

## Reference Architecture

Developing Storage  
Solutions with Intel Cloud  
Edition for Lustre\* and  
Amazon Web Services

# Developing High-Performance, Scalable, Cost-Effective Storage Solutions with Intel Cloud Edition Lustre\* and Amazon Web Services

Designed specifically for high performance computing, the open source Lustre parallel file system is one of the most popular, powerful, and scalable data storage systems available. It is used in supercomputing scenarios that require high performance and enormous storage capacity. Sixty percent of the largest 100 clusters in the world<sup>1</sup> are currently running Lustre. Amazon Web Services (AWS) is a leading provider of cloud computing infrastructure that allows scientists and engineers to solve problems that require fast computation coupled with high-bandwidth, low-latency networking.

Intel Cloud Edition for Lustre\* software provides a high-performance Lustre file system on AWS using AWS resources. It includes CentOS, Lustre, Ganglia, and Lustre Monitoring Tool (LMT). The product is delivered in the form of an Amazon Machine Image (AMI) available on the AWS Marketplace.

### Authors:

**Gabriele Paciucci**  
Intel High Performance Data  
Division Solution Architect

**Steve Paper**  
Intel Technical Account  
Manager

**Ian Meyers**  
Amazon Principal Solution  
Architect

**Dougal Ballantyne**  
Amazon HPC Solution Lead  
Architect

### Table of Contents

- Typical NFS HPC File System ..... 2
- Lustre Architecture .....3
- Intel Cloud Edition for Lustre .....4
- How to Create a Lustre File System .....5
- Building a Compute Cluster with CfnCluster .....10
- Instance Types and Performance .....13
- Summary.....15
- Legal Notices and Disclaimers.....16

### Typical NFS HPC File System

Scale-up storage solutions and other traditional network file systems, such as Network File System Version 3 (NFSv3), designate a single node to function as the I/O server for the storage cluster. All I/O data reads and writes go through that single node.

Figure 1 shows a typical NFS configuration. Although this system is simple to manage in a single cluster deployment, pushing all of an enterprise’s I/O through one server node creates a bottleneck for data-intensive workloads and for workloads that need a high number of threads and processes.

When scaling up an NFS-based environment, each NFS cluster must be managed individually, which adds to data bottlenecks as well as management overhead and costs.

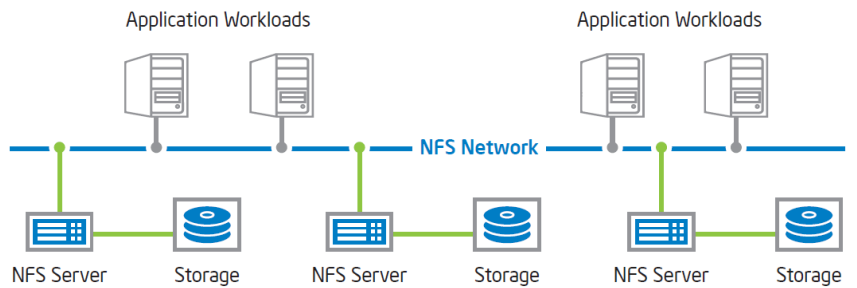


Figure 1: Typical NFS Configuration

## Lustre Architecture

Lustre is a Portable Operating System Interface (POSIX) object-based file system that splits file metadata, such as the file system namespace, file ownership, and access permission, from the file data and stores each on different servers. File metadata is stored on a metadata server. File data is split into multiple objects and stored in parallel across several object storage targets (OSTs). Figure 2 shows a typical Lustre file system configuration. The Lustre network, a very powerful and fast abstraction layer, makes it possible for the Lustre file system to run on different heterogeneous networks. Lustre Networking (LNET) provides the communications infrastructure required by the Lustre file system. It enables highly available cluster communication across a variety of networking technologies and supports transparent recovery during failures.

Lustre is designed to achieve maximum performance and scalability for POSIX applications that require outstanding streamed I/O. Users can create a single POSIX namespace of up to 512 petabytes (PB) and very large files up to 32 PB. Several sites with a Lustre cluster scale beyond one terabyte (TB) per second<sup>2</sup> and have metadata operation rates of 800,000 statistics per second.

