Lustre Deployment Guide using Pavilion Hyperparallel Flash Array

Version 2.0

Last Updated: October 2020
Document Version: Version 2.0

Legal Notice:

Technical Support:

Technical Support maintains support centre's globally. If you have questions regarding an existing support agreement, please email/phone the support agreement administration team as follows:

Phone:

USA & Canada: 1-888-342-0461

United Kingdom: 0800–69-8055

Netherlands: 0800-022-2832

International: 1-408-684-4958

Email:

support@pavilion.io

Documentation:
Make sure that you have the current version of the documentation. Each document displays the date of the last update on page 2.

Documentation Feedback:
Your feedback is important for us.
For documentation feedback send your comments to your Pavilion field support contact.

Table of Contents

1. About This Guide

The purpose of this document is to guide through the installation and configuration of **Lustre solution using Pavilion Hyperparallel Flash Array** and to describe the general set of configurations that have been validated.

The guide is intended for audience familiar with **Lustre File System.** This guide is not intended to serve as a comprehensive **Lustre Scale Guide**.

Note: For more detailed implementation assistance, please contact your **Pavilion** Support/ Sales representative

2. Key Recommendations

This section lists the key recommendations.

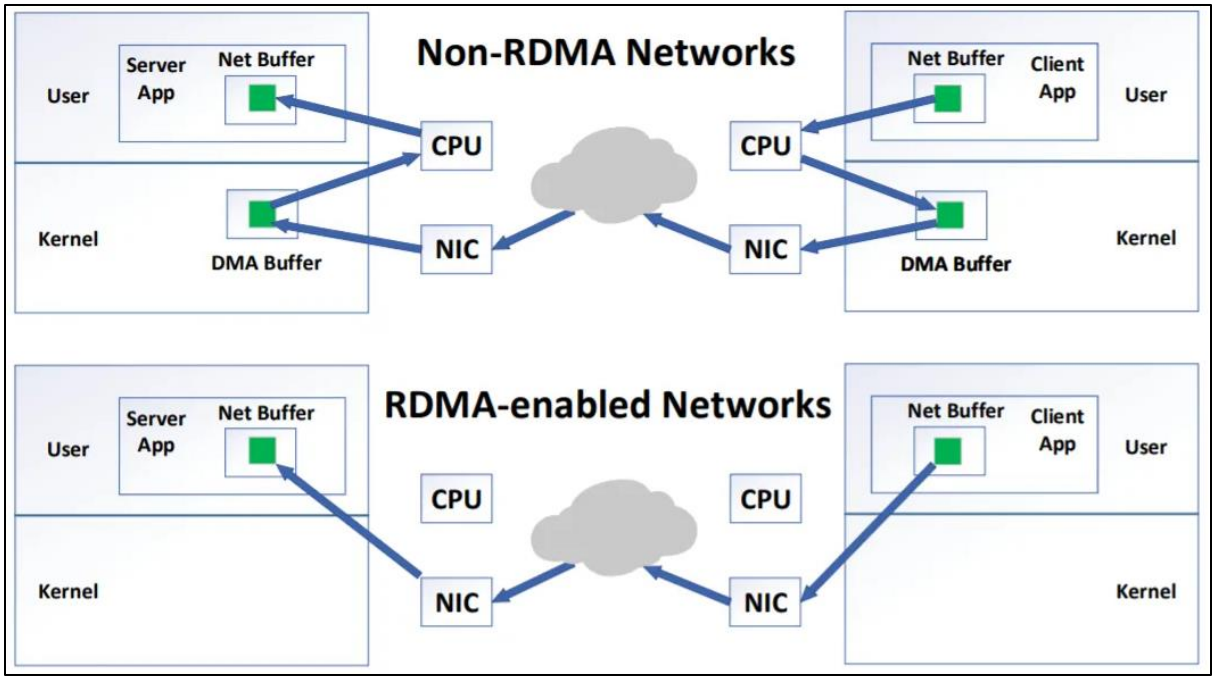
- Infiniband or RDMA capable Ethernet switch
- Ethernet
  - 100, 40, or 25 GbE Ethernet switch (100GbE for best performance), MTU 9216
  - Converged Ethernet NIC (Mellanox ConnectX-4 or newer) with MTU of 9000
- IB
  - EDR, FDR or FDR-10 IB switch (EDR for best performance)
  - Mellanox ConnectX-4 or newer
- RedHat or CentOS Linux 7.5 or newer for Lustre client & server nodes
- Lustre packages version 2.12.4 installed
- Use Linux in-box driver or Mellanox OFED (3.4.2 or newer)
- Configure Pavilion NVMe-oF controllers to use NVMe-oF RoCE or IB, via GUI or CLI
- Create volumes and assign to controllers on the **Pavilion** HFA
- Install the “nvme-cli” package on clients and use it to attach volumes
- Use volumes on the client as if they were local NVMe devices
- Make the **Pavilion** volumes persistent across reboots with the PDS RPM
- Build HA volume connections with Linux multipath, **Pavilions** primary and standby controllers, and network infrastructure without any SPOFs
- Dedicated network/subnets between
  - Lustre server nodes and clients
  - Lustre server nodes and Pavilion IO controllers

3. NVMe and NVMe over Fabrics

Non-Volatile Memory Express (NVMe) is an advanced protocol used to access flash storage on PCI Express bus. “Non-Volatile” stands for persistent storage, while “Express” refers to the fact that the data travels over the PCI Express (PCIe) interface on the computer's motherboard. This gives the drive a direct connection with the CPU and memory and eliminates many latency-inducing layers of traditional storage stacks such as SAS or SATA controllers. All modern servers and operating systems support NVMe out-of-the-box, and for enterprise and cloud scale Solid State Drives (SSDs) it is the interface of choice.

Because NVMe removes the intermediate legacy storage controllers and storage stack from the OS, it can provide significantly reduced latency (on the order of tens of microseconds per 4K I/O on enterprise NVMe SSDs). Thanks to its use of the PCI Express interface, it is also able to support individual drive bandwidths of up to 4 GB/s per individual NVMe SSD. NVMe over Fabrics (**NVMe-oF**) is an industry standard extension of the NVMe protocol which allows remote NVMe storage to be attached to servers. This is like how legacy Fibre Channel and iSCSI protocols enabled servers to utilize disks in SAN arrays for data storage, but with massively higher bandwidth and an order of magnitude lower latency. Modern Linux operating systems support it out-of-the-box today, with work ongoing for Microsoft Windows and VMWare cloud operating systems.

The trick that enables **NVMe-oF**, NVMe remote storage, to provide as good or better performance than local direct-attached (DAS) NVMe storage is something called Remote DMA (RDMA). DMA is used in NVMe to allow the SSD to directly load or store data to server memory, without CPU intervention (just like 0-copy accelerated TCP on high-performance NICs). RDMA allows for this same kind of direct memory access, but from outside of the server itself. An RDMA enabled storage array like the **Pavilion** HFA can handle I/O requests without the server’s CPU needing to copy any data whatsoever.
Refer to the below image for reference:



RDMA uses the same link layer but is a separate protocol with its own requirements separate from the more common IP with UDP or TCP on top. It is supported on the two standard networking infrastructures deployed today: InfiniBand and Ethernet.

4. Pavilion Lustre Solution

4.1 Introduction

The solution presented in this document highlights the simplicity and benefits of deploying Lustre with **Pavilion** HFA. This document provides detailed insights into the configuration, performance parameters, and scalability of the solution. It also intends to provide an overview of both **Lustre** and **Pavilion** getting deployed in such an environment and the associated best practices.

The **Pavilion Lustre solution** is designed to showcase the **Pavilion** compatibility with performance for **Lustre File System**. The flexibility in **Pavilion** HFA architecture enables the storage to be provided across multiple zones. 20 x IO Controllers in the **Pavilion** HFA ensures that each volume gets a dedicated controller to drive performance.

This section describes the solution in detail providing insights into the components used. This also details the network connectivity and establishes the communication paths among Lustre server nodes to storage targets and Lustre server nodes and Lustre clients.

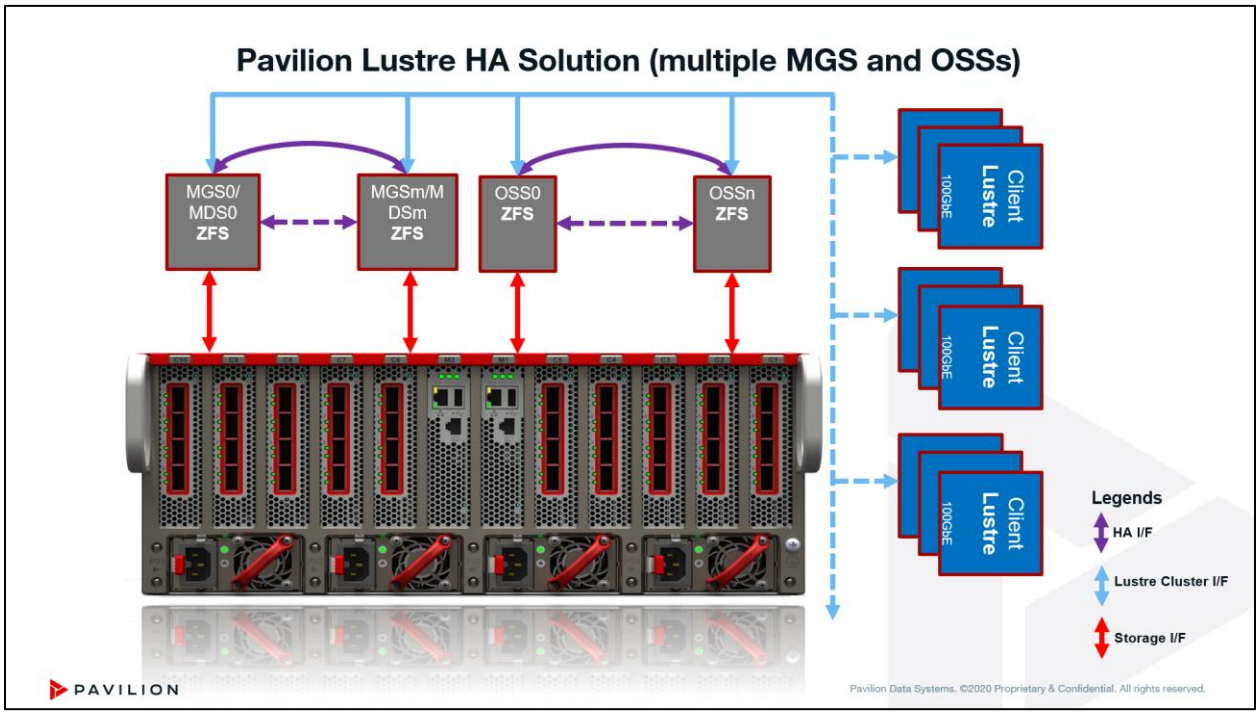


Image: Generic Pavilion HFA Lustre Solution Block Diagram

4.2 Architecture

Lustre is an open source distributed file system. It has a client server architecture and can cater a large number of clients as the solution grows over time. Lustre has a scalable architecture and can be separated into mutually exclusive components each responsible for a specific function. The components can be scaled up separately as the solution grows over time and becomes more demanding.

In general, Lustre architecture comprise of the following components and they are described in detail in the upcoming sections:

- Metadata Servers
- Data Servers
- Network
- Clients

Following **Pavilion Lustre HA Solution** diagram gives a detailed view of the solution being presented and elaborates on the components listed in **Image: Pavilion HFA Lustre Generic Solution** and describes the topology of the solution presented.

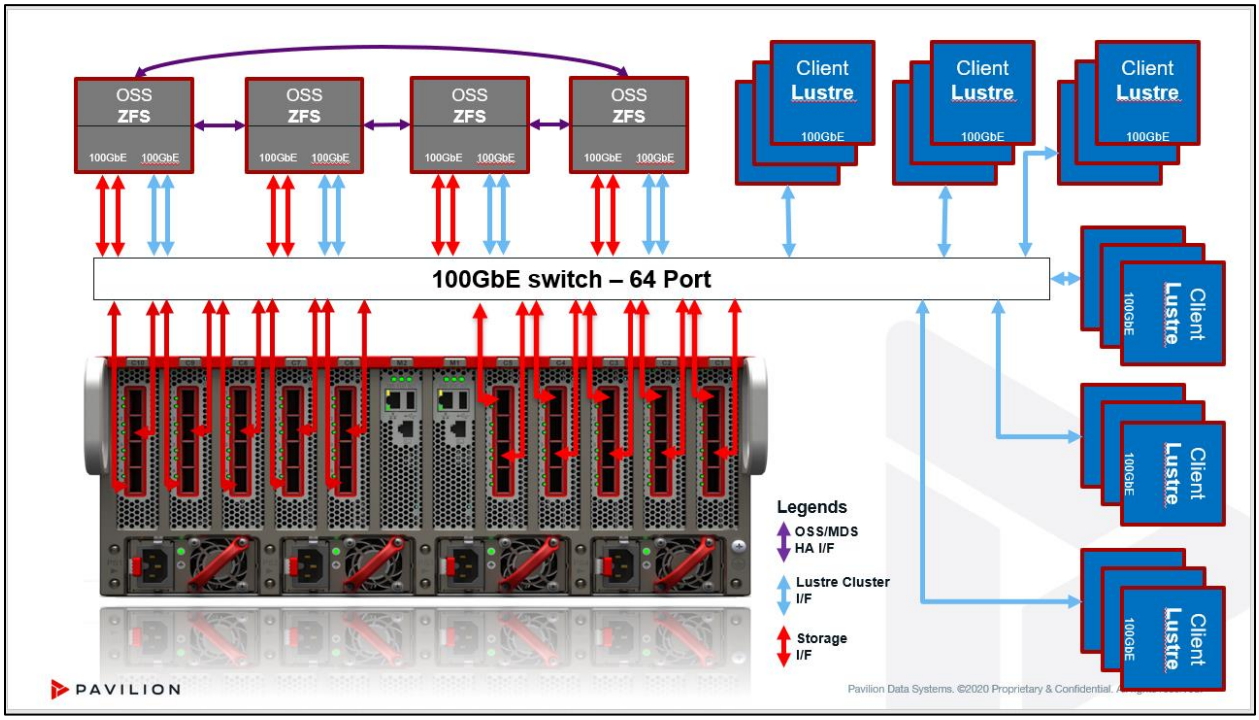


Image: Pavilion Lustre HA Solution

The **Pavilion** HFA with 20 x IO Controllers each having one port connected to 100 GbE or EDR network and accessible to the MDS/OSS Lustre server nodes as marked by **Red links.**

In turn, the MDS/OSS nodes are connected to the 100GbE or EDR network. Lustre clients can access Lustre server nodes and their services via separate network interfaces as represented with **Blue links**.

