Helpful HyperFlex General Knowledge

Contents

Introduction

This document describes general knowledge about Cisco HyperFlex (HX) that administrators should have at their fingertips.

Commonly Used Acronyms

SCVM = **S**torage **C**ontroller **V**irtual **M**achine

VMNIC = **V**irtual **M**achine **N**etwork **I**nterface **C**ard

VNIC = **V**irtual **N**etwork **I**nterface **C**ard

SED = **S**elf **E**ncrypting **D**rive

VM = **V**irtual **M**achine

HX = **H**yperFle**x**

HyperFlex VMware VMNIC Ordering

VMNIC placement has been revised in HX version 3.5 and later.

Pre 3.5 Ordering

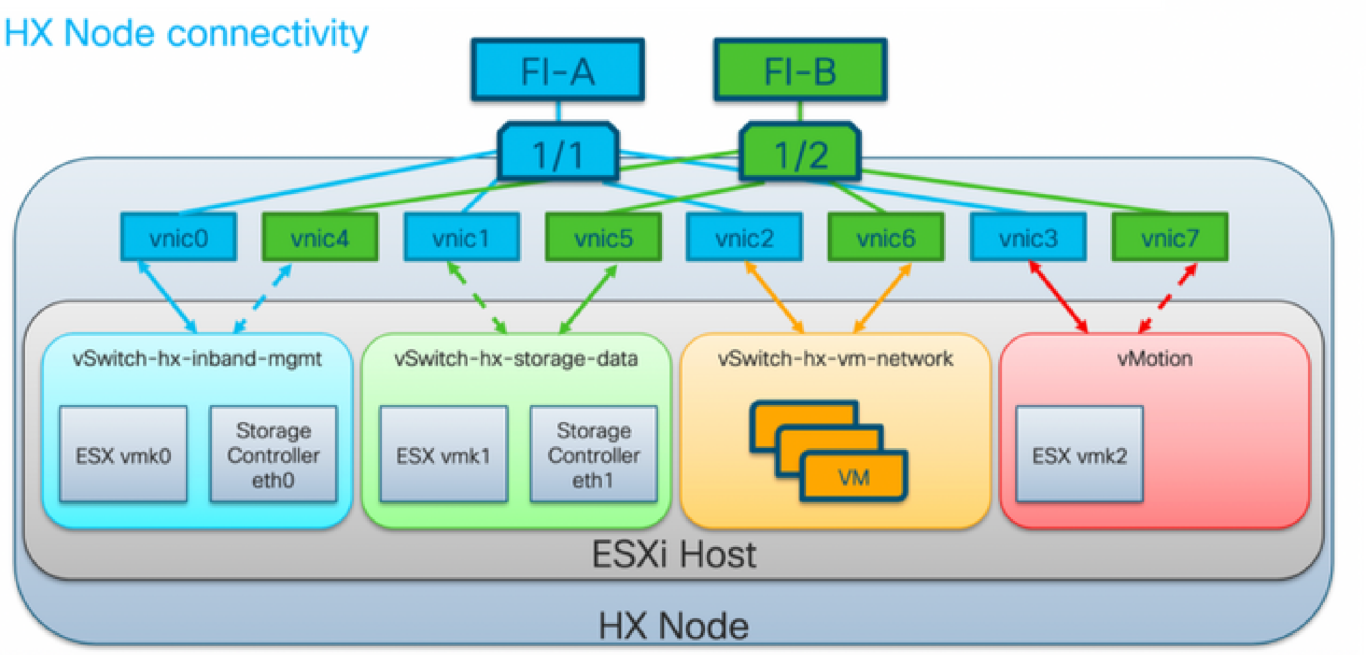
Prior to version 3.5, the VNICs were assigned based on the VNIC numbers.

