FlexSDS Scale-out Storage Deployment

Wednesday, February 15, 2023

FlexSDS Software Limited.

www.flexsds.com

Table of Contents

Overview

FlexSDS is a reliable, high-availability, high-performance, and distributed block-level storage solution that provides true scale-out capabilities ranging from 1 to 1024 nodes, making it the lowest TCO option available. It is a Server-SAN solution offers full functionality of storage service over iSCSI, iSER, and NVMe over Fabrics transport protocols, and is fully compatible with various virtualization platforms such as VMware, Hyper-V, Citrix XenServer, and OpenStack-QEMU-KVM based cloud and HCI systems.

FlexSDS is optimized for all-flash appliances and fully supports NVMe kernel-bypass and RDMA, enabling 100% utilization of hardware performance. It also supports traditional hardware such as TCP and SATA/SAS disks.

Topology

FlexSDS to be a scale-out software defined storage, that supports three topology modes:

- Single node scale up SDS, this mode supports only in node data redundancy
- Dual-nodes HA SDS, this mode is the minimum requires to support HA service, that provides not only in-node data redundancy, but also cross-node data redundancy.
- Tree or more nodes scale-out clustered SDS, this mode is true scale-out/scale-up clustered storage, that supports all types of data redundancy.

Single node mode

When working in single node mode, FlexSDS supports n-way replication and EC protection between in-node devices. However, since single node does not offer high availability, the storage service will fail if the single node fails. Therefore, single node mode is not recommended for most production cases. However, it can be used to test and prove solutions, and is a good starting point for upgrading to a two node HA or three node clustered setup. By adding new server nodes, it's easy to expand the storage service to a dual node HA or 3+ node cluster.

