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 currently available. It is widely used in super-computing scenarios that require high performance and enormous storage capacity. Sixty percent of the largest 100 clusters in the world[1] 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

A Typical HPC File System

Scale-up storage solutions and other traditional network file systems such as 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/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.

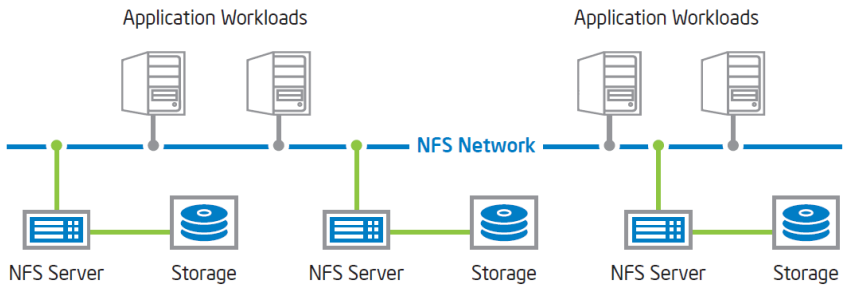


Figure 1: Typical NFS Configuration

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 (OST). 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[2] and have metadata operation rates of 800,000 statistics per second.

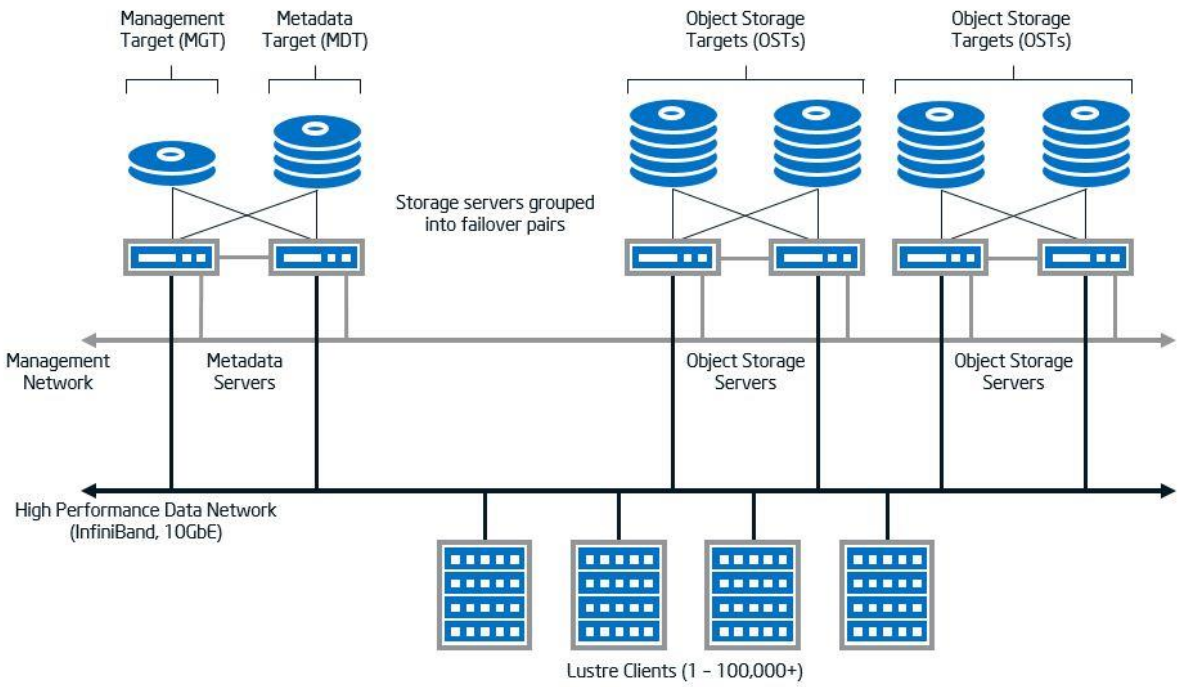


Figure 2: Typical Lustre File System Configuration

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.

- The Amazon Virtual Private Cloud (Amazon VPC) 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 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 resource (network interface and Amazon Elastic Block Store [Amazon EBS] volumes) and restart the target.

- In some end-user environments, direct access to non-VPC network resources is not allowed. As Intel Cloud Edition for Lustre* requires access to AWS service endpoints in order to perform configuration tasks on behalf of the user, this can present a deployment challenge.

  Fortunately, Intel Cloud Edition for Lustre* does support proxied access to the AWS endpoints, using a special cut-down CloudFormation template. This proxy template does not create a NAT instance to act as a gateway, and instead accepts input parameters for a HTPP proxy hostname and port. Using this information, the template is able to create a Lustre cluster without direct access to the AWS endpoints.