VNIC	Virtual Switch (vSwitch)
VNIC 0 & VNIC 1	vSwitch-hx-inband-mgmt
VNIC 2 & VNIC 3	vSwitch-hx-storage-data
VNIC 4 & VNIC 5	vSwitch-hx-vm-network
VNIC 6 & VNIC 7	vMotion

Post 3.5 Ordering

In version 3.5 and later, the VNICs are assigned based on Media Access Control (MAC) address. Therefore, there is no particular order of assignment.

If an upgrade from a version older than 3.5 to 3.5 or higher is performed, the VMNIC ordering is maintained.



Note: For HX Hyper-V, this will not be applicable since Hyper-V uses Consistent Device Naming (CDN).

SCVM on Converged Node versus Compute Node

SCVMs reside on both Converged and Compute Nodes and there are differences between them.

Converged Node

CPU Resource Reservations

Since the SCVMs provide critical functionality of the Cisco HX Distributed Data Platform, the HyperFlex installer will configure CPU resource reservations for the controller VMs. This reservation guarantees that the controller VMs will have central processing unit (CPU) resources at a minimum level, in situations where the physical CPU resources of the ESXi hypervisor host are being heavily consumed by the guest VMs. This is a soft guarantee, meaning in most situations the SCVMs are not using all of the CPU resources reserved, therefore allowing the guest VMs to use them. The following table details the CPU resource reservation of the storage controller VMs:

Number of vCPU	Shares	Reservation	Limit
8	Low	10800 MHZ	Unlimited

Memory Resource Reservations

Since the SCVMs provide critical functionality of the Cisco HX Distributed Data Platform, the HyperFlex installer will configure memory resource reservations for the controller VMs. This reservation guarantees that the controller VMs will have memory resources at a minimum level, in situations where the physical memory resources of the ESXi hypervisor host are being heavily consumed by the guest VMs. The following table details the memory resource reservation of the storage controller VMs:

Server Models	Amount of Guest Memory	Reserve All Guest Memory
HX 220c-M5SX HXAF 220c-M5SX HX 220c-M4S HXAF220c-M4S	48 GB	Yes
HX 240c-M5SX HXAF 240c-M5SX HX240c-M4SX HXAF240c-M4SX	72 GB	Yes
HX240c-M5L	78 GB	Yes

Compute Node

The compute-only nodes have a lightweight SCVM. It is configured with only 1 vCPU of 1024MHz and 512 MB of memory reservation.

The purpose of having the compute node is mainly for maintaining the vCluster Distributed Resource Scheduler™ (DRS) settings, to ensure that DRS does not move the user VMs back to converged nodes.

Unhealthy Cluster Scenarios

A HX cluster can be rendered unhealthy in the following scenarios.

Scenario 1: Node Down

A cluster goes into an unhealthy state when a node goes down. A node is expected to be down during a cluster upgrade or when a server is put into maintenance mode.

```
root@SpringpathController:~# stcli cluster storage-summary --detail
<snip>
current ensemble size:3
```

```
# of caching failures before cluster shuts down:2
minimum cache copies remaining:2
minimum data copies available for some user data:2
current healing status:rebuilding/healing is needed, but not in progress yet. warning:
insufficient node or space resources may prevent healing. storage node 10.197.252.99is either
down or initializing disks.
minimum metadata copies available for cluster metadata:2
# of unavailable nodes:1
# of nodes failure tolerable for cluster to be available:0
health state reason:storage cluster is unhealthy.storage node 10.197.252.99 is unavailable.
# of node failures before cluster shuts down:2
# of node failures before cluster goes into readonly:2
# of persistent devices failures tolerable for cluster to be available:1
# of node failures before cluster goes to enospace warn trying to move the existing data:na
# of persistent devices failures before cluster shuts down:2
# of persistent devices failures before cluster goes into readonly:2
# of caching failures before cluster goes into readonly:na
# of caching devices failures tolerable for cluster to be available:1
resiliencyInfo:
messages:
----------------------------------------
Storage cluster is unhealthy.
----------------------------------------
Storage node 10.197.252.99 is unavailable.
----------------------------------------
state: 2
nodeFailuresTolerable: 0
cachingDeviceFailuresTolerable: 1
persistentDeviceFailuresTolerable: 1
zoneResInfoList: None
spaceStatus: normal
totalCapacity: 3.0T
totalSavings: 5.17%
usedCapacity: 45.9G
zkHealth: online
clusterAccessPolicy: lenient
dataReplicationCompliance: non_compliant
dataReplicationFactor: 3
```