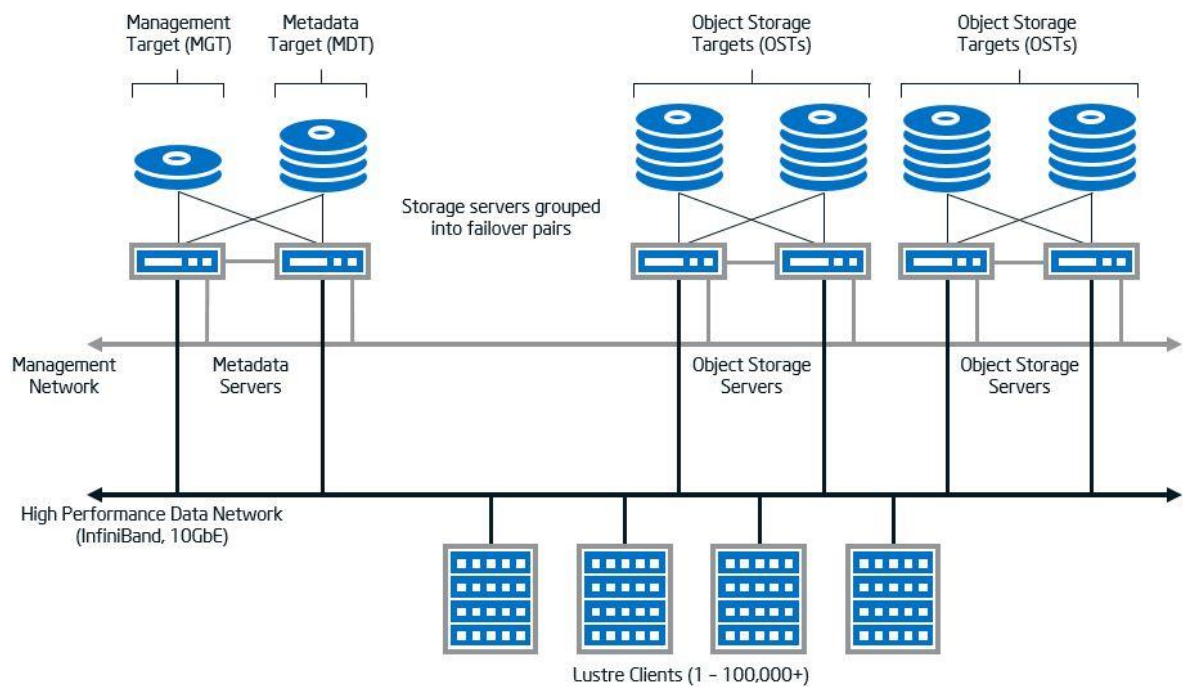


Figure 2: Typical Lustre File System Configuration

## Intel Cloud Edition for Lustre\*

Intel Cloud Edition for Lustre\* is available through the AWS Marketplace. This product provides a high performance Lustre file system on the AWS cloud using AWS compute, storage, and I/O resources supported by Intel. Intel Cloud Edition for Lustre\* is intended to be used as the working file system for High Performance Computing (HPC) or other I/O intensive workloads. It is not intended to be used as long-term storage or as an alternative to cloud storage options, such as Amazon Simple Storage Service (Amazon S3). Amazon S3 is recommended for long-term data storage on AWS; Lustre is recommended wherever a high-performance shared file system is required. With the latest edition of Intel Cloud Edition for Lustre\*, Amazon S3 storage can be used to import data into the Lustre file system.

### Available Versions

Intel Cloud Edition for Lustre\* supports several advanced AWS capabilities.

- Amazon Virtual Private Cloud (Amazon VPC), available on the Global Support and Global Support (HVM) versions, lets you provision a logically isolated section of the AWS cloud where you can launch AWS resources in a virtual network that you define. Amazon VPC is now the default mode of networking in AWS deployments. It allows for full control over addressing and access.
- The Lustre high-availability solution automatically configures Amazon Elastic Compute Cloud (Amazon EC2) Auto Scaling, which adds support for restarting unhealthy Amazon EC2 instances. If an instance becomes unhealthy, the preconfigured Auto Scaling feature will detect the failure and start a new instance. After the new instance is online, it will reattach the orphaned target's resources (network interface and Amazon Elastic Block Store [Amazon EBS] volumes) and restart the target.

The following table lists the three Intel Cloud Edition for Lustre\* versions and their features available on AWS.

| Features                               | <a href="#">Community Version</a> | <a href="#">Global Support Version</a> | <a href="#">Global Support HVM Version</a> |
|--|-----------------------------------|--|--|
| Instance types                         | M3                                | M3                                     | C3,C4                                      |
| Ephemeral storage                      | yes                               | yes                                    | yes  |
| EBS storage                            | yes                               | yes                                    | yes  |
| High availability through Auto Scaling | no                                | yes                                    | yes  |
| VPC                                    | no                                | yes                                    | yes  |
| Enhanced networking                    | no                                | no                                     | yes  |

| Features           | <a href="#">Community Version</a> | <a href="#">Global Support Version</a> | <a href="#">Global Support HVM Version</a> |
|--------------------|-----------------------------------|--|--|
| Raw initialization | no                                | yes                                    | yes  |

## Community Version

The Community version is an entry-level product that can be used for proof-of-concept development and testing. The storage templates for this product use the local disk, so it is recommended that you limit your use to scratch file systems; data will be lost if one of the Lustre servers is terminated.

## Global Support Version

The Global Support version is intended for use with HPC workloads. With high availability and storage on Amazon EBS volumes, Lustre can recover the file system after a Lustre server is terminated. In addition to supporting Amazon EBS for more resilient storage, this version also supports higher-end instance types (M3), which enable a higher performance file system.

## Global Support (HVM) Version

Hardware assisted virtualization (HVM) instances, with support for enhanced networking in the C3 and C4 compute-optimized instances, provide high performance on AWS and are recommended for Intel Cloud Edition for Lustre\*. Enhanced networking uses single root I/O virtualization (SR-IOV), which allows a physical device to be virtualized and connected directly to a virtual machine, which provides lower latency and more consistent performance.

