414985594)

# Introduction

# Background

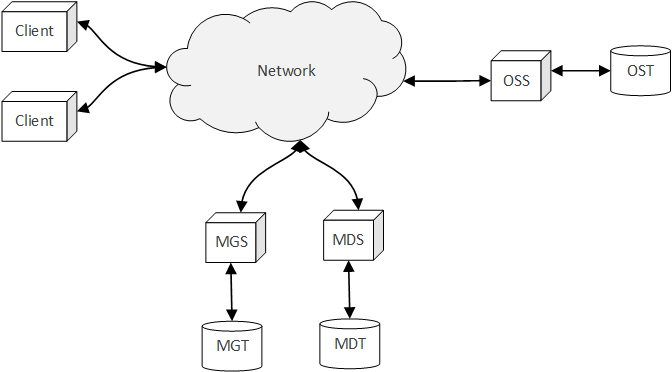
## An Overview of the Lustre File System

A Lustre file system is a highly distributed cluster of nodes, consisting of two distinct portions: (1) the server portion and (2) the client portion. The server portion of a Lustre cluster is responsible for persisting and storing the files within the file system, managing and controlling the file system, and presenting the client portion of the file system with an interface through which to interact with the files within the Lustre file system. The client portion is conversely responsible for using the server portion interface and providing end-users with access to the Lustre file system. Using this scheme, it is important that the client portion of a Lustre file system appear as though it were a local file system, ensuring that the distribution of the file system is abstracted from the end-user.

To fulfill this responsibility, the Lustre file system divides the server portion into three pairs of components: (1) object storage components, (2) management components, and (3) metadata components. These pairs are further divided into two components classifications that make up each pair: (1) servers, responsible for providing services to the file system and (2) targets, responsible for storing persistent data used by its accompanying server. This results in a total of six component classifications:

1. Object Storage Servers (OSSs)
2. Object Storage Targets (OSTs)
3. Management Servers (MGSs)
4. Management Targets (MGTs)
5. Metadata Servers (MDSs)
6. Metadata Targets (MDTs)

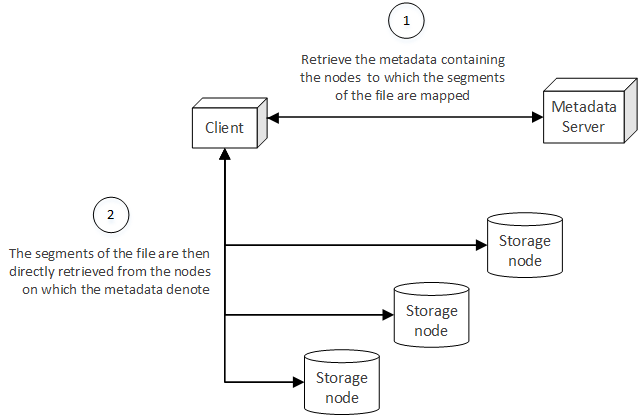
A simple, example Lustre topology is illustrated below in **Figure 1**.



**Figure 1.** Lustre file systems consist of client nodes and server nodes (composed of OSSs, OSTs, MGSs, MGTs, MDSs, and MDTs) which work in concert to provide the end-user with a seemingly local interface to the distributed file system.

An OSS in the Lustre file system is responsible for providing the services needed by the Lustre cluster to access the persistent storage of its OSTs. Each OSS typically manages and serves two to eight OSTs, with each OST having a maximum of of storage [1]. Thus, the OSS, paired with its multiple OSTs, make up the backbone of the storage capability of a Lustre file system: During operation, the data that makes up a file resides on the OSTs.

A MDS in a Lustre file system is responsible for managing the metadata about the file system. Lustre is a metadata-based file system, as opposed to a Controlled Replication under Scalable Hashing (CRUSH) file system, where the data about the files in the file system are stored in a centralized metadata server (or set of servers). In order for a client in the Lustre file system to find the location of the segments that make up a file, the client must first contact the metadata server to obtain the location of these segments. Once these locations have been obtained, the client can directly contact the nodes containing these segments, and upon gathering the required segments, can reconstruct the file. This process is illustrated below in **Figure 2**.



**Figure 2.** Once the metadata associated with a file has been retrieved from the metadata server, the client then directly accesses the storage nodes of the file system to obtain the segments of a file, ultimately reconstructing the file from the segments.

In the context of a Lustre file system, the MDS acts as the metadata server, storing the persistent metadata information on its accompanying MDT. This MDS stores the location of the segments that make up a file in the Lustre file system; in the case of Lustre, these locations amount to a mapping from a segment to a specific OST. Therefore, in order to construct a file distributed across a Lustre file system, a client requests the metadata, or mappings from segment to OST, from the MDS. Once the metadata has been gathered, the client then directly contacts the OSTs and obtains the segments of the file. Upon receiving all of the segments for a file, the client then reconstructs the file and presents the file to the end-user [1], [2].

A CRUSH file system, on the other hand, does not maintain a centralized metadata server. Rather, the location of a file segment in the file system is computed through a deterministic algorithm, as a function of the topology of the network and a rule set [3]. The advantage of a CRUSH file system is that the location of the segments of a file can be calculated by any node in the cluster, removing the need for a centralized metadata service and ultimately resulting in a truly distributed file system. While this is a desirable end-goal for a file system, a more practical technique (and more widely adopted by many of the most commonly used distributed file systems) is to use a centralized metadata server, as is the case with the Lustre file system[[1]](#footnote-1).

An MDS in a Lustre file system is responsible for maintaining the configuration data associated with the file system. In a working Lustre file system, the target nodes of the cluster will contact the MGS, providing it with relevant configuration data and information about the target, and clients will contact the MGS in order to obtain this data [1]. Similar to the OST and MDT, the MGT acts as the persistent storage device used by the MGS to persist the configuration and usage information passed to it.

While the Lustre cluster depicted in **Figure 1** is a simple example of the server and client portions of Lustre, enterprise Lustre file systems contain many more nodes than simply illustrated. As of Lustre 2.4, multiple MDTs can be supported by a single file system, but only a single MDS can interface with the MDT on behalf of a client at a given time [1]. Therefore, the metadata portion of the Lustre file system can be established to serve a fail-over capacity, where a failure in one MDS results in another MDS providing clients with access to the MDTs previously associated with the now-failed MDS. Theoretically, a Lustre file system can support the following [1]:

* 4,096 MDTs
* 8,150 OSTs

While this is a theoretical upper bound, the following has been defined as the practical range for a Lustre file system (not yet tested in a practical environment) [1]:

* 1 primary MDT with 1 MDT backup
* 500 OSSs with a total of 4,000 OSTs[[2]](#footnote-2)

Likewise, the connections between the components in a Lustre cluster are more complex than those presented in **Figure 1**. In reality, the nodes in a Lustre file system are connected through the Lustre Network (LNET). LNET is a networking Application Programming Interface (API) specific to Lustre that provides many of the capabilities required by Lustre to perform large-scale input/output (I/O) aggregation and transfers over disparate networks.

For example, if some of the nodes are connected to the file system using an InfiniBand network, while others are connected using an Ethernet network, LNET abstracts these differences and provides a consistent, aggregate network interface over which each of the components in the file system interact. This abstract network also provides routers to connect disparate networks to one another, in much the same way that a router in a physical telecommunication network operates. While LNET is a complex interface, the details of its implementation and configuration are generally abstract enough that a simple Lustre file system will not require a great deal of LNET configuration. For more information on LNET, see [1].

With an understanding of the topology and components of a Lustre cluster, it is important to the likewise understand how and where files on a Lustre file system exist. The following section delves into the details of the representation of files in a Lustre file system, including the mechanisms used to store the files on OSTs.

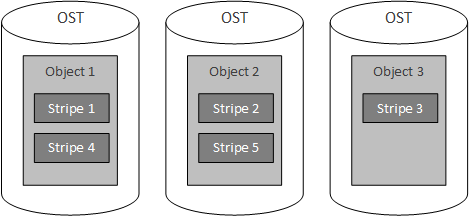
## Object Storage & Striping

Lustre is an object-based distributed file system, and therefore divides a file into numerous objects and stores these objects on OSTs through the cluster. In general, an object can be thought of as a logical portion of a file and is analogous to a file segment, as described in the previous section. The main advantage to storing the objects that make up a file on separate OSTs is that the I/O requires to retrieve a file from the file system become parallelized: Instead of a client reading the objects in a serial fashion from a single storage device (as is the case on a local file system), the client and simultaneously read the objects from different OSTs in the cluster. Thus, the distribution of objects in the file system closely mimics the behavior of a Redundant Array of Independent Disks (RAID) setup for the disks of a single machine.

