

# ScaleIO/VxFlex OS IP Fabric Best Practice and Deployment Guide with Dell EMC Networking OS10 Enterprise Edition

\*Dell EMC VxFlex OS – formerly ScaleIO

Dell EMC Networking Leaf-Spine Architecture with the S4248FB-ON

Dell EMC Networking Infrastructure Solutions  
May 2018

## Revisions

Date	Version	Description	Authors
May 2018	1.0	Initial release with OS10 content. (Based on <i>ScaleIO IP Fabric Best Practice and Deployment Guide</i> version 2.0)	Gerald Myres, Rutvij Shah, Curtis Bunch, Colin King

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2018 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be the property of their respective owners. Published in the USA May 2018.

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

# Table of contents

- 1 Introduction to network virtualization and hyper-converged infrastructure and networking ..... 6
  - 1.1 VxFlex OS..... 7
    - 1.1.1 Benefits of VxFlex OS ..... 7
    - 1.1.2 Components ..... 8
  - 1.2 Traffic model for the modern data center ..... 9
  - 1.3 Attachments..... 9
- 2 Building a Leaf-Spine topology..... 10
  - 2.1 Management network ..... 11
  - 2.2 IP-based storage ..... 11
  - 2.3 Routing protocol selection for Leaf-Spine ..... 12
    - 2.3.1 BGP ..... 12
    - 2.3.2 OSPF ..... 13
- 3 Configuration and Deployment..... 14
  - 3.1 VxFlex OS solution example ..... 14
    - 3.1.1 VxFlex OS nodes..... 14
    - 3.1.2 VxFlex OS deployment options ..... 14
  - 3.2 Physical switch configuration..... 14
    - 3.2.1 BGP ASN configuration ..... 15
    - 3.2.2 BGP neighbor fall-over ..... 15
    - 3.2.3 Loopback addresses ..... 15
    - 3.2.4 Point-to-point interfaces..... 16
    - 3.2.5 Interface/IP configuration..... 17
    - 3.2.6 ECMP ..... 18
    - 3.2.7 VRRP ..... 18
    - 3.2.8 Flow control ..... 18
  - 3.3 VMware virtual networking configuration..... 19
    - 3.3.1 Create a datacenter object and add hosts ..... 19
    - 3.3.2 Create clusters and add hosts..... 20
    - 3.3.3 Deploy vSphere distributed switch ..... 20
    - 3.3.4 Create the VDS..... 23
    - 3.3.5 Add distributed port groups ..... 24
    - 3.3.6 Create LACP LAGs ..... 26

3.3.7	Associate hosts and assign uplinks.....	28
3.3.8	Configure teaming and failover on LAGs .....	31
3.3.9	Add VMkernel adapters for MDM, SDS-SDC, and vMotion .....	32
3.3.10	Verify VDS configuration .....	34
3.3.11	Enable LLDP .....	35
4	Scaling and tuning guidance .....	37
4.1	Decisions on scaling .....	37
4.2	Examples of scaling, port count, and oversubscription .....	37
4.3	Scaling beyond 16 racks .....	38
4.4	Configure Bandwidth Allocation for System Traffic .....	38
4.5	Tuning Jumbo frames .....	39
4.6	Quality of Service (QoS).....	40
4.6.1	DSCP marking on virtual distributed switches.....	40
4.6.2	Switch QoS configuration .....	41
4.6.3	QoS validation .....	43
A	Additional resources.....	44
A.1	Virtualization components .....	44
A.2	Dell EMC servers and switches.....	44
A.3	Server and switch component details.....	45
A.4	PowerEdge R730xd server.....	45
A.5	S4248FB-ON switch .....	46
A.6	Z9100-ON switch.....	47
A.7	S3048-ON switch.....	47
B	Prepare your environment.....	48
B.1	Confirm CPU virtualization is enabled in BIOS .....	48
B.2	Confirm network adapters are at factory default settings.....	48
B.3	Configure the PERC H730 Controller.....	49
B.4	Install ESXi .....	50
B.5	Configure the ESXi management network connection.....	50
C	VxFlex OS with routed leaf - multiple rack considerations.....	52
C.1	VMkernel configuration - add static routes for default gateways.....	52
C.2	Deploying VxFlex OS .....	52
C.2.1	Registering the VxFlex OS Plug-in and uploading the OVA template.....	53

C.2.2 Installing the SDC on ESXi hosts .....	54
C.2.3 VxFlex OS deployment.....	54
C.2.4 Deployment wizard modifications .....	56
Routed leaf-spine.....	56
Modification steps .....	57

**Note:**

Dell EMC ScaleIO is now named Dell EMC VxFlex OS. Both names are utilized within the body and configuration examples within this document. Although the brand name has changed, the functionality of the software remains the same. The information contained in this document, irrespective of the name, applies to the software release version listed in Appendix A.

# 1 Introduction to network virtualization and hyper-converged infrastructure and networking

With the advent of network virtualization and hyper-converged infrastructure (HCI), the nature of network traffic is undergoing a significant transformation. A dedicated Fibre Channel (FC) network with many of the advanced storage features located on a dedicated storage array is no longer the norm. Storage and application traffic both can reside on Ethernet networks with many advanced storage features executing in a distributed fashion across a network.

Figure 1 shows a relatively simple, four-node/server HCI solution with examples of the expected network traffic running internally and externally to the cluster.

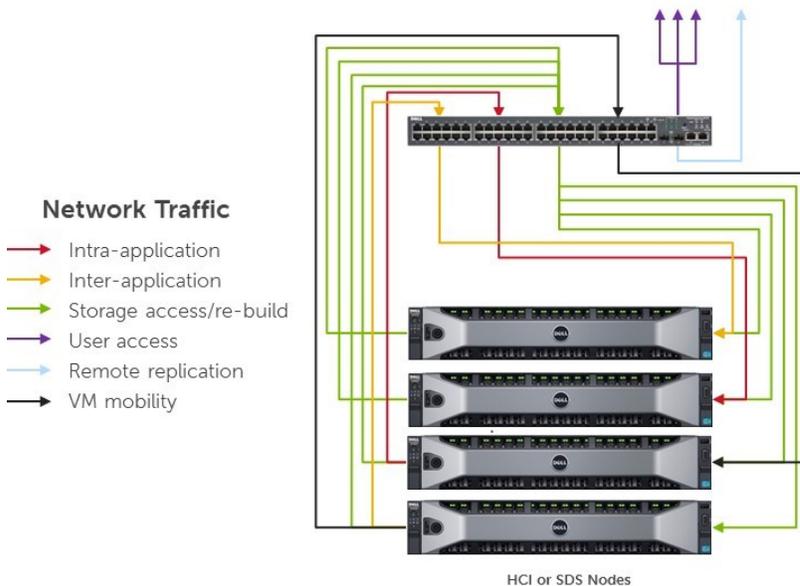


Figure 1 HCI cluster with typical network traffic types

In the early days of software-defined storage and HCI, when most of the workloads were CPU-intensive or memory-intensive, the network was often an afterthought. The increasing I/O and storage-intensive workloads within software-defined storage/HCI solutions place stresses on the network that ultimately impact performance on inappropriately designed networks. Gartner, in recognition of this trend, forecasts 10% of hyper-converged solutions suffering from unavoidable, network-induced performance problems by 2018. Gartner further predicts 25% of hyper-converged solutions being plugged into the wrong tier of the data center network in that same timeframe.