Scenario 2: Disk Down

A cluster goes into an unhealthy state when a disk is unavailable. The condition should clear when the data is distributed to other disks.

```
root@SpringpathController:~# stcli cluster storage-summary --detail
<snip>
current ensemble size:3
# of caching failures before cluster shuts down:2
minimum cache copies remaining:2
minimum data copies available for some user data:2
current healing status:rebuilding/healing is needed, but not in progress yet. warning:
insufficient node or space resources may prevent healing. storage node is either down or
initializing disks.
minimum metadata copies available for cluster metadata:2
# of unavailable nodes:1
# of nodes failure tolerable for cluster to be available:0
health state reason:storage cluster is unhealthy. persistent device disk
[5000c5007e113d8b:0000000000000000] on node 10.197.252.99 is unavailable.
# of node failures before cluster shuts down:2
# of node failures before cluster goes into readonly:2
# of persistent devices failures tolerable for cluster to be available:1
# of node failures before cluster goes to enospace warn trying to move the existing data:na
# of persistent devices failures before cluster shuts down:2
# of persistent devices failures before cluster goes into readonly:2
# of caching failures before cluster goes into readonly:na
# of caching devices failures tolerable for cluster to be available:1
resiliencyInfo:
messages:
----------------------------------------
Storage cluster is unhealthy.
----------------------------------------
Persistent Device Disk [5000c5007e113d8b:0000000000000000] on node 10.197.252.99 is unavailable.
----------------------------------------
state: 2
nodeFailuresTolerable: 0
cachingDeviceFailuresTolerable: 1
persistentDeviceFailuresTolerable: 1
zoneResInfoList: None
spaceStatus: normal
totalCapacity: 3.0T
totalSavings: 8.82%
usedCapacity: 45.9G
zkHealth: online
clusterAccessPolicy: lenient
dataReplicationCompliance: non_compliant
dataReplicationFactor: 3
```

Scenario 3: Neither Node Nor Disk Down

A cluster can go into an unhealthy state when neither a node nor a disk is down. This condition occurs if rebuilding is in progress.

```
root@SpringpathController:~# stcli cluster storage-summary --detail
<snip>
resiliencyDetails:
        current ensemble size:5
        # of caching failures before cluster shuts down:3
        minimum cache copies remaining:3
        minimum data copies available for some user data:2
current healing status:rebuilding is in progress, 98% completed.         minimum metadata copies
available for cluster metadata:2
        time remaining before current healing operation finishes:7 hr(s), 15 min(s), and 34
sec(s)
        # of unavailable nodes:0
        # of nodes failure tolerable for cluster to be available:1
        health state reason:storage cluster is unhealthy.
        # of node failures before cluster shuts down:2
        # of node failures before cluster goes into readonly:2
        # of persistent devices failures tolerable for cluster to be available:1
        # of node failures before cluster goes to enospace warn trying to move the existing
data:na
        # of persistent devices failures before cluster shuts down:2
        # of persistent devices failures before cluster goes into readonly:2
       # of caching failures before cluster goes into readonly:na
        # of caching devices failures tolerable for cluster to be available:2
resiliencyInfo:
    messages:
        Storage cluster is unhealthy.
    state: 2
    nodeFailuresTolerable: 1
    cachingDeviceFailuresTolerable: 2
    persistentDeviceFailuresTolerable: 1
    zoneResInfoList: None
spaceStatus: normal
totalCapacity: 225.0T
totalSavings: 42.93%
usedCapacity: 67.7T
clusterAccessPolicy: lenient
dataReplicationCompliance: non_compliant
dataReplicationFactor: 3
```