Moreover, a Lustre file system can distribute the objects of a file over topologically diverse portions of a network (or an aggregate network, as in the case of LNET), ensuring that the I/O induced by the transfer of the objects is spread over different portions of a network, rather than concentrated or focused on a singular region in a network. In this sense, the distribution of objects not only provides for the parallelization of object retrieval, but also aids in mitigating possible bottlenecks induced in a network by the transfer of a large number of object.

Using this approach, a Lustre file system matches a RAID 0 arrangement, where data is sequentially distributed, or striped, across a collection of disks [5]. While there is currently work being done on the Lustre file system to support other types of RAID arrangements, particularly RAID 0+1 and RAID 5/6 (logical arrangements, in the same sense that the currently arrangement mimics that of a RAID 0 arrangement in a physical disk organization), Lustre currently supports RAID 0 and is focused on the increased I/O performance of this setup, rather than the redundancy benefits of alternative types of RAID setups[[3]](#footnote-3). Instead, it is assumed that the disks of a single OST can be configured to use a RAID selection that supports redundancy, ensuring that data on this OST is less likely to be lost, in turn reducing the likelihood of file loss over the entirety of the Lustre file system [1].

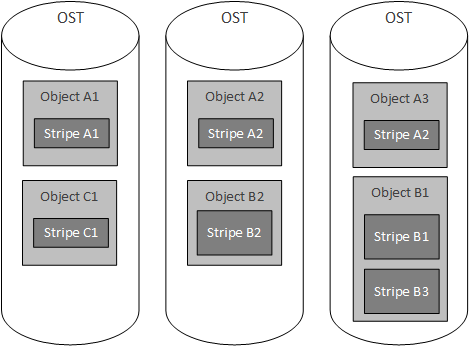
In a Lustre file system, striping occurs in a round-robin fashion, where strips are written according to a defined stripe size. When a file is stored, bits are written to an object on an OST until the stripe size is reached, whereupon bits are written to an object on the next OST. This process continues until the entirety of the file has been written to the selected OSTs. If a file is large enough that the stripes of the file have been written to the object on each of the OSTs once, the next stripe is written after the stripe within the object of the first OST [1]. This process is illustrated below in **Figure 3**.



**Figure 3.** Parts of a file exist on an OST as an object, which contains one or more stripes that are written to the object in a round-robin fashion.

This round-robin scheme is parameterized by two factors: (1) the stripe count and (2) the stripe size. The stripe count denotes the number of objects over which the file will be written (this can be logically thought of as the number of OSTs used to store the file)[[4]](#footnote-4). In **Figure 3**, the stripe count is set to 3. The stripe size is the number of bits written for each stripe. By default, the stripe count for a Lustre file system is 1, while the stripe size is [1].

Note that multiple objects, storing data for different files, exist on a single OST, each with possibly varying stripe counts and stripe sizes. **Figure 4** illustrates a more complex storage configuration, where three files are stored with different stripe counts and strip sizes. In this example, file A is written with a stripe count of 3 over each of the OSTs, while file B is written with a stripe count of 2 (and larger stripe size than file A and file C) over only two of the OSTs, starting with the right-most OST. Lastly, file C is written with a stripe count of 1, and therefore only exists on a single OST.

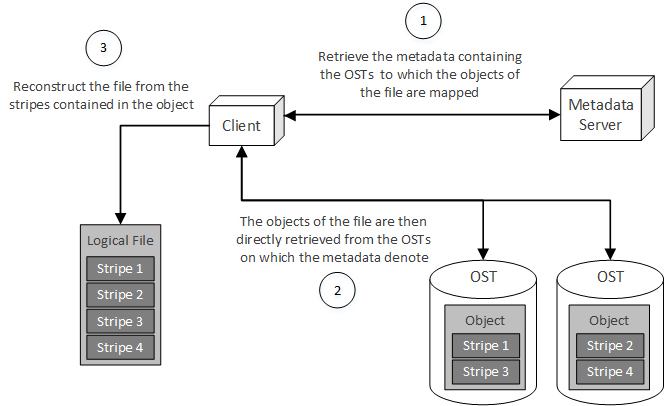


**Figure 4.** Various objects, representing parts of files, concurrently exist on a single OST, each with possibly varying stripe count and strip size.

With this understanding, it is possible to reconstruct a file given the location of the objects and the stipe size (note that the stripe count is superfluous information, as the number of OSTs found from the locations of the objects is equal to the stripe count). For example, if the objects, object 1 and object 2, for a file are known to exist on OST 1 and OST 2, respectively, and the stripe size is known, then object 1 and object 2 can be retrieved by a client. Upon retrieving the objects, object 1 is read for bits, where is equal to the stripe size. Once bits are read from the first object (and it known that there are more bits in the file, such as with knowledge of the file size), then bits are read from the second object. If there are more bits to read, bits are read from first object, starting at bit , and this process continues until the entire file is read.

Note that the last read from an object may not be equal to bits. Instead, the last number of bits, , to read from the object is bound by and is equal to

where is equal to size of the file in bits, is the number of stripes read so far, is the stripe size, and is the integer floor of (truncation of the decimal value for a given number ). The resulting process of obtaining the object locations from the MDS and reconstructing a file is illustrated in **Figure 5**. This figure is an expansion of **Figure 2**, and includes the details described in this section.



**Figure 5.** After obtaining the location of the objects from the MDS, the client then retrieves the objects from the OSTs and reconstructs a logical file from the stripes.

fasfasdfasfasfasfasfdasfasfsadf

## Client Interface to a Lustre File System

# Installation Procedures

This section contains the installation instructions used to create the Virtual Machine (VM) image used to install the Lustre server and client software. While [1] describes the installation procedures for Lustre, this source does not describe the process through which the VMs for the Lustre file system are created. Therefore, the following walkthrough describes the installation and configuration process, from start to finish, for both creating the VMs as well as installing the Lustre software. For more detailed information on how to install the Lustre software and create a Lustre cluster, see **Part II: Installing and Configuring Lustre** of [1].

Section 1 describes the process used to install the Lustre server software on a VM and configure the VM to act as a server node in a Lustre cluster. Section 2 describes the procedures required to configure a series of Lustre server VMs, as created in section 1, to act as the server-side nodes (MGS, MDS, and OSSs) in a Lustre cluster. Section 3 describes the installation procedures for installing and configuring a VM to act as a client in a Lustre cluster. Section 4 describes the configuration procedures required to connect the client or clients created in section 3 to the Lustre cluster created in section 1 and 2.

Note that each of these sections contain information on any issues encountered during the procedures. Likewise, at the time of writing, the server-side portion of the Lustre cluster (MGS, MDS, and OSSs) could not be successfully connect to one another. Therefore, the sections pertaining to this shortfall contain detailed description of the errors encountered and information relating to possible solutions and the solution approach taken thus far to mitigate or solve any of these issues.

## Lustre Server Installation & Configuration Procedures

This section contains the detailed procedures for creating a base server VM used to create the MGS, MDS, and OSSs in the Lustre cluster. This section includes information on how to create the VM image for the MGS, MDS, and OSSs, as well as how to configure this image to be included in a Lustre cluster (the scope of this configuration stops at network configuration; subsequent sections cover the configuration required to create a MGS, MDS, or OSS from this base-image).

### Creating Virtual Machine Image

The VM used to run the Lustre software was created and executing using VMWare Player 7[[5]](#footnote-5) using the auto-installer for CentOS 6.6[[6]](#footnote-6) 64-bit. In order to create the VM for CentOS 6.6, complete the following steps:

1. Open VMWare Player
2. Under the **Welcome to VMWare Player** heading, press the **Create a New Virtual Machine**
3. In the **New Virtual Machine Wizard** window, select the **Installer disc image (iso)** option
4. Select the **Browse** button and select the International Organization for Standardization (ISO) file representing the CentOS 6.6 installation image
5. Press the **Next** button
6. Enter the personalized information for the CentOS installation, such as the full name of the user, the login username, and the password for the login user
7. Press the **Next** button
8. Enter the **Virtual Machine name**, which will displayed in the list of VMs in VMWare Player
9. Select a location to store the VM files on the local machine
10. Press the **Next** button
11. Select a **Maximum disk size**
12. Select the **Split disk into multiple files** option
13. Press the **Next** button
14. Ensure that the **Power on this virtual machine after creation** option is checked
15. Press the **Finish** button

VMWare Player will then execute the auto-installer for CentOS 6.6, installing a Graphical User Interface (GUI) for CentOS. While this GUI is not required, some of the tools needed, such as Wireshark, are arguably easier to use with a GUI, and therefore, a GUI for CentOS is installed. The auto-installer will take a few minutes to install the operating system; once this installation is completed, CentOS will automatically boot. When presented with the CentOS login screen, enter the login username and password specified in step (6) above. Once logged into the CentOS VM, the Lustre software and supporting tools can be installed.