As such, the overall network design plays a larger role in defining the performance of software-defined storage/HCI-based solutions and the applications that run on them. The goal of this document is to identify best practices and apply them on a deployment example for the software-defined storage/HCI solution with Dell EMC VxFlex OS. This best practice guide not only mitigates network-induced performance problems, but also sets the foundation for a network with inherent scalability. This provides a stable and predictable foundation for the future growth of any software-defined storage/HCI solution.

This document defines a VxFlex OS deployment on a Leaf-Spine topology where deployed network-management technologies, advanced routing, and link aggregations are implemented. The paper focuses on a Leaf-Spine network with greater cluster size and workloads, requiring high-speed network ports on leaf switches and interconnection ports for large workloads.

## 1.1 VxFlex OS

Dell EMC VxFlex OS is a software solution that uses existing server storage in application servers to create a server-based storage area network (SAN). This software-defined storage environment gives all member servers access to all unused storage in the environment, regardless of which server the storage is on. VxFlex OS combines different types of storage (hard disk drives, solid-state disks and Peripheral Component Interconnect Express [PCIe] flash cards) to create shared block storage. VxFlex OS is hardware-agnostic and supports physical and/or virtual application servers. By not requiring an FC fabric to connect the servers and storage, VxFlex OS reduces the cost and complexity of a traditional FC SAN.

**Note:** VxFlex OS supports all Ethernet speeds (100Mb, 1Gb, 10Gb, 25Gb, 40Gb, 100Gb), and the InfiniBand (IB) communications standard.

This document supplements existing VxFlex OS documentation. It provides details on how to implement a VxFlex OS solution using Dell EMC products. The following list shows a sample of the VxFlex OS documentation available:

- Quick Start Guide
- Deployment Guide
- Security Configuration Guide
- User Guide

**Note:** Navigate to [Dell EMC ScaleIO/VxFlex OS Software Only: Documentation Library](#) to download the full documentation package and [Dell EMC ScaleIO/VxFlex OS Documentation Hub](#) for additional VxFlex OS documentation. You may need to enter EMC Community Network (ECN) login credentials or create an ECN account to access the documentation package.

### 1.1.1 Benefits of VxFlex OS

The benefits of VxFlex OS include the following:

#### **Scalability**

VxFlex OS scales from a minimum of 3 nodes to a maximum of 1024 nodes. It can add more compute or storage resources when needed without incurring downtime. This allows resources to grow individually or together to maintain balance.

#### **Extreme Performance**

All servers in the VxFlex OS cluster process I/O operations, making all I/O and throughput accessible to any application in the cluster. Throughput and Input/output Operations per Second (IOPS) scale linearly with the number of servers and local storage devices added to the environment. This allows the cost/performance ratio to improve as the environment grows. VxFlex OS automatically performs rebuild and rebalance

operations in the background so that, when necessary, optimization has minimal or no impact to applications and users.

### Compelling Economics

VxFlex OS can reduce the cost and complexity of a typical SAN by allowing users to exploit unused local storage capacity on the servers. This eliminates the need for an FC fabric between servers and storage, as well as additional hardware like Host Bus Adapters.

### Unparalleled Flexibility

The hyper-converged deployment includes the application and storage installed on the same server in the VxFlex OS cluster. This creates a single-tier architecture with the lowest footprint and cost profile. VxFlex OS provides additional flexibility by supporting mixed server brands, operating systems and storage media.

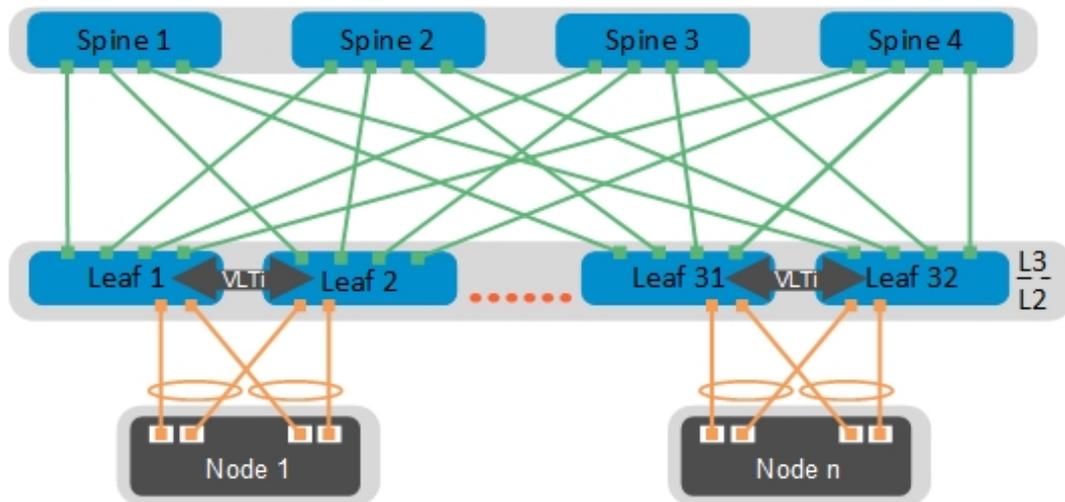


Figure 2 Hyper-converged topology design

### Supreme Elasticity

VxFlex OS enables you to add or remove storage or compute resources dynamically at any time with no downtime. The system automatically rebalances the data "on the fly."

## 1.1.2 Components

VxFlex OS is comprised of three main components.

### VxFlex OS Data Client (SDC)

The SDC is a lightweight, block device-driver that presents VxFlex OS shared block volumes to applications. The SDC runs on the same server as the application. This enables the SDC to fulfill IO requests issued by the application regardless of where the particular blocks physically reside.

### VxFlex OS Data Server (SDS)

The SDS manages the local storage that contributes to the VxFlex OS storage pools. The SDS runs on each of the servers that contribute storage to the VxFlex OS system. The SDS performs the back-end operations that SDCs request.