4.3 Components

Any Lustre solution is composed of following major components:

- Lustre Nodes - MDS/OSS
- Storage targets - MDT/OST
- Network (Ethernet or InfiniBand based)
- Clients
- Software (File System and cluster management utilities.)

The solution can be visualized as depicted in the following image:
(reference https://lustre.org/about/):

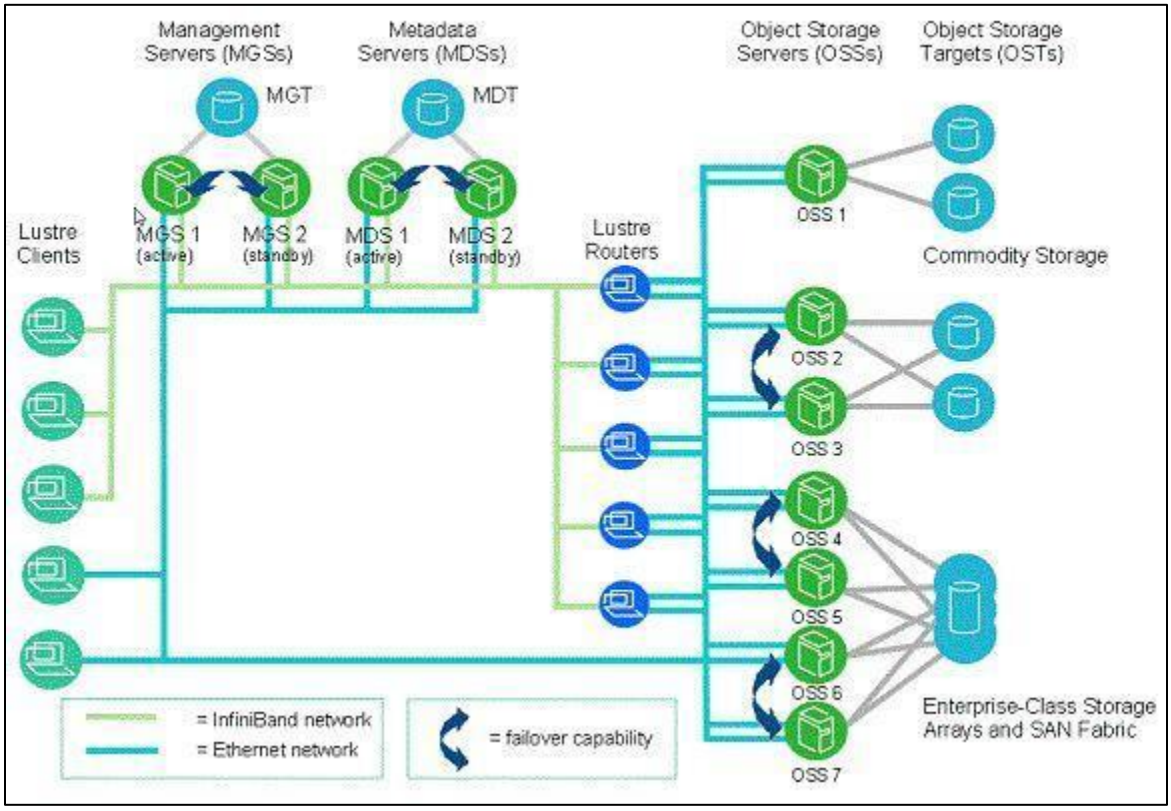


Image: Generic Lustre HA solution Architecture

4.4 Lustre server nodes

MGS, **MDS**, and **OSS** form the basic building blocks of any **Lustre solution architecture**. Each has a specific role in the architecture, and each performs a unique function in the Lustre configuration.

An MDS is a Metadata server that is responsible for storing all the metadata and recording all file creation, deletion operations along with information about the files such as access permissions, striping of the file, location of the file etc.

The MDS needs dedicated storage for storing all the metadata information, this storage is made available to the MDS via one or more MDT or Metadata targets. This storage requirement is typically not exceptionally large and for starting off with the Lustre solution, can be satisfied with a single volume. Following is the calculation of the MDT size:

```
Total Object Storage space: 10PB; Average file size: 10MB
Total number of inodes: 10PB/10MB = 10^9 inodes
Lustre takes ~2KB for 1 inode
Total space required for inodes = (2KB* 10^9 *2)/2^30=3.8TB
```

As the File System grows, and more clients start accessing the files, there might be a need to add more MDT, which essentially is another **Pavilion** volume. The additional target for an MDS can be created and attached online.

There can be multiple MDS servers as well in any given Lustre configuration, however, for the purposes of this solution we present a single MDT to a single instance of MDS.

The MDT is backed by ZFS. The **ZFS pool** is created on top of a **Pavilion** block volume which is connected to the MDS server with NVMe over RoCE and is available for use with dm-mapper block device for with multipath. This solution uses the ZFS and backfs in the Lustre File System keeping in mind the scalability and distribution of IO offered by zfs against ldiskfs.

An MDS provides all the necessary metadata information to the clients wanting to access the files on the Lustre, but the MDS does not participate in the actual data access. So, IO workload is very minimal for MDT.

Like the MDS, OSS i.e. Object Storage Server, has a scale out design, which means there can be one or more OSSs in any Lustre solution configuration.

The OSS is responsible for accessing storage for actual data and serving that to clients would access. The metadata for the user files is stored on the MDS while the actual files and data are stored on the OSS. OSS provides this storage in terms of Object storage targets or OSTs. An OST can be any storage target such as a Zpool, LVM or directory etc. which is to be used as the datastore.

In the presented solution there are 4 OSSs that connect to the **Pavilion** HFA NVMe-oF Target via NVMe over RoCE over 100GbE links. These imported NVMe-oF drives act as the OSTs and are exported to applications as dm-multipath devices.

The OSS servers have multiple NIC cards with separate subnets to provide exclusive connectivity to Lustre clients and **Pavilion** target storage. The Mellanox adapters support RoCE protocol and all the storage to the Lustre server nodes is provided by **Pavilion** over NVMe over RoCE protocol.

The **MDS and OSS** in the solution have a total of 4 network interfaces; 2 dedicated for Lustre cluster connections, and 2 dedicated for storage connections.

Four Penguin servers were used as Lustre server nodes for this solution with following details:

- `Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz (Dual sockets, 72 cores), 192GiB DDR4 Memory 2666MHz`

Out of the 4 servers used, 1 server served 3 roles of MGS, MDS as well as OSS, while the other 3 servers acted as dedicated OSS nodes.

The volume connectivity for the Lustre Server nodes is tabulated in the following table:

Lustre Server Node w/ Roles	Pavilion Volumes
MGS/MDS/OSS1	1 x MDT with 4 X Block Devices - 100GB each 1 X OST with 5 X Block Devices - 3TB each
OSS2	1 X OST with 5 X Block Devices - 3TB each
OSS3	1 X OST with 5 X Block Devices - 3TB each
OSS4	1 X OST with 5 X Block Devices - 3TB each

The network connectivity for the Lustre Nodes was provided by the Mellanox NIC cards as detailed below:

- `Mellanox MT27800 Family [ConnectX-5/ConnectX-4/ConnectX-3] 100GbE ports`

Following is the Lustre version and kernel used in the Lustre cluster (including OSS and MGS):

- `Centos 7.7 kernel-3.10.0-1062.9.1.el7_lustre.x86_64`
- `Lustre Version: 2.12.4`

4.5 Storage Targets

The storage needed for any Lustre deployment is classified based on which type of OSD uses the storage.

For instance:

1. MDS needs a volume of comparatively small size to store all the metadata of the Lustre fs and can be backed by one or multiple MDTs.
2. OSS connects to large capacity volumes and these volumes provide the pool to store all the files.

In this solution, **Pavilion** storage is used as a backing for both MDT and OSTs. The volume configuration is described in the earlier figure.

The Volumes are block devices carved out from RAID-6 Media groups, which span across 18 drives in a Media group. Pavilion HFA has 4 such zones/Media group.

For this solution 20 x 3TB volumes were created for OSTs and 4 x 100GB volumes were created for MDT. Each zone on **Pavilion** HFA, therefore, had 5 x 3TB NVMe-oF devices for OST and 1 x 100GB NVMe-oF devices per zone for MDT.

The IO controllers are configured for NVMe over RoCE protocol to enable the volumes to be available over RoCE. Above volumes are assigned to an Active/Standby pair of controller ports, where each port in the pair comes from different IO Controllers for high availability (HA) of Volume through dm-multipath. This allows handing off the HA and performance benefits of NVMe over RoCE to the OSS.

Since the ZFS pool on each OSS spans across all the volumes, the incoming IO requests are distributed across the OSTs/volumes. Since each OST/volume is backed by a separate IO controller port, each OST gets a dedicated IO controller and in turn performance leading to higher throughput. The data protection used in this solution is RAID 6 inherently supported at **Pavilion** HFA.

4.6 Pavilion Hyperparallel Flash Array

Pavilion HFA provides NVMe-oF devices from RAID groups as the units of usable storage for Lustre known as MDTs and OSTs and that can be connected to the OSS nodes in the Lustre cluster. These volumes are highly available using multipath from **Pavilion HFA**, which in-turn provides fault tolerance across multiple controllers.

For this solution, following configuration has been used with 1 volume active on each controller. (Refer volume mapping).

- **Pavilion** HFA Config
  - 20 controllers
  - 2 RAID groups of 16+2 drives
- Volume Config
  - 20 x 3TB volumes as 4 OSTs
  - 4 x 100GB volume as 1 MDT

Image: Detailed View Of Pavilion HFA Configuration displayed below depicts the same:

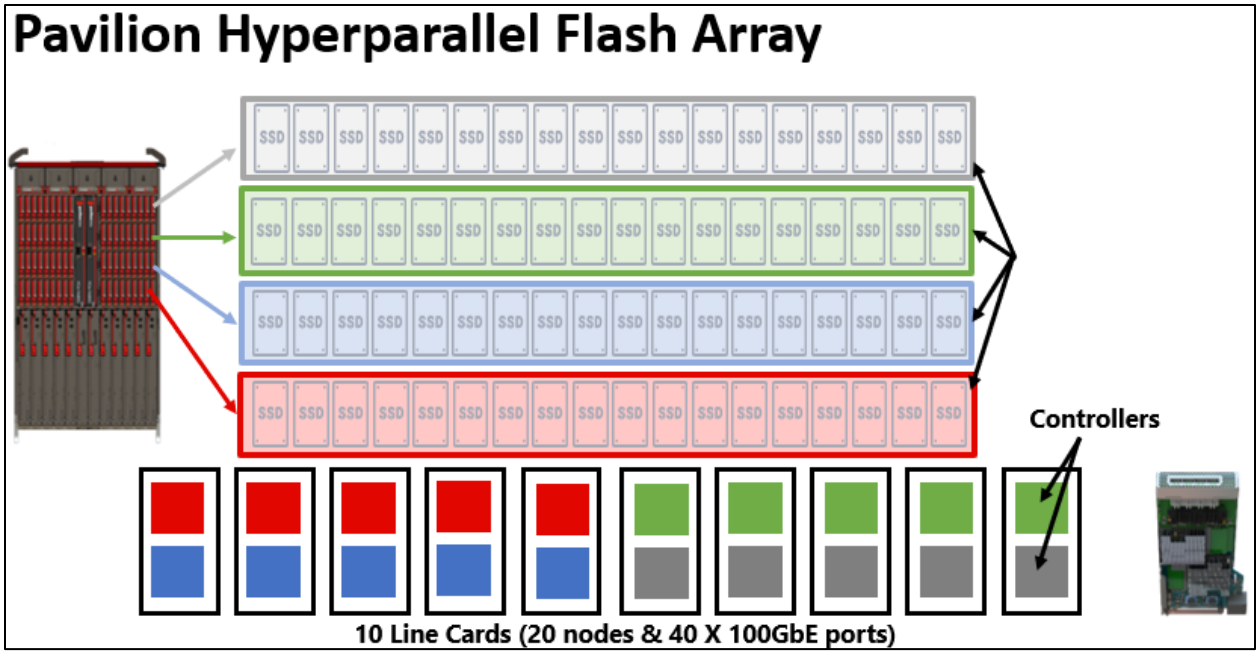


Image: Detailed View Of Pavilion HFA Configuration

4.7 Network Interconnects

Network plays a vital role in bringing together Lustre OSDs, Lustre clients and the storage. All Lustre OSSs have a network component installed called Lnet. Lnet is responsible for processing the data requests by forwarding them to the network drivers and by returning the requested data to the clients.

In this solution, the clients communicate with the OSS nodes via TCP. The clients need an Lnet interface to mount the Lustre File System. When a client requests data over the network the request goes to the MDS server, it checks the metadata and presents the OSTs across which the file is stripped, the client then communicates with the OSS directly without intervention of the MDS until the file operation completes. In the current solution entire communication happens over TCP via Lustre Network infrastructure referred as Lnet.