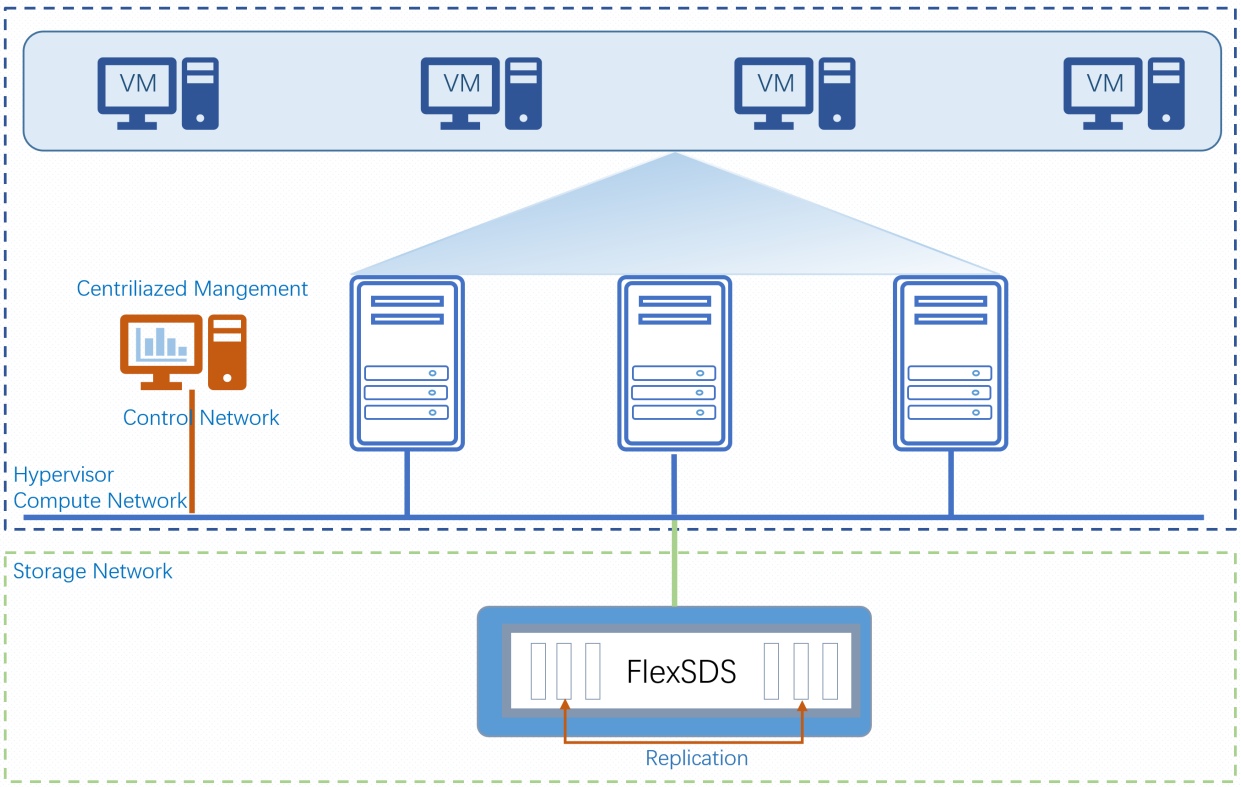


Figure 1. Single Node SDS

Dual nodes HA mode

FlexSDS does not work in true 2-node mode as a software-only solution can't prevent brain-split completely. Therefore, a third-party arbitration node is required. The arbitration node can be another server node within the same network or a virtual machine running on business servers like ESX, QEMU-KVM, and even a public web service. The arbitration node requires network function to communicate with the two nodes in HA, but no other requirements (CPU, storage, etc.) are necessary.

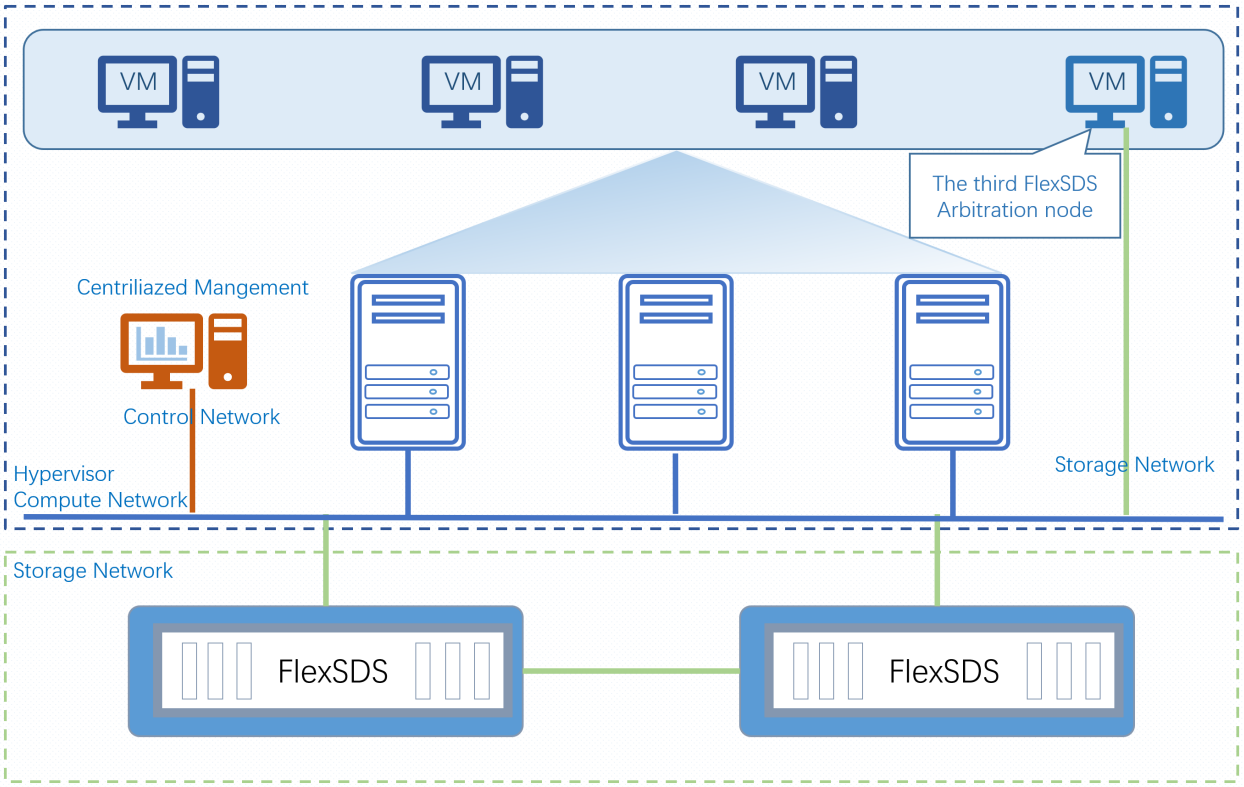


Figure 2. Tow Nodes HA SDS

3(+) Nodes Scale-out SDS mode

Since FlexSDS was designed as a scale-out, distributed storage service, the most commonly used working mode is with 3 or more nodes. With this configuration, users can dynamically add or remove nodes to expand or shrink the storage cluster.