How to check for a SED Cluster using the Command Line Interface (CLI)

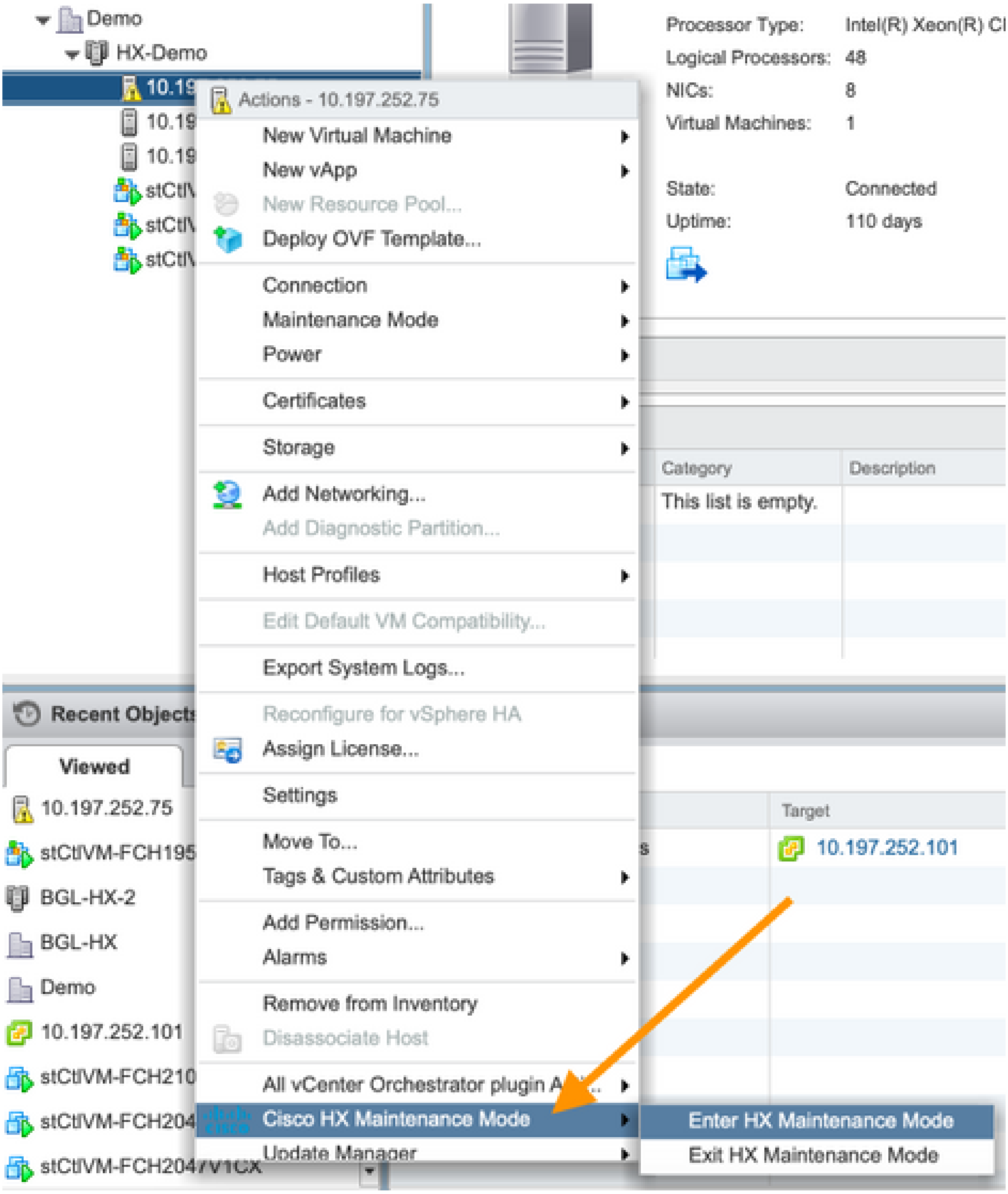
If access to HX Connect is not available, the CLI can be used to check if the cluster is SED.

```
# Check if the cluster is SED capable
root@SpringpathController:~# cat /etc/springpath/sed_capability.conf
sed_capable_cluster=False

# Check if the cluster is SED enabled root@SpringpathController:~# cat /etc/springpath/sed.conf
sed_encryption_state=unknown

root@SpringpathController:~# /usr/share/springpath/storfs-appliance/sed-client.sh -l
WWN,Slot,Supported,Enabled,Locked,Vendor,Model,Serial,Size
5002538c40a42d38,1,0,0,0,Samsung,SAMSUNG_MZ7LM240HMHQ-00003,S3LKNX0K406548,228936
5000c50030278d83,25,1,1,0,MICRON,S650DC-800FIPS,ZAZ15QDM0000822150Z3,763097
500a07511d38cd36,2,1,1,0,MICRON,Micron_5100_MTFDDAK960TCB_SED,17261D38CD36,915715
500a07511d38efbe,4,1,1,0,MICRON,Micron_5100_MTFDDAK960TCB_SED,17261D38EFBE,915715
500a07511d38f350,7,1,1,0,MICRON,Micron_5100_MTFDDAK960TCB_SED,17261D38F350,915715
500a07511d38eaa6,3,1,1,0,MICRON,Micron_5100_MTFDDAK960TCB_SED,17261D38EAA6,915715
500a07511d38ce80,6,1,1,0,MICRON,Micron_5100_MTFDDAK960TCB_SED,17261D38CE80,915715
500a07511d38e4fc,5,1,1,0,MICRON,Micron_5100_MTFDDAK960TCB_SED,17261D38E4FC,915715
```