When the MDS directs Lustre clients to the OSTs where the file is striped; and then clients start communicating directly to the OSS nodes, that is when the network between OSS and Pavilion HFA comes into play.
In this solution, **Pavilion** volumes are assigned to IO Controllers and exported over RoCE protocol as described earlier. The IO Controllers have IP addresses for each of its

connected ports and the OSS nodes connect with the volumes using these IP addresses using `nvme-cli` utility.

Once the connection between the OSS and **Pavilion** IO Controller is established and ZFS volumes is up, then OSS upon request from a client, reads or writes data from underlying Pavilion NVMe-oF block devices.

The data read and write requests are sent over the underlying TCP/RoCE connection. For the Lustre solution to work the MGS, MDS and OSS servers need network connectivity which are as under:

1. OSS to OST connectivity was over RoCE (Connecting to **Pavilion** HFA IO Controllers)
2. OSS to Lustre clients the connectivity used is TCP
3. Management interface to Pavilion HFA, is regular ethernet over 1GbE NIC

In this solution both cluster network and storage connectivity were established using the following switch:

- Dell Z9264F-ON (64 ports)

4.8 Lustre Client Nodes

The Lustre clients can be any Linux server running the Lustre client software, the Lustre File System can also be made available to containers by mapping the Lustre mount point on a physical machine into the container.

Lustre clients mount the **Lustre File System** just like any other FS and the applications accessing the data can be agnostic of the underlying Lustre presence.

The clients connect to the MDS and OSS for accessing the data. The client to MDS connection is done when a file metadata being requested (i.e. opening and requesting metadata for a file for further operations) and MDS has to validate the metadata such as file permissions, file striping across OSTs/OSS, file size etc. Post initial connection client will be directed to the OSS directly until the end of the operation pertaining to requested file.

The clients need to configure Lnet interfaces before they can mount the Lustre file system. The Lnet interfaces thus configured to work on TCP protocol. The clients while requesting files or during actual IO operations route the requests via Lnet.

Lustre clients can be physical as well as virtual, for the purpose of this solution the following physical clients were used:

- 8 x Supermicro Servers - Lustre Clients
  - Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz (Dual sockets, 64 cores), 64GiB DDR4 Memory
- 12 x Supermicro servers - Lustre Clients
  - Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz (Dual sockets, 48 cores), 64GiB DDR4 Memory

Communication among Lustre cluster happens via IB/TCP protocol (for the purpose of this solution we used TCP). The clients can be virtual machines, containers, or physical machines. All clients need to have Lustre client packages installed and also need to have a network for Lustre created in order to mount the **Lustre File System**.

4.9 Lustre File System and Cluster Management Utilities

Following are the Lustre management utilities to manage a **Lustre File System** across nodes:

1. lfs (Lustre monitoring and configuration utilities)
2. lnetctl (Lustre LNet configuration management)
3. lctl (Low level Lustre File System configuration utility)
4. Mkfs.lustre
5. ZFS utilities

5. Configuring the Pavilion HFA for NVMe-oF

The **Pavilion HFA** can be configured with between 2 and 20 controllers. These controllers handle the connection between SSDs in the array (and all data services such as snapshots and clones) and the wider network and servers using that storage. Each one of these controllers can be configured to use a specific protocol, such as NFS, iSCSI, RoCE or IB. All exported volumes (LUNs) from any single controller will use the same protocol.

This section will describe how to configure the controller to use NVMe over RoCE/IB, a process which is often done only once per controller on installation. After configuring it to use the RoCE/IB protocol, standard volume creation and management can be performed without having to reset this setting. As such, these steps are often only run one time, on array installation.

Note: The **Pavilion** HFA fully supports a CLI, a web-based GUI, and a RESTful interface for management. For more details, it is recommended user refers to *Pavilion GUI Reference Guide, CLI Reference Guide and REST API Reference Guide.*

Note: CLI and GUI interactions will be demonstrated in the upcoming sections. Only one method needs to be used, of course, to configure the volumes and array. For one-off configurations, the GUI has better ease-of-use, but for larger deployments where automation is required, the CLI or RESTful interfaces are a better choice.

5.1 Configuring a controller to use the RoCE/IB protocol

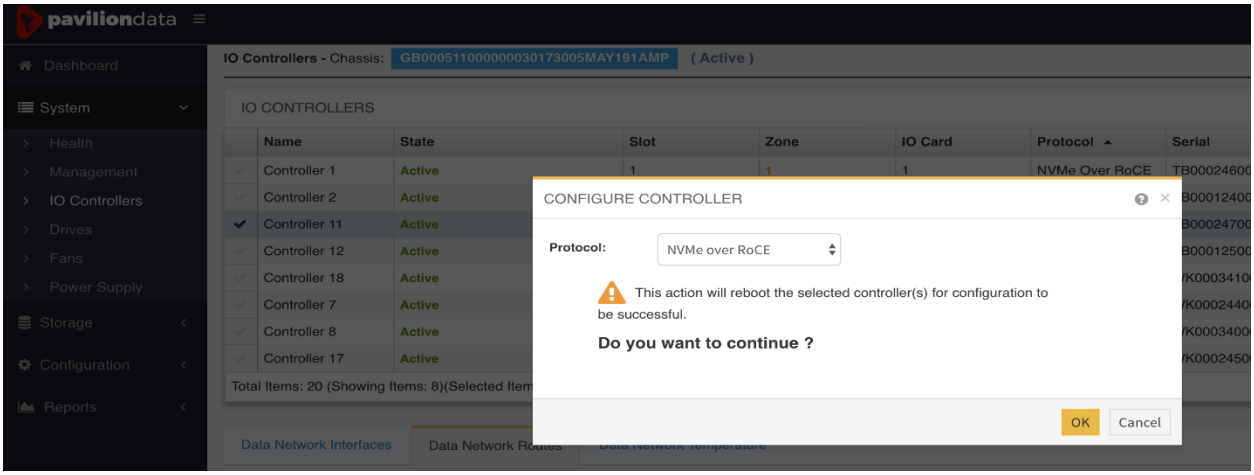
The following snippet shows the sequence of commands used in the CLI to enable the RoCE/IB protocol on a controller. Your controller numbering may be different, please refer to your purchased configuration. Note that if a controller is actively serving volumes, it cannot change its protocol. In that case, you will need to disconnect any volumes prior to configuring.

```
admin@GB0...AMP> switch config
Switched to config namespace

(config) configure controller id 11 protocol roce

OR
```

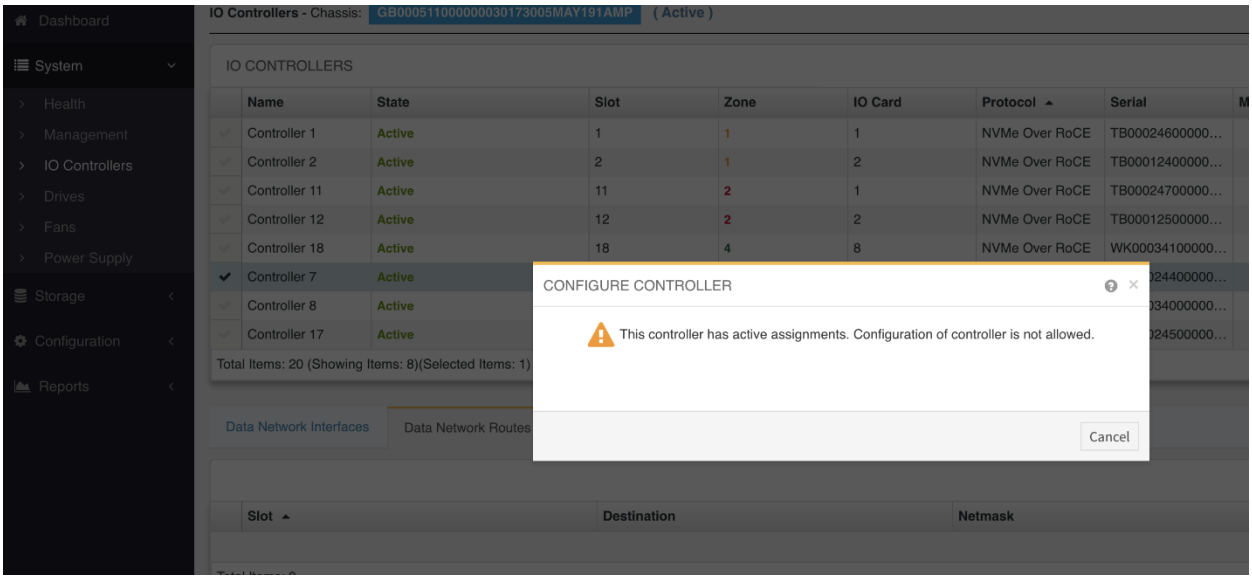
```
(config) configure controller id 11 protocol ib

  Creation of task was successful
  Monitor task : 2d04159f-a094-4e86-93de-cd935e240f93
  Command format : show task id [id]
```

Note: The **Pavilion** HFA utilizes an asynchronous queue for configuration changes. When a long-running configuration change is requested, a background process (“task”) is created to perform the request and control returned to the user immediately. The “Creation of task was successful” messages indicate that the command is ongoing in the background. See the CLI User's’ Guide for more in-depth information.

Using **Pavilion** GUI, simply log in to the array and click on the System->I/O Controllers side menu. Check the desired controller, and then use the “Configure Protocol” button and change the protocol to “NVMe over RoCE” or “NVMe over IB”:



Press “OK” to configure protocol, which will internally reboot the controller. If there are volumes assigned to the controller, the following error message will be displayed. You will need to remove the volume assignments, in this case, to change the protocol.



5.2 Creating and Assigning Volumes on the Pavilion HFA

Once the one-time configuration of controllers is done, volumes may be created and then assigned to RoCE/IB interfaces. This process is similar to standard storage management processes, with several additions. Each volume to be carved from an available “media group” which is equivalent to a typical “drive shelf” or “LUN group.” Media groups are typically configured by the factory during the installation process, but if necessary, more info is available in the full User’s Guide. Controllers have multiple network interfaces whose configuration should also have been performed (typical IP and netmask) on system installation. Again, more detailed interface configuration is available in the User’s Guide.

Volume workflow is a simple two-step process which can be completed from the CLI, the GUI, or the RESTful interface:

- Create volume of desired size, name, and options
- Assign newly created volume to a controller to enable access by clients

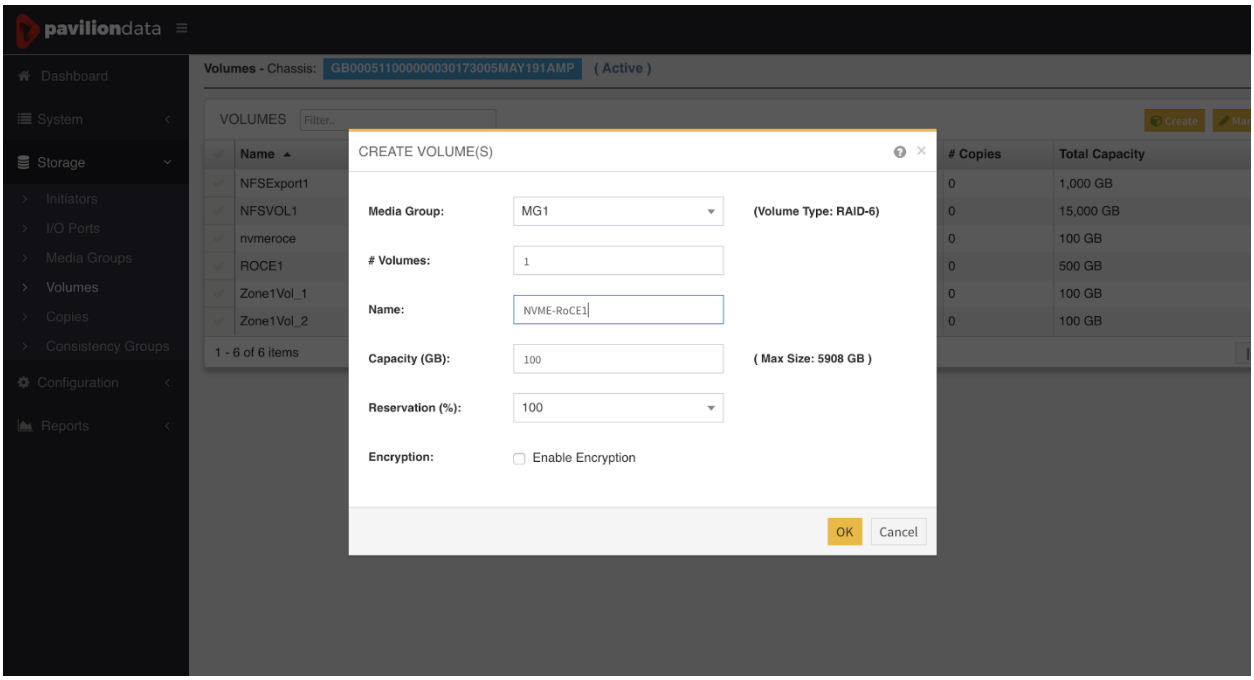
Creating a volume

Create a volume using the single CLI command:

```
admin@GB0...AMP> switch config
Switched to config namespace
```

```
(config) create volume name <VOLNAME> size <SIZE> GB grpname
<ZONE>

  Creation of task was successful
  Monitor task : ac63c2a2-a1ad-4930-aa18-510b5859bcf3
  Command format : show task id [id]
```

Using the **Pavilion** GUI to create an array is also simple. Use the left-hand menu Storage->Volumes, then the “Create” and fill out the form presented.



Assigning volume to a controller and RoCE interface

Once a volume is created (a process which may take several seconds while it is initialized), it is available for connecting to a specific controller and interface. Unless this step is performed, the volume will not be accessible to external clients.