## Support

The Intel Cloud Edition for Lustre\* software is supported by Intel. Product support includes the latest software updates, patches, and fixes to ensure a stable, flexible, and robust storage environment that leverages the benefits of cloud-based infrastructure.

## How to Create a Lustre Cluster on AWS

Intel Cloud Edition for Lustre\* is designed to create a scalable, very fast parallel Lustre file system to be attached to an external cluster of compute nodes. During the creation of the Lustre cluster, a single client will be created. This is used for test purposes only. The compute cluster can be created using a variety of cluster managers. AWS has simplified this process with an easy-to-use tool called [CfnCluster](#), which is discussed later in this paper.

## Step 1: Subscribe to a Product Version

Choose the version of Intel Cloud Edition for Lustre\* that meets your requirements and then

subscribe using the AWS Marketplace or the [Intel web page](#).

## Step 2: Launch a Cloud Formation Template

After you receive confirmation email, you are ready to use the templates to create your cluster. On the Intel web page, click the link that corresponds to your product version.

Each version has several templates to choose from. Select a cluster configuration that meets your requirements.

Templates have been created for the following AWS regions: US East (N. Virginia), US West (Oregon, N. California), Asia Pacific (Tokyo, Singapore, Sydney), South America (Sao Paulo), EU (Ireland). Choose the template for your preferred Availability Zone.

Amazon VPC and high availability (HA) templates will require additional parameters, as described in the VPC Templates section of the Launch a Cloud Formation Template page.

## Step 3: Customize Your Cluster

You can open the template files. For example, in the Template column, you can click vpc-c3 (used to deploy on the C3 instance type), modify it, and save it to a location of your choice. This gives you the flexibility to customize your cluster: to define the instance types you want to use, for example, or to include Amazon VPC settings. If you have your own modified version, select **Choose File** (shown in Figure 3), and browse to your template location. Click **Next** to continue.

Select Template

Specify a stack name and then select the template that describes the stack that you want to create.

**Stack**

An AWS CloudFormation stack is a collection of related resources that you provision and update as a single unit.

**Name**

**Template**

A template is a JSON-formatted text file that describes your stack's resources and their properties. AWS CloudFormation stores the stack's template in an Amazon S3 bucket. [Learn more.](#)

**Source**

- Select a sample template
- Upload a template to Amazon S3
  - No file chosen
- Specify an Amazon S3 template URL
  -

Figure 3: Select Template Screen

Figure 4 shows the parameters required to build a Lustre file system cluster using the templates available in the Global Support HVM version. AWS CloudFormation templates are stored on Amazon S3, and the path is filled in automatically when you click **Launch Stack**.

### Step 4: Pass the Private Key Used for SSH Connections

Enter the name of a private key to be used for SSH connections, as shown in Figure 4. The key must be created before you use the templates. For more information, see [Amazon EC2 Key Pairs](#). At this stage, you can change a number of parameters, including the number of object storage servers. (The default is 4.)

The screenshot shows a 'Parameters' form with the following fields and descriptions:

- AccessFrom:** 0.0.0.0/0. Lockdown access to Lustre services (default is accessible for 0.0.0.0/0)
- FsName:** scratch. Name of the lustre filesystem.
- HTTPFrom:** 0.0.0.0/0. Lockdown access to Lustre Ganglia on MGS (default is accessible for 0.0.0.0/0)
- ImportBucket:** [Optional] Bucket to import data from.
- ImportDest:** [Optional] Subdirectory in Lustre filesystem to import data into. Will default to ImportPrefix, if specified.
- ImportPrefix:** [Optional] Import all keys below prefix in ImportBucket. If unspecified, all keys in ImportBucket will be imported.
- KeyName:** bill. Name of and existing EC2 KeyPair to enable SSH access to the instance.
- NATInstanceType:** m3.medium. NAT Device EC2 instance type
- OssCount:** 8. Number of OSS instances.
- OstRaid:** stripe. Configure how storage is used by the lustre target. Stripe mode creates a single Lustre target on a RAID0 volume containing all available storage volumes. JBOD mode creates one Lustre target per available volume.
- OstVolumeSize:** 100. Size of EBS volumes to use for OSTs.
- SSHFrom:** 0.0.0.0/0. Lockdown SSH access to the NAT host (default can be accessed from anywhere)
- VpcId:** vpc-f250ce97. Id of an existing VPC that contains a public subnet i.e. vpc-d54ebeb7.
- VpcPrivateCIDR:** 172.31.64.0/20. CIDR for new private subnet i.e. 10.0.2.0/24.
- VpcPublicSubnetId:** subnet-83be78a8. Id of an existing public VPC subnet i.e. subnet-ae4ebec4.
- WebServerPort:** 80. The TCP port for the Web Server