HX Maintenance Mode versus ESXi Maintenance Mode

When maintenance activities need to be performed on a server that is part of a HX cluster, HX Maintenance Mode should be used instead of ESXi Maintenance Mode. The SCVM is gracefully shut down when HX Maintenance Mode is used while it is abruptly shut when ESXi Maintenance Mode is used.

While a node is in maintenance mode it will be considered down, that is, 1 node failure.



Processor Type: Intel(R) Xeon(R) CI
Logical Processors: 48
NICs: 8
Virtual Machines: 1
State: Connected
Uptime: 110 days

Ensure the cluster shows as healthy before moving another node into maintenance mode.

```
root@SpringpathController:~# stcli cluster storage-summary --detail
<snip>
current ensemble size:3
# of caching failures before cluster shuts down:3
minimum cache copies remaining:3
minimum data copies available for some user data:3
minimum metadata copies available for cluster metadata:3
# of unavailable nodes:0
# of nodes failure tolerable for cluster to be available:1
health state reason:storage cluster is healthy.
# of node failures before cluster shuts down:3
# of node failures before cluster goes into readonly:3
# of persistent devices failures tolerable for cluster to be available:2
# of node failures before cluster goes to enospace warn trying to move the existing data:na
# of persistent devices failures before cluster shuts down:3
# of persistent devices failures before cluster goes into readonly:3
# of caching failures before cluster goes into readonly:na
# of caching devices failures tolerable for cluster to be available:2
resiliencyInfo:
messages:
Storage cluster is healthy.
state: 1
nodeFailuresTolerable: 1
cachingDeviceFailuresTolerable: 2
<snip>
```

Frequently Asked Questions

Where are the SCVMs installed on Cisco HyperFlex M4 and M5 Servers?

The SCVM location is different between Cisco Hyperflex M4 and M5 Servers. The table below lists the location of the SCVM and provides other useful information.