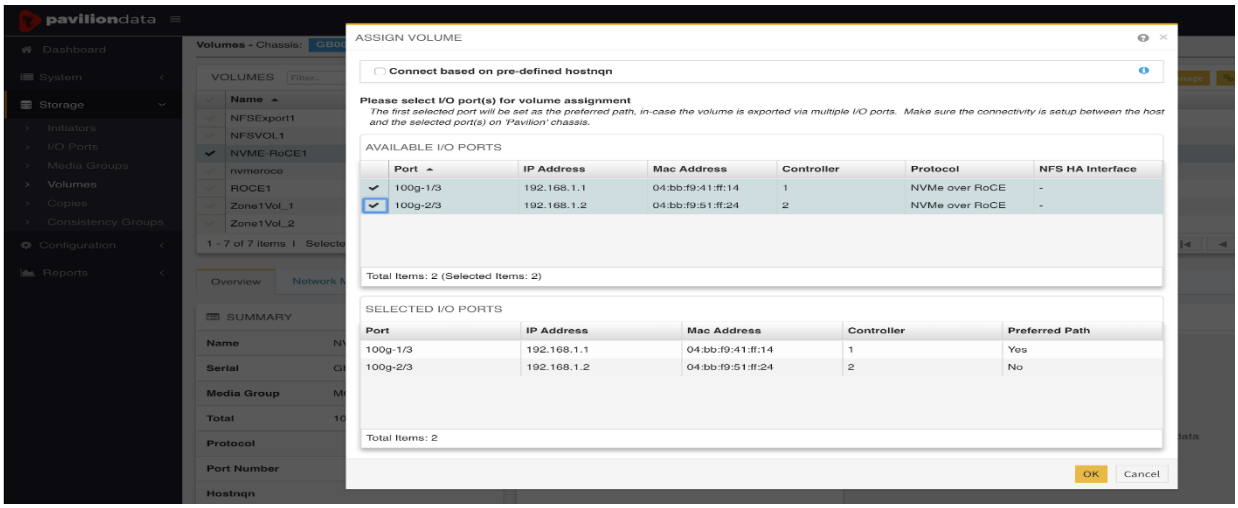
To assign a volume via the CLI, the volume name (given previously) and the interface “port-name” (of the form 100g-<port>/<zone> like “100g-1/3"). Make a note of the “Device Serial” as it will be used by the clients to connect to it later:

```
admin@GB0...AMP> switch config
Switched to config namespace
```

```
(config) assign volume name <VOLNAME> port-name <PORTNAME>
preferred

  Creation of task was successful
  Monitor task : 7bc94647-118d-46fc-a3ed-467d21efcb64
  Command format : show task id [id]

Block Device: dms12
Device Serial: GB00051104bbf912
```

Note: Make a note of the “Device Serial” as it will be required when connecting clients to this newly generated NVMe-over-RoCE volume. The “Device Name” shown is only for internal array use.

The same process can be done more simply using the GUI. Select the left-hand menu Storage->Volumes, then click on the configured volume and press the “Assign” button to bring up the assignment dialog. In this case, the “NVMe-oF Device Serial Number” will be presented in the updated GUI list after connection:



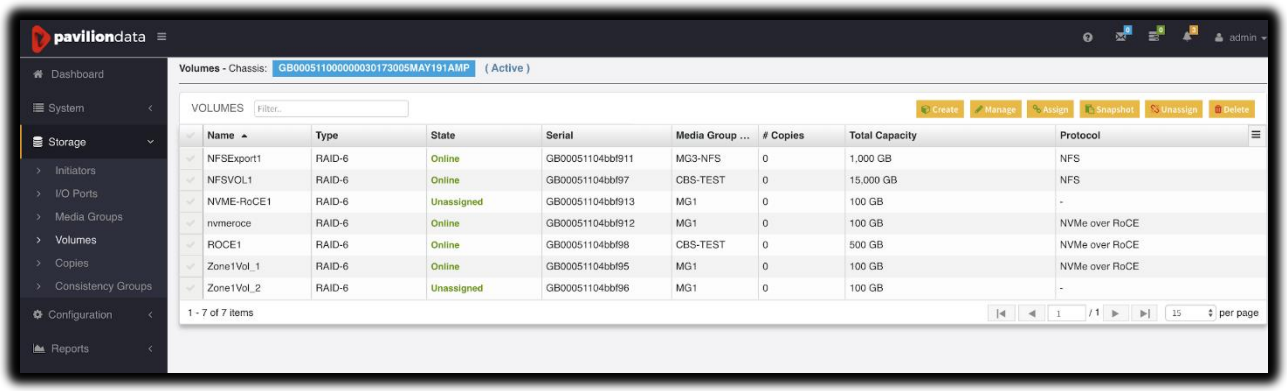
Verifying volume assignment

Once configured on the Pavilion HFA, volume status can be checked from both CLI and GUI. The CLI uses the following command (once in the “`show`” namespace):

```
admin@GB0...AMP> switch show
```

```
Switched to show namespace

(show) show volume-mapped-network name nvmeroce

Volume Mapped Network list
| Mac Address  | ... | IP Address   | Slot     | Controller | ...
+==============+=...=+==============+==========+============+=...
| 04:bb:f9:... | ... | 192.168.1.1  | 100g-1/3 |          1 | ...
+--------------+-...-+--------------+----------+------------+-...
```

In the GUI, simply navigate using the left-hand menu to Storage->Volumes:



6. Configuring Lustre server nodes

6.1 Setting up Local Lustre Repository

This step is especially important as Open Source Luster source is in continuous development and default latest stable Lustre build binary location keep changing for stable build. Select any node where we can create the local repository (current solution used se-lustre-repo). Follow the steps mentioned here:

Step 1: Create temporary **repo config** file to create repo on **se-lustre-repo,** see below sample:

```
se-lustre-repo# cat >/tmp/lustre-repo.conf <<\__EOF
[lustre-server]
name=lustre-server
baseurl=https://downloads.whamcloud.com/public/lustre/lustre-
2.12.4/el7/server
# exclude=*debuginfo*
gpgcheck=0

[lustre-client]
name=lustre-client
baseurl=https://downloads.whamcloud.com/public/lustre/lustre-
2.12.4/el7/client
# exclude=*debuginfo*
gpgcheck=0

[lustre-lsdisk-server]
name=lustre-ldiskfs-server
baseurl=https://downloads.whamcloud.com/public/lustre/lustre-
2.12.4/el7/patchless-ldiskfs-server/
# exclude=*debuginfo*
gpgcheck=0
[e2fsprogs-wc]
name=e2fsprogs-wc
baseurl=https://downloads.whamcloud.com/public/e2fsprogs/latest/e
l7
# exclude=*debuginfo*
```

```
gpgcheck=0
__EOF
```

Step 2: Create Directory for Lustre repo and sync the repo to created directory

```
se-lustre-repo# mkdir -p /var/www/html/repo
se-lustre-repo# cd /var/www/html/repo
se-lustre-repo# reposync -c /tmp/lustre-repo.conf -n -r lustre-
server patchless-ldiskfs-server lustre-client e2fsprogs-wc
```

Step 3: Create repo metadata for each folder fetched from Lustre online repository

```
se-lustre-repo# cd /var/www/html/repo
se-lustre-repo# for i in e2fsprogs-wc lustre-client lustre-server
patchless-ldiskfs-server; do
(cd $i && createrepo .)
done
```

Step 4: Apply configuration changes to get the repository folder accessible over network. Additional packages to be installed on all Lustre cluster nodes (including MGSs, MDSs, OSSs and Clients)

```
bash-4.2# yum install -y asciidoc audit-libs-devel automake bc
binutils-devel bison device-mapper-devel elfutils-devel elfutils-
libelf-devel expect flex gcc gcc-c++ git glib2 glib2-devel
hmaccalc keyutils-libs-devel krb5-devel ksh libattr-devel
libblkid-devel libselinux-devel libtool libuuid-devel libyaml-
devel lsscsi make ncurses-devel net-snmp-devel net-tools newt-
devel numactl-devel parted patchutils pciutils-devel perl-
ExtUtils-Embed pesign python-devel redhat-rpm-config rpm-build
systemd-devel tcl tcl-devel tk tk-devel wget xmlto yum-utils
zlib-devel
```

6.2 Installing Lustre Using Local Repository

Add following lustre.repo file at yum.repo.d directory

```
bash-4.2# cat lustre.repo
[lustre-server]
name=lustre-server
baseurl=http://se-lustre-repo/repo/lustre-server
enabled=1
gpgcheck=0
#proxy=_none_

[lustre-lsdisk-server]
name=lustre-ldiskfs-server
baseurl=http://se-lustre-repo/repo/lustre-ldiskfs-server
enabled=1
gpgcheck=0

[lustre-client]
name=lustre-client
baseurl=http://se-lustre-repo/repo/lustre-client
enabled=1
gpgcheck=0

[e2fsprogs-wc]
name=e2fsprogs-wc
baseurl=http://se-lustre-repo/repo/e2fsprogs-wc
enabled=1
gpgcheck=0
```

6.3 Installing Lustre ZFS and LDISKFS on MDS/OSS

This section lists the steps required to install Lustre ZFA and LDISKFS on MDS/OSS.

Step 1: Install e2fsprogs:

```
bash-4.2# yum --nogpgcheck --disablerepo=* --
enablerepo=e2fsprogs-wc install -y e2fsprogs
```

Step 2: Install EPEL repository support:

```
bash-4.2# yum --nogpgcheck --disablerepo=* --
enablerepo=e2fsprogs-wc install -y e2fsprogs
```

Step 3: Install ZFS from ZFS repository depending on the RHEL distro
(select ZFS for RHEL distro at:
https://openzfs.github.io/openzfsdocs/Getting%20Started/RHEL%20and%20CentOS.html)

```
bash-4.2# yum install -y http://download.zfsonlinux.org/epel/zfs-
release.el7_7.noarch.rpm
```

Step 4: Install Lustre patched kernel from local repository of se-lustre-repo:

```
bash-4.2# yum --nogpgcheck --disablerepo=base,extras,updates --
enablerepo=lustre-server install -y kernel kernel-devel kernel-
headers kernel-tools kernel-tools-libs kernel-tools-libs-devel
```

Step 5: Generate host ID and persist it:

```
bash-4.2# hid=`[ -f /etc/hostid ] && od -An -tx /etc/hostid|sed
's/ //g'`
bash-4.2# [ "$hid" = `hostid` ] || genhostid
```

Step 6: Reboot system:

```
bash-4.2# init 6
```

Step 7: Server: After system comes up install LDISKFS and ZFS kmod packages for server nodes:

```
bash-4.2# yum --nogpgcheck --enablerepo=lustre-server install -y
kmod-lustre-osd-ldiskfs lustre-dkms lustre-osd-ldiskfs-mount
lustre-osd-zfs-mount lustre lustre-resource-agents zfs

bash-4.2# yum --nogpgcheck --enablerepo=lustre-client install
lustre-client-dkms lustre-client
```

Step 8: Setup Lnet interface:

```
bash-4.2# -bash-4.2# cat /etc/modprobe.d/pds_lnet.conf
options lnet networks="tcp2 (enp216s0f0), tcp1 (enp94s0f0)"
```

Step 9: Load ZFS and Lustre kernel modules:

```
bash-4.2# modprobe -v zfs  # issue this command on Lustre server
nodes only
bash-4.2# modprobe -v lustre
```

Step 10: To unload Lustre modules use following command:

```
bash-4.2# lustre_rmmod
```

7.Configuring NVMe-oF devices for MDTs and OSTs

7.1 Connecting to NVMe-oF MDT storage

For example, by running following nvme-cli commands NVMe-oF Devices can be connected:

```
bash-4.2# nvme discover -t rdma -a 192.168.1.11
Discovery Log Number of Records 2, Generation counter 20
=====Discovery Log Entry 0======
trtype:  rdma
adrfam:  ipv4
subtype: nvme subsystem
treq:    not specified
portid:  28
trsvcid: 4420
subnqn:  GB00040804bbf9149
traddr:  192.168.1.11
rdma_prtype: unrecognized
rdma_qptype: unrecognized
rdma_cms:    unrecognized
rdma_pkey: 0x0000
=====Discovery Log Entry 1======
trtype:  rdma
adrfam:  ipv4
subtype: nvme subsystem
treq:    not specified
portid:  28
trsvcid: 4420
subnqn:  GB00040804bbf9153
traddr:  192.168.1.11
rdma_prtype: unrecognized
rdma_qptype: unrecognized
rdma_cms:    unrecognized
rdma_pkey: 0x0000
-bash-4.2# nvme connect -t rdma -a 192.168.1.11 -n
GB00040804bbf9153
```

Verify the volumes are connected by running nvme list exposed as DM devices:

```
bash-4.2# nvme list
Node             SN                   Model
Namespace Usage                      Format           FW Rev
---------------- -------------------- --------------------------
------------ --------- ------------------------- -------------
-- --------
/dev/nvme0n1     960ea4e923d0347d     PVL-MX18S0P2L2C1-F100TP0TY1
1          536.87  GB / 536.87  GB      4 KiB +  0 B   2330
/dev/nvme1n1     266326518e80c554     PVL-MX18S0P2L2C1-F100TP0TY1
1          536.87  GB / 536.87  GB      4 KiB +  0 B   2330
```

7.2 Configuring Network for MDS/OSS

The MDS or OSS needs an Lnet interface configured to enable it to communicate with OSS and the client nodes. For example, MDS and OSS has 2 interfaces *enp94s0f0* and *enp216s0f0* configured with TCP; this can be seen on MDS and OSS using following CLI:

```
bash-4.2# lnetctl net show
net:
    - net type: lo
      local NI(s):
        - nid: 0@lo
          status: up
    - net type: tcp2
      local NI(s):
        - nid: 192.168.10.220@tcp2
          status: up
          interfaces:
              0: enp216s0f0
    - net type: tcp1
      local NI(s):
        - nid: 192.168.20.220@tcp1
          status: up
          interfaces:
              0: enp94s0f0
```

7.3 Configuring MDTs for MDS/MGS

The MDS needs to be configured to provide connectivity to the clients as well as for the access to Pavilion HFA NVMe-oF Devices. After successful connection of the NVMe volume, format the MSTs for MDS server for that follow steps as described under:

Step 1: After figuring out the DM devices to be used for MSTs (for e.g. dm-0 and dm-1); run following command to create ZFS pool for MST

```
bash-4.2# zpool create pvlmst /dev/dm-0
```

Step 2: Format MST with Lustre file system name “pavlstr”

```
bash-4.2# mkfs.lustre --fsname=pavlstr --backfstype=zfs --verbose
--reformat --index=0 --mgs --mdt pvlmst/mst0
```

Step 3: Mount MDT to be accessed by MDS server to serve clients’ metadata requests:

```
bash-4.2# mount pvlmst /lustre/pvlmst
```

7.4 Configuring OSTs for OSS

The OSS node needs to have all Pavilion NVMe-oF devices connected (refer section “Pavilion exported NVMe-oF devices for MDTs and OSTs” to connect NVMe-oF devices), which are designated as OSTs for current OSS node. For example, OSS0 has dm-2, dm3, dm-4 and dm-5 devices designated for OSTs.