Figure 4: Parameters Screen

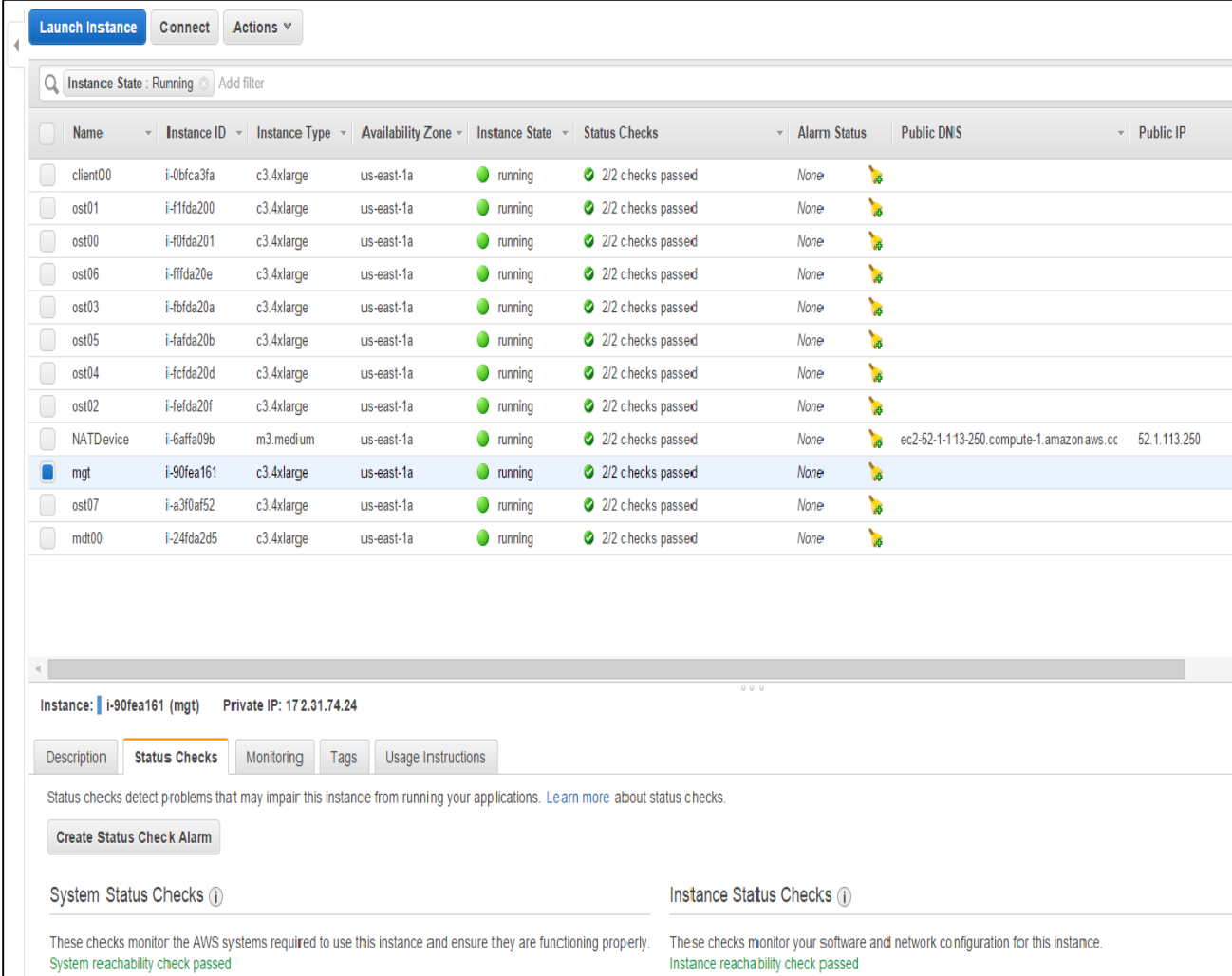


## Step 5: Launch the Instance

To launch the instance, review and acknowledge the selections at the bottom of the page, and then click **Create**. The AWS CloudFormation stack process will begin. You can use the AWS CloudFormation console to check the creation status, as shown in Figure 5.

### Important:

After the AWS CloudFormation stack process is complete, Amazon EC2 resources will be running; Lustre resources will have automatically started; the Lustre file system might be mounted by Lustre clients; and billing for the use of newly created resources will have begun.



The screenshot displays the AWS CloudFormation console interface. At the top, there are buttons for 'Launch Instance', 'Connect', and 'Actions'. Below this is a search bar for 'Instance State: Running' and a table of instances. The table has columns for Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, Public DNS, and Public IP. The 'mgt' instance is selected and highlighted in blue. Below the table, there are tabs for 'Description', 'Status Checks', 'Monitoring', 'Tags', and 'Usage Instructions'. The 'Status Checks' tab is active, showing 'System Status Checks' and 'Instance Status Checks' both with a 'passed' status.