### Meta Data Manager (MDM)

The MDM configures and monitors VxFlex OS. It contains all the metadata required for VxFlex OS operation. Configure the MDM in Single Mode on a single server or redundant Cluster Mode – three members on three servers or five members on five servers.

**Note:** Use Cluster Mode for all production environments. Dell EMC does not recommend Single Mode because it exposes the system to a single point of failure.

## 1.2 Traffic model for the modern data center

The legacy traffic pattern for data centers has been the classic client-server path and model. The end user sends a request to a resource inside the data center and that resource computes and responds to the end-user traffic. The data center is designed more for the expected North-to-South traffic that travels to and from the data center, rather than the possible East-to-West traffic that traverses between the racks in the data center. The design focuses on transport into and out of the data center and not residential applications. This results in traffic patterns that do not always follow a predictable path due to asymmetrical bandwidth between the networking layers.

The evolution that has taken place is the increased machine-to-machine traffic inside the data center. The reasons for this increased East-West traffic are:

- Applications are much more tiered where web, database and storage interact with each other.
- Increased virtualization where applications easily move to available compute resources.
- Increased server-to-storage traffic due to storage solutions like Dell EMC VxFlex OS that involve hundreds of deployed nodes and require higher bandwidth and scalability.

In large data clusters, VxFlex OS can achieve distribution and replication of all the data. This places a heavy burden on the data center infrastructure since all the data is replicated and that replication happens continuously. This is a clear difference from the old model of one interaction being limited to one North-South communication. Instead, there is a noticeable increase of East-West traffic before any messages are delivered to the end user.

A well thought out design includes predictable traffic paths and well-used resources. By industry consensus, a Leaf-Spine design with an ECMP, load-balanced mesh creates a well-built fabric that both scales and provides efficient, predictable traffic flows. A two-tier Leaf-Spine design provides paths that are never more than two hops away, which results in consistent round-trip time.

## 1.3 Attachments

This main document includes multiple attachment files for spine-switch configurations and leaf-switch configurations. The example configurations are based on a solution assembled in the Dell EMC Networking lab and require editing to fit any other infrastructure.

## 2 Building a Leaf-Spine topology

This paper describes a general-purpose virtualization infrastructure suitable for a modern data center. The solution is based on a Leaf-Spine topology utilizing Dell EMC Networking S4248FB-ON switches for the leaf switches, and Dell EMC Networking Z9100-ON switches for the spine switches. This example topology uses four spine switches to maximize throughput between leaf switches and spine switches.

Figure 3 shows a high-level diagram of the Leaf-Spine topology used in this guide. As a best practice, each new rack added to the data center contains two leaf switches. Join these two switches using a VLT interconnect (VLTi) so the other layer 2 downstream devices see them as a single logical device.

The connections between spine switches and leaf switches can be layer 2 (switched) or layer 3 (routed). The deployment scenario in this guide uses layer 3 connections. The main benefit is layer 2 broadcast domains are limited, resulting in improved network stability and scalability.

The Dell EMC Networking Z9100-ON supports a maximum of 32 leaf switches per pod. However, the example in this document uses only two leaf switches. A single rack can provide VxFlex OS management, VMware management, and several VxFlex OS compute/storage nodes. As administrators add racks to the data center, they add two leaf switches to each rack. As bandwidth requirements increase, administrators add spine switches.

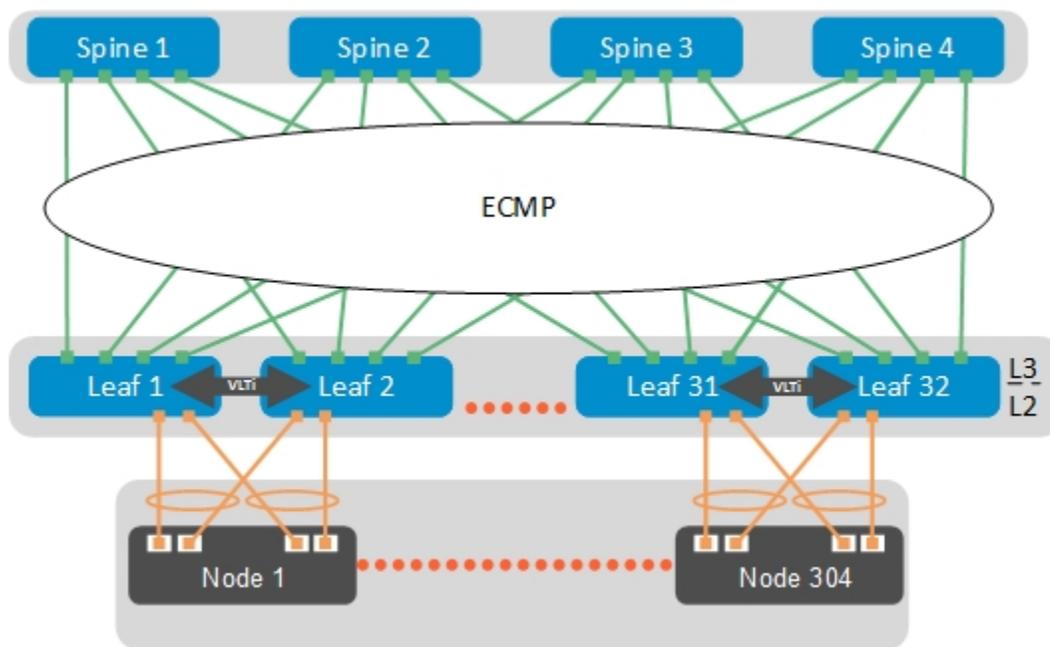


Figure 3 Leaf-Spine topology example

The following physical concepts apply to all routed Leaf-Spine topologies:

- Each leaf switch connects to every spine switch in the topology.
- Spine switches only connect to leaf switches.
- Leaf switches connect to spine switches and other devices such as servers, storage arrays, and edge routers.
- Servers, storage arrays, edge routers and other non-leaf-switch devices never connect to spine switches.
- It is a best practice to use VLT for connecting leaf switch pairs. This provides redundancy at the layer 2 level for devices that use an active-active link aggregation group (LAG) to attach to both switches.

## 2.1 Management network

A management network is not a requirement to setup or configure the Leaf-Spine network. However, Dell EMC recommends using a management network in larger network topologies for efficient management of several devices. Typical deployments use a single management-traffic network isolated from the production network. A Dell EMC S3048-ON switch in each rack can provide connectivity to the management network. Out-of-band (OOB) ports on each production switch connect them to the management network.