Use following steps to create OSTs accessible to Lustre client:

Step 1: After figuring out the DM devices to be used for OSTs (for e.g. dm-0, dm-1, dm-2, dm-3 and dm-4); run following command to create ZFS pool for OST

```
bash-4.2# zpool create pvlost /dev/dm-0 /dev/dm-1 /dev/dm-2
/dev/dm-3 /dev/dm-4
```

Step 2: Format MST with Lustre file system without HA and add this OST to a Luster file system named “pavlstr”

```
bash-4.2# mkfs.lustre --fsname=pavlstr --backfstype=zfs --verbose
--reformat --index=1 --mkfsoptions "recordsize=1024K" --mgsnode
192.168.20.220@tcp1 --mgsnode 192.168.10.220@tcp2 --ost
pvlost/ost0
```

Step 3: Mount OST to be accessed by OSS and MDS server to serve clients' data requests:

```
bash-4.2# mount -t lustre pvlost/ost0 /lustre/ost0/
```

8. Configuring Lustre Clients

Clients should have the Lustre components pre-installed to qualify the servers as Lustre clients (refer section “Install Lustre Using Local Repository”).

Clients must have to have LNet interface configured as mentioned in the earlier sections.

Clients can simply mount the Lustre FS using the following command:

```
mount -t 192.168.20.220@tcp1:/lustre/pavlstr /mnt
```

9. Configuring Lustre for high availability

There can be following types of failover scenarios with Lustre:

- Lustre node maintenance
  - OSS failure
  - MGS failure
  - MDS failure
- Storage controller failure

9.1 Configuring High Availability for Lustre server nodes

Lustre has no single service like NFS to start or stop access to Lustre File System. The start of Lustre storage availability is synonymous to mount and the unavailability of Lustre storage corresponds to an unmount of the Lustre FS on a particular client node.

Since the Lustre availability is defined by the ability of a server to access and mount OST and MSTs, in order to make a Lustre File System highly available, it has to be accessible across various OSS and MDS identified as a part of the failover group. This means, the OSTs should be accessible across different OSS designated to act as failover servers for the specified OSS.

To make Lustre Highly available we need to create file system using following commands (currently the fail over hosts are selected on the round-robin basis):

Making MDS & MGS highly available

```
bash-4.2# mkfs.lustre --fsname=pavlstr --backfstype=zfs --verbose
--reformat --index=0 --mgsnode 192.168.20.220@tcp1 --
mgsnode=192.168.10.220@tcp2 --failnode 192.168.20.220@tcp1--
failnode 192.168.20.224@tcp1 --mgs --mdt pvlmst/mst0
```

Making OSS highly available

```
bash-4.2# mkfs.lustre --fsname=pavlstr --backfstype=zfs --verbose
--reformat --index=1 --mkfsoptions "recordsize=1024K" --mgsnode
192.168.20.220@tcp1 --mgsnode=192.168.10.220@tcp2 --
failnode=192.168.20.161@tcp1 --failnode=192.168.10.161@tcp2 --ost
pvlost/ost0
```

OSS, MDS & MGS failures can be dealt using the alternate server nodes (minimum 2 are required) mentioned while creating the MDT and OST.

9.2 Configuring High Availability for Storage target Controllers

This scenario will be taken care by DM-multipath path failure handling.

10. Configuring Lustre for high performance

In order to get the best performance from the Lustre File System, the following points should be considered:

- Independent and separate storage and client access

- Apply following network buffer modification commands on Lustre Server Nodes

```
bash-4.2# sysctl -w net.core.netdev_max_backlog=300000
bash-4.2# sysctl -w sunrpc.tcp_slot_table_entries=128
bash-4.2# sysctl -w net.ipv4.tcp_synack_retries=2
bash-4.2# sysctl -w net.ipv4.tcp_rfc1337=1
bash-4.2# sysctl -w net.ipv4.tcp_fin_timeout=15
bash-4.2# sysctl -w net.ipv4.tcp_keepalive_time=300
bash-4.2# sysctl -w net.ipv4.tcp_keepalive_probes=5
bash-4.2# sysctl -w net.ipv4.tcp_keepalive_intvl=15
bash-4.2# sysctl -w net.ipv4.tcp_low_latency=1
bash-4.2# sysctl -w net.ipv4.tcp_timestamps=1
bash-4.2# sysctl -w net.core.rmem_default=31457280
bash-4.2# sysctl -w net.core.rmem_max=12582912
bash-4.2# sysctl -w net.core.wmem_default=31457280
bash-4.2# sysctl -w net.core.wmem_max=12582912
bash-4.2# sysctl -w net.core.netdev_max_backlog=65536
bash-4.2# sysctl -w net.core.optmem_max=25165824
bash-4.2# sysctl -w net.core.somaxconn=65535
bash-4.2# sysctl -w net.ipv4.tcp_mem='8388608 8388608 8388608'
bash-4.2# sysctl -w net.ipv4.udp_mem='65536 131072 262144'
bash-4.2# sysctl -w net.ipv4.tcp_rmem='8192 87380 16777216'
bash-4.2# sysctl -w net.ipv4.tcp_wmem='8192 65536 16777216'
bash-4.2# sysctl -w net.ipv4.udp_wmem_min=16384
bash-4.2# sysctl -w net.ipv4.tcp_tw_reuse=1
```

- Set the stripe width for the Lustre File System to 4M for Write major workload.
    - `lfs setstripe -S 4M -c -1` <respective directory created on Lustre FS>. "-c -1" parameter will let Lustre to stripe data across all OSTs equally.

- All the OSTs should be made active on different target controller ports

11. Performance Measurements

In order to baseline the performance characteristics, a series of IO benchmarks were performed using the tool “FIO” with a setup of up to 20 Lustre Clients, 4 OSS Servers and 1 Fully configured Pavilion HFAs. The linearity of performance was very evident as we saw the performance scale up as we added new clients to the cluster. The tests were conducted on a File System using the buffered IO option, that was more than 90% full. Here are the numbers from those tests.

Half Populated Chassis:

Block size	Read (GB/s)	Write (GB/s)
4096KB	38.6	24.34
8192KB	39.12	24.56

Fully Populated Chassis:

Block size	Read (GB/s)	Write (GB/s)
4096KB	79.90	50.30
8192KB	80.22	50.50

In a big Lustre deployment with 20 of **Pavilions** fully populated HFAs, overall cluster throughput is expected to be greater than 1.5 TB/s read & 1TB/s write.