Cisco HX Server	ESXi	SCVM sda	Caching Solid State Drive (SSD)	Housekeeping SSD sdb1 and sdb2
HX 220 M4	Secure Digital(SD cards)	3.5G on SD cards	Slot 2	Slot 1
HX 240 M4	SD cards	On PCH controlled SSD (esxi has control of this)	Slot 1	On PCH controlled SSD
HX 220 M5	M.2 Drive	M.2 Drive	Slot 2	Slot 1
HX 240 M5	M.2 Drive	M.2 Drive	Rear slot SSD	Slot 1

How many failed nodes can a cluster tolerate?

The number of failures a cluster can tolerate will depend on the Replication Factor and Access Policy.

Clusters with 5 or more Nodes

When the Replication Factor (RF) is 3 and Access Policy is set to Lenient, if 2 nodes fail the cluster will still be in a Read/Write state. If 3 nodes were to fail, then the cluster will shutdown.

Replication Factor	Access Policy	Number of Failed Nodes		
		Read/Write	Read Only	Shutdo
3	Lenient	2	--	3
3	Strict	1	2	3
2	Lenient	1	--	2
2	Strict	--	1	2

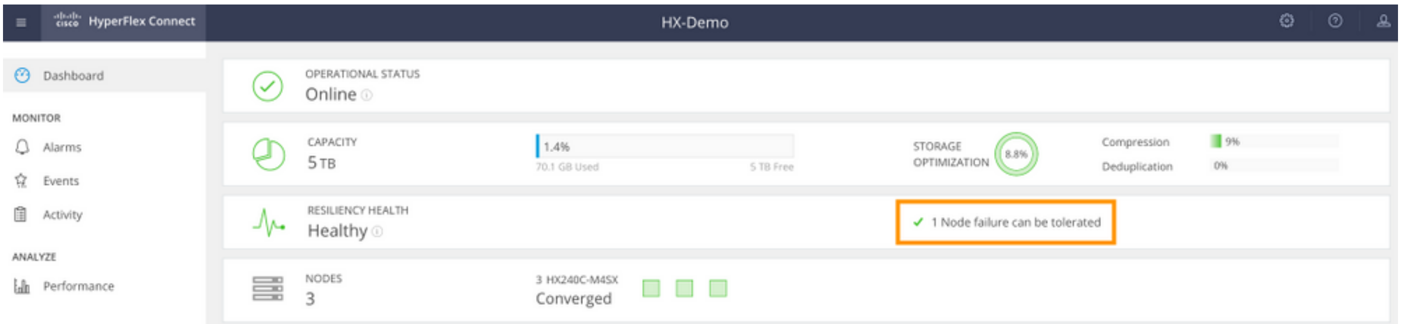
Clusters with 3 and 4 Nodes

When the RF is 3 and Access Policy is set to Lenient or Strict, if a single node fails, the cluster is still be in a Read/Write state. If 2 nodes fail, the cluster will shut down.

Replication Factor	Access Policy	Number of Failed Nodes		
		Read/Write	Read Only	Shutdo
3	Lenient or Strict	1	--	2
2	Lenient	1	--	2
2	Strict	--	1	2

Example of a 3 Node Cluster (RF: 3, Access Policy: Lenient)

Graphical User Interface (GUI) Example



OPERATIONAL STATUS
Online

CAPACITY
5 TB
1.4%
70.1 GB Used
5 TB Free
STORAGE OPTIMIZATION 8.8%
Compression 9%
Deduplication 0%

RESILIENCY HEALTH
Healthy
✓ 1 Node failure can be tolerated

NODES
3
3 HX240C-M4SX
Converged

CLI Example

```
root@SpringpathController:~# stcli cluster storage-summary --detail
<snip>
current ensemble size:3
# of caching failures before cluster shuts down:3
minimum cache copies remaining:3
minimum data copies available for some user data:3
minimum metadata copies available for cluster metadata:3
# of unavailable nodes:0
# of nodes failure tolerable for cluster to be available:1
health state reason:storage cluster is healthy.
# of node failures before cluster shuts down:3
# of node failures before cluster goes into readonly:3
# of persistent devices failures tolerable for cluster to be available:2
# of node failures before cluster goes to enospace warn trying to move the existing data:na
# of persistent devices failures before cluster shuts down:3
# of persistent devices failures before cluster goes into readonly:3
# of caching failures before cluster goes into readonly:na
# of caching devices failures tolerable for cluster to be available:2
resiliencyInfo:
messages:
Storage cluster is healthy.
state: 1
<snip>
clusterAccessPolicy: lenient
```

What happens if one of the SCVMs is shutdown? Do VMs continue to function?