| Name       | Instance ID       | Instance Type     | Availability Zone | Instance State | Status Checks            | Alarm Status | Public DNS                               | Public IP    |
|------------|-------------------|-------------------|-------------------|----------------|--------------------------|--------------|--|--------------|
| client00   | i-0bfca3fa        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost01      | i-f1fda200        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost00      | i-f0fda201        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost06      | i-ffffa20e        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost03      | i-fbfda20a        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost05      | i-fafda20b        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost04      | i-fcfda20d        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| ost02      | i-fefda20f        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| NATDevice  | i-6affa09b        | m3.medium         | us-east-1a        | running        | 2/2 checks passed        | None         | ec2-52-1-113-250.compute-1.amazonaws.com | 52.1.113.250 |
| <b>mgt</b> | <b>i-90fea161</b> | <b>c3.4xlarge</b> | <b>us-east-1a</b> | <b>running</b> | <b>2/2 checks passed</b> | <b>None</b>  |  |              |
| ost07      | i-a3f0af52        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |
| mdt00      | i-24fda2d5        | c3.4xlarge        | us-east-1a        | running        | 2/2 checks passed        | None         |  |              |

Instance: **i-90fea161 (mgt)** Private IP: 172.31.74.24

System Status Checks (1): System reachability check passed

Instance Status Checks (1): Instance reachability check passed

Figure 5: AWS CloudFormation Console

## Using CfnCluster to Build an HPC Compute Cluster Using a Lustre File System on AWS

Intel Cloud Edition for Lustre\* is designed to create storage nodes, but not compute nodes. Fortunately, [CfnCluster](#) can be used to create HPC compute nodes tailored to Message Passing Interface (MPI)-based applications in AWS. It does not matter what the cluster is used for and can easily be extended to support different frameworks. The command line interface (CLI) is stateless and all operations are performed using AWS CloudFormation or other AWS services. The CfnCluster tool includes a Lustre client. Be sure to verify the availability of a compatible Lustre client in distinct Amazon Machine Images (AMIs).

### Install CfnCluster and Edit the Config File

To install CfnCluster, follow [these instructions](#).

Before you can use CfnCluster, you must edit the config file, which is divided into several sections. This is where you can customize your cluster with details, such as Amazon EC2 instance types (the default is t2.micro) and the initial number of compute nodes to create (the default is 2).

In the VPC Settings section shown below, type the settings used in step 4. Otherwise, you will not be able to connect the Lustre file system and Lustre clients available on CfnCluster. For more information about the high-level network configurations CfnCluster supports, see [Network Configurations](#).

At a minimum, you will need to update the following sections of the config file:

```
[aws]
# This is the AWS credentials section (required).
# These settings apply to all clusters
# replace these with your AWS keys
# If not defined, boto will attempt to use a) environment
# or b) EC2 IAM role.
aws_access_key_id = "enter your key"
aws_secret_access_key = "enter your key"

[cluster default]
# Name of an existing EC2 KeyPair to enable SSH access to the
instances.
key_name = bill (Replace with your key name)

## VPC Settings
[vpc public]
# ID of the VPC you want to provision cluster into.
vpc_id = vpc-f250ce97 (Replace with your vpc id)
# ID of the Subnet you want to provision the Master server into
master_subnet_id = subnet-83be78a8 (replace with your subnet id)
```

### Important:

After you have updated the parameters in the config file, follow the installation instructions to

create the cluster. After the cluster is created, Amazon EC2 resources will be running and billing will have begun.