Appendix A: Sample Lustre Configuration

```
========================================================================
OSS-0 (with MGS & MDS)
========================================================================
-bash-4.2# lnetctl net show
net:
    - net type: lo
      local NI(s):
        - nid: 0@lo
          status: up
    - net type: tcp2
      local NI(s):
        - nid: 192.168.10.220@tcp2
          status: up
          interfaces:
              0: enp216s0f0
    - net type: tcp1
      local NI(s):
        - nid: 192.168.20.220@tcp1
          status: up
          interfaces:
              0: enp94s0f0
-bash-4.2# lnetctl peer show
peer:
    - primary nid: 0@lo
      Multi-Rail: False
      peer ni:
        - nid: 0@lo
          state: NA
    - primary nid: 192.168.20.103@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.103@tcp1
          state: NA
    - primary nid: 192.168.10.201@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.201@tcp2
          state: NA
    - primary nid: 192.168.10.104@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.104@tcp2
          state: NA
    - primary nid: 192.168.20.150@tcp1
```

```
  Multi-Rail: True
  peer ni:
    - nid: 192.168.20.150@tcp1
      state: NA
- primary nid: 192.168.10.31@tcp2
  Multi-Rail: True
  peer ni:
    - nid: 192.168.10.31@tcp2
      state: NA
- primary nid: 192.168.10.161@tcp2
  Multi-Rail: True
  peer ni:
    - nid: 192.168.20.161@tcp1
      state: NA
    - nid: 192.168.10.161@tcp2
      state: NA
- primary nid: 192.168.10.224@tcp2
  Multi-Rail: True
  peer ni:
    - nid: 192.168.10.224@tcp2
      state: NA
- primary nid: 192.168.20.206@tcp1
  Multi-Rail: True
  peer ni:
    - nid: 192.168.20.206@tcp1
      state: NA
- primary nid: 192.168.10.164@tcp2
  Multi-Rail: True
  peer ni:
    - nid: 192.168.10.164@tcp2
      state: NA
    - nid: 192.168.20.164@tcp1
      state: NA
- primary nid: 192.168.20.104@tcp1
  Multi-Rail: True
  peer ni:
    - nid: 192.168.20.104@tcp1
      state: NA
- primary nid: 192.168.20.230@tcp1
  Multi-Rail: True
  peer ni:
    - nid: 192.168.20.230@tcp1
      state: NA
- primary nid: 192.168.10.122@tcp2
  Multi-Rail: True
  peer ni:
```

```
        - nid: 192.168.10.122@tcp2
          state: NA
    - primary nid: 192.168.20.122@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.122@tcp1
          state: NA
    - primary nid: 192.168.10.101@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.101@tcp2
          state: NA
-bash-4.2# lctl dl
  0 UP osd-zfs pavlstr-MDT0000-osd pavlstr-MDT0000-osd_UUID 10
  1 UP mgs MGS MGS 30
  2 UP mgc MGC192.168.10.220@tcp2 5c5d4ead-349d-a145-4046-d7672cf2bbe9 4
  3 UP mds MDS MDS_uuid 2
  4 UP lod pavlstr-MDT0000-mdtlov pavlstr-MDT0000-mdtlov_UUID 3
  5 UP mdt pavlstr-MDT0000 pavlstr-MDT0000_UUID 48
  6 UP mdd pavlstr-MDD0000 pavlstr-MDD0000_UUID 3
  7 UP qmt pavlstr-QMT0000 pavlstr-QMT0000_UUID 3
  8 UP lwp pavlstr-MDT0000-lwp-MDT0000 pavlstr-MDT0000-lwp-MDT0000_UUID 4
  9 UP osd-zfs pavlstr-OST0001-osd pavlstr-OST0001-osd_UUID 4
 10 UP mgc MGC192.168.20.220@tcp1 efc5973c-0cce-e0ca-250c-4a949e06d2c0 4
 11 UP ost OSS OSS_uuid 2
 12 UP obdfilter pavlstr-OST0001 pavlstr-OST0001_UUID 24
 13 UP lwp pavlstr-MDT0000-lwp-OST0001 pavlstr-MDT0000-lwp-OST0001_UUID 4
 14 UP osp pavlstr-OST0001-osc-MDT0000 pavlstr-MDT0000-mdtlov_UUID 4
 15 UP osp pavlstr-OST0002-osc-MDT0000 pavlstr-MDT0000-mdtlov_UUID 4
 16 UP osp pavlstr-OST0003-osc-MDT0000 pavlstr-MDT0000-mdtlov_UUID 4
 17 UP lov pavlstr-clilov-ffff8ea79fdfe000 9eb37eb9-70d6-7d57-8483-7db0fb94a360 3
 18 UP lmv pavlstr-clilmv-ffff8ea79fdfe000 9eb37eb9-70d6-7d57-8483-7db0fb94a360 4
 19 UP mdc pavlstr-MDT0000-mdc-ffff8ea79fdfe000 9eb37eb9-70d6-7d57-8483-
    7db0fb94a360 4
 20 UP osc pavlstr-OST0001-osc-ffff8ea79fdfe000 9eb37eb9-70d6-7d57-8483-
    7db0fb94a360 4
 21 UP osc pavlstr-OST0002-osc-ffff8ea79fdfe000 9eb37eb9-70d6-7d57-8483-
    7db0fb94a360 4
 22 UP osc pavlstr-OST0003-osc-ffff8ea79fdfe000 9eb37eb9-70d6-7d57-8483-
    7db0fb94a360 4
-bash-4.2# lnetctl peer list
peer list:
    - nid: 0@lo
    - nid: 192.168.20.103@tcp1
    - nid: 192.168.10.201@tcp2
    - nid: 192.168.10.104@tcp2
```

```
    - nid: 192.168.20.150@tcp1
    - nid: 192.168.10.31@tcp2
    - nid: 192.168.10.161@tcp2
    - nid: 192.168.10.224@tcp2
    - nid: 192.168.20.206@tcp1
    - nid: 192.168.10.164@tcp2
    - nid: 192.168.20.104@tcp1
    - nid: 192.168.20.230@tcp1
    - nid: 192.168.10.122@tcp2
    - nid: 192.168.20.122@tcp1
    - nid: 192.168.10.101@tcp2
-bash-4.2# zpool get all
NAME    PROPERTY                       VALUE                          SOURCE
pvlmst  size                           496G                           -
pvlmst  capacity                       0%                             -
pvlmst  altroot                        -                              default
pvlmst  health                         ONLINE                         -
pvlmst  guid                           12180659381911837860           -
pvlmst  version                        -                              default
pvlmst  bootfs                         -                              default
pvlmst  delegation                     on                             default
pvlmst  autoreplace                    off                            default
pvlmst  cachefile                      -                              default
pvlmst  failmode                       wait                           default
pvlmst  listsnapshots                  off                            default
pvlmst  autoexpand                     off                            default
pvlmst  dedupditto                     0                              default
pvlmst  dedupratio                     1.00x                          -
pvlmst  free                           496G                           -
pvlmst  allocated                      10.3M                          -
pvlmst  readonly                       off                            -
pvlmst  ashift                         12                             local
pvlmst  comment                        -                              default
pvlmst  expandsize                     -                              -
pvlmst  freeing                        0                              -
pvlmst  fragmentation                  0%                             -
pvlmst  leaked                         0                              -
pvlmst  multihost                      on                             local
pvlmst  feature@async_destroy          enabled                        local
pvlmst  feature@empty_bpobj            active                         local
pvlmst  feature@lz4_compress           active                         local
pvlmst  feature@multi_vdev_crash_dump  enabled                        local
pvlmst  feature@spacemap_histogram     active                         local
pvlmst  feature@enabled_txg            active                         local
pvlmst  feature@hole_birth             active                         local
pvlmst  feature@extensible_dataset     active                         local
```

```
pvlmst  feature@embedded_data          active                    local
pvlmst  feature@bookmarks              enabled                   local
pvlmst  feature@filesystem_limits      enabled                   local
pvlmst  feature@large_blocks           enabled                   local
pvlmst  feature@large_dnode            active                    local
pvlmst  feature@sha512                 enabled                   local
pvlmst  feature@skein                  enabled                   local
pvlmst  feature@edonr                  enabled                   local
pvlmst  feature@userobj_accounting     active                    local
pvlost  size                           11.7T                     -
pvlost  capacity                       23%                       -
pvlost  altroot                        -                         default
pvlost  health                         ONLINE                    -
pvlost  guid                           4864548095743480396       -
pvlost  version                        -                         default
pvlost  bootfs                         -                         default
pvlost  delegation                     on                        default
pvlost  autoreplace                    off                       default
pvlost  cachefile                      -                         default
pvlost  failmode                       wait                      default
pvlost  listsnapshots                  off                       default
pvlost  autoexpand                     off                       default
pvlost  dedupditto                     0                         default
pvlost  dedupratio                     1.00x                     -
pvlost  free                           8.90T                     -
pvlost  allocated                      2.79T                     -
pvlost  readonly                       off                       -
pvlost  ashift                         12                        local
pvlost  comment                        -                         default
pvlost  expandsize                     -                         -
pvlost  freeing                        0                         -
pvlost  fragmentation                  6%                        -
pvlost  leaked                         0                         -
pvlost  multihost                      on                        local
pvlost  feature@async_destroy          enabled                   local
pvlost  feature@empty_bpobj            active                    local
pvlost  feature@lz4_compress           active                    local
pvlost  feature@multi_vdev_crash_dump  enabled                   local
pvlost  feature@spacemap_histogram     active                    local
pvlost  feature@enabled_txg            active                    local
pvlost  feature@hole_birth             active                    local
pvlost  feature@extensible_dataset     active                    local
pvlost  feature@embedded_data          active                    local
pvlost  feature@bookmarks              enabled                   local
pvlost  feature@filesystem_limits      enabled                   local
pvlost  feature@large_blocks           active                    local
```

```
pvlost  feature@large_dnode            active                         local
pvlost  feature@sha512                 enabled                        local
pvlost  feature@skein                  enabled                        local
pvlost  feature@edonr                  enabled                        local
pvlost  feature@userobj_accounting     active                         local
-bash-4.2# zfs get all
NAME         PROPERTY              VALUE                                  SOURCE
pvlmst       type                  filesystem                             -
pvlmst       creation              Mon Aug 10 14:59 2020                  -
pvlmst       used                  10.2M                                  -
pvlmst       available             480G                                   -
pvlmst       referenced            96K                                    -
pvlmst       compressratio         1.00x                                  -
pvlmst       mounted               yes                                    -
pvlmst       quota                 none                                   default
pvlmst       reservation           none                                   default
pvlmst       recordsize            128K                                   default
pvlmst       mountpoint            /pvlmst                                default
pvlmst       sharenfs              off                                    default
pvlmst       checksum              on                                     default
pvlmst       compression           off                                    default
pvlmst       atime                 on                                     default
pvlmst       devices               on                                     default
pvlmst       exec                  on                                     default
pvlmst       setuid                on                                     default
pvlmst       readonly              off                                    default
pvlmst       zoned                 off                                    default
pvlmst       snapdir               hidden                                 default
pvlmst       aclinherit            restricted                             default
pvlmst       createtxg             1                                      -
pvlmst       canmount              on                                     default
pvlmst       xattr                 on                                     default
pvlmst       copies                1                                      default
pvlmst       version               5                                      -
pvlmst       utf8only              off                                    -
pvlmst       normalization         none                                   -
pvlmst       casesensitivity       sensitive                              -
pvlmst       vscan                 off                                    default
pvlmst       nbmand                off                                    default
pvlmst       sharesmb              off                                    default
pvlmst       refquota              none                                   default
pvlmst       refreservation        none                                   default
pvlmst       guid                  3398100655848891719                    -
pvlmst       primarycache          all                                    default
pvlmst       secondarycache        all                                    default
pvlmst       usedbysnapshots       0B                                     -
```

```
pvlmst       usedbydataset         96K                              -
pvlmst       usedbychildren        10.1M                            -
pvlmst       usedbyrefreservation  0B                               -
pvlmst       logbias               latency                          default
pvlmst       dedup                 off                              default
pvlmst       mlslabel              none                             default
pvlmst       sync                  standard                         default
pvlmst       dnodesize             legacy                           default
pvlmst       refcompressratio      1.00x                            -
pvlmst       written               96K                              -
pvlmst       logicalused           4.34M                            -
pvlmst       logicalreferenced     40K                              -
pvlmst       volmode               default                          default
pvlmst       filesystem_limit      none                             default
pvlmst       snapshot_limit        none                             default
pvlmst       filesystem_count      none                             default
pvlmst       snapshot_count        none                             default
pvlmst       snapdev               hidden                           default
pvlmst       acltype               off                              default
pvlmst       context               none                             default
pvlmst       fscontext             none                             default
pvlmst       defcontext            none                             default
pvlmst       rootcontext           none                             default
pvlmst       relatime              off                              default
pvlmst       redundant_metadata    all                              default
pvlmst       overlay               off                              default
pvlmst/mst0  type                  filesystem                       -
pvlmst/mst0  creation              Mon Aug 10 15:00 2020            -
pvlmst/mst0  used                  9.68M                            -
pvlmst/mst0  available             480G                             -
pvlmst/mst0  referenced            9.68M                            -
pvlmst/mst0  compressratio         1.00x                            -
pvlmst/mst0  mounted               no                               -
pvlmst/mst0  quota                 none                             default
pvlmst/mst0  reservation           none                             default
pvlmst/mst0  recordsize            128K                             default
pvlmst/mst0  mountpoint            /pvlmst/mst0                     default
pvlmst/mst0  sharenfs              off                              default
pvlmst/mst0  checksum              on                               default
pvlmst/mst0  compression           off                              default
pvlmst/mst0  atime                 on                               default
pvlmst/mst0  devices               on                               default
pvlmst/mst0  exec                  on                               default
pvlmst/mst0  setuid                on                               default
pvlmst/mst0  readonly              off                              default
pvlmst/mst0  zoned                 off                              default
```

```
pvlmst/mst0  snapdir               hidden                          default
pvlmst/mst0  aclinherit            restricted                      default
pvlmst/mst0  createtxg             10                              -
pvlmst/mst0  canmount              off                             local
pvlmst/mst0  xattr                 sa                              local
pvlmst/mst0  copies                1                               default
pvlmst/mst0  version               5                               -
pvlmst/mst0  utf8only              off                             -
pvlmst/mst0  normalization         none                            -
pvlmst/mst0  casesensitivity       sensitive                       -
pvlmst/mst0  vscan                 off                             default
pvlmst/mst0  nbmand                off                             default
pvlmst/mst0  sharesmb              off                             default
pvlmst/mst0  refquota              none                            default
pvlmst/mst0  refreservation        none                            default
pvlmst/mst0  guid                  11605027522660399998            -
pvlmst/mst0  primarycache          all                             default
pvlmst/mst0  secondarycache        all                             default
pvlmst/mst0  usedbysnapshots       0B                              -
pvlmst/mst0  usedbydataset         9.68M                           -
pvlmst/mst0  usedbychildren        0B                              -
pvlmst/mst0  usedbyrefreservation  0B                              -
pvlmst/mst0  logbias               latency                         default
pvlmst/mst0  dedup                 off                             default
pvlmst/mst0  mlslabel              none                            default
pvlmst/mst0  sync                  standard                        default
pvlmst/mst0  dnodesize             auto                            local
pvlmst/mst0  refcompressratio      1.00x                           -
pvlmst/mst0  written               9.68M                           -
pvlmst/mst0  logicalused           4.18M                           -
pvlmst/mst0  logicalreferenced     4.18M                           -
pvlmst/mst0  volmode               default                         default
pvlmst/mst0  filesystem_limit      none                            default
pvlmst/mst0  snapshot_limit        none                            default
pvlmst/mst0  filesystem_count      none                            default
pvlmst/mst0  snapshot_count        none                            default
pvlmst/mst0  snapdev               hidden                          default
pvlmst/mst0  acltype               off                             default
pvlmst/mst0  context               none                            default
pvlmst/mst0  fscontext             none                            default
pvlmst/mst0  defcontext            none                            default
pvlmst/mst0  rootcontext           none                            default
pvlmst/mst0  relatime              off                             default
pvlmst/mst0  redundant_metadata    all                             default