For a detailed configuration example see [ScaleIO IP Fabric Best Practice and Deployment Guide](#).

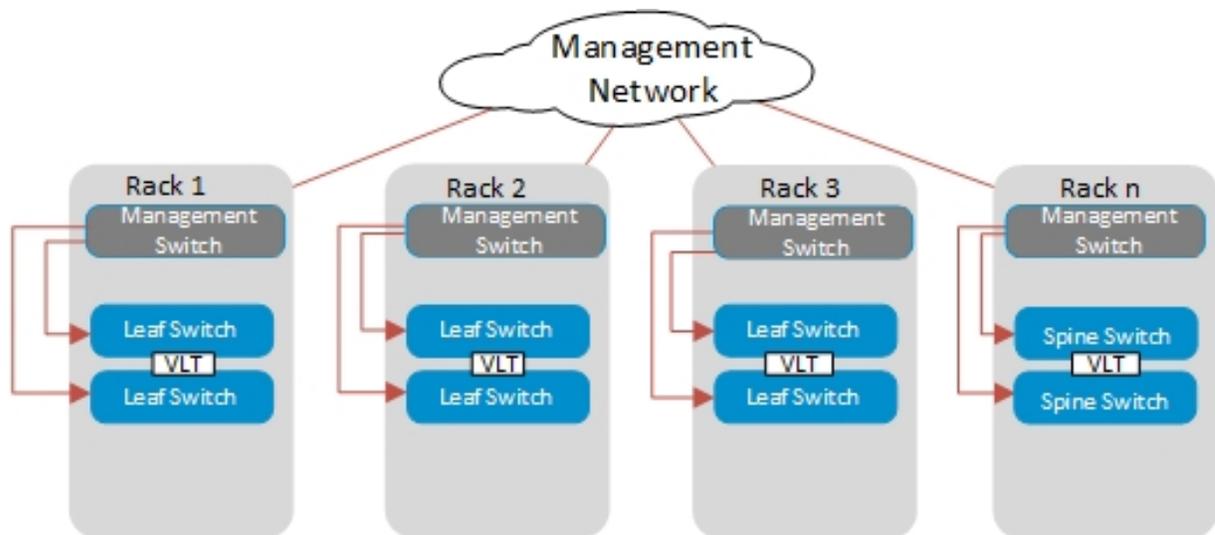


Figure 4 Management network

## 2.2 IP-based storage

One of the benefits of using VxFlex OS in a Leaf-Spine infrastructure is its reachability. Any rack in the data center can reach the solution using IP routing. This allows VxFlex OS to reach hundreds of SDS-SDC nodes with nearly seamless scale-out for growth.

## 2.3 Routing protocol selection for Leaf-Spine

Typical routing protocols used when designing a Leaf-Spine network.

- Border Gateway Protocol (BGP) - External (EBGP) or Internal (IBGP)
- Open Shortest Path First (OSPF)

This guide includes examples for both OSPF and EBGP configurations.

Table 1 lists items to consider when choosing between OSPF and BGP protocols.

Table 1 OSPF and BGP table

OSPF	BGP
OSPF is the choice of protocol for networks under same administration or internal networks. OSPF is an internal gateway protocol.	BGP is the choice of protocol for networks under different administration or different ISPs. BGP is an external gateway protocol.
OSPF is a link state protocol. OSPF chooses the best path among all available paths and requires no additional configurations as the protocol dynamically reacts to any changes in paths.	BGP is a path vector routing protocol. It makes routing decisions based on paths, network policies or rule-sets configured by a network administrator and is involved in making core routing decisions.
An OSPF network can be divided into sub-domains called areas. An area is a logical collection of OSPF networks, routers and links that have the same area identification. A router within an area must maintain a topological database for the area to which it belongs. The router does not have detailed information about network topology outside of its area, thereby reducing the size of its database.	When BGP runs between two peers in the same autonomous system (AS), it is referred to as Internal BGP (iBGP or Interior Border Gateway Protocol). When it runs between different autonomous systems, it is called External BGP (eBGP or Exterior Border Gateway Protocol).
OSPF works within a single autonomous system.	BGP works with multiple autonomous systems.
Route filtering is not possible.	Route filtering is possible in BGP through Network Layer Reachability, AS_Path and Community attributes.
OSPF cannot scale easily in large networks.	BGP can scale in very large networks.

### 2.3.1 BGP

BGP provides scalability whether configured as EBGP or IBGP. See section 3.2.3 for an EBGP routing configuration example on a Leaf-Spine network. The nature of a Leaf-Spine network is based on the use of ECMP. EBGP and IBGP handle ECMP differently. By default, EBGP supports ECMP without any adjustments. However, IBGP requires a BGP route reflector and the use of the AddPath feature to support ECMP.

### 2.3.2 OSPF

OSPF provides routing inside a company's autonomous network, or a network that a single organization controls. While generally more memory and CPU-intensive than BGP, it offers a faster convergence without any tuning. An example for configuring OSPF routing on the Leaf-Spine network is also provided in the guide.

**Note:** For detailed information on Leaf-Spine design, see [Dell EMC Networking L3 Design for Leaf-Spine with OS10EE](#).

## 3 Configuration and Deployment

This section describes how to configure the physical and virtual networking environment for this hyper-converged VxFlex OS deployment. The following items are discussed:

- Key networking protocols
- Network IP addressing
- Physical switch configuration
- Virtual datacenter, clusters, and hosts
- Virtual networking configuration

### 3.1 VxFlex OS solution example

There are several options available when deploying VxFlex OS. This section provides a summary of the options used in this deployment example.

#### 3.1.1 VxFlex OS nodes

The VxFlex OS Data nodes in this deployment example use the following:

**Hardware:** Dell PowerEdge R730xd (4 qty.)

**Hardware setup:** PERC H730 configured to enable RDM devices (Appendix B.3)

**Cluster name:** atx01-w02-ScaleIO

**Software:** ESXi 6.5

#### 3.1.2 VxFlex OS deployment options

The following options were utilized during the installation of VxFlex OS:

**Protection domain:** A single protection domain used for simplicity. Protection domains are not within the scope of this document and are not contingent on the network design.

**Storage Pools:** Single storage pool.

**Fault sets:** No fault sets are configured. Fault sets are not within the scope of this document and are not contingent on the network design.

### 3.2 Physical switch configuration

This section provides details on configuring the leaf-spine switches that form the physical network.

This document includes multiple attachment files for spine-switch configurations and leaf-switch configurations. The example configurations are based on a solution assembled in the Dell EMC Networking lab and require editing to fit your network.