### Installing Lustre Software

Throughout the following steps, it is assumed that CentOS VM is configured to use a Network Address Translation (NAT) network configuration. In order to change the network configuration for the VM,

1. Select the **Player** dropdown at the top-left of the VMWare window executing the CentOS VM
2. Select the **Manage** option
3. Select the **Virtual Machine Settings…** option
4. Select the **Network Adapter** option under the **Hardware** tab
5. Change the network configuration options under the **Network Configuration** heading on the right
6. Press the **OK** button once the desire configuration is set

In order to install the software required to run the Lustre file system, a shared directory is used and mounted in the CentOS VM, thus allowing the needed RedHat Package Management (RPM) files to be transferred to and installed on the CentOS VM. To create the shared directory,

1. Select the **Player** dropdown at the top-left of the VMWare window executing the CentOS VM
2. Select the **Manage** option
3. Select the **Virtual Machine Settings…** option
4. Select the **Options** tab (next to the **Hardware** tab used when configuring the VM network)
5. Select the **Shared Folders** option on the left column
6. Check the **Always Enabled** option under the **Folder sharing** section in the right column
7. Press the **Add…** button at the bottom of the **Folders** section below the **Folder sharing** section
8. Press the **Next** button
9. Press the **Browse…** button under the **Host path** section
10. Select the directory to be shared between the host machine and the CentOS VM
11. Change the name of the shared directory, if desired, under the **Name** section (the name of this directory will be referenced as <shared\_dir> for the remainder of the installation procedures)
12. Press the **Next** button
13. Ensure that the **Enable this share** checkbox is checked under the **Additional attributes** section
14. Press the **Finish** button

To verify that the shared directory has been properly mounted in the CentOS VM, open a shell in the VM and execute the following command,

|  |  |
| --- | --- |
| $ | ls -l /mnt/hgfs/<shared\_dir> |

where <shared\_dir> is the name of the directory selected in step (11) when creating the shared directory. Once the shared directory has been established, the needed packages can be moved into this directory and installed. In order to install the Lustre file system on the CentOS VM, the following packages are required:

* kernel-2.6.32-431.20.3.el6\_lustre.x86\_64.rpm
* lustre-2.6.0-2.6.32\_431.20.3.el6\_lustre.x86\_64.x86\_64.rpm
* lustre-iokit-2.6.0-2.6.32\_431.20.3.el6\_lustre.x86\_64.x86\_64.rpm
* lustre-modules-2.6.0-2.6.32\_431.20.3.el6\_lustre.x86\_64.x86\_64.rpm
* lustre-osd-ldiskfs-2.6.0-2.6.32\_431.20.3.el6\_lustre.x86\_64.x86\_64.rpm
* lustre-tests-2.6.0-2.6.32\_431.20.3.el6\_lustre.x86\_64.x86\_64.rpm
* libcom\_err-1.42.12.wc1-7.el6.x86\_64.rpm
* libss-1.42.12.wc1-7.el6.x86\_64.rpm
* e2fsprogs-1.42.12.wc1-7.el6.x86\_64.rpm
* e2fsprogs-libs-1.42.12.wc1-7.el6.x86\_64.rpm
* compat-openmpi-1.4.3-1.2.el6.x86\_64.rpm
* environment-modules-3.2.10-1.el6\_5.x86\_64.rpm
* libesmtp-1.0.4-15.el6.x86\_64.rpm
* libgfortran-4.4.7-11.el6.x86\_64.rpm
* libgssglue-0.1-11.el6.x86\_64.rpm
* libibverbs-1.1.8-3.el6.x86\_64.rpm
* librdmacm-1.0.18.1-1.el6.x86\_64.rpm
* plpa-libs-1.3.2-2.1.el6.x86\_64.rpm
* sg3\_utils-1.28-6.el6.x86\_64.rpm
* tcl-8.5.7-6.el6.x86\_64.rpm

While not all of the files listed above are required directly for a Lustre installation, this list includes all dependencies of the core Lustre packages, as well, allowing a user to install the complete Lustre file system server files without the need for an internet connection (which may not be present in the environment of the VM). Each of these files can be downloaded directly from https://github.com/albanoj2/grp/tree/master/lustre-packages/server. Apart from the core Lustre server files, the following RPMs should also be installed:

* wireshark-gnome-1.8.10-7.el6\_5.x86\_64.rpm

This package provides the Wireshark application, and its associated GUI. This application will be used to analyze the network traffic originating from the Lustre file system. The non-core packages can likewise be found at https://github.com/albanoj2/grp/tree/master/lustre-packages/tools.

To install these packages, login as the root user using the following command:

|  |  |
| --- | --- |
| $ | su |

When prompted, enter the login password selected during the creation of the CentOS VM (the default root password is the login password selected during the creation of the CentOS VM). Once logged in as the root user, change directory to the shared directory containing the RPMs to be installed and execute the following command:

|  |  |
| --- | --- |
| # | yum --nogpgcheck install \* |

This command assumes that all of the packages to be installed (both the core Luster server packages, as well as the non-core packages) reside in the same directory. If this is not the same, simply change directory to any directory containing packages to be installed and execute the following command:

|  |  |
| --- | --- |
| # | yum --nogpgcheck install <rpm\_1> <rpm\_2> ... <rpm\_n> |

where <rpm\_1>, <rpm\_2>, etc. are the names of the RPMs to install, including the .rpm file extension. The

--nogpgcheck flag disables the GNU Privacy Guard (GPG) check, which allows unsigned packages to be installed (note that the authenticity of unsigned packages cannot be determined). While this is not a suggested practice when downloading packages from unknown or unsafe locations (such as from an unknown repository), the authenticity of these files is known *a priori*, since they were obtained from the official Lustre repository at [10] and [11].

Upon executing this command, the installation process with begin. When prompted to confirm the installation of the packages, enter y. The installation may take a few minutes. Once the installation is complete, a restart is required for the installation of the Lustre kernel to complete (the Lustre kernel will not be loaded until the CentOS VM is restarted). Therefore, reboot the system using the following command as the root user:

|  |  |
| --- | --- |
| # | reboot |

Although logging in as the root user is required to install packages, changing user to the root user using the su command is not always advised. Instead, the user created during the creation process for the CentOS VM can be given sudo rights. Once given sudo rights, this user will no longer be required to switch to the root user. Instead, the user can simply prepend the sudo command to each of the commands requiring root access. For example, sudo echo “Hello, world!” For more information on granting sudo rights to a user, see [12]. The remainder of these installation procedures will assume that user executing commands has sudo rights and therefore, the su command will not be used to switch to the root user.

Once the system has restarted, login to the CentOS VM. To ensure that the Lustre kernel has properly installed, open a terminal and execute the following command:

|  |  |
| --- | --- |
| $ | uname -r |

This prints the release name of the kernel used on the system. After the previous reboot, the CentOS VM should have loaded the Lustre kernel previously installed. Therefore, the output from this command should be

|  |
| --- |
| 2.6.32-431.20.3.el6\_lustre.x86\_64 |

Once the Lustre kernel installation has been confirmed, the CentOS VM can be configured.

### Configuring the Server VM

The configuration of the CentOS VM can be divided into two main parts: (1) configuring Security-Enhanced Linux (SELinux) and (2) configuring the VM hostname and Internet Protocol (IP) address. In order for Lustre to run on a Linux machine, SELinux must be disabled. In order to do this, open the /etc/selinux/config file and change the line

|  |
| --- |
| SELINUX=enforcing |

to

|  |
| --- |
| SELINUX=disabled |

For this change to take effect, the system must be rebooted. However, before rebooting, the hostname can also be changed (saving time by only rebooting the system once, after the hostname has been configured).[[7]](#footnote-7)

In order to configure the hostname, the IP address of the system must be made static. To obtain the current IP address used by the CentOS VM, execute the ifconfig command. This command should return output similar to the following:

|  |
| --- |
| eth0 Link encap:Ethernet HWaddr <some\_mac\_address>  inet addr:192.168.aaa.bbb Bcast:192.168.aaa.255 Mask:255.255.255.0  inet6 addr: fe80::20c:29ff:fe6e:ef4a/64 Scope:Link  UP BROADCAST RUNNING MULTICAST MTU:1500 Metric:1  RX packets:73 errors:0 dropped:0 overruns:0 frame:0  TX packets:51 errors:0 dropped:0 overruns:0 carrier:0  collisions:0 txqueuelen:1000  RX bytes:20412 (19.9 KiB) TX bytes:4570 (4.4 KiB)  lo Link encap:Local Loopback  inet addr:127.0.0.1 Mask:255.0.0.0  inet6 addr: ::1/128 Scope:Host  UP LOOPBACK RUNNING MTU:16436 Metric:1  RX packets:8 errors:0 dropped:0 overruns:0 frame:0  TX packets:8 errors:0 dropped:0 overruns:0 carrier:0  collisions:0 txqueuelen:0  RX bytes:480 (480.0 b) TX bytes:480 (480.0 b) |