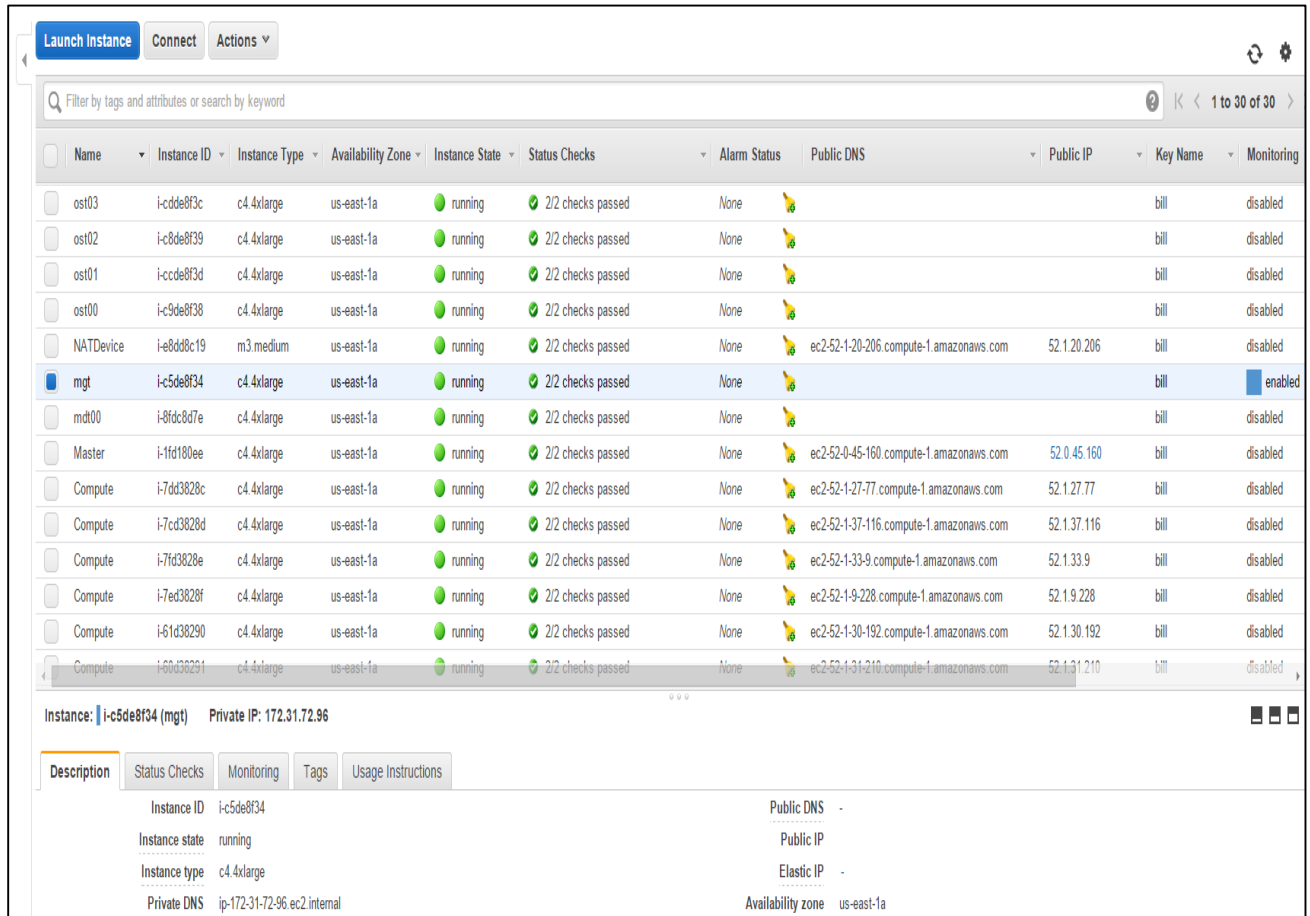


Figure 6: AWS CloudFormation Console Showing Monitoring Enabled

You can use the Amazon EC2 Management console to obtain the public IP address of your master server and the private IP address for the mgt instance shown in Figure 5 and Figure 6.

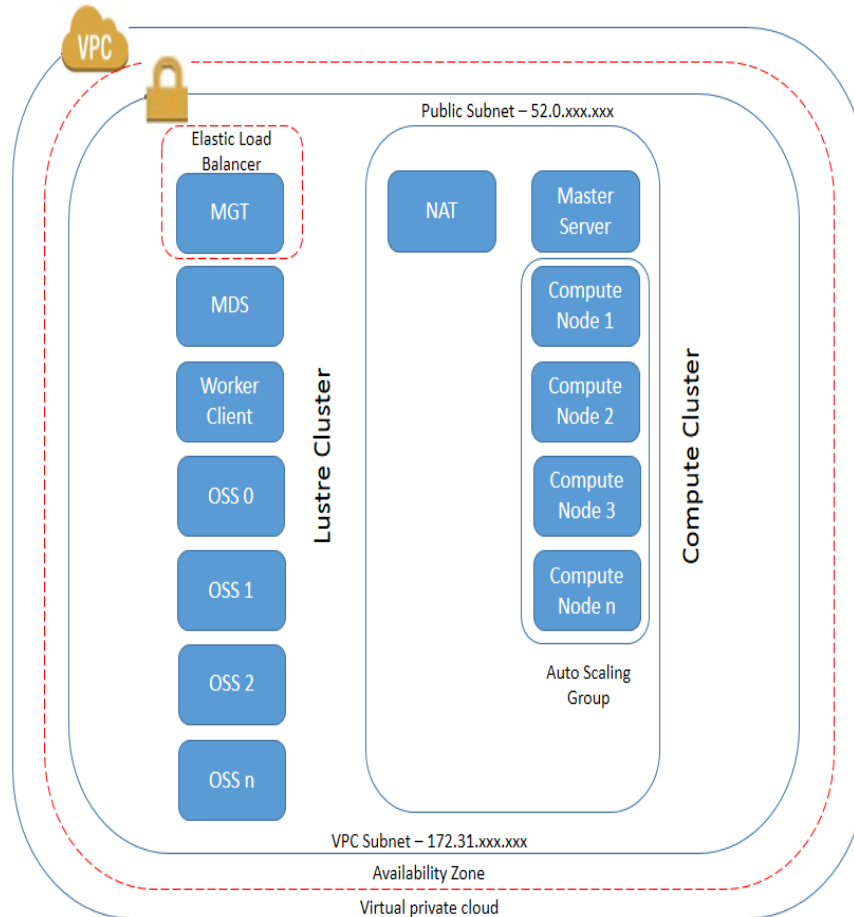


Figure 7: Resulting Client Cluster and Lustre Cluster Model

## Mount the Lustre File System

Using an SSH connection, connect to the master server, and then mount the Lustre file system on all of the compute nodes.