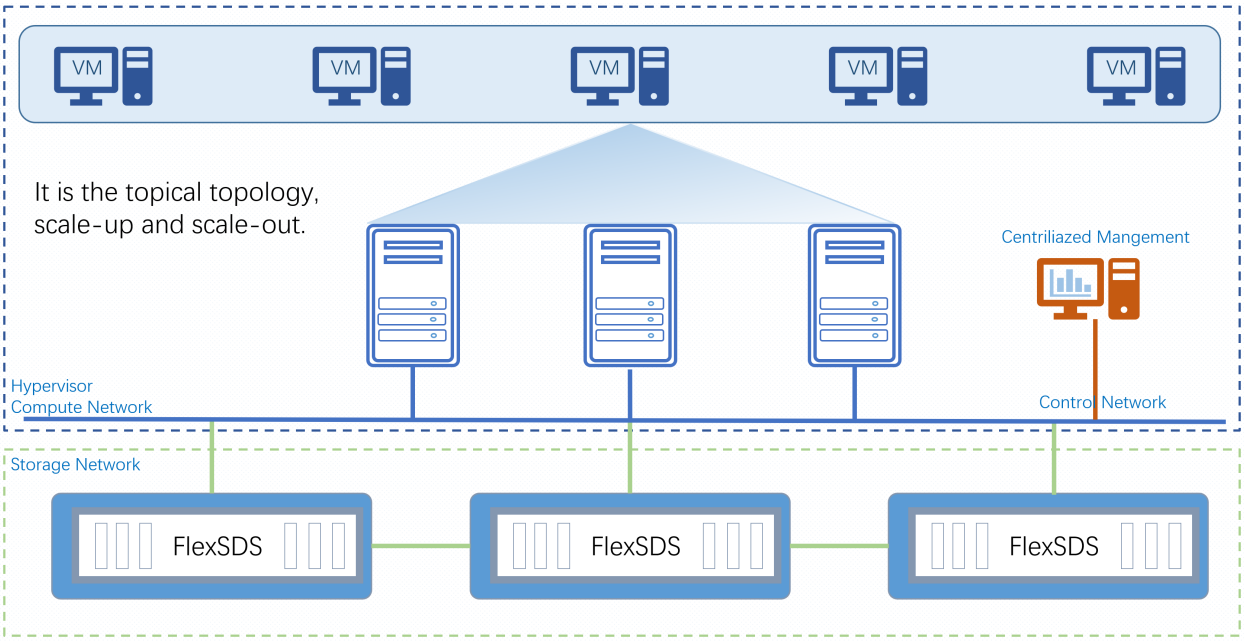


Figure 3. 3 Nodes Scale-out Clustered SDS

Install the FlexSDS Software

FlexSDS is Linux based, the CentOS and Ubuntu Server are the recommended OS to install FlexSDS Scale-out Storage, install FlexSDS software in Linux is very easy.

Prepare Environment

Configure firewall to allow the following ports:

FlexSDS Cluster: TCP 6260, UDP 6260 (Inbound, Outbound)

FlexSDS Manage: TCP 6261 (Inbound, Outbound)

iSCSI: TCP 3260 (Inbound, Outbound for remote mirror)

iSER: TCP and RDMA 3260 (Inbound, Outbound for remote mirror)

NVMe-oF: TCP 4420 (Inbound, Outbound for remote mirror)

or stop it in testing mode:

service firewalld stop

Install depended on packages, we need some packages to be installed like libibverbs:

yum install -y librdmacm

yum install -y libibverbs

In Debian series system, run:

apt-get install -y librdmacm

apt-get install -y libibverbs

Install flexsds software stack

rpm -ivh flexsds-5.6-1.el7.centos.x86_64.rpm

In Debian series system, run:

dpkg -i flexsds-5.6-1.el7.deb

Note, after installed FlexSDS Software, user must configure it before to use it.

Deploying the FlexSDS

Deploy the Scale-out storage cluster, deployment scripts:
Executable: /opt/flexsds/bin/deploy.sh
Parameters:
-m, to specify working mode, “cluster” for multiple nodes storage cluster, “self” for single mode, “cluster” for preparing a member for adding it to a cluster as well.
-h, host list, i.e., 192.168.80.101,192.168.80.102,192.168.80.102, the IP list should be in storage (not management) network.