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

Features	Evaluation	Self-Support	Premier Support (8x5)
Intel Support	no	no	yes
Instance Types	T2	C3, C4, M4	C3, C4, M4
IPSec Encryption	yes	yes	yes
EBS Ecryption	yes	yes	yes
EBS Storage	yes	yes	yes
VPC*	yes	yes	yes
Enhanced Networking**	no	yes	yes

* Contact Intel support for information on using these features

** Enhanced Networking is only supported with C3 and C4 instances

Premier Support (8x5)

This product offering is our high-end option with support from the Lustre* experts at Intel. It includes IPSec over-the-wire encryption, EBS encryption, Enhanced Networking, and C3 and C4 compute optimized instances as well as M4.10xlarge instances. Enhanced Networking provides SRIOV, which allows a physical device to be virtualized and connected directly to a virtual machine. This provides lower latency and more consistent performance.

Self-Support

This product offering provides the same features as the Premier Support product, but without the included support from Intel. It includes IPSec over-the-wire encryption, EBS encryption, Enhanced Networking, and C3, C4, and M4.10xlarge instances. Also includes Enhanced Networking (SRIOV), and provides lower latency and consistent performance.

Evaluation

The "Evaluation" offering is an entry level product offered on Marketplace which can be used for proof-of-concept and development testing. Support is not included with this version, so we recommend moving to a product which includes "Premier Support" once you have evaluated the product.

Support

Product versions containing **Premier Support** are supported by the Lustre* experts at Intel®. Product support includes live 8x5 PST phone & email support as well as the latest software updates, patches, and fixes to ensure a stable, flexible, and robust storage environment that leverages the benefits of cloud-based infrastructure. The Evaluation and Self-Support versions of the Intel Cloud Edition for Lustre* software does not include Support. If you would like more information on this product or are interested in adding Support please contact hpdd-cloud-lustre@intel.com for more information.

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 shown in Figure 4 or the Intel web page shown in Figure 5

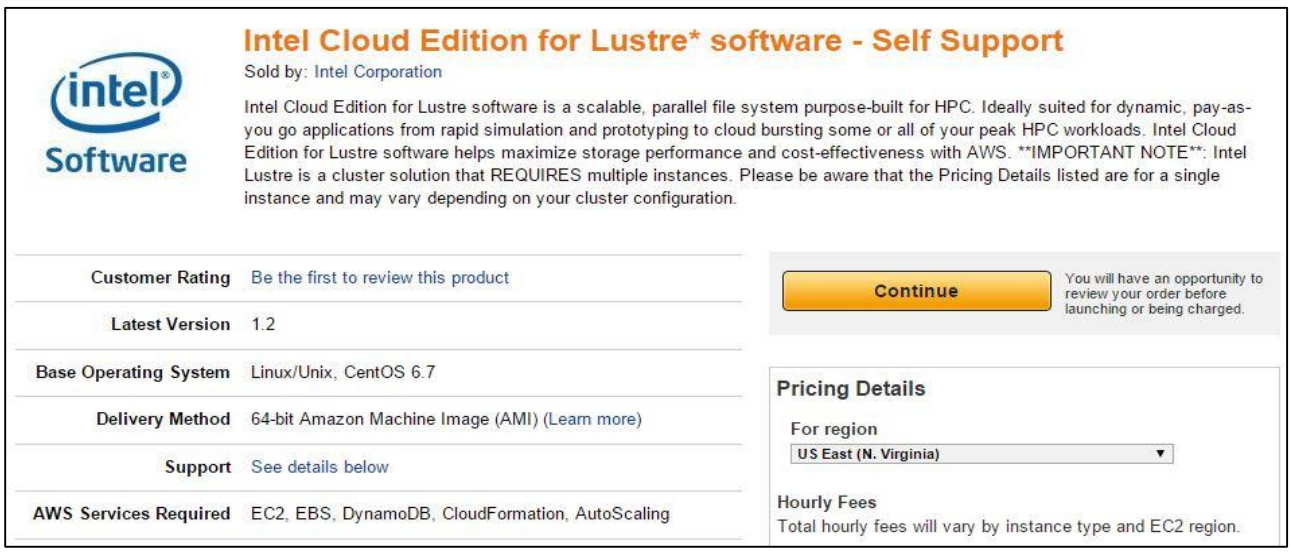


Figure 4: AWS Marketplace Page for Subscribing to a Product Version

Figure 5: Intel Web Page for Global (HVM) Version

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, as shown in Figure 6.

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.

Templates will require additional configuration parameters.