The Lustre client is already available on all of the compute nodes, so you can run the following command as root:

```
# mount -t lustre <Private IP of MGT>@tcp0:/scratch /mnt/lustrefs
```

To simplify the administration of all the compute nodes, use `pdsh` as `ec2-user`:

```
$ pdsh -g clients sudo -u root mount -t lustre <Private IP of MGT>@tcp0:/scratch /mnt/lustrefs
```

Open MPI libraries are also included in the `CfnCluster` software stack. MPI-based applications can be easily rebuilt to run in this environment.

To measure the I/O performance of the cluster, we compiled IOR (<https://github.com/chaos/ior>) version 2.10 with the MPI libraries.

## Instance Type and Performance Measurements

To establish the I/O performance of the file system, we created an example Lustre file system using 8x c4.4x large instances as object storage servers and 16x c4.4x large instances.

We used IOR, a parallel file system test developed by the [Scalable/IO Project \(SIOP\)](#) at Lawrence Livermore National Laboratory (LLNL). This program performs parallel writes and reads to and from a file using MPI-IO and reports the throughput rates. MPI is used for process synchronization.

We ran IOR using 256 threads across 16 nodes with xfersize of 1 MiB block size for each thread of 4 GiB and an aggregate file size of 1024 GiB, which resulted in:

- Max Write: 1818.42 MiB/sec (1906.75 MB/sec)\*
- Max Read: 1810.10 MiB/sec (1898.03 MB/sec)\*

The [Lustre Monitoring Tool](#) (LMT) is installed with Intel Cloud Edition for Lustre. Ltop is a command-line utility that gathers I/O statistics from Lustre file system servers. We used LTOP to record the file system activity during the IOR experiment:

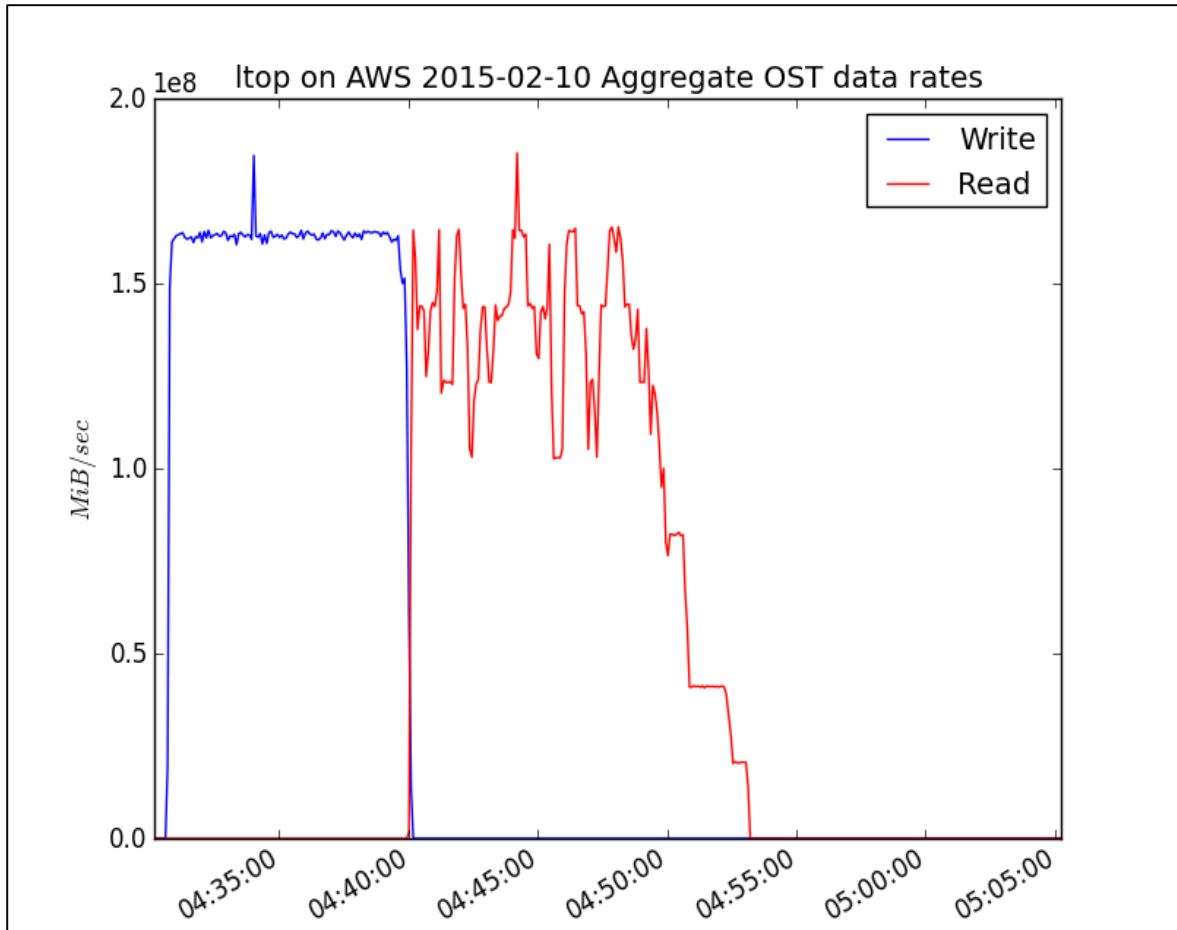


Figure 8: Lustre File System I/O Performance

Figure 9 shows the testing results. The same parameters used in steps 3 and 4 were used to show the effects of scaling the Lustre file system by increasing the number of Amazon EC2 object storage server instances.