The commands within each configuration file can be modified to apply to the reader's network. Network interfaces, VLANs, and IP schemes can be easily changed and adapted using a text editor. Once modified, copy/paste the commands directly into the switch CLI of the appropriate switch.

The following subsections provide information to assist in the deployment examples detailed in this guide.

### 3.2.1 BGP ASN configuration

BGP has a reserved, private, two-byte Autonomous System Number (ASN) ranging from 64,512 to 65,535.

Each switch is assigned a separate ASN. Figure 5 below shows an example of ASN numbering. The deployment steps use the ASN numbers in Table 2.

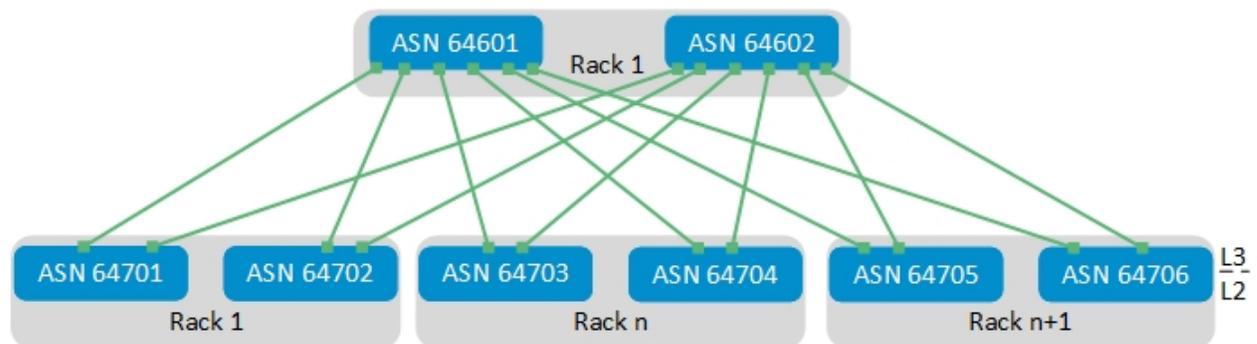


Figure 5 BGP ASN assignments

### 3.2.2 BGP neighbor fall-over

BGP tracks IP reachability to the peer remote address and the peer local address. If either becomes unreachable (for example, no active route exists in the routing table for the peer IPv4 destination/local address), BGP brings down the session with the peer. This feature is called neighbor fall-over. Dell EMC recommends enabling neighbor fall-over for EBGP settings.

### 3.2.3 Loopback addresses

Figure 6 shows part of the created topology. This figure shows the point-to-point IP address and loopback addresses for switches in the topology. All of the point-to-point addresses come from the same base IP prefix, 192.168.0.0/16. The third octet represents the appropriate topology layer, 1 for spine (top of Leaf-Spine topology) and 2 for leaf (bottom of Leaf-Spine topology). All loopback addresses are part of the 10.0.0.0/8 address space in this example with each switch using a 32-bit mask. This address scheme helps with establishing BGP neighbor adjacencies, as well as troubleshooting connectivity.

As illustrated, scaling horizontally requires following the IP scheme below, depending on whether the device is a spine or leaf switch.

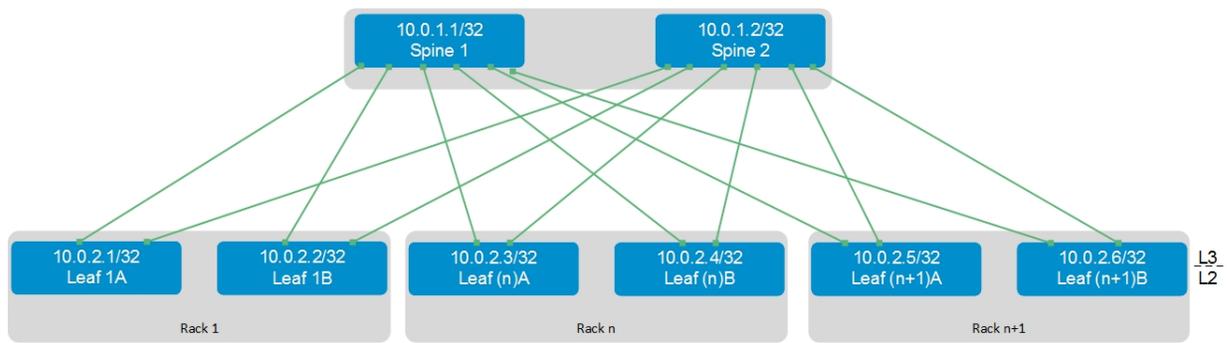


Figure 6 Loopback IP addressing

Each leaf connects to each spine, but note that the spines do not connect to one another. In a Leaf-Spine, topology there is no requirement for the spines to have any interconnectivity. Given any single-link failure scenario, all leaf switches retain connectivity to one another.

Table 2 shows loopback addressing and BGP ASN numbering associations. Building BGP neighbor relationships require this information along with the point-to-point information in Table 3 the next section.

Table 2 Loopback interfaces and ASN associations

Switch Name	Loopback	BGP ASN
Spine 1	10.0.1.1/32	64601
Spine 2	10.0.1.2/32	64602
Leaf 1A	10.0.2.1/32	64701
Leaf 1B	10.0.2.2/32	64702

### 3.2.4 Point-to-point interfaces

Below Table 3 lists physical connection details from each Leaf-Spine switch. The table presents the switch name, source interface, source IP network and network IP addresses. The IP scheme below easily extends to account for additional Leaf-Spine switches.

All addresses come from the same base IP prefix, 192.168.0.0/16 with the third octet representing the spine number. For instance, 192.168.1.0/31 is a two-host subnet that ties to Spine 1 while 192.168.2.0/31 ties to Spine 2.