pvlmst/mst0  overlay               off                             default
pvlmst/mst0  lustre:fsname         pavlstr                         local
```

```
pvlmst/mst0  lustre:flags          37                          local
pvlmst/mst0  lustre:svname         pavlstr-MDT0000             local
pvlmst/mst0  lustre:version        1                           local
pvlmst/mst0  lustre:index          0                           local
pvlost       type                  filesystem                  -
pvlost       creation              Mon Aug 10 14:59 2020       -
pvlost       used                  2.78T                       -
pvlost       available             8.54T                       -
pvlost       referenced            96K                         -
pvlost       compressratio         1.00x                       -
pvlost       mounted               yes                         -
pvlost       quota                 none                        default
pvlost       reservation           none                        default
pvlost       recordsize            128K                        default
pvlost       mountpoint            /pvlost                     default
pvlost       sharenfs              off                         default
pvlost       checksum              on                          default
pvlost       compression           off                         default
pvlost       atime                 on                          default
pvlost       devices               on                          default
pvlost       exec                  on                          default
pvlost       setuid                on                          default
pvlost       readonly              off                         default
pvlost       zoned                 off                         default
pvlost       snapdir               hidden                      default
pvlost       aclinherit            restricted                  default
pvlost       createtxg             1                           -
pvlost       canmount              on                          default
pvlost       xattr                 on                          default
pvlost       copies                1                           default
pvlost       version               5                           -
pvlost       utf8only              off                         -
pvlost       normalization         none                        -
pvlost       casesensitivity       sensitive                   -
pvlost       vscan                 off                         default
pvlost       nbmand                off                         default
pvlost       sharesmb              off                         default
pvlost       refquota              none                        default
pvlost       refreservation        none                        default
pvlost       guid                  5082540486041219791         -
pvlost       primarycache          all                         default
pvlost       secondarycache        all                         default
pvlost       usedbysnapshots       0B                          -
pvlost       usedbydataset         96K                         -
pvlost       usedbychildren        2.78T                       -
pvlost       usedbyrefreservation  0B                          -
```

```
pvlost       logbias               latency                        default
pvlost       dedup                 off                            default
pvlost       mlslabel              none                           default
pvlost       sync                  standard                       default
pvlost       dnodesize             legacy                         default
pvlost       refcompressratio      1.00x                          -
pvlost       written               96K                            -
pvlost       logicalused           2.78T                          -
pvlost       logicalreferenced     40K                            -
pvlost       volmode               default                        default
pvlost       filesystem_limit      none                           default
pvlost       snapshot_limit        none                           default
pvlost       filesystem_count      none                           default
pvlost       snapshot_count        none                           default
pvlost       snapdev               hidden                         default
pvlost       acltype               off                            default
pvlost       context               none                           default
pvlost       fscontext             none                           default
pvlost       defcontext            none                           default
pvlost       rootcontext           none                           default
pvlost       relatime              off                            default
pvlost       redundant_metadata    all                            default
pvlost       overlay               off                            default
pvlost/ost0  type                  filesystem                     -
pvlost/ost0  creation              Mon Aug 10 15:00 2020          -
pvlost/ost0  used                  2.78T                          -
pvlost/ost0  available             8.54T                          -
pvlost/ost0  referenced            2.78T                          -
pvlost/ost0  compressratio         1.00x                          -
pvlost/ost0  mounted               no                             -
pvlost/ost0  quota                 none                           default
pvlost/ost0  reservation           none                           default
pvlost/ost0  recordsize            1M                             local
pvlost/ost0  mountpoint            /pvlost/ost0                   default
pvlost/ost0  sharenfs              off                            default
pvlost/ost0  checksum              on                             default
pvlost/ost0  compression           off                            default
pvlost/ost0  atime                 on                             default
pvlost/ost0  devices               on                             default
pvlost/ost0  exec                  on                             default
pvlost/ost0  setuid                on                             default
pvlost/ost0  readonly              off                            default
pvlost/ost0  zoned                 off                            default
pvlost/ost0  snapdir               hidden                         default
pvlost/ost0  aclinherit            restricted                     default
pvlost/ost0  createtxg             14                             -
```

```
pvlost/ost0  canmount             off                              local
pvlost/ost0  xattr                sa                               local
pvlost/ost0  copies               1                                default
pvlost/ost0  version              5                                -
pvlost/ost0  utf8only             off                              -
pvlost/ost0  normalization        none                             -
pvlost/ost0  casesensitivity      sensitive                        -
pvlost/ost0  vscan                off                              default
pvlost/ost0  nbmand               off                              default
pvlost/ost0  sharesmb             off                              default
pvlost/ost0  refquota             none                             default
pvlost/ost0  refreservation       none                             default
pvlost/ost0  guid                 16937295366784941238             -
pvlost/ost0  primarycache         all                              default
pvlost/ost0  secondarycache       all                              default
pvlost/ost0  usedbysnapshots      0B                               -
pvlost/ost0  usedbydataset        2.78T                            -
pvlost/ost0  usedbychildren       0B                               -
pvlost/ost0  usedbyrefreservation 0B                               -
pvlost/ost0  logbias              latency                          default
pvlost/ost0  dedup                off                              default
pvlost/ost0  mlslabel             none                             default
pvlost/ost0  sync                 standard                         default
pvlost/ost0  dnodesize            auto                             local
pvlost/ost0  refcompressratio     1.00x                            -
pvlost/ost0  written              2.78T                            -
pvlost/ost0  logicalused          2.78T                            -
pvlost/ost0  logicalreferenced    2.78T                            -
pvlost/ost0  volmode              default                          default
pvlost/ost0  filesystem_limit     none                             default
pvlost/ost0  snapshot_limit       none                             default
pvlost/ost0  filesystem_count     none                             default
pvlost/ost0  snapshot_count       none                             default
pvlost/ost0  snapdev              hidden                           default
pvlost/ost0  acltype              off                              default
pvlost/ost0  context              none                             default
pvlost/ost0  fscontext            none                             default
pvlost/ost0  defcontext           none                             default
pvlost/ost0  rootcontext          none                             default
pvlost/ost0  relatime             off                              default
pvlost/ost0  redundant_metadata   all                              default
pvlost/ost0  overlay              off                              default
pvlost/ost0  lustre:fsname        pavlstr                          local
pvlost/ost0  lustre:flags         34                               local
pvlost/ost0  lustre:svname        pavlstr-OST0001                  local
pvlost/ost0  lustre:version       1                                local
```

```
pvlost/ost0  lustre:index          1                                       local
pvlost/ost0  lustre:mgsnode        192.168.20.220@tcp1:192.168.10.220@tcp2  local
pvlost/ost0  lustre:failover.node  192.168.20.161@tcp1:192.168.10.161@tcp2  local
-bash-4.2# zpool status
  pool: pvlmst
 state: ONLINE
  scan: none requested
config:

        NAME        STATE     READ WRITE CKSUM
        pvlmst      ONLINE       0     0     0
          dm-0      ONLINE       0     0     0

errors: No known data errors

  pool: pvlost
 state: ONLINE
  scan: none requested
config:

        NAME        STATE     READ WRITE CKSUM
        pvlost      ONLINE       0     0     0
          dm-11     ONLINE       0     0     0
          dm-1      ONLINE       0     0     0
          dm-2      ONLINE       0     0     0
          dm-12     ONLINE       0     0     0

errors: No known data errors
-bash-4.2# cat /sys/block/dm-*/queue/scheduler
noop [deadline] cfq
noop [deadline] cfq
noop [deadline] cfq
noop [deadline] cfq
noop [deadline] cfq
-bash-4.2# tuned-adm active
-bash-4.2# tuned-adm profile  throughput-performance
-bash-4.2# lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                72
On-line CPU(s) list:   0-71
Thread(s) per core:    2
Core(s) per socket:    18
Socket(s):             2
NUMA node(s):          2
```

```
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz
Stepping:              4
CPU MHz:               2714.685
CPU max MHz:           3700.0000
CPU min MHz:           1000.0000
BogoMIPS:              4600.00
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              25344K
NUMA node0 CPU(s):     0-17,36-53
NUMA node1 CPU(s):     18-35,54-71
Flags:                 fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov
pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp
lm constant_tsc art arch_perfmon pebs bts rep_good nopl xtopology nonstop_tsc
aperfmperf eagerfpu pni pclmulqdq dtes64 monitor ds_cpl vmx smx est tm2 ssse3 sdbg
fma cx16 xtpr pdcm pcid dca sse4_1 sse4_2 x2apic movbe popcnt tsc_deadline_timer aes
xsave avx f16c rdrand lahf_lm abm 3dnowprefetch epb cat_l3 cdp_l3 invpcid_single
intel_ppin intel_pt ssbd mba ibrs ibpb stibp tpr_shadow vnmi flexpriority ept vpid
fsgsbase tsc_adjust bmi1 hle avx2 smep bmi2 erms invpcid rtm cqm mpx rdt_a avx512f
avx512dq rdseed adx smap clflushopt clwb avx512cd avx512bw avx512vl xsaveopt xsavec
xgetbv1 cqm_llc cqm_occup_llc cqm_mbm_total cqm_mbm_local dtherm ida arat pln pts hwp
hwp_act_window hwp_epp hwp_pkg_req pku ospke md_clear spec_ctrl intel_stibp flush_l1d
-bash-4.2# free -h
              total        used        free      shared  buff/cache   available
Mem:           187G        101G         83G         27M        1.5G         84G
Swap:          2.0G        1.2M        2.0G
-bash-4.2# vmstat -s
    196491696 K total memory
    107080656 K used memory
       678584 K active memory
       740304 K inactive memory
     87836688 K free memory
       233820 K buffer memory
      1340528 K swap cache
      2088956 K total swap
         1280 K used swap
      2087676 K free swap
       334704 non-nice user cpu ticks
          630 nice user cpu ticks
    161925699 system cpu ticks
   5508097895 idle cpu ticks
```

```
     65767893 IO-wait cpu ticks
            0 IRQ cpu ticks
     19048621 softirq cpu ticks
            0 stolen cpu ticks
 642761797269 pages paged in
    124558440 pages paged out
          130 pages swapped in
          326 pages swapped out
   4023529799 interrupts
   1471253441 CPU context switches
   1596307914 boot time
    205941621 forks
-bash-4.2# vmstat -d
disk- ------------reads------------ ------------writes----------- -----IO------
       total merged sectors      ms  total merged sectors      ms    cur    sec
sda   835750    959 192419627  812032 350548 1622888 249083256 88072483      0   1397
nvme0n1 217970      0 1973832   41529 1130120      0 16830768  280451      0    215
dm-0  217940      0 1973160   50543 1130120      0 16830768  362297      0    287
nvme1n1     11      0      88      31      0      0       0       0      0      0
nvme2n1 93709260      0 46913185256 255407351 110834448      0 50836938456 347734282
0  29231
nvme3n1     13      0     104      39      0      0       0       0      0      0
dm-1  93709172      0 46913181096 260382073 110834448      0 50836938456 364482489
0  29677
nvme4n1 94027373      0 47094821216 130630435 110666336      0 50745098712 346544965
0  28864
nvme5n1     13      0     104      31      0      0       0       0      0      0
dm-11 94027287      0 47094818080 135734997 110666336      0 50745098712 363332034
0  29339
nvme6n1 100447806      0 50304453504 102927273 119764967      0 55417951064 331176669
0  27840
nvme7n1     13      0     104      35      0      0       0       0      0      0
dm-12 100447718      0 50304449344 108257186 119764967      0 55417951064 348477175
0  28398
nvme8n1 60777059      0 30109999352 183331549 82051540      0 36950387000 263600297
0  23273
nvme9n1     13      0     104      31      0      0       0       0      0      0
dm-2  60776971      0 30109995192 186639440 82051540      0 36950387000 276340969
0  23632
-bash-4.2# vmstat 2 6
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
11 52   1280 87893720 233828 1340560    0    0 11170     2    1    0
0  3 96  1  0
10 70   1280 87966984 233828 1340560    0    0 9376922     0 274477 300263  0 13 49
38  0
```

```
17 53   1280 87917000 233828 1340560    0    0 9291910     0 268641 298906  0 13 49
38  0
12 60   1280 87702480 233828 1340560    0    0 9432646     0 275850 305137  0 14 49
37  0
10 54   1280 88035536 233828 1340560    0    0 9350146     0 275953 302169  0 14 49
38  0
17 62   1280 87800576 233828 1340560    0    0 9270412     0 274985 300202  0 14 49
37  0
-bash-4.2#  vmstat -S k 1 10
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
13 54   1310 89979256 239439 1372733    0    0 11177     2    0    1  0  3 96  1  0
 9 69   1310 89770472 239439 1372733    0    0 9483968     0 284586 306276  0 14 48
38  0
13 42   1310 89776952 239439 1372749    0    0 9300312     0 270181 301217  0 14 49
37  0
12 44   1310 90143160 239439 1372749    0    0 9351024     0 278842 308482  0 14 50
36  0
11 56   1310 89931232 239439 1372749    0    0 9290400     0 275459 297570  0 14 49
38  0
16 60   1310 89777856 239439 1372749    0    0 9319176     0 271100 297986  0 13 48
38  0
14 58   1310 89876040 239439 1372749    0    0 9496812     0 286332 309808  0 14 48
38  0
10 65   1310 90110152 239443 1372749    0    0 9316320    56 273684 299861  0 14 49
38  0
 7 58   1310 90010208 239443 1372749    0    0 9375408     0 277335 301898  0 14 49
37  0
12 61   1310 89945240 239443 1372749    0    0 9260828     0 271840 301748  0 14 49
37  0
```

```
========================================================================
OSS-1
========================================================================
-bash-4.2# lnetctl net show
net:
    - net type: lo
      local NI(s):
        - nid: 0@lo
          status: up
    - net type: tcp2
      local NI(s):
        - nid: 192.168.10.161@tcp2
          status: up
          interfaces:
              0: enp216s0f0
    - net type: tcp1
      local NI(s):
        - nid: 192.168.20.161@tcp1
          status: up
          interfaces:
              0: enp94s0f0
-bash-4.2# lnetctl peer show
peer:
    - primary nid: 192.168.10.220@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.220@tcp1
          state: NA
        - nid: 192.168.10.220@tcp2
          state: NA
    - primary nid: 192.168.20.103@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.103@tcp1
          state: NA
    - primary nid: 192.168.10.201@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.201@tcp2
          state: NA
    - primary nid: 192.168.10.104@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.104@tcp2
          state: NA
    - primary nid: 192.168.20.150@tcp1
```

```
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.150@tcp1
          state: NA
    - primary nid: 192.168.10.31@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.31@tcp2
          state: NA
    - primary nid: 192.168.20.206@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.206@tcp1
          state: NA
    - primary nid: 192.168.10.224@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.224@tcp2
          state: NA
    - primary nid: 192.168.20.104@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.104@tcp1
          state: NA
    - primary nid: 192.168.20.230@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.230@tcp1
          state: NA
    - primary nid: 192.168.10.122@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.122@tcp2
          state: NA
    - primary nid: 192.168.20.122@tcp1
      Multi-Rail: True
      peer ni:
        - nid: 192.168.20.122@tcp1
          state: NA
    - primary nid: 192.168.10.101@tcp2
      Multi-Rail: True
      peer ni:
        - nid: 192.168.10.101@tcp2
          state: NA
-bash-4.2# lctl dl
  0 UP osd-zfs pavlstr-OST0002-osd pavlstr-OST0002-osd_UUID 4
```

```
  1 UP mgc MGC192.168.20.220@tcp1 2bb5042e-9cac-8ccd-d2bd-f80673399c89 4
  2 UP ost OSS OSS_uuid 2
  3 UP obdfilter pavlstr-OST0002 pavlstr-OST0002_UUID 24
  4 UP lwp pavlstr-MDT0000-lwp-OST0002 pavlstr-MDT0000-lwp-OST0002_UUID 4
-bash-4.2# lnetctl peer list
peer list:
    - nid: 192.168.10.220@tcp2
    - nid: 192.168.20.103@tcp1
    - nid: 192.168.10.201@tcp2
    - nid: 192.168.10.104@tcp2
    - nid: 192.168.20.150@tcp1
    - nid: 192.168.10.31@tcp2
    - nid: 192.168.20.206@tcp1
    - nid: 192.168.10.224@tcp2
    - nid: 192.168.20.104@tcp1
    - nid: 192.168.20.230@tcp1
    - nid: 192.168.10.122@tcp2
    - nid: 192.168.20.122@tcp1
    - nid: 192.168.10.101@tcp2
-bash-4.2# zpool get all
NAME    PROPERTY                       VALUE                          SOURCE
pvlost  size                           8.77T                          -
pvlost  capacity                       27%                            -
pvlost  altroot                        -                              default
pvlost  health                         ONLINE                         -
pvlost  guid                           13322126776979571432           -
pvlost  version                        -                              default
pvlost  bootfs                         -                              default
pvlost  delegation                     on                             default
pvlost  autoreplace                    off                            default
pvlost  cachefile                      -                              default