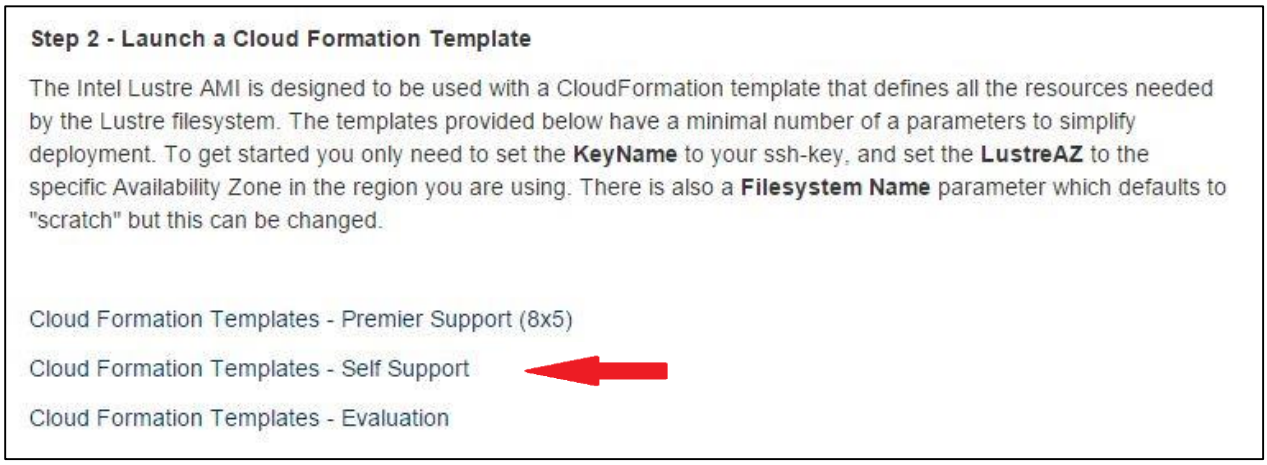


Figure 6a: Launch a Cloud Formation Template Screen

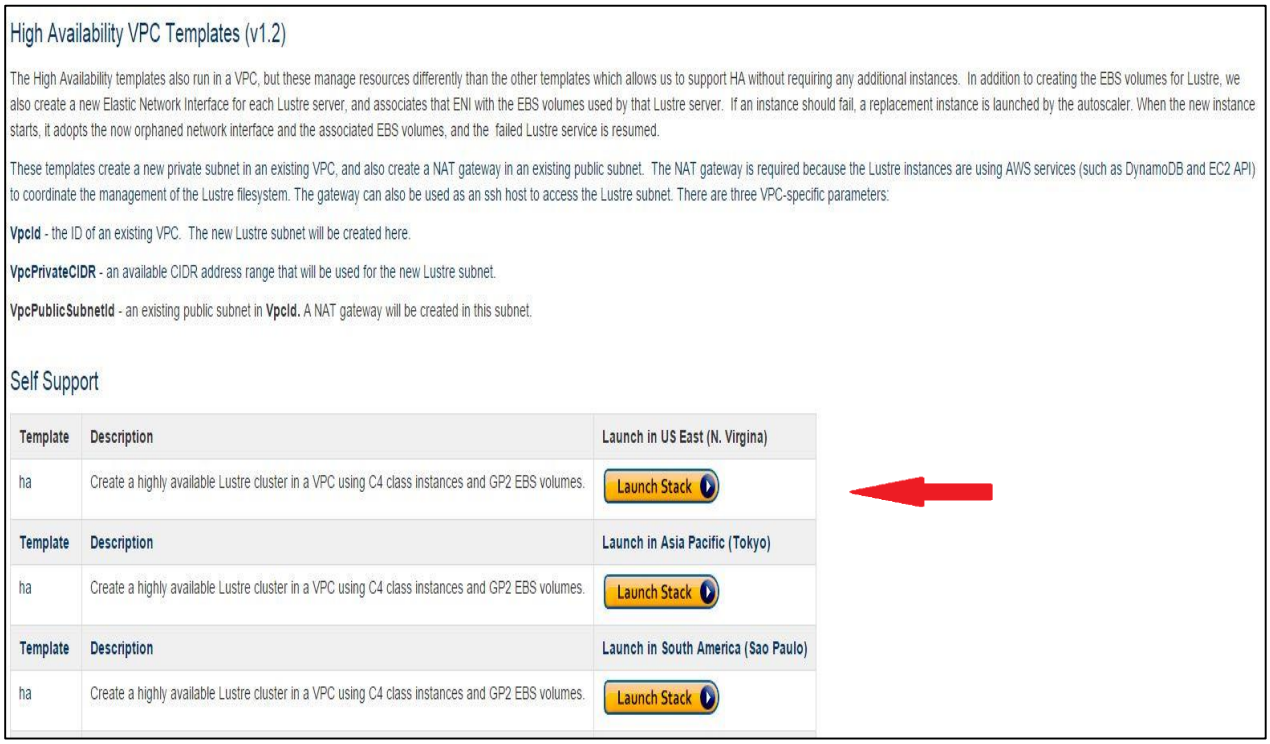


Figure 6b: Launch a Cloud Formation Template Screen

Step 3: Customize Your Cluster

You can open the template files. For example, in Figure 6b, in the Template column, you can click HA (used to deploy on the C4 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 7), and browse to your template location. Click **Next** to continue.