> Warning: This is not a supported operation on a SCVM. This is only for demonstration purposes.

> **Note**: Ensure that only one SCVM is down at a time. Also, ensure that the cluster is healthy before a SCVM is shut down. This scenario is only meant to demonstrate that the VMs and the data stores are expected to function even if a SCVM is down or unavailable.

VMs will continue to work normally. Below is an output example where the SCVM was shutdown, but the datastores remained mounted and available.

```
[root@node1:~] vim-cmd vmsvc/getallvms
Vmid Name File Guest OS Version Annotation
1 stCtlVM-F 9H [SpringpathDS-F 9H] stCtlVM-F 9H/stCtlVM-F 9H.vmx ubuntu64Guest vmx-13

[root@node1:~] vim-cmd vmsvc/power.off 1
Powering off VM:

[root@node1:~] vim-cmd vmsvc/power.getstate 1
Retrieved runtime info
Powered off

[root@node1:~] esxcfg-nas -l
Test is 10.197.252.106:Test from 3203172317343203629-5043383143428344954 mounted available
ReplSec is 10.197.252.106:ReplSec from 3203172317343203629-5043383143428344954 mounted available
New_DS is 10.197.252.106:New_DS from 3203172317343203629-5043383143428344954 mounted available
```

The VMware hardware version on the SCVM has been updated. Now what?

> Warning: This is not a supported operation on a SCVM. This is only for demonstration purposes.

Upgrading the VMware hardware version by editing VM settings in **Compatibility** > **Upgrade VM Compatibility** is the vSphere Web Client is NOT a supported operation on a SCVM. The SCVM will report as Offline in HX Connect.

```
root@SpringpathController0      UE:~# lsblk
NAME   MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
sda      8:0    0  2.5G  0 disk
`-sda1   8:1    0  2.5G  0 part /
sdb      8:16   0  100G  0 disk
|-sdb1   8:17   0   64G  0 part /var/stv
`-sdb2   8:18   0   24G  0 part /var/zookeeper


root@SpringpathController0      UE:~# lsscsi
[2:0:0:0]    disk    VMware   Virtual disk     2.0   /dev/sda
[2:0:1:0]    disk    VMware   Virtual disk     2.0   /dev/sdb


root@SpringpathController0      UE:~# cat /var/log/springpath/diskslotmap-v2.txt
1.11.1:5002538a17221ab0:SAMSUNG:MZIES800HMHP/003:S1N2NY0J201389:EM19:SAS:SSD:763097:Inactive:/de
v/sdc
1.11.2:5002538c405537e0:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
98:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdd
```

```
1.11.3:5002538c4055383a:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
88:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sde
1.11.4:5002538c40553813:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
49:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdf
1.11.5:5002538c4055380e:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
44:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdg
1.11.6:5002538c40553818:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
54:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdh
1.11.7:5002538c405537d1:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
83:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdi
1.11.8:5002538c405537d8:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
90:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdj
1.11.9:5002538c4055383b:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
89:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdk
1.11.10:5002538c4055381f:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
61:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdl
1.11.11:5002538c40553823:Samsung:SAMSUNG_MZ7LM3T8HMLP-00003:S
65:GXT51F3Q:SATA:SSD:3662830:Inactive:/dev/sdm
```

Caution: If this operation was accidentally performed, please call Cisco Support for further assistance. The SCVM will need to be redeployed.