-c, CPU core count, default is 1, user can set to specific count of CPU core to gain more performance, please confirm if the machine have enough CPU cores.
-r, enable RDMA in crossing nodes data transfer.
-s, configuring ssh, only need once in a storage cluster setting up. If not specify this parameter, password inputting will be required in every time.

Deploying Single Node Scale-up SDS

/opt/flexsds/bin/deploy.sh -m cluster -h 192.168.80.104 -c 1 -r

To deploy single node, mode could be self or cluster, cluster means can be expanded into multiple nodes in future, therefor cluster mode is recommended.

Deploying two nodes HA SDS

Deploying a two-node HA configuration in FlexSDS involves utilizing a lightweight witness service. This service can operate within a resource-efficient VM or on any machine within the Ethernet or internet. FlexSDS witness supports x86 and ARM-based systems running Linux or Windows.

To deploy cluster nodes (execute the following command on one node):

/opt/flexsds/bin/deploy.sh -m cluster -h 192.168.80.101,192.168.80.102 -c 1 -r

Install witness node, we take 192.168.80.103 as an example.

rpm -ivh flexwitness-5.6-1.el7.x86_64.rpm

In Debian series system, run:

dpkg -i flexwitness-5.6_amd64.deb

After those setups finished, user can check storage cluster status and witness service status.

flexsds node list

service flexwitness status (on witness node)

Then user can issue the command to enable witness on the two nodes cluster:

flexsds witness enable –url https://192.168.80.103:8080

The default port of witness is 8080, user can modify it by editing /etc/flexwitness.conf

Check witness is working correctly, run the follow command to see search nodes' ttl is keep updating every a few seconds.

flexsds witness status

Deploying 3(+) nodes Scale-out SDS

deploy cluster member node, the new node is 192.168.80.104 as an example, and need to specify current cluster nodes list: 192.168.80.101,192.168.80.102,192.168.80.103

/opt/flexsds/bin/deploy.sh -m cluster -h 192.168.80.101,192.168.80.102,192.168.80.103 -c 1 -r

Waiting for finish and checking status

it will take in less than 1 minute, the following message will be shown one or more times (for multiple nodes):

Installing flexsdsd...

**********setup flexsds**********

Copying files...

Setting up service...

Starting service...

Starting flexsdsd (via systemctl): [OK]

**********finished**********

After deployment finished, user can check its working status by:

service flexsdsd status

to check service is setting up correctly.

To check cluster nodes online status:

flexsds node list

Note, for single mode, the node list will show the only self-node.

Install WEB management platform (management node)

The WEB management platform can be installed to any one of the storage cluster nodes and it can be installed into separated node which is outside of the cluster as well.

rpm -ivh flexsds-web-5.6-1.el7.centos.x86_64.rpm

In Debian series system, run:

dpkg -i flexsds-web-5.6-1.el7.deb

Step 7. Launch the WEB management platform

Launch any Internet Browser and navigate to the URL:
http://server_ip_of_web_platform/

Default credentials are:
root
flexsds

System Requires

Software requires:

- Linux with kernel >= 3.10
- Recommend OS: CentOS 7.x and Ubuntu Server 18+.
- x64 based system.
- Network Connections: RDMA (InfiniBand, RoCE, iWarp) or Ethernet (TCP/IP).

Hardware requires:

Intel Xeon class processor or similar.

4 cores and more.

16 GB of RAM.

At least one additional disk for storage.

Recommend Configuration (All Flash):

Intel Xeon 2680 v4 2.4GHZ x 2, 28 cores in total.

64 GB DDR4 memory.

NVMe Disks x 8

CentOS 7.2, with 3.10 kernel.

InfiniBand Network.

Initially Setup

Users can use this guide for a cost-effective initial setup (creating storage pool and volumes) of FlexSDS. For detailed setup instructions, please refer to the FlexSDS user's manual or the knowledge base.

Add Disks to Backends

User can add disk to backends via normal interface or user mode (kernel-bypass) interface:

Normal interface, traditional SCSI/SAS/ATA device, PCI NVMe device is supported but it is slower than user mode interface.

```
#flexsds backend add –disk /dev/sdb
```

User mode interface, only support PCI NVMe disks.