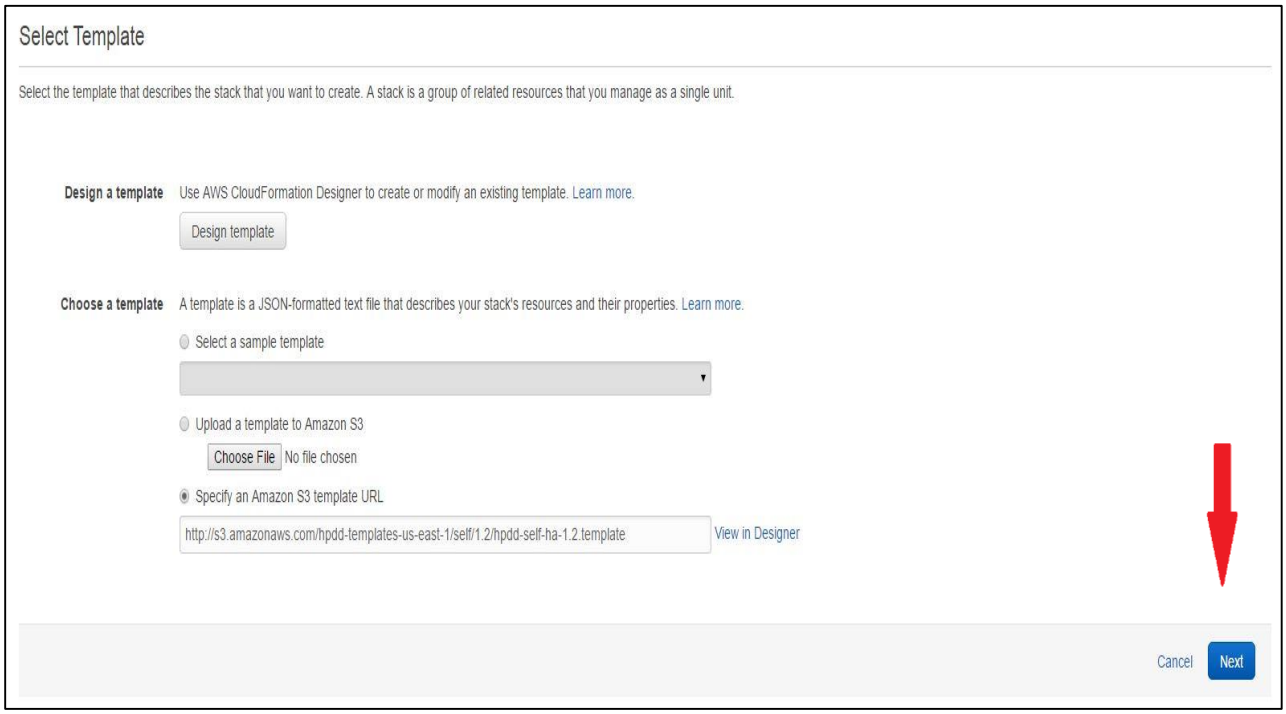


Figure 7: Select Template Screen

Figure 8 shows the parameters required to build a Lustre file system cluster using the templates available in the Self-Service version. AWS CloudFormation templates are stored on Amazon S3, and the path is filled in automatically when you press **Launch Stack** (shown in Figure 6b).

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 8. 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.)