Table 3 Interface and IP configuration

Link	Source switch	Source interface	Source IP	Network	Destination switch	Destination interface	Destination IP
A	Leaf 1A	eth1/1/45	.1	192.168.1.0/31	Spine 1	eth1/1/1	.0
B	Leaf 1A	eth1/1/46	.1	192.168.2.0/31	Spine 2	eth1/1/1	.0
C	Leaf 1B	eth1/1/45	.3	192.168.1.2/31	Spine 1	eth1/1/2	.2
D	Leaf 1B	eth1/1/46	.3	192.168.2.2/31	Spine 2	eth1/1/2	.2

Figure 7 shows the links from Table 3

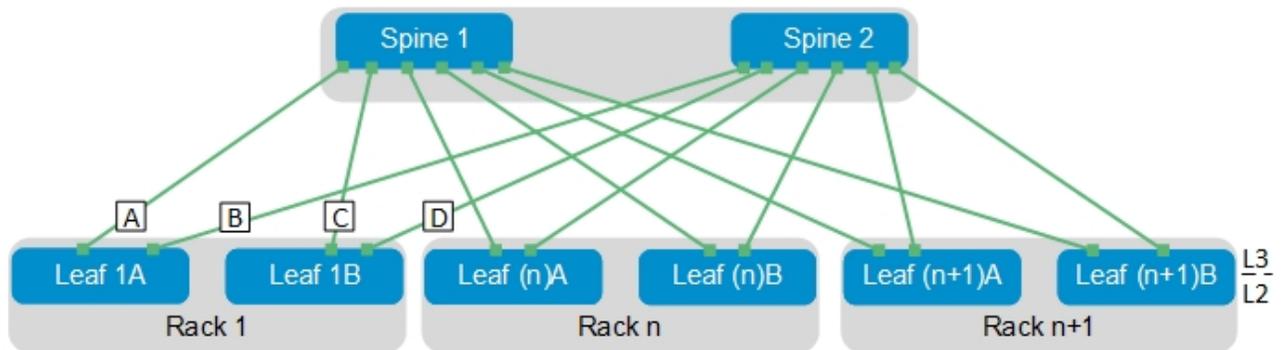


Figure 7 Point-to-point interface IP addressing

**Note:** The example point-to-point addresses use a 31-bit mask to prevent unnecessary sprawling of internal addresses. This IP scheme is optional and covered in [RFC 3021](#).

### 3.2.5 Interface/IP configuration

Table 4 outlines the subnets, VLANs and default gateways for the broadcast networks. Notice that the same VLAN IDs, with different networks, repeat in each rack. The VLANs and subnets are configured on both leaf switches and advertised through the routing instance at the same cost. ECMP spreads server traffic flow across all uplinks.

Table 4 VLAN and subnet examples

Rack ID	Network Name	Subnet	VLAN	Gateway
1	ESXi management	172.17.31.0/24	1731	172.17.31.253
1	vmotion	172.17.32.0/24	1732	172.17.32.253
1	ScaleIO-management	172.17.33.0/24	1733	172.17.33.253
1	ScaleIO-data01	172.17.34.0/24	1734	172.17.34.253

### 3.2.6 ECMP

ECMP is the core protocol facilitating the deployment of a layer 3 leaf-spine topology. ECMP gives each spine and leaf switch the ability to load balance flows across a set of equal next-hops. For example, when using two spine switches, each leaf has a connection to each spine. For every flow egressing a leaf switch, there exists two equal next-hops, one to each spine.

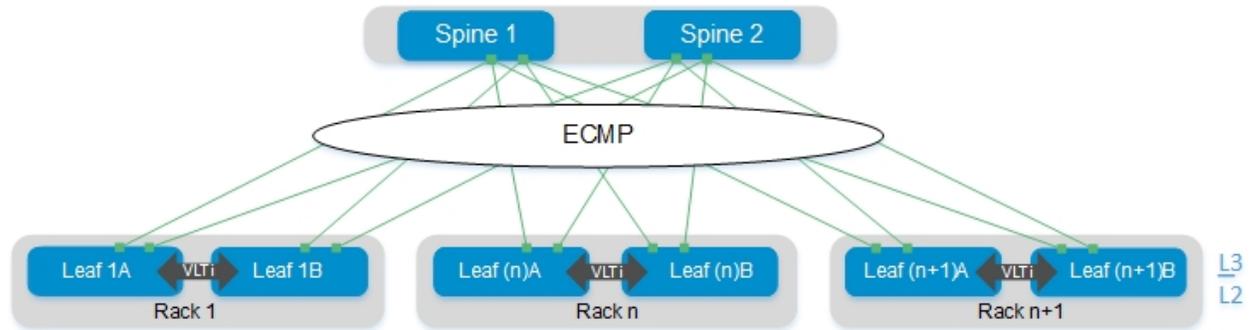


Figure 8 ECMP

### 3.2.7 VRRP

A VRRP instance is created for each VLAN/network in Table 4. As illustrated in Figure 9 below, Node 1 participates in the vMotion VLAN 32 broadcast domain. The host's configuration sends traffic for 172.17.32.0/24 to the Virtual IP (VIP) 172.17.32.253 provided by the VRRP instance running between leaf switches 1A and 1B.

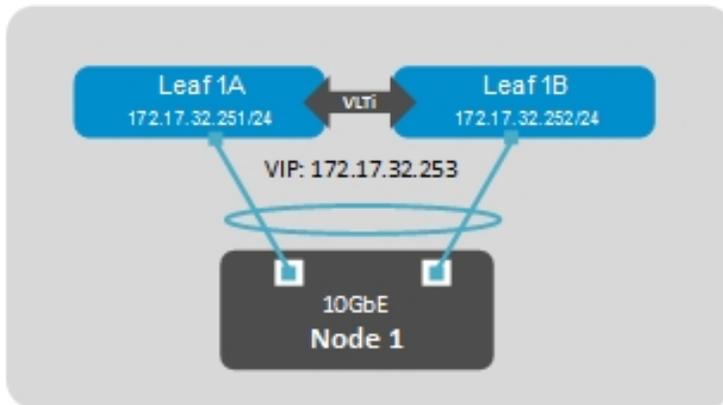


Figure 9 VRRP example

### 3.2.8 Flow control

The sample deployment in this document utilizes the Dell EMC S4248FB-ON network switch which provides deep buffers for egress traffic. Dell EMC recommends the use of flow control on the host interfaces to fully utilize this feature to minimize packet drops due to incasts or host network congestion. Flow control is enabled on ESXi by default but receive flow control must be enabled on the network switch interfaces.

## 3.3 VMware virtual networking configuration

This section provides details on configuration of virtual networking settings within VMware vCenter. All steps provided in the next sections are completed with the vCenter Web Client.

**Note:** ESXi has been installed on all hosts, vCenter has been deployed, and all hosts have been added to vCenter. See Appendix B for information on preparing servers for VxFlex OS deployment.

### 3.3.1 Create a datacenter object and add hosts

A datacenter object needs to be created before hosts can be added. This guide uses a single datacenter object named Datacenter.

To create a datacenter object, complete the following steps:

1. On the web client Home screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right-click on the vCenter Server object and select **New Datacenter**.
3. Provide a name (ScaleIO) and click **OK** (Figure 10).