Under the eth0 section, the IPv4 address can be found under inet addr:. In the case of this example output, the IPv4 address is 192.168.aaa.bbb. Using this existing address, a static address can be selected from range 192.168.aaa. In the case of this example, the static IP address 192.168.aaa.140 is chosen (this IP address will be referenced by <chosen\_ip> for the remainder of this installation procedure). To set this static IP address, open the /etc/sysconfig/network-scripts/ifcfg-eth0 file in a text editor and perform the following actions:

1. Comment out the line containing UUID=”<some\_uuid4>” by placing a # at the beginning of the line (i.e., change the line to #UUID=”<some\_uuid4>”)
2. Change the line BOOTPROTO="dhcp" to BOOTPROTO="static"
3. Add the line IPADDR=”<chosen\_ip>” (replacing <chosen\_ip> with the IP address selected in the previous paragraph, not literally <chosen\_ip>)
4. Add the line NETMASK=”255.255.255.0” (using the literal value 255.255.255.0)

Save the file. The resulting configuration should resemble the following:

|  |
| --- |
| DEVICE="eth0"  BOOTPROTO="static"  IPADDR="<chosen\_ip>"  NETMASK="255.255.255.0"  HWADDR="<some\_mac\_address>"  IPV6INIT="yes"  NM\_CONTROLLED="yes"  ONBOOT="yes"  TYPE="Ethernet"  #UUID="<some\_uuid4>" |

In order for this new static IP configuration to take effect, the eth0 network adapter must be restarted. To do this, execute the following commands:

|  |  |
| --- | --- |
| $ | ifdown eth0 |
| $ | ifup eth0 |

Once the network adapter is brought up (using the ifup command), the IP address of the machine can be verified using the ifconfig command. Upon running this command, the new IP address should be set to <chosen\_ip>. With the static IP address of the CentOS VM set, the hostname of the VM must be referenced to this address. This step must be performed, since the hostname of a machine running the Lustre server software cannot resolve to localhost (for more information, see the **Troubleshooting llmount.sh** section of [14]).

To change the hostname of the CentOS VM, open the /etc/sysconfig/network file in a text editor, and change the value of the HOSTNAME= key to the new hostname (for example, lustre-vm). The resulting file contents should resemble the following:

|  |
| --- |
| NETWORKING=yes  HOSTNAME=lustre-vm |

Save and close the file. The next step is to map the hostname to the static IP previously set. To do this, open the /etc/hosts file in a text editor and add the following line to the end of the file (be sure to add the following on its own line within the file):

|  |
| --- |
| <chosen\_ip> lustre-vm |

Note that lustre-vm should be replaced with the hostname selected in the /etc/sysconfig/network file. The file contents of the hosts file should resemble the following:

|  |
| --- |
| 127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain4  ::1 localhost localhost.localdomain localhost6 localhost6.localdomain6  192.168.44.140 lustre-vm |

Where lustre-vm is the hostname selected in the /etc/sysconfig/network file. For these changes to take effect, reboot the CentOS VM using the command sudo reboot.

### Creating Copies of the Server VM

With the base server image created, copies of this image can be used to create the server nodes (MGS, MDS, and OSS) in the Lustre cluster. In order to copy the base image, the VM must be shutdown. Therefore, if the VM created in the previous section is still running, shutdown the VM by completing the following steps:

1. Select the **Player** dropdown at the top-left of the VMWare window executing the CentOS VM
2. Select the **Power** option
3. Click the **Shut Down Guest** option

Be sure not to simply suspend the guest, as suspending the guest saves the state of the VM and may cause issues when copying the VM. To create a copy of the server image, complete the following steps:

1. Open a file explorer in the directory in containing the virtual machines used by VMWare Player[[8]](#footnote-8)
2. Duplicate (copy and paste) the directory containing the VM files representing the VM created in the previous section (the directory will have a name similar to that of the VM name set in the previous section, with spaces replaced by underscores)
3. Rename the duplicated directory to the desired name of the new VM directory (for example, the desired name of the VM, replacing spaces with underscores)
4. Open VMWare Player
5. Select the **Player** dropdown at the top-left of the VMWare window
6. Select the **File** option
7. Select the **Open…** option
8. Navigate to the newly copied directory (the duplicate directory created in step 2 and renamed in step 3)
9. Open this duplicated directory
10. Select the .vmx file in this duplicated directory
11. Press the **Open** button
12. Right click the newly added VM in the list of VMs on the left (the name of the VM will match the name of the original VM from which the copy was made)
13. Click the **Settings…** option
14. Select the **Options** tab
15. Change the name of the VM under the **Virtual machine name** section on the right column to the desired name of the new VM
16. Press the **OK** button at the bottom of the settings window
17. Play the renamed VM
18. Press **I Copied It** option from the window warning *This virtual machine might have been moved or copied* after playing the duplicated VM

With these steps completed, the duplicated VM is a direct copy of the VM from which it was duplicated. This process should be repeated for each of the server nodes desired. In the case of this walkthrough, the MGS and MDS are combined into a single VM, and only one OSS will be created. Therefore, one copy of the original CentOS VM is sufficient (providing two VMs: the original VM and the copied VM). It is highly suggested that the base server image (the CentOS VM at this point in the walkthrough) is copied or archived. Archiving this VM will allow new server VMs to be created at will by copying this archived VM using the steps above.

## Configuring Server Nodes

In order to configure the server nodes, the one of the server VMs must be configured to act as a MGS and MDS. While the MGS and MDS can be configured as separate nodes in a larger Lustre cluster, for the purposes of this research, a combined MGS/MDS will suffice to support the cluster. Once the MGS/MDS node has been configured, the remaining server VM must be configured as an OSS.

In the case of the MGS/MDS VM, the configuration of the node entails creating a virtual block device (representing the disk that will act as the MGT/MDT) and mounting this block device. Likewise, in the case of the OSS, a block device must be created and mounted for the OST.

### Creating & Mounting MGT/MDT Block Device

To create the block device used as the MGT/MDT disk, the MGS/MDS VM must be shutdown. Either of the server VMs created in the previous steps may be used as the MGS/MDS; whichever is selected, power down the VM and complete the following steps:

1. Right click on the VM selected as the MGS/MDS VM
2. Click the **Settings…** option
3. Click the **Add…** option at the button of the **Hardware** section on the left (if prompted to approve administrative access, agree)
4. Click **Hard Disk** in the left menu
5. Click the **Next** button
6. Select the **SCSI** option
7. Click the **Next** button
8. Select **Create a new virtual disk**
9. Click the **Next** button
10. Select an appropriate disk size (for the sake of this research, will suffice)[[9]](#footnote-9)
11. Select the **Split virtual disk into multiple files** options at the bottom of the window
12. Click the **Next** button
13. Change the name of the disk image in the **File** field under the **Disk file** section
14. Click the **Finish** button

Once the virtual hard disk is created, it will appear under the MGS/MDS VM under /dev/sdb (where /dev/sda is the primary block device on which the CentOS operating system is installed). With the block device created, the device must be formatted and mounted in order for the MGS/MDS to serve the Lustre cluster. To format the block device, play the MGS/MDS VM, and open a terminal. Once the terminal has opened, format the block device with the following command:

|  |  |
| --- | --- |
| $ | sudo mkfs.lustre --fsname=lustre --mgs --mdt --index=0 /dev/sdb |

This command assumes that the newly created block device is located at /dev/sdb; if this is not the case, simply replace /dev/sdb with the location of the block device. Note that the above command assumes that the mkfs.lustre has not been previously run on the device. If the disk is being reformatted (the mkfs.lustre must be run on a disk that has already been formatted using the mkfs.lustre command), include the --reformat flag before the location of the block device. For example,

|  |  |
| --- | --- |
| $ | sudo mkfs.lustre --fsname=lustre --mgs --mdt --index=0 --reformat /dev/sdb |