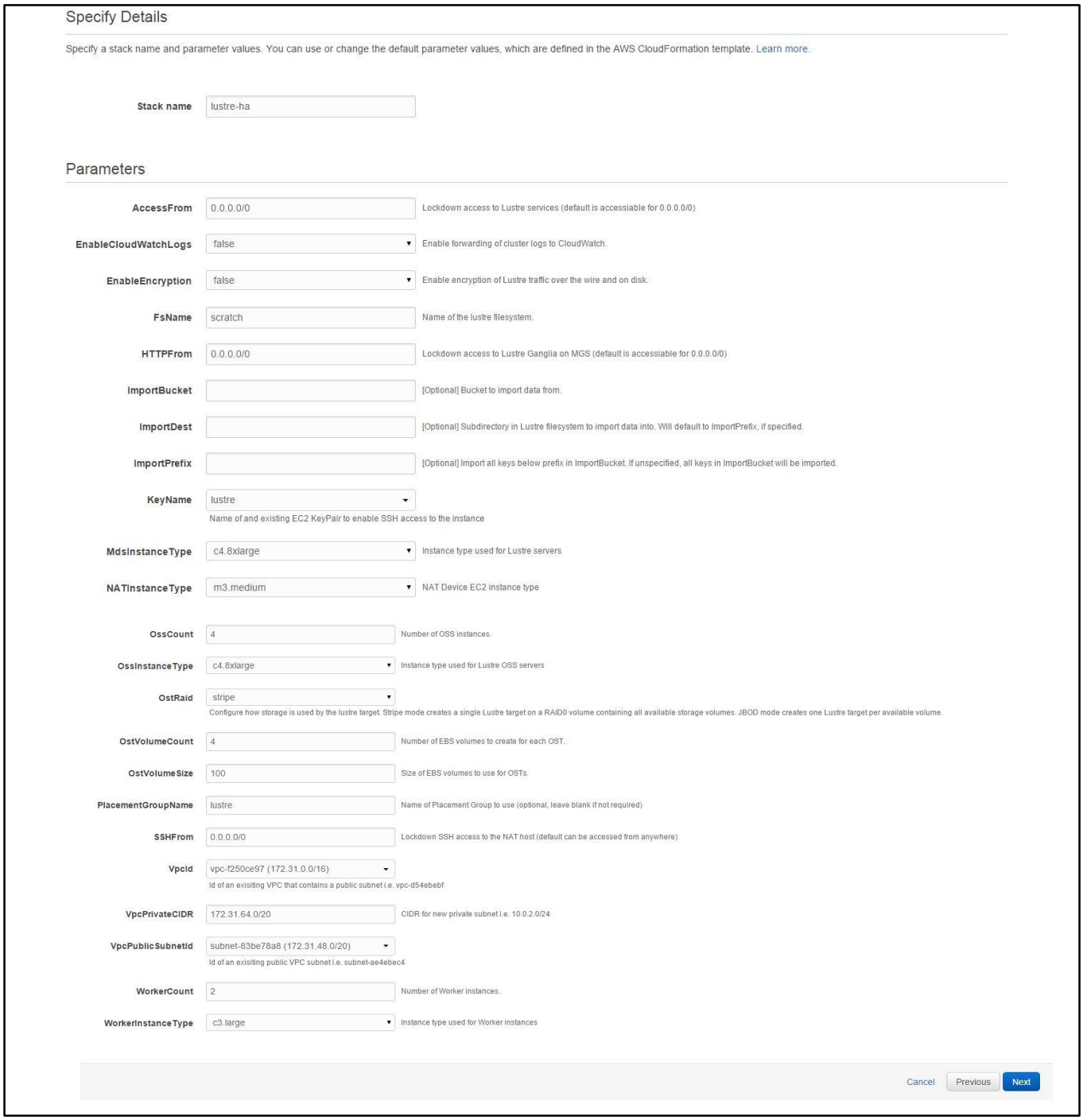


Figure 8: Parameters Screen

5: Launch the Instance

To launch the instance you will need to review and acknowledge the selections at the bottom of the screen, and then click **Create**. These steps are not shown. The AWS CloudFormation stack process will begin. You can use the AWS CloudFormation console to check the creation status, as shown in Figure 9.

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.