Check device's PCI path:

```
#lspci | grep Non
```

Add NVMe disk to backend.

```
#flexsds backend add –disk nvme://0000:86:00.0
```

Create Storage Pool

Storage Pool is clustered ant floating over the whole cluster, it is distributed and scale-out. User can create multiply copies (n-ways replication) and EC (erase codes) pool.

Multiple copies pool, create 2 copies storage pool,

```
#flexsds stor_pool create -c 2 -n data-pool
```

Requires of creating N copies storage pool:

1, disk count is N or more on single node mode.

2. Nodes (has free spaces) count is equal or more than N.

EC pool, create a 4+2 storage pool

```
#flexsds stor_pool create -c 4+2 -n ec-pool
```

And waiting for the storage pool become healthy.

```
#flexsds stor_pool list
```

Requires of creating N+M EC storage pool:

1, disk count is N+M or more on single node mode.

2. Nodes (has free spaces) count is equal or more than N+M

Create Volume

FlexSDS supports 3 type of volumes, RAW, Thin and Log Structured, we suggest user to create RAW volumes at the beginning.

Create a RAW volume and check:

```
#flexsds volume create -p data-pool -n vol1 -s 100G -f raw
#flexsds volume list -p data-pool
```

RAW volume will be exported to all available nodes (for HA) and automatically full filled, use all its capacity of data spaces.

Attach Volume Interface

Attach interface, volumes need to be exported over NVMe-oF, iSER, iSCSI or vHost.

Export via NVMe-oF

```
#flexsds volume attach -p data-pool -v vol1 -i nvmf -n nqn.2016-12.com.flexsds:data-pool.vol1
```

Export via iSER and iSCSI

```
#flexsds volume attach -p data-pool -v vol1 -i iser -n iqn.2016-12.com.flexsds:data-pool.vol1

#flexsds volume attach -p data-pool -v vol1 -i iscsi -n iqn.2016-12.com.flexsds:data-pool.vol1
```

Setup High Availability

FlexSDS is a distributed, scale-out software defined storage solution with built-in high availability as a key feature. This allows for creating volumes that can be exported over multiple nodes. In the event of a node failure, other nodes will continue to operate, and data recovery will automatically occur when the failed node becomes ready, all without any human intervention required.

Setting up high availability in FlexSDS is simple, as all volumes support it automatically, there is nothing additional configuration required.

Multipath Mode

Until version 5.x, FlexSDS supported three types of volumes: RAW volumes, thin and logged volumes. And two type of storage pools: n-ways replicated (multi-copies) and EC (erase code) protected.

We can split them as two types:

Multi-copies RAW volumes:

They do not have a volume controller (a component to control data path) and use spaces of all their capacity, making them the fastest. They support **active-active** high availability.

metadata-managed volumes

Thin and logged volumes, as well as volumes in the EC storage pool, are metadata-managed volumes that have primary and secondary controllers assigned to them. The controllers are switched automatically in response to heavy client loading, and ANA-optimized and non-optimized paths are also adjusted automatically. When the primary and secondary controllers are switched, I/O is pending for a short time, so **active-passive** mode is recommended to avoid I/O pending.

FlexSDS supports iSCSI, iSER, and NVMe-oF targets, with some differences between using NVMe-oF and iSCSI/iSER. High availability can be easily set up via NVMe-oF by ensuring compatibility with the kernel version (>= 5.5) and taking advantage of FlexSDS's support for NVMe ANA. On the other hand, setting up high availability for iSCSI/iSER requires installation of the multi-path (device-mapper-multipath) package.

NVMe-oF ANA High Availability

FlexSDS utilizes NVMe-oF technology for high-performance storage. It leverages Asymmetric Namespace Access (ANA) to optimize namespace access for initiators. With active-active mode, all paths are advertised as "Optimized" for fast service times, and there are no performance penalties as there is no significant difference between primary and secondary controllers. With active-passive mode, the FlexSDS system reports "Non-Optimized" paths on the passive path, where secondary controllers are working. This ensures seamless and efficient storage access, resulting in high availability and reliability for your data.

Prepare

Requirements:

Linux Kernel 4.8 and newer.