Once the format process has completed, the following output (or similar output) should be seen:

|  |
| --- |
| Permanent disk data:  Target: lustre:MDT0000  Index: 0  Lustre FS: lustre  Mount type: ldiskfs  Flags: 0x65  (MDT MGS first\_time update )  Persistent mount opts: user\_xattr,errors=remount-ro  Parameters:  device size = 2048MB  formatting backing filesystem ldiskfs on /dev/sdb  target name lustre:MDT0000  4k blocks 524288  options -J size=81 -I 512 -i 2048 -q -O dirdata,uninit\_bg,^extents,dir\_nlink,quota,huge\_file,flex\_bg -E lazy\_journal\_init -F  mkfs\_cmd = mke2fs -j -b 4096 -L lustre:MDT0000 -J size=81 -I 512 -i 2048 -q -O dirdata,uninit\_bg,^extents,dir\_nlink,quota,huge\_file,flex\_bg -E lazy\_journal\_init -F /dev/sdb 524288  Writing CONFIGS/mountdata |

With the block device formatted, the disk must be mounted in order to start the MGS/MDS service in the Lustre cluster. To mount the disk, first create a mount point and then mount the Lustre file system, using the following set of commands:

|  |  |
| --- | --- |
| $ | sudo mkdir -p /mnt/mgs-mds |
| $ | sudo mount -t lustre /dev/sdb /mnt/mgs-mds |

To ensure that the MGS/MDS has been successfully added to the Lustre cluster, execute

|  |  |
| --- | --- |
| $ | sudo cat /proc/fs/lustre/mgs/MGS/live/\* |

This command should produce the following output [1][[10]](#footnote-10):

|  |
| --- |
| fsname: lustre  flags: 0x20 gen: 7  lustre-MDT0000  Secure RPC Config Rules:  imperative\_recovery\_state:  state: startup  nonir\_clients: 0  nidtbl\_version: 3  notify\_duration\_total: 0.000000  notify\_duation\_max: 0.000000  notify\_count: 1  fsname: params  flags: 0x21 gen: 1  Secure RPC Config Rules:  imperative\_recovery\_state:  state: startup  nonir\_clients: 0  nidtbl\_version: 2  notify\_duration\_total: 0.000000  notify\_duation\_max: 0.000000  notify\_count: 0 |

If the MGT/MDT block device must be unmounted, execute the following command:

|  |  |
| --- | --- |
| $ | sudo umount /dev/sdb |

Note that the command is umount (without the *n*), not unmount. Secondly, note that if the MGT/MDT block device must be reformatted, the block device must be unmounted prior to running the reformat command. With the block device mounted, the MGS/MDS is now running in the Lustre cluster. In order to complete the server-side portion of the Lustre cluster, at least one OSS, with an accompanying OST, must be connected to the cluster.

### Creating & Mounting OST Block Device

Using the remaining server VM, create a virtual hard disk using the process presented in above in **Creating & Mounting MGT/MDT Block Device**. This new virtual hard disk will act as the block device for the OST associated with the OSS. Before formatting this hard disk, the static IP configuration of this OSS VM must be changed: Because the OSS VM is a copy of the MGS/MDS VM, they will have the same static IP configuration. Leaving this duplicate static IP configuration will result in an IP clash on the NAT network that both VMs are connected to. In order to resolve this conflict, play the OSS VM and open the /etc/sysconfig/network-scripts/ifcfg-eth0 file and change the IPADDR value to a different IP address (called <oss\_ip> for the remainder of this document). Save and close the file.

Once the static IP has been changed, the IP-host mapping must also be changed. To change this mapping, open the /etc/sysconfig/network file and change the IP address to <oss\_ip> for the lustre-vm host name. Save and close this file. In order for these changes to take effect, restart the OSS VM. Note that through the remainder of this document, the static IP address of the MGS/MDS VM (the unchanged IP address originally established for the server VM, called <chosen\_ip> in section **Configuring the Server VM** above) will be called <mgs\_ip> and the changed static IP for the OSS will be called <oss\_ip>.

To format the block device as an OST, execute the following command:

|  |  |
| --- | --- |
| $ | sudo mkfs.lustre --fsname=lustre --mgsnode=<mgs\_ip>@tcp0 --ost --index=0 /dev/sdb |

Note that <mgs\_ip> should be replaced with the static IP address of the MGS/MDS VM. Note that if the OST must be reformatted, the --reformat flag must be included in the above command. Once this command completes, the OST must be mounted. To mount the newly formatted OST, a mount must be created and the OST block device must be mounted to this mount point. To accomplish this, execute the following commands:

|  |  |
| --- | --- |
| $ | sudo mkdir -p /mnt/ost0 |
| $ | sudo mount -t lustre /dev/sdb /mnt/ost0 |

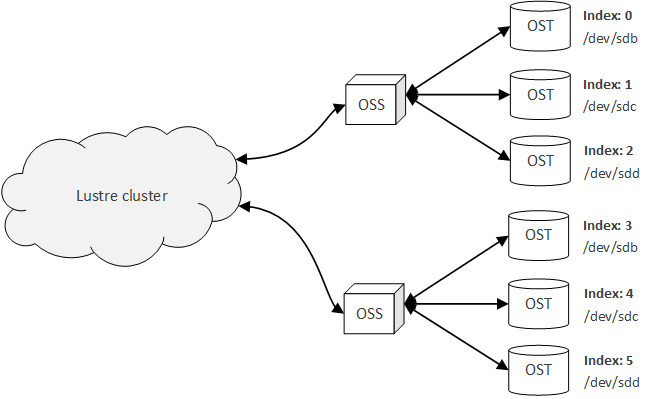
Once the OST is mounted, the creation and configuration of the server-side portion of the Lustre cluster is completed. The resulting Lustre cluster contains a single MGS/MDS node, with an accompanying MGT/MDT block device, and a single OSS, with an accompanying OST block device. Note that this is essentially a minimum Lustre cluster, and is not common of enterprise Lustre file systems. To expand the cluster, simply add more OSTs to the OSS, and likewise, increase the number of OSSs (and accompanying OSTs) to the file system using the methods descried about.

For example, to add another OST to the OSS VM, simply create another virtual hard disk (using the procedure presented at the begin of section **Creating & Mounting MGT/MDT Block Device**), format the hard disk using the same command as the original hard disk, but incrementing the index to 1. For example,

|  |  |
| --- | --- |
| $ | sudo mkfs.lustre --fsname=lustre --mgsnode=<mgs\_ip>@tcp0 --ost --index=1 /dev/sdc |

Note that the device location is no longer /dev/sdb, but rather, /dev/sdc. As more OSTs are added to the OSS, both the index and the location will increment (/dev/sdd, /dev/sde, etc.). Also note that the supplied index is not local to the OSS. Therefore, if a second OSS were added, and three OSTs are mounted on the first OSS (OST0, OST1, and OST2), an OST added to the second OSS would have an index of 4, not 0. Since there are three OSTs existing in the file system prior to the mounting of the OSTs on the second OSS, the index must be incremented to 4.

The same pattern does not apply for the block device location on each of the OSS: The location of the OST device blocks on each of the OSSs is unique to the OSS. Therefore, the virtual hard disk representing the OST on the second OSS would be located at /dev/sdb, even though there are three existing OSTs on the first OSS. Since the locations of the virtual hard disks are local to the machine mounting the hard disk, the locations of the three existing virtual hard disks representing OST0, OST1, and OST2 are unknown to the second OSS. This scheme is illustrated below in **Figure 5**.

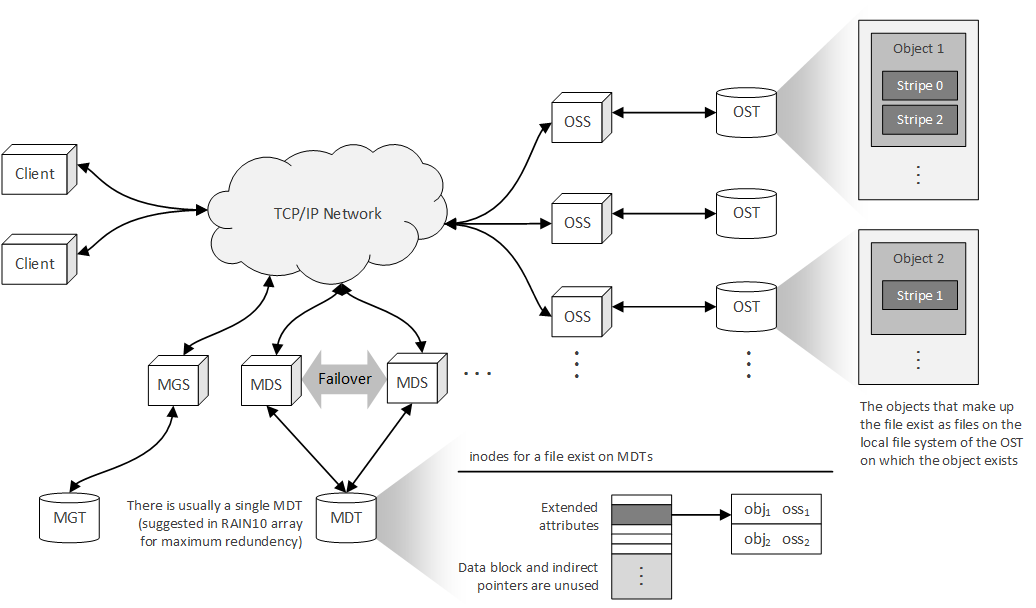


**Figure 6.** The OST indices are global in scope and are therefore sequential, even when associated with different OSSs, while the device locations of the OSTs are local to each OSS, and therefore are sequential only within the scope of each OSS.