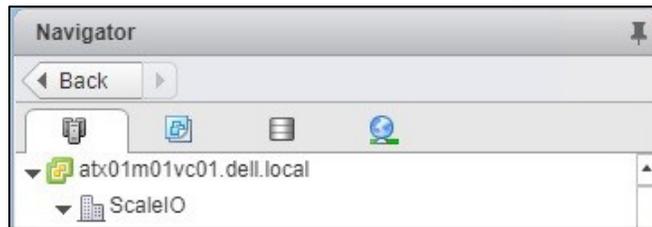


Figure 10 Datacenter created

To add ESXi hosts to the datacenter, complete the following steps:

1. On the web client Home screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right click on **Datacenter** and select **Add Host**.
3. Specify the **IP address** of an ESXi host (or the **hostname** if DNS is configured on your network). Click **Next**.
4. Enter the credentials for the ESXi host and click **Next**. If a security certificate-warning box displays, click **Yes** to proceed.
5. On the Host summary screen. Click **Next**.
6. Assign a license or select the evaluation license. This guide uses a VMware vSphere 6 Enterprise Plus license for ESXi hosts. Click **Next**.
7. Select a **Lockdown mode**. This guide uses the default setting, **Disabled**. Click **Next**.
8. For the VM location, select **Datacenter**. Click **Next**.
9. On the **Ready to complete** screen, select **Finish**.

Repeat for all servers running ESXi that will be part of the deployment. This deployment example uses four R730xd servers running ESXi.

### 3.3.2 Create clusters and add hosts

When a host is added to a cluster, the host's resources become part of the cluster's resources. The cluster manages the resources of all hosts within it. This section shows how to create a cluster.

All ESXi hosts are added to the cluster. The cluster name in this example is for identification purposes only.

To add clusters to the datacenter, complete the following steps:

1. On the web-client Home screen, select **Hosts and Clusters**.
2. In the **Navigator** pane, right-click the datacenter object and select **New Cluster**.
3. Name the cluster. For this example, the cluster is named **atx01-w02-ScaleIO**.
4. Leave **DRS**, **vSphere HA**, **EVC** and **Virtual SAN** at their default settings (**Off/Disabled**). Click **OK**.

**Note:** vSphere DRS, HA, and EVC cluster features are outside the scope of this guide. For more information on these features, see the [VMware vSphere 6.5 documentation](#).

In the Navigator pane, drag and drop ESXi hosts into the cluster.

When complete, each cluster (📁) should contain its assigned hosts (📱) as shown in Figure 11.

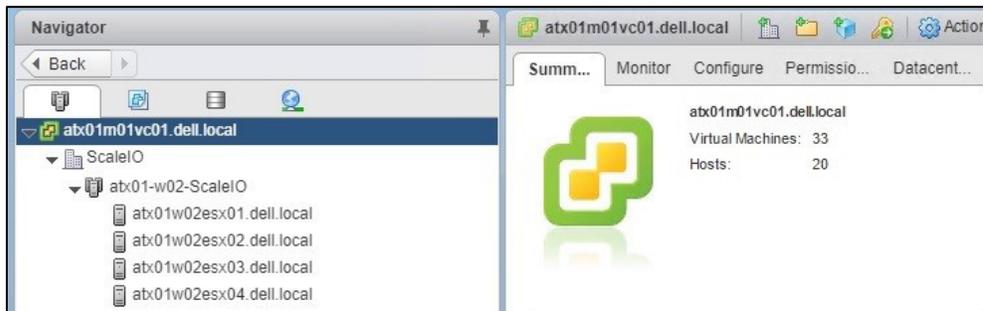


Figure 11 Clusters and hosts after initial configuration

### 3.3.3 Deploy vSphere distributed switch

A single vSphere distributed switch (VDS) is created in vCenter and deployed to each host. The VDS contains port groups named **management**, **ScaleIO-data01**, **ScaleIO-management**, and **vmotion**. The **management** port group supports management traffic to the ESXi hosts. The **ScaleIO-data01** port group supports all VxFlex OS traffic to include SDS-SDC and MDM traffic. The **ScaleIO-management** port group support management traffic to VxFlex OS. Finally, the **vmotion** port group supports all vMotion traffic.

The VDS is assigned four vmnics from each host. Each host from the VxFlex OS cluster combine **vmnic0** and **vmnic5** into an LACP-enabled port channel, **lag 1**. This lag is mapped to the **management**, **ScaleIO-management**, and **vmotion** port groups. Each host from the VxFlex OS cluster combines **vmnic1** and **vmnic4** into an LACP-enabled port channel, **lag 2**. This lag is mapped to the **ScaleIO-data01** port group. In this configuration, compute, or application traffic is forwarded over **lag1**, while storage traffic is forwarded over **lag2**.

**Note:** The steps to migrate an ESXi management VMkernel adapter from a standard switch to a VDS configured for a LAG is not covered in this document. The following VMware documents provide details to assist with the migration:

<https://kb.vmware.com/s/article/1010614>

<https://docs.vmware.com/en/VMware-vSphere/6.5/com.vmware.vsphere.networking.doc/GUID-34A96848-5930-4417-9BEB-CEF487C6F8B6.html>

The following diagrams show the configuration of each port group type:

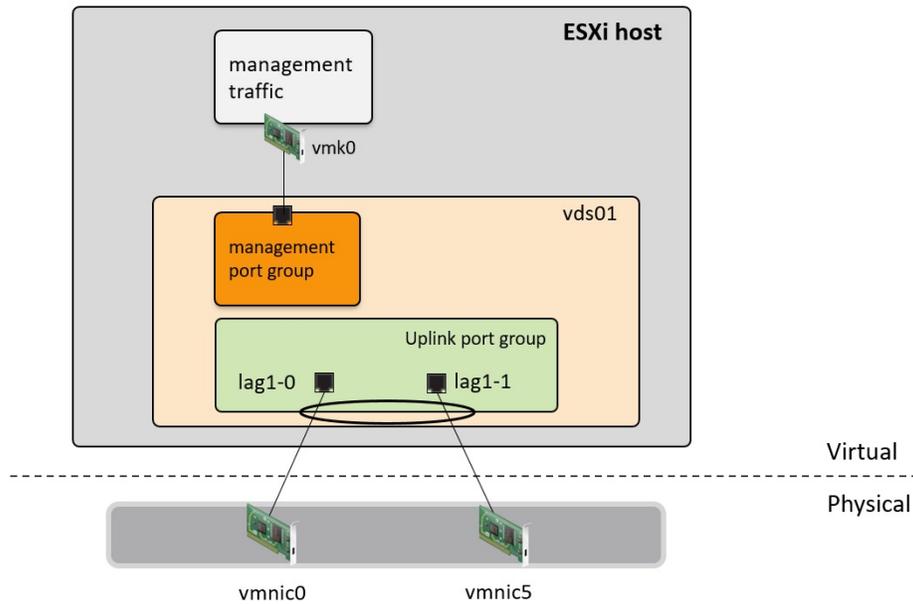


Figure 12 Management port group

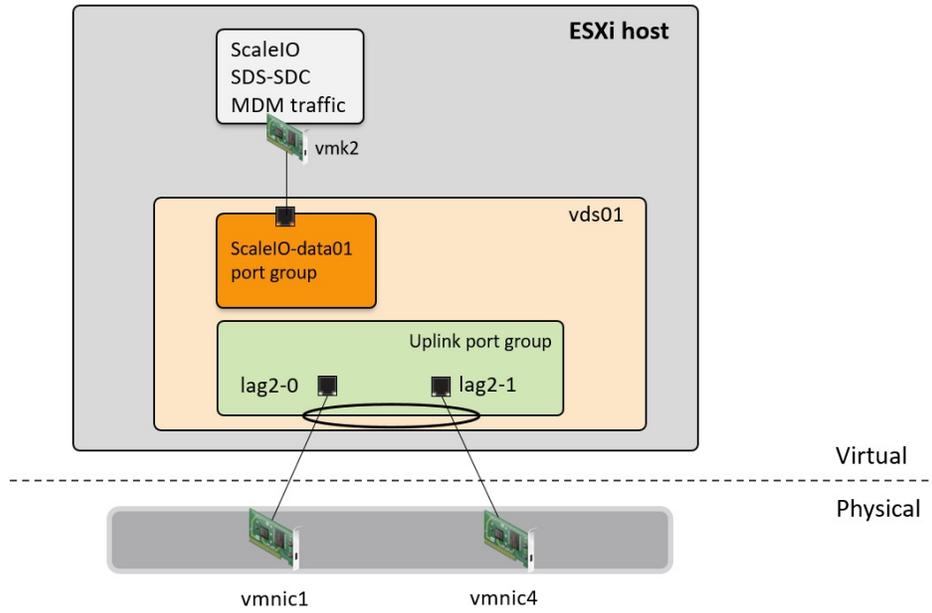


Figure 13 ScaleIO-data01 port group

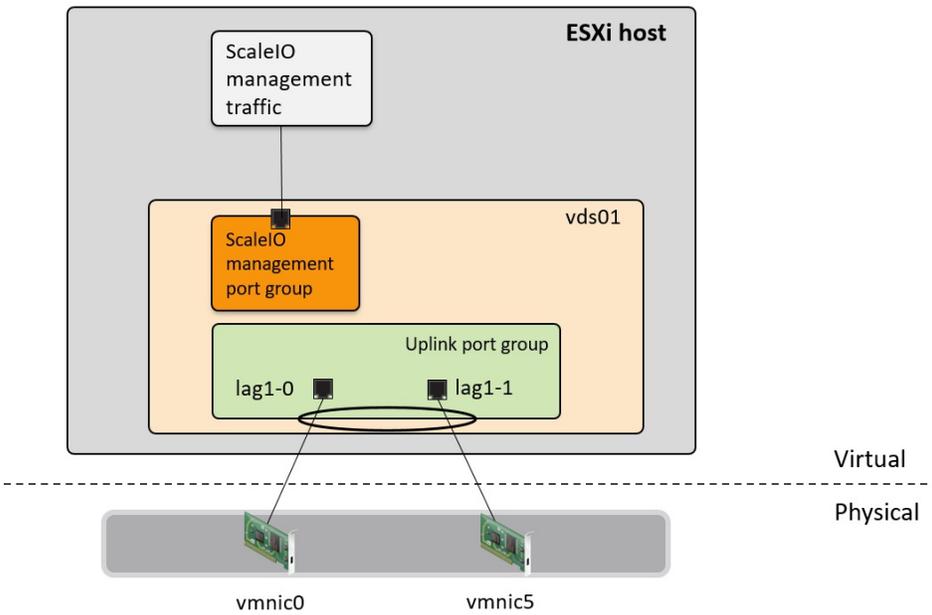


Figure 14 ScaleIO management port group

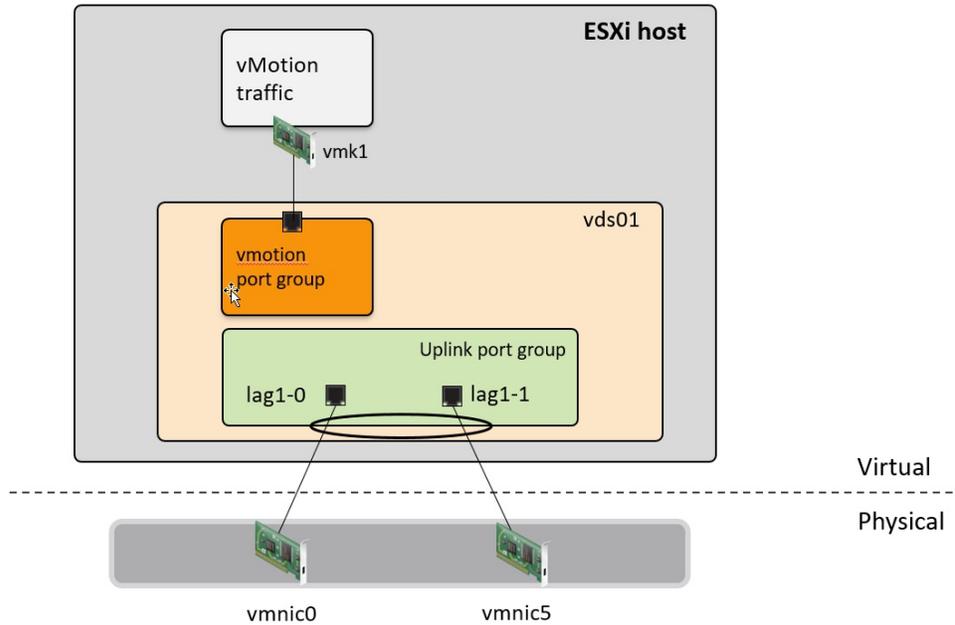


Figure 15 vMotion port group

### 3.3.4 Create the VDS

To create the VDS complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right-click Datacenter. Select **Distributed switch > New Distributed Switch**.
3. Provide a name for the VDS. Click **Next**.
4. On the Select version page, select **Distributed switch: 6.5.0 > Next**.
5. On the **Edit settings** page:
  - a. Set the **Number of uplinks** to **4**.
  - b. Leave **Network I/O Control** set to **Enabled**.
  - c. **Uncheck** the **Create a default port group** box.
6. Click **Next** followed by **Finish**.
7. The VDS is created with the dvUplink port group shown beneath it.

When complete, the Navigator pane should look similar to Figure 16.