Package: nvme-cli

Issue the command to install them:

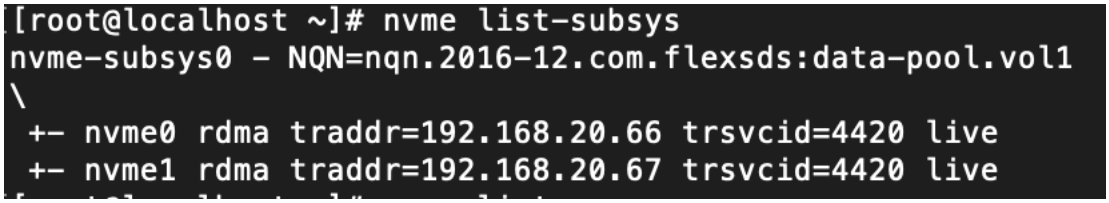
```
#yum install nvme-cli
```

Connect to FlexSDS's NVMe-oF targets

```
#nvme connect -t rdma -a 192.168.20.120 -s 4420 -n nqn.2016-12.com.flexsds:data-pool.vol1
```

Issue the command to check multipath status, multiple NVMe devices should be bound to one subsys.

```
#nvme list
#nvme list-subsys
```



Multipath High Availability

As FlexSDS iSCSI/iSER target can only use multipath package for performing high availability, and NVMe-oF target can of course work with multipath as well, here are the steps to set up multipath.

Prepare

Requirements:

Linux Kernel 4.8 and newer.

Package: multipath, nvme-cli (required only for nvme-of)

Issue the command to install them:

```
#yum install device-mapper-multipath nvme-cli
```

Connect to FlexSDS's any targets

User can use iSCSI, iSER or NVMe-oF. User can follow above step to connect to NVMe-oF target, here use iSER as an example (-I iser for iSER, otherwise iSCSI).

```
#iscsiadm -m discovery -t st -p 192.168.20.66 -I iser
#iscsiadm -m discovery -t st -p 192.168.20.67 -I iser
# iscsiadm -m node -T iqn.2016-12.com.flexsds:data-pool.vol1 -l -p 192.168.20.66 -I iser
# iscsiadm -m node -T iqn.2016-12.com.flexsds:data-pool.vol1 -l -p 192.168.20.67 -I iser
```

Configure the multipath.

create or modify /etc/multipath.conf, make its content to be:

```
defaults {
user_friendly_names yes
find_multipaths no
path_grouping_policy multibus
failback immediate
no_path_retry fail
uid_attribute ID_WWN
}

blacklist_exceptions {
property “(ID_WWN|SCSI_IDENT_.*|ID_SERIAL|DEVTYPE)”
devnode “nvme*”  #for nvme-of
}
```

For FlexSDS NVMe-oF devise, user need to do the following step to configure udev.

1. Add the following to udev rules under /etc/udev/rules.d/<any_udev_rules_file> or /lib/udev/rules.d/60-persistent-storage.rules:

```
KERNEL==”nvme*[0-9]n*[0-9]”, ENV{DEVTYPE}==”disk”, ATTRS{wwid}==”?*”, ENV{ID_WWN}=”$attr{wwid}”
```

2. load the new udev rules run:

udevadm control –reload-rules && udevadm trigger

Restart multipathd or reconfigure it to take effect:

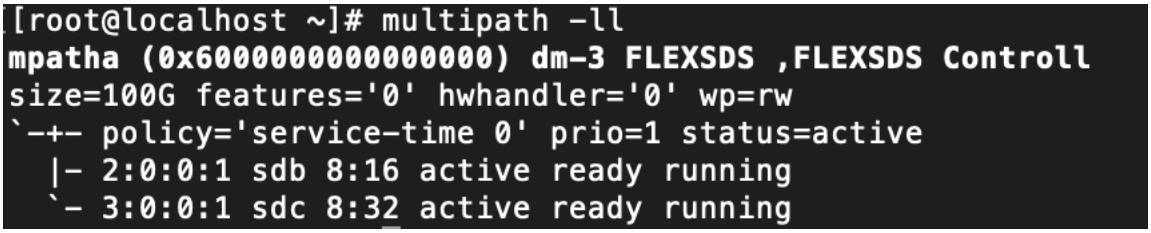
#service multipathd restart

or

#multipath && multipathd reconfigure

Now user should see the new device mapper by running:

multipath -ll



Contact

Support:	support@flexsds.com
Sales:	sales@flexsds.com
Home Page:	http://www.flexsds.com/
Product Page:	https://www.flexsds.com/scale-out-storage/
Purchase	https://www.flexsds.com/pricing/
Knowledge Base:	https://www.flexsds.com/support/kb/

FlexSDS Software Limited.

www.flexsds.com