In general, the f capacity of a Lustre file system is equal to the aggregate storage provided by each of the OSTs in the file system. Therefore, if more storage space is needed, it is advised that more OSTs are created, rather than increasing the size of the existing OSTs.

At this point in the installation process, an unresolved issue was discovered. Due to the nature of the issue, the OSTs associated with the OSS were unable to mount, and therefore, the server-side portion of the Lustre file system could not be constructed. The nature of this issue, as well as the approaches taken thus far to resolve the issue is document in the **Failure to Connect OSS to MGS/MDS Node** section of this document.

# Solution



# Outstanding & Unresolved Issues

This section contains a detailed description of all outstanding issues, of which a complete solution has not been found. Any information or possible approaches discovered during the research conducted within this document is described and referenced below. This section is intended to provide a background and context to the roadblocks discovered during research and provide the reader with a possible solution to these problem, without having to experience many of the pitfalls his or herself.

## Failure to Connect OSS to MGS/MDS Node

Upon completion of the steps outlined in the **Creating & Mounting OST Block Device** section of this document, the singular OST device of the OSS failed to mount to the OSS. Upon completion of the mount command, the following error was received through the command line,

|  |
| --- |
| mount.lustre: mount /dev/sdb at /mnt/ost0 failed: Input/output error  Is the MGS running? |

where /dev/sdb is the location of the OST block device on the OSS and /mnt/ost0 is the mount point for the formatted OST block device. In order to remove the possibility of network connection errors as a possible solution, the ping command was issued from the OSS VM to the MGS/MDS VM, and vice versa; both resulted in a successful ping, with relatively low latency (less than ). The results of the ping command from the OSS VM to the MGS/MDS VM were

|  |
| --- |
| PING 192.168.44.130 (192.168.44.130) 56(84) bytes of data.  64 bytes from 192.168.44.130: icmp\_seq=1 ttl=64 time=0.582 ms  64 bytes from 192.168.44.130: icmp\_seq=2 ttl=64 time=0.771 ms  64 bytes from 192.168.44.130: icmp\_seq=3 ttl=64 time=0.183 ms  ^C  --- 192.168.44.130 ping statistics ---  3 packets transmitted, 3 received, 0% packet loss, time 2749ms  rtt min/avg/max/mdev = 0.183/0.512/0.771/0.245 ms |

Note that 192.168.44.130 is the static IP address of the MGS/MDS VM. Likewise, the output received by executing the ping command from the MGS/MDS VM, with the OSS VM as the target, was

|  |
| --- |
| PING 192.168.44.200 (192.168.44.200) 56(84) bytes of data.  64 bytes from 192.168.44.200: icmp\_seq=1 ttl=64 time=0.211 ms  64 bytes from 192.168.44.200: icmp\_seq=2 ttl=64 time=0.803 ms  64 bytes from 192.168.44.200: icmp\_seq=3 ttl=64 time=0.304 ms  ^C  --- 192.168.44.200 ping statistics ---  3 packets transmitted, 3 received, 0% packet loss, time 2343ms  rtt min/avg/max/mdev = 0.211/0.439/0.803/0.260 ms |

Likewise note that 192.168.44.200 is the static IP address of the OSS VM. Based on the output of the two ping commands, it is clear that there was a network connection between the two machines (using a NAT network, as established through VMWare Player). In order to check if the MGS was indeed running on the MGS/MDS VM, the mounted targets in the Lustre cluster were displayed using the following command:

|  |  |
| --- | --- |
| $ | sudo cat /proc/fs/lustre/mgs/MGS/live/\* |

The results of this command were:

|  |
| --- |
| fsname: lustre  flags: 0x20 gen: 7  lustre-MDT0000  Secure RPC Config Rules:  imperative\_recovery\_state:  state: full  nonir\_clients: 0  nidtbl\_version: 3  notify\_duration\_total: 0.000000  notify\_duation\_max: 0.000000  notify\_count: 1  fsname: params  flags: 0x21 gen: 1  Secure RPC Config Rules:  imperative\_recovery\_state:  state: full  nonir\_clients: 0  nidtbl\_version: 2  notify\_duration\_total: 0.000000  notify\_duation\_max: 0.000000  notify\_count: 0 |

These results show that the MDT (lustre-MDT0000) mounted successfully. A subsequent command was issued to ensure that the MGS was started[[11]](#footnote-11):

|  |  |
| --- | --- |
| $ | sudo cat /proc/fs/lustre/devices |

The results of this command were as follows:

|  |
| --- |
| 0 UP osd-ldiskfs lustre-MDT0000-osd lustre-MDT0000-osd\_UUID 8  1 UP mgs MGS MGS 5  2 UP mgc MGC192.168.44.130@tcp 87b95ad5-7792-e71f-2b63-32b1981ee0ce 5  3 UP mds MDS MDS\_uuid 3  4 UP lod lustre-MDT0000-mdtlov lustre-MDT0000-mdtlov\_UUID 4  5 UP mdt lustre-MDT0000 lustre-MDT0000\_UUID 5  6 UP mdd lustre-MDD0000 lustre-MDD0000\_UUID 4  7 UP qmt lustre-QMT0000 lustre-QMT0000\_UUID 4  8 UP lwp lustre-MDT0000-lwp-MDT0000 lustre-MDT0000-lwp-MDT0000\_UUID 5 |

According to this output, the MGS is in fact running on the MGS/MDS VM. In order to test that the MGS/MDS VM could communicate through Lustre with the OSS VM, and vice versa, the lctl ping command was executed from the OSS VM, with the MGS/MDS VM as the target:

|  |  |
| --- | --- |
| $ | sudo lctl ping 192.168.44.130 |

Note that 192.168.44.130 is the static IP address of the MGS/MDS VM. This command resulted in the following output:

|  |
| --- |
| failed to ping 192.168.44.130@tcp: Input/output error |

Pinging the OSS from the MGS/MDS node using the lctl ping command resulted in similar output:

|  |  |
| --- | --- |
| $ | sudo lctl ping 192.168.44.200 |
| failed to ping 192.168.44.200@tcp: Input/output error | |

In order to discover any issues in the network connection between the OSS and MGS/MDS VMs, the /var/log/messages file on the OSS VM was scanned. The following pertinent output relating to the issue was found within this log file:

|  |
| --- |
| Mar 25 20:08:11 oss0 kernel: LDISKFS-fs (sdb): mounted filesystem with ordered data mode. quota=on. Opts:  Mar 25 20:08:16 oss0 kernel: Lustre: 2667:0:(client.c:1926:ptlrpc\_expire\_one\_request()) @@@ Request sent has timed out for slow reply: [sent 1427339291/real 1427339291] req@ffff880021e34c00 x1496173068157008/t0(0) o250->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 400/544 e 0 to 1 dl 1427339296 ref 1 fl Rpc:XN/0/ffffffff rc 0/-1  Mar 25 20:08:21 oss0 kernel: LustreError: 6084:0:(client.c:1083:ptlrpc\_import\_delay\_req()) @@@ send limit expired req@ffff880021e34800 x1496173068157012/t0(0) o253->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 4768/4768 e 0 to 0 dl 0 ref 2 fl Rpc:W/0/ffffffff rc 0/-1  Mar 25 20:08:21 oss0 kernel: LustreError: 6084:0:(obd\_mount\_server.c:1165:server\_register\_target()) lustre-OST0000: error registering with the MGS: rc = -5 (not fatal)  Mar 25 20:08:26 oss0 kernel: LustreError: 6084:0:(client.c:1083:ptlrpc\_import\_delay\_req()) @@@ send limit expired req@ffff880021e34800 x1496173068157016/t0(0) o101->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 328/344 e 0 to 0 dl 0 ref 2 fl Rpc:W/0/ffffffff rc 0/-1  Mar 25 20:08:31 oss0 kernel: LustreError: 6084:0:(client.c:1083:ptlrpc\_import\_delay\_req()) @@@ send limit expired req@ffff880021e34800 x1496173068157020/t0(0) o101->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 328/344 e 0 to 0 dl 0 ref 2 fl Rpc:W/0/ffffffff rc 0/-1  Mar 25 20:08:31 oss0 kernel: LustreError: 13a-8: Failed to get MGS log lustre-OST0000 and no local copy.  Mar 25 20:08:31 oss0 kernel: LustreError: 15c-8: MGC192.168.44.130@tcp: The configuration from log 'lustre-OST0000' failed (-2). This may be the result of communication errors between this node and the MGS, a bad configuration, or other errors. See the syslog for more information.  Mar 25 20:08:31 oss0 kernel: LustreError: 6084:0:(obd\_mount\_server.c:1297:server\_start\_targets()) failed to start server lustre-OST0000: -2  Mar 25 20:08:31 oss0 kernel: LustreError: 6084:0:(obd\_mount\_server.c:1769:server\_fill\_super()) Unable to start targets: -2  Mar 25 20:08:31 oss0 kernel: LustreError: 6084:0:(obd\_mount\_server.c:1496:server\_put\_super()) no obd lustre-OST0000  Mar 25 20:08:31 oss0 kernel: Lustre: server umount lustre-OST0000 complete  Mar 25 20:08:31 oss0 kernel: LustreError: 6084:0:(obd\_mount.c:1342:lustre\_fill\_super()) Unable to mount (-2)  Mar 25 20:08:51 oss0 kernel: LDISKFS-fs (sdb): mounted filesystem with ordered data mode. quota=on. Opts:  Mar 25 20:08:51 oss0 kernel: Lustre: 2667:0:(client.c:1926:ptlrpc\_expire\_one\_request()) @@@ Request sent has failed due to network error: [sent 1427339331/real 1427339331] req@ffff88003b82ec00 x1496173068157024/t0(0) o250->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 400/544 e 0 to 1 dl 1427339336 ref 1 fl Rpc:XN/0/ffffffff rc 0/-1  Mar 25 20:09:01 oss0 kernel: LustreError: 6124:0:(client.c:1083:ptlrpc\_import\_delay\_req()) @@@ send limit expired req@ffff88003b82ec00 x1496173068157028/t0(0) o253->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 4768/4768 e 0 to 0 dl 0 ref 2 fl Rpc:W/0/ffffffff rc 0/-1  Mar 25 20:09:01 oss0 kernel: LustreError: 6124:0:(obd\_mount\_server.c:1165:server\_register\_target()) lustre-OST0000: error registering with the MGS: rc = -5 (not fatal)  Mar 25 20:09:06 oss0 kernel: LustreError: 6124:0:(client.c:1083:ptlrpc\_import\_delay\_req()) @@@ send limit expired req@ffff88003b82ec00 x1496173068157032/t0(0) o101->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 328/344 e 0 to 0 dl 0 ref 2 fl Rpc:W/0/ffffffff rc 0/-1  Mar 25 20:09:11 oss0 kernel: LustreError: 13a-8: Failed to get MGS log lustre-OST0000 and no local copy.  Mar 25 20:09:11 oss0 kernel: LustreError: 15c-8: MGC192.168.44.130@tcp: The configuration from log 'lustre-OST0000' failed (-2). This may be the result of communication errors between this node and the MGS, a bad configuration, or other errors. See the syslog for more information.  Mar 25 20:09:11 oss0 kernel: LustreError: 6124:0:(obd\_mount\_server.c:1297:server\_start\_targets()) failed to start server lustre-OST0000: -2  Mar 25 20:09:11 oss0 kernel: LustreError: 6124:0:(obd\_mount\_server.c:1769:server\_fill\_super()) Unable to start targets: -2  Mar 25 20:09:11 oss0 kernel: LustreError: 6124:0:(obd\_mount\_server.c:1496:server\_put\_super()) no obd lustre-OST0000  Mar 25 20:09:12 oss0 kernel: Lustre: server umount lustre-OST0000 complete  Mar 25 20:09:12 oss0 kernel: LustreError: 6124:0:(obd\_mount.c:1342:lustre\_fill\_super()) Unable to mount (-2) |

Of particular importance is the line

|  |
| --- |
| Mar 25 20:08:16 oss0 kernel: Lustre: 2667:0:(client.c:1926:ptlrpc\_expire\_one\_request()) @@@ Request sent has timed out for slow reply: [sent 1427339291/real 1427339291] req@ffff880021e34c00 x1496173068157008/t0(0) o250->MGC192.168.44.130@tcp@192.168.44.130@tcp:26/25 lens 400/544 e 0 to 1 dl 1427339296 ref 1 fl Rpc:XN/0/ffffffff rc 0/-1 |

This line states that a request sent from the OSS to the MGS/MDS node has timed out. This is likely caused by the inability of the OSS to connect to the MGS/MDS, with respect to its Lustre network, rather than the IP network connection. In an attempt to remedy this time out error (to ensure that it was in fact not an issue of a premature time out, which given enough time, would complete), the timeout for the OST was set to 100 using the --param="sys.timeout=100" flag to the mkfs.lustre command.[[12]](#footnote-12) Again, an attempt was made to mount the OST block device to the OSS, but this mount attempt failed as well. Due to the less-than-1-millisecond latency between the OSS VM and the MGS/MDS VM, it is not likely that this timeout was caused by any delay in the connections between the two nodes.

Upon further investigation, others had been found to experience the same problems. In particular, [17] and [18] suggested possible solutions to a problem description matching the issue documented in this section; both of these solutions were attempted, but to no avail. Likewise, the steps suggested in [16] were also tried, but likewise, did not result in a solution to this issue.

Approaching the problem from a different perspective, both [19] and [20] suggest that opening port 988 in the IPTables would solve similar issues with a simple Lustre cluster. This approach was attempted by opening port 988 in the IPTables, using the following command:

|  |  |
| --- | --- |
| $ | sudo iptables -A INPUT -p tcp --dport 988 -j ACCEPT |

This command was executed on both the OSS and MGS/MDS VMs. To ensure that port 988 was in fact opened after the execution of the above command, the following command was executed[[13]](#footnote-13):

|  |  |
| --- | --- |
| $ | sudo iptables -L |

This resulted in the following output:

|  |
| --- |
| Chain INPUT (policy ACCEPT)  target prot opt source destination  ACCEPT all -- anywhere anywhere state RELATED,ESTABLISHED  ACCEPT icmp -- anywhere anywhere  ACCEPT all -- anywhere anywhere  ACCEPT tcp -- anywhere anywhere state NEW tcp dpt:ssh  REJECT all -- anywhere anywhere reject-with icmp-host-prohibited  ACCEPT tcp -- anywhere anywhere tcp dpt:988  Chain FORWARD (policy ACCEPT)  target prot opt source destination  REJECT all -- anywhere anywhere reject-with icmp-host-prohibited  Chain OUTPUT (policy ACCEPT)  target prot opt source destination |

Based on this output, it can be seen that the Transmission Control Protocol (TCP) port 988 is open from any source to any destination. In order to ensure that no other network-based services conflicted with this configuration, or that no other network-based services required a restart prior to the changes to the IP taking effect, the network service was restarted. After restarting the network service, an attempt was again made to mount the OST block device to the OSS, but again, this attempt failed.

It is worth noting that when an attempt was made to mount the OST to the OSS, after the first failure (without reformatting the OST block device using the mkfs.lustre command), the resulting error changed to

|  |
| --- |
| mount.lustre: mount /dev/sdb at /mnt/ost0 failed: No such file or directory  Is the MGS specification correct?  Is the filesystem name correct?  If upgrading, is the copied client log valid? (see upgrade docs) |

Even with this change to the reported error, the result remained the same: The OST block device was unable to mount. Also, the /var/log/messages file did not reveal any new information (apart from what was seen when the mounting process with the original error message). At the time of writing, this issue still remains unresolved.