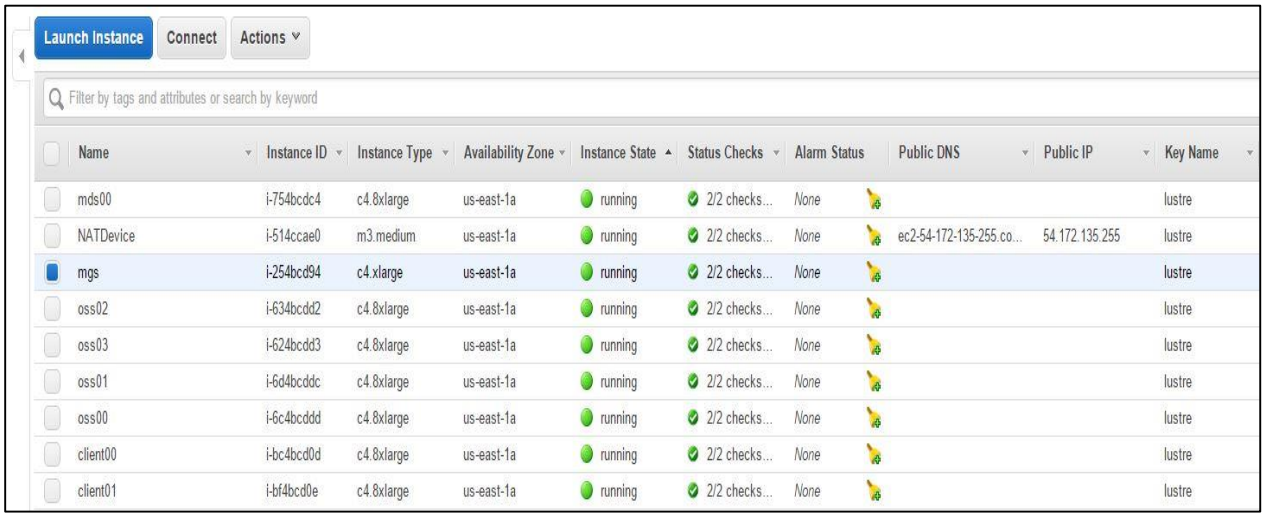


Figure 9: AWS CloudFormation Console

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 (AMI).

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:

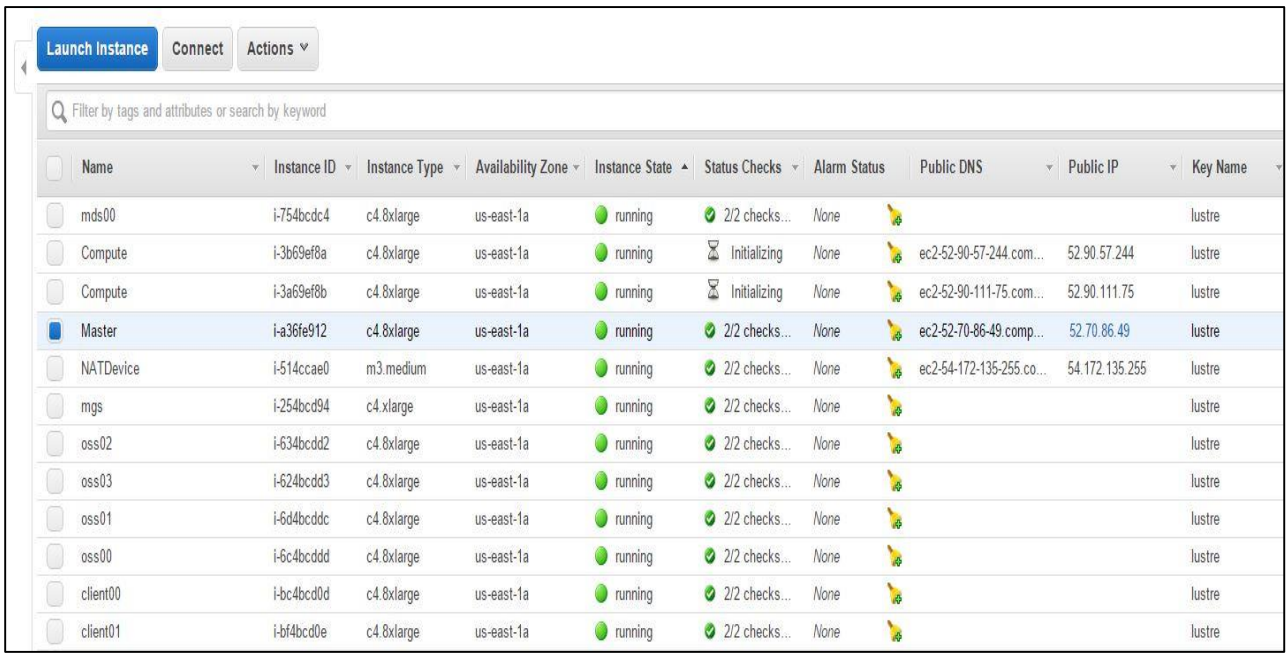
```
[aws]
# This is the AWS credentials section (required).
# These settings apply to all clusters
# replace these with your AWS keys
# If not defined, boto will attempt to use a) environment
# or b) EC2 IAM role.
aws_access_key_id = ''enter your key''
aws_secret_access_key = ''enter your key''

[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 Elastic Compute Cloud (Amazon EC2) resources will be running and billing will have begun.



Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name
mds00	i-754bcdc4	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre
Compute	i-3b69ef8a	c4.8xlarge	us-east-1a	running	Initializing	None	ec2-52-90-57-244.com...	52.90.57.244	lustre
Compute	i-3a69ef8b	c4.8xlarge	us-east-1a	running	Initializing	None	ec2-52-90-111-75.com...	52.90.111.75	lustre
Master	i-a36fe912	c4.8xlarge	us-east-1a	running	2/2 checks...	None	ec2-52-70-86-49.comp...	52.70.86.49	lustre
NATDevice	i-514ccae0	m3.medium	us-east-1a	running	2/2 checks...	None	ec2-54-172-135-255.co...	54.172.135.255	lustre
mgs	i-254bcd94	c4.xlarge	us-east-1a	running	2/2 checks...	None			lustre
oss02	i-634bcdd2	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre
oss03	i-624bcdd3	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre
oss01	i-6d4bcddc	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre
oss00	i-6c4bcddd	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre
client00	i-bc4bcd0d	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre
client01	i-bf4bcd0e	c4.8xlarge	us-east-1a	running	2/2 checks...	None			lustre

Figure 10: 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 9 and Figure 10.