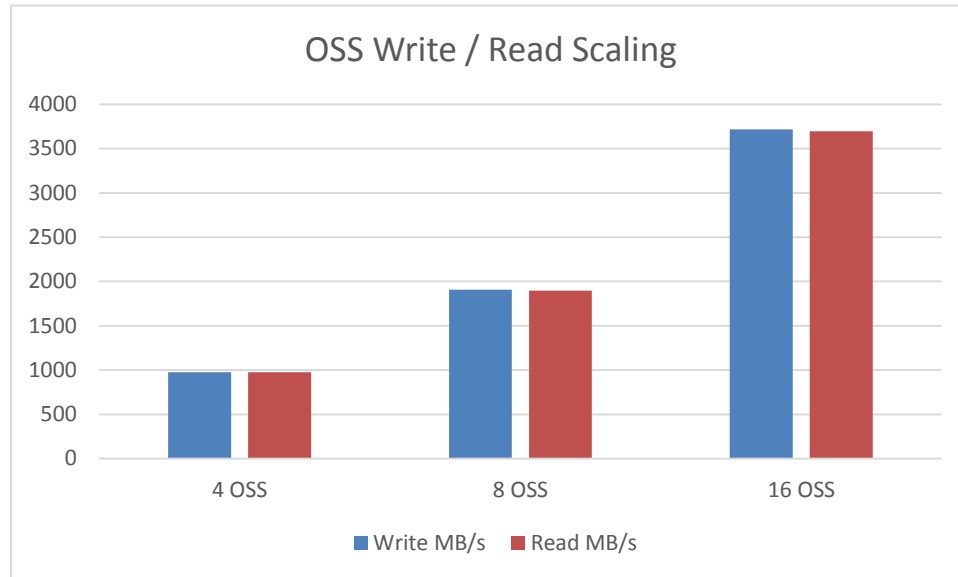


Figure 9: Testing Results

As the number of object storage server Amazon EC2 instances in the cluster increased, both read and write performance increased at a near-linear rate.

## Summary

The Intel Lustre solution is a fast, scalable storage platform positioned to accelerate application performance, even with complex workloads. Intel Cloud Edition for Lustre\* software is an ideal foundation for dynamic AWS-based workloads that require fast, scalable, and cost-effective storage. Using the resources and templates described in this document, you can innovate on your problem, not your infrastructure.

### For more information

Amazon Web Services Instance Types:

<http://aws.amazon.com/ec2/>

Intel Cloud Edition for Lustre:

<https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page>

CfnCluster Getting Started:

[http://cfncluster.readthedocs.org/en/latest/getting\\_started.html](http://cfncluster.readthedocs.org/en/latest/getting_started.html)

Configuring CfnCluster

[http://cfncluster.readthedocs.org/en/latest/getting\\_started.html#configuring-cfncluster](http://cfncluster.readthedocs.org/en/latest/getting_started.html#configuring-cfncluster)

Network Configurations Supported by CfnCluster:

<http://cfncluster.readthedocs.org/en/latest/networking.html>

IOR HPC Benchmark Source:

<https://github.com/chaos/ior>

<sup>1</sup> [www.top500.org](http://www.top500.org)

<sup>2</sup> [http://zfsonlinux.org/docs/LUG12\\_ZFS\\_Lustre\\_for\\_Sequoia.pdf](http://zfsonlinux.org/docs/LUG12_ZFS_Lustre_for_Sequoia.pdf), results in presentation by LLNL at Lustre User Group 2012, April 23, 2012

## Legal Notices & Disclaimers

© 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.

This document is provided for informational purposes only. It represents AWS's current product offerings and practices as of the date of issue of this document, which are subject to change without notice. Customers are responsible for making their own independent assessment of the information in this document and any use of AWS's products or services, each of which is provided "as is" without warranty of any kind, whether express or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from AWS, its affiliates, suppliers or licensors. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at Intel.com, or from the OEM or retailer

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

Test and System Configurations: All tests were performed by Intel using cfnccluster Version 18 to create the compute stack. Version 1.1 (01-29-2015) of the Intel Cloud Edition for Lustre GlobalSupport (HVM) was used to create the Lustre Stack. For more complete information about performance and benchmark results, visit <http://www.intel.com/performance>.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request

For more information on performance tests and on the performance of Intel products, reference [www.intel.com/procs/perf/limits.htm](http://www.intel.com/procs/perf/limits.htm) and any Intel source materials such as [performance briefs](#) or white papers.

Copyright © 2015 Intel Corporation. All rights reserved. Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

\*Other names and brands may be claimed as the property of others.