# Glossary

|  |  |  |
| --- | --- | --- |
| **Entry** | **Definition** | **Aliases** |
|  |  |  |

# Acronyms & Abbreviations

|  |  |
| --- | --- |
| **Entry** | **Expanded Phrase** |
| API | Application Programming Interface |
| CRUSH | Controlled Replication Under Scalable Hashing |
| GPG | GNU Privacy Guard |
| GUI | Graphical User Interface |
| I/O | Input/Output |
| IP | Internet Protocol |
| ISO | International Organization for Standardization |
| LNET | Lustre Network (or Lustre Networking) |
| MDS | Metadata Server |
| MDT | Metadata Target |
| MDT | Management Target |
| MGS | Management Server |
| OSS | Object Storage Server |
| OST | Object Storage Target |
| RAID | Redundant Array of Independent Disks (or Redundant Array of Inexpensive Disks) |
| RPM | RedHat Package Manager |
| SELinux | Security-Enhanced Linux |
| TCP | Transmission Control Protocol |
| VM | Virtual Machine |

# References

1. "Lustre Software Release 2.x Operations Manual." (n.d.): n. pag. *HPDD Community Space Documentation*. Intel Corporation, 19 Mar. 2015. Web. 19 Mar. 2015. <https://build.hpdd.intel.com/job/lustre-manual/lastSuccessfulBuild/artifact/lustre\_manual.pdf>.
2. Petersen, Torben K. *Inside The Lustre File System* (n.d.): n. pag. Seagate. Seagate. Web. 26 Mar. 2015. <http://www.seagate.com/files/www-content/solutions-content/cloud-systems-and-solutions/high-performance-computing/\_shared/docs/clusterstor-inside-the-lustre-file-system-ti.pdf>.
3. Weil, Sage A., Scott A. Brandt, Ethan Miller, and Carlos Maltzahn. CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data. Tampa, FL: IEEE, 2006. Storage Systems Research Center, University of California Santa Cruz. The Institute of Electrical and Electronics Engineers, Nov. 2006. Web. 26 Mar. 2015. <http://www.ssrc.ucsc.edu/Papers/weil-sc06.pdf>.
4. Honicky, R. J., and Ethan L. Miller. Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS 2004) (2004): n. pag. Storage Systems Research Center. Jack Baskin School of Engineering, University of California, Santa Cruz, Apr. 2004. Web. 26 Mar. 2015. <http://www.ssrc.ucsc.edu/Papers/honicky-ipdps04.pdf>.
5. Stallings, William. Computer Organization & Architecture: Designing for Performance (9th Edition). Upper Saddle River, NJ: Prentice Hall, 2013. Print.
6. "Project 1: Layout Enhancement Design." Scope of Work: n. pag. OpenSFS: The Lustre File System Community. OpenFSF. Web. 26 Mar. 2015.
7. "Layout Enhancement Scope Statement." (n.d.): n. pag. OpenSFS: The Lustre File System Community. OpenFSF, 10 Oct. 2013. Web. 26 Mar. 2015.
8. “Layout Enhancement Solution Architecture.” OpenSFS: The Lustre File System Community. Open Scalable File Systems, Inc., 20 Dec. 2013. Web. 26 Mar. 2015.
9. Hammond, John. “Layout Enhancement High Level Design.” Ed. Richard Henwood. OpenSFS: The Lustre File System Community. Open Scalable File Systems, Inc., 7 Feb. 2014. Web. 26 Mar. 2015.
10. “Lustre Enterprise Linux 6.6 x86\_64 Server Packages.” *HPDD Community*. Intel Corporation, 19 Mar. 2015. Web. 19 Mar. 2015. <https://downloads.hpdd.intel.com/public/lustre/latest-feature-release/el6.6/server/RPMS/x86\_64/>.
11. “Lustre Enterprise Linux 6.6 x86\_64 e2fsprogs Packages.” *HPDD Community*. Intel Corporation, 19 Mar. 2015. Web. 19 Mar. 2015. <https://downloads.hpdd.intel.com/public/e2fsprogs/1.42.12.wc1/el6/RPMS/x86\_64/>.
12. Ellingwood, Justin. "How To Edit the Sudoers File on Ubuntu and CentOS." *How To Edit the Sudoers File on Ubuntu and CentOS*. DigitalOcean, 17 Sept. 2013. Web. 19 Mar. 2015. <https://www.digitalocean.com/community/tutorials/how-to-edit-the-sudoers-file-on-ubuntu-and-centos>.
13. Thompson, Kerry. "Crypt.Gen.NZ." *How to Disable SELinux*. Crypt.Gen.NZ, n.d. Web. 19 Mar. 2015. <http://www.crypt.gen.nz/selinux/disable\_selinux.html>.
14. Henwood,, Richard. "Testing a Lustre Filesystem." *HPDD Community Space Documentation*. Intel Corporation, 29 Oct. 2013. Web. 19 Mar. 2015. <https%3A%2F%2Fwiki.hpdd.intel.com%2Fdisplay%2FPUB%2FTesting%2Ba%2BLustre%2Bfilesystem>.
15. "Locating a Hosted Virtual Machine's Files." VMWare Knowledge Base. VMWare, 29 July 2014. Web. 24 Mar. 2015. <http%3A%2F%2Fkb.vmware.com%2Fselfservice%2Fmicrosites%2Fsearch.do%3Flanguage%3Den\_US%26cmd%3DdisplayKC%26externalId%3D1003880>.
16. "How to Create a Lustre File System." How to Create a Lustre File System. New York University: Stern School of Business, n.d. Web. 25 Mar. 2015. <http://pages.stern.nyu.edu/~nwhite/scrc/createlustre.html>.
17. Marks, Dusty. "[Lustre-discuss] Unable to Activate OST." Google Groups. Lustre-discuss Google Groups, 14 Jan. 2010. Web. 26 Mar. 2015. <https://groups.google.com/forum/#!topic/lustre-discuss-list/7NPC130LYXw>.
18. G., Sunny, and Dmitry Eremin. "Compiling, Configuring and Running Lustre on Intel® Xeon Phi™ Coprocessor." Intel Developer Zone. Intel Corporation, 25 Nov. 2014. Web. 26 Mar. 2015. <https://software.intel.com/en-us/blogs/2014/11/06/lustre-on-intel-xeon-phi>.
19. Atchley, Scott. "[Lustre-discuss] Lctl Ping Fails To/from the Client." Lustre Mailing Lists. Lustre, 14 Apr. 2007. Web. 26 Mar. 2015. <http%3A%2F%2Flists.lustre.org%2Fpipermail%2Flustre-discuss%2F2007-April%2F003223.html>.
20. Ethrbunny. "Lustre - Issues with Simple Setup." ServerFault. StackExchange, 14 Sept. 2012. Web. 26 Mar. 2015. <http://serverfault.com/questions/427447/lustre-issues-with-simple-setup>.
21. Eldar-aliyev8. "Ubuntu Documentation." IptablesHowTo. Ubuntu, 8 Feb. 2015. Web. 26 Mar. 2015. <https://help.ubuntu.com/community/IptablesHowTo>.

# Index

airport staff 1

BAS Development Team 1

BASDT 2

Birdinator Autonomous System 1

Electrical, Computer, Software & Systems Engineering i, 2, 23, 24

ECSSE 2, 23

Embry-Riddle Aeronautical University i, 2, 23, 24

ERAU 2, 23

European Union 4, 23

event 5, 15, 17, 21

Global Positioning System 4, 21, 23

GPS 4, 8, 9, 10, 21, 23

IEEE Std. 830-1998 i, 1, 24

latitude 9, 21

longitude 9, 21

use case 5, 6, 8, 11, 12

diagram 5

vehicle parameters 21

waypoint 21

Waypoint Navigation 8

World Geodetic System 1984 18, 21, 23, 24

WGS84 18, 21, 23

1. For more information on CRUSH file systems, see [3] and [4]. [↑](#footnote-ref-1)
2. Within a single Lustre file system, not per OSS [↑](#footnote-ref-2)
3. For more information on the future plans for supporting other types of logical RAID configurations, refer to Lustre Contract SFS-DEV-003. See also [6], [7], [8], and [9]. [↑](#footnote-ref-3)
4. Note that the stripe size of a file is *not* the number of stripes that make up the file, but rather, the number of objects that make up the file [1]. [↑](#footnote-ref-4)
5. VMWare Player 7.1.0 build-2496824 [↑](#footnote-ref-5)
6. CentOS-6.6-x86\_64 [↑](#footnote-ref-6)
7. For more information on disabling SELinux see [13]. [↑](#footnote-ref-7)
8. In Windows, the directory containing the virtual machine images for VMWare Player is C:\Users\<username>\Documents\Virtual Machines. For more information on location the directory containing the files that make up a VM, see [15]. [↑](#footnote-ref-8)
9. For more information on selecting an appropriate size for the MGT/MDT block device, see section **Determining Hardware Configuration Requirements and Formatting Options** of [1]. [↑](#footnote-ref-9)
10. For more information on maintaining the Lustre file system, see section Lustre Maintenance of [1]. [↑](#footnote-ref-10)
11. This command was found in the walkthrough presented in [16]. [↑](#footnote-ref-11)
12. For more information on the timeout settings for a Lustre file system, including the timeout configurations possible when creating the Lustre file system, see the **Lustre Operations** and **LustreProc** chapters of [1] (chapter 13 and 31, respectively). [↑](#footnote-ref-12)
13. For more information on altering the IPTables, see [21]. [↑](#footnote-ref-13)