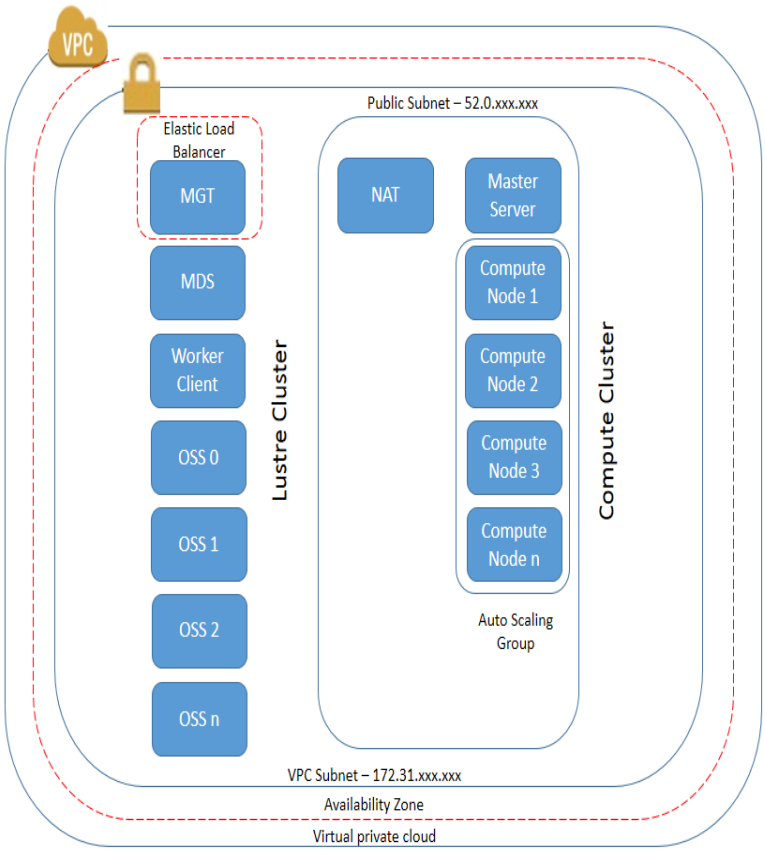


Figure 11: Resulting Client Cluster and Lustre Cluster Model

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 a baseline for I/O performance of the file system, we created an example Lustre file system using four c4.8x large instances as object storage servers and four c4.8x large client compute instances.

We used IOR, a parallel file system test developed by the Scalable/IO Project (SIOP) at LLNL (https://computing.llnl.gov/?set=code&page=sio). 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 144 threads across 2 nodes with xfersize of 1 MiB block size for each thread of 4 GiB and an aggregate file size of 576 GiB, which resulted in:

- Max Write: 2023.97 MiB/sec (2122.29 MB/sec)*
- Max Read: 1520.97 MiB/sec (1594.86 MB/sec)*

LMT or the Lustre Monitoring Tool (https://github.com/chaos/lmt) is installed with Intel Cloud Edition for Lustre. Ltop is a part of LMT and is a command line utility which gathers I/O statistics from Lustre filesystem servers. We used LTOP to record the filesystem activity during the IOR experiment:

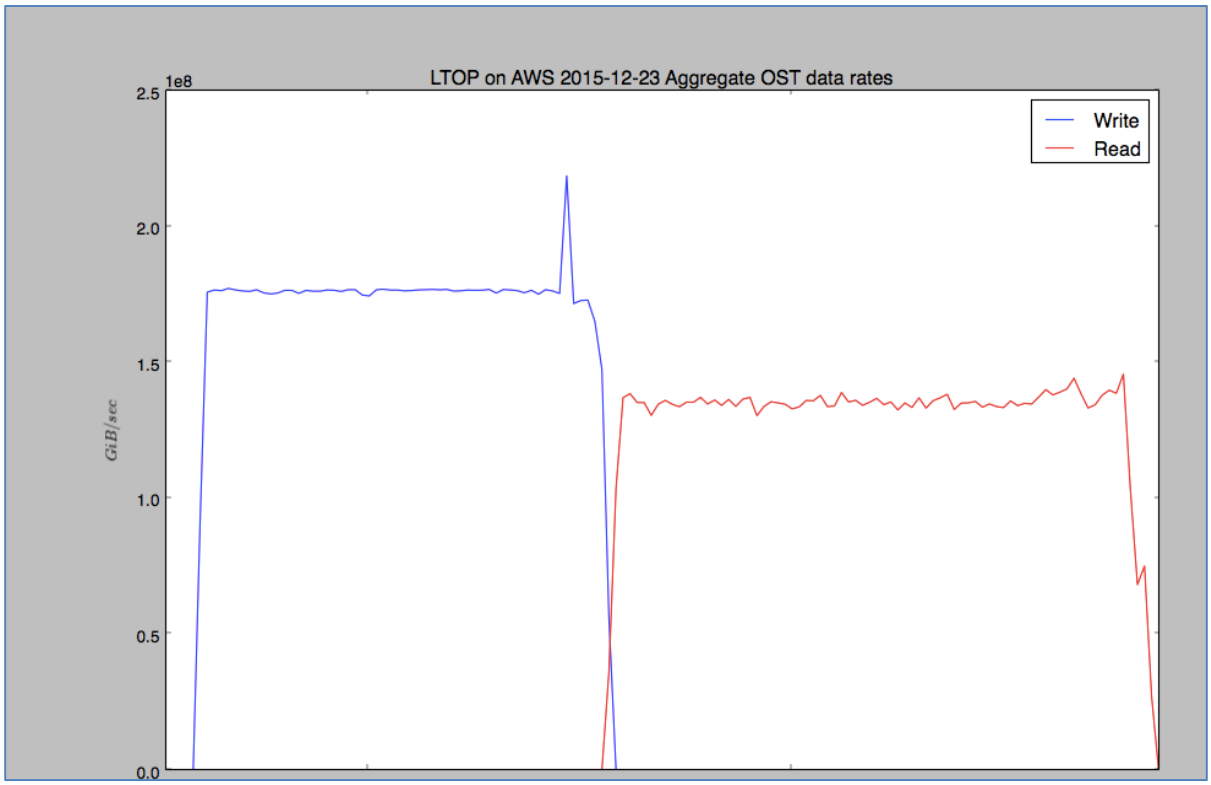


Figure 12: Lustre filesystem I/O performance

Figure 13 shows the testing results. The same parameters as used in Steps 3 and 4 were then used to show the effects of scaling the Lustre file system by increasing the number of Amazon EC2 object storage server instances.

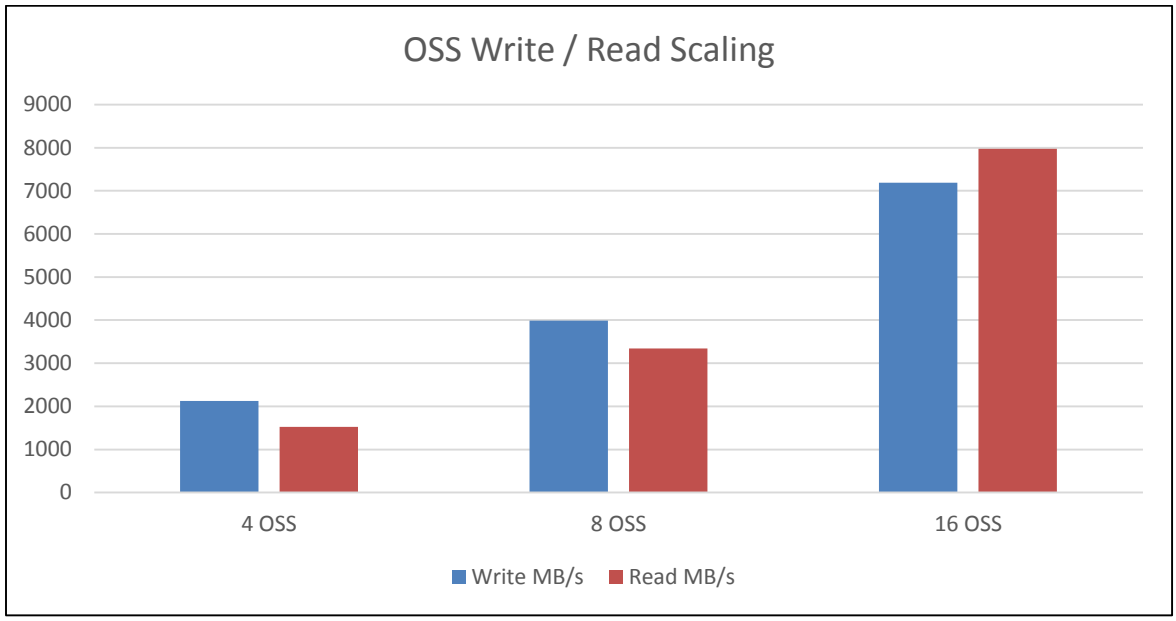


Figure 13: Testing Results

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

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

[1] www.top500.org

[2] 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