pvlost  failmode                       wait                           default
pvlost  listsnapshots                  off                            default
pvlost  autoexpand                     off                            default
pvlost  dedupditto                     0                              default
pvlost  dedupratio                     1.00x                          -
pvlost  free                           6.37T                          -
pvlost  allocated                      2.39T                          -
pvlost  readonly                       off                            -
pvlost  ashift                         12                             local
pvlost  comment                        -                              default
pvlost  expandsize                     -                              -
pvlost  freeing                        0                              -
pvlost  fragmentation                  2%                             -
pvlost  leaked                         0                              -
pvlost  multihost                      on                             local
```

```
pvlost  checkpoint                     -                              -
pvlost  load_guid                      17782960997735285430           -
pvlost  autotrim                       off                            default
pvlost  feature@async_destroy          enabled                        local
pvlost  feature@empty_bpobj            active                         local
pvlost  feature@lz4_compress           active                         local
pvlost  feature@multi_vdev_crash_dump  enabled                        local
pvlost  feature@spacemap_histogram     active                         local
pvlost  feature@enabled_txg            active                         local
pvlost  feature@hole_birth             active                         local
pvlost  feature@extensible_dataset     active                         local
pvlost  feature@embedded_data          active                         local
pvlost  feature@bookmarks              enabled                        local
pvlost  feature@filesystem_limits      enabled                        local
pvlost  feature@large_blocks           active                         local
pvlost  feature@large_dnode            active                         local
pvlost  feature@sha512                 enabled                        local
pvlost  feature@skein                  enabled                        local
pvlost  feature@edonr                  enabled                        local
pvlost  feature@userobj_accounting     active                         local
pvlost  feature@encryption             enabled                        local
pvlost  feature@project_quota          active                         local
pvlost  feature@device_removal         enabled                        local
pvlost  feature@obsolete_counts        enabled                        local
pvlost  feature@zpool_checkpoint       enabled                        local
pvlost  feature@spacemap_v2            active                         local
pvlost  feature@allocation_classes     enabled                        local
pvlost  feature@resilver_defer         enabled                        local
pvlost  feature@bookmark_v2            enabled                        local
-bash-4.2# zfs get all
NAME         PROPERTY              VALUE                                  SOURCE
pvlost       type                  filesystem                             -
pvlost       creation              Mon Aug 10 15:00 2020                  -
pvlost       used                  2.39T                                  -
pvlost       available             6.10T                                  -
pvlost       referenced            96K                                    -
pvlost       compressratio         1.00x                                  -
pvlost       mounted               yes                                    -
pvlost       quota                 none                                   default
pvlost       reservation           none                                   default
pvlost       recordsize            128K                                   default
pvlost       mountpoint            /pvlost                                default
pvlost       sharenfs              off                                    default
pvlost       checksum              on                                     default
pvlost       compression           off                                    default
pvlost       atime                 on                                     default
```

```
pvlost       devices               on                                  default
pvlost       exec                  on                                  default
pvlost       setuid                on                                  default
pvlost       readonly              off                                 default
pvlost       zoned                 off                                 default
pvlost       snapdir               hidden                              default
pvlost       aclinherit            restricted                          default
pvlost       createtxg             1                                   -
pvlost       canmount              on                                  default
pvlost       xattr                 on                                  default
pvlost       copies                1                                   default
pvlost       version               5                                   -
pvlost       utf8only              off                                 -
pvlost       normalization         none                                -
pvlost       casesensitivity       sensitive                           -
pvlost       vscan                 off                                 default
pvlost       nbmand                off                                 default
pvlost       sharesmb              off                                 default
pvlost       refquota              none                                default
pvlost       refreservation        none                                default
pvlost       guid                  13061051670594097O4                 -
pvlost       primarycache          all                                 default
pvlost       secondarycache        all                                 default
pvlost       usedbysnapshots       0B                                  -
pvlost       usedbydataset         96K                                 -
pvlost       usedbychildren        2.39T                               -
pvlost       usedbyrefreservation  0B                                  -
pvlost       logbias               latency                             default
pvlost       objsetid              54                                  -
pvlost       dedup                 off                                 default
pvlost       mlslabel              none                                default
pvlost       sync                  standard                            default
pvlost       dnodesize             legacy                              default
pvlost       refcompressratio      1.00x                               -
pvlost       written               96K                                 -
pvlost       logicalused           2.39T                               -
pvlost       logicalreferenced     42K                                 -
pvlost       volmode               default                             default
pvlost       filesystem_limit      none                                default
pvlost       snapshot_limit        none                                default
pvlost       filesystem_count      none                                default
pvlost       snapshot_count        none                                default
pvlost       snapdev               hidden                              default
pvlost       acltype               off                                 default
pvlost       context               none                                default
pvlost       fscontext             none                                default
```

```
pvlost       defcontext            none                                default
pvlost       rootcontext           none                                default
pvlost       relatime              off                                 default
pvlost       redundant_metadata    all                                 default
pvlost       overlay               off                                 default
pvlost       encryption            off                                 default
pvlost       keylocation           none                                default
pvlost       keyformat             none                                default
pvlost       pbkdf2iters           0                                   default
pvlost       special_small_blocks  0                                   default
pvlost/ost1  type                  filesystem                          -
pvlost/ost1  creation              Mon Aug 10 15:01 2020               -
pvlost/ost1  used                  2.39T                               -
pvlost/ost1  available             6.10T                               -
pvlost/ost1  referenced            2.39T                               -
pvlost/ost1  compressratio         1.00x                               -
pvlost/ost1  mounted               no                                  -
pvlost/ost1  quota                 none                                default
pvlost/ost1  reservation           none                                default
pvlost/ost1  recordsize            1M                                  local
pvlost/ost1  mountpoint            /pvlost/ost1                        default
pvlost/ost1  sharenfs              off                                 default
pvlost/ost1  checksum              on                                  default
pvlost/ost1  compression           off                                 default
pvlost/ost1  atime                 on                                  default
pvlost/ost1  devices               on                                  default
pvlost/ost1  exec                  on                                  default
pvlost/ost1  setuid                on                                  default
pvlost/ost1  readonly              off                                 default
pvlost/ost1  zoned                 off                                 default
pvlost/ost1  snapdir               hidden                              default
pvlost/ost1  aclinherit            restricted                          default
pvlost/ost1  createtxg             16                                  -
pvlost/ost1  canmount              off                                 local
pvlost/ost1  xattr                 sa                                  local
pvlost/ost1  copies                1                                   default
pvlost/ost1  version               5                                   -
pvlost/ost1  utf8only              off                                 -
pvlost/ost1  normalization         none                                -
pvlost/ost1  casesensitivity       sensitive                           -
pvlost/ost1  vscan                 off                                 default
pvlost/ost1  nbmand                off                                 default
pvlost/ost1  sharesmb              off                                 default
pvlost/ost1  refquota              none                                default
pvlost/ost1  refreservation        none                                default
pvlost/ost1  guid                  5334884274952026580                 -
```

```
pvlost/ost1  primarycache          all                                    default
pvlost/ost1  secondarycache        all                                    default
pvlost/ost1  usedbysnapshots       0B                                     -
pvlost/ost1  usedbydataset         2.39T                                  -
pvlost/ost1  usedbychildren        0B                                     -
pvlost/ost1  usedbyrefreservation  0B                                     -
pvlost/ost1  logbias               latency                                default
pvlost/ost1  objsetid              153                                    -
pvlost/ost1  dedup                 off                                    default
pvlost/ost1  mlslabel              none                                   default
pvlost/ost1  sync                  standard                               default
pvlost/ost1  dnodesize             auto                                   local
pvlost/ost1  refcompressratio      1.00x                                  -
pvlost/ost1  written               2.39T                                  -
pvlost/ost1  logicalused           2.39T                                  -
pvlost/ost1  logicalreferenced     2.39T                                  -
pvlost/ost1  volmode               default                                default
pvlost/ost1  filesystem_limit      none                                   default
pvlost/ost1  snapshot_limit        none                                   default
pvlost/ost1  filesystem_count      none                                   default
pvlost/ost1  snapshot_count        none                                   default
pvlost/ost1  snapdev               hidden                                 default
pvlost/ost1  acltype               off                                    default
pvlost/ost1  context               none                                   default
pvlost/ost1  fscontext             none                                   default
pvlost/ost1  defcontext            none                                   default
pvlost/ost1  rootcontext           none                                   default
pvlost/ost1  relatime              off                                    default
pvlost/ost1  redundant_metadata    all                                    default
pvlost/ost1  overlay               off                                    default
pvlost/ost1  encryption            off                                    default
pvlost/ost1  keylocation           none                                   default
pvlost/ost1  keyformat             none                                   default
pvlost/ost1  pbkdf2iters           0                                      default
pvlost/ost1  special_small_blocks  0                                      default
pvlost/ost1  lustre:index          2                                      local
pvlost/ost1  lustre:fsname         pavlstr                                local
pvlost/ost1  lustre:version        1                                      local
pvlost/ost1  lustre:flags          34                                     local
pvlost/ost1  lustre:svname         pavlstr-OST0002                        local
pvlost/ost1  lustre:mgsnode        192.168.20.220@tcp1:192.168.10.220@tcp2  local
pvlost/ost1  lustre:failover.node  192.168.20.164@tcp1:192.168.10.164@tcp2  local
-bash-4.2# zpool status
  pool: pvlost
 state: ONLINE
  scan: none requested
```

```
config:

        NAME        STATE     READ WRITE CKSUM
        pvlost      ONLINE       0     0     0
          dm-0      ONLINE       0     0     0
          dm-1      ONLINE       0     0     0
          dm-2      ONLINE       0     0     0

errors: No known data errors
-bash-4.2# cat /sys/block/dm-*/queue/scheduler
noop [deadline] cfq
noop [deadline] cfq
noop [deadline] cfq
-bash-4.2# lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                72
On-line CPU(s) list:   0-71
Thread(s) per core:    2
Core(s) per socket:    18
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz
Stepping:              4
CPU MHz:               1000.073
CPU max MHz:           3700.0000
CPU min MHz:           1000.0000
BogoMIPS:              4600.00
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              25344K
NUMA node0 CPU(s):     0-17,36-53
NUMA node1 CPU(s):     18-35,54-71
Flags:                 fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov
pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp
lm constant_tsc art arch_perfmon pebs bts rep_good nopl xtopology nonstop_tsc
aperfmperf eagerfpu pni pclmulqdq dtes64 monitor ds_cpl vmx smx est tm2 ssse3 sdbg
fma cx16 xtpr pdcm pcid dca sse4_1 sse4_2 x2apic movbe popcnt tsc_deadline_timer aes
xsave avx f16c rdrand lahf_lm abm 3dnowprefetch epb cat_l3 cdp_l3 invpcid_single
intel_ppin intel_pt ssbd mba ibrs ibpb stibp tpr_shadow vnmi flexpriority ept vpid
```

```
fsgsbase tsc_adjust bmi1 hle avx2 smep bmi2 erms invpcid rtm cqm mpx rdt_a avx512f
avx512dq rdseed adx smap clflushopt clwb avx512cd avx512bw avx512vl xsaveopt xsavec
xgetbv1 cqm_llc cqm_occup_llc cqm_mbm_total cqm_mbm_local dtherm ida arat pln pts hwp
hwp_act_window hwp_epp hwp_pkg_req pku ospke md_clear spec_ctrl intel_stibp flush_l1d
-bash-4.2# free -h
              total        used        free      shared  buff/cache   available
Mem:           187G        104G         76G         26M        6.2G         81G
Swap:          2.0G          0B        2.0G
-bash-4.2# vmstat -s
    196491680 K total memory
    109773976 K used memory
      4767680 K active memory
      1557412 K inactive memory
     80209424 K free memory
       168968 K buffer memory
      6339320 K swap cache
      2088956 K total swap
            0 K used swap
      2088956 K free swap
        63957 non-nice user cpu ticks
          453 nice user cpu ticks
    152141200 system cpu ticks
   4120210648 idle cpu ticks
     94064149 IO-wait cpu ticks
            0 IRQ cpu ticks
     14754829 softirq cpu ticks
            0 stolen cpu ticks
 568605147842 pages paged in
      1732174 pages paged out
            0 pages swapped in
            0 pages swapped out
   2029175042 interrupts
    511479492 CPU context switches
   1596499546 boot time
    157496114 forks
-bash-4.2# vmstat  -d
disk- ------------reads------------ ------------writes----------- -----IO------
       total merged sectors      ms  total merged sectors      ms    cur    sec
sda    88028   1062 10050341   58080  56715  45830 3443116  767716      0     39
nvme0n1 128674820      0 65068846024 144808949 132348455      0 63125218392 360251571
0  26969
nvme1n1     13      0     104      32      0      0       0       0      0      0
dm-0  128674732      0 65068841864 153025265 132348455      0 63125218392 378003296
0  27384
nvme2n1 122901458      0 62110075824 225460695 126158641      0 60172560272 370097237
0  27648
```

```
nvme3n1     13      0     104      31      0      0       0       0      0      0
dm-1  122901370      0 62110071664 233052112 126158641      0 60172560272 387762100
0  27977
nvme4n1 112167345      0 56659895288 203765878 115353973      0 54803717480 388520855
0  28676
nvme5n1     13      0     104      37      0      0       0       0      0      0
dm-2  112167257      0 56659891128 210720613 115353973      0 54803717480 404723786
0  28985
-bash-4.2# vmstat 2 6
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 1  0      0 80209456 168972 6339320    0    0 12978     0    0    1  0  4 94  2  0
 0  0      0 80209520 168972 6339320    0    0    12     0  246  577  0  0 100  0  0
 0  0      0 80209520 168972 6339320    0    0    12     0  177  485  0  0 100  0  0
 0  0      0 80209552 168972 6339320    0    0    12     0  224  562  0  0 100  0  0
 0  0      0 80209536 168972 6339320    0    0    12     0  202  500  0  0 100  0  0
 0  0      0 80209552 168972 6339320    0    0    12     0  226  569  0  0 100  0  0
-bash-4.2# vmstat -S k 1 10
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 1  0      0 82135008 173027 6491463    0    0 12978     0    0    1  0  4 94  2  0
 0  0      0 82134840 173027 6491463    0    0    12     0  220  517  0  0 100  0  0
 0  0      0 82134576 173027 6491463    0    0    12     0  208  549  0  0 100  0  0
 0  0      0 82134592 173027 6491463    0    0    12     0  224  581  0  0 100  0  0
 0  0      0 82134584 173027 6491463    0    0    12     0  193  510  0  0 100  0  0
 0  0      0 82134584 173027 6491463    0    0    12     0  225  562  0  0 100  0  0
 0  0      0 82134616 173027 6491463    0    0    12     0  244  602  0  0 100  0  0
 0  0      0 82134632 173027 6491463    0    0    12     0  277  622  0  0 100  0  0
 0  0      0 82134648 173027 6491463    0    0    12     0  168  484  0  0 100  0  0
 0  0      0 82134632 173027 6491463    0    0    12     0  186  488  0  0 100  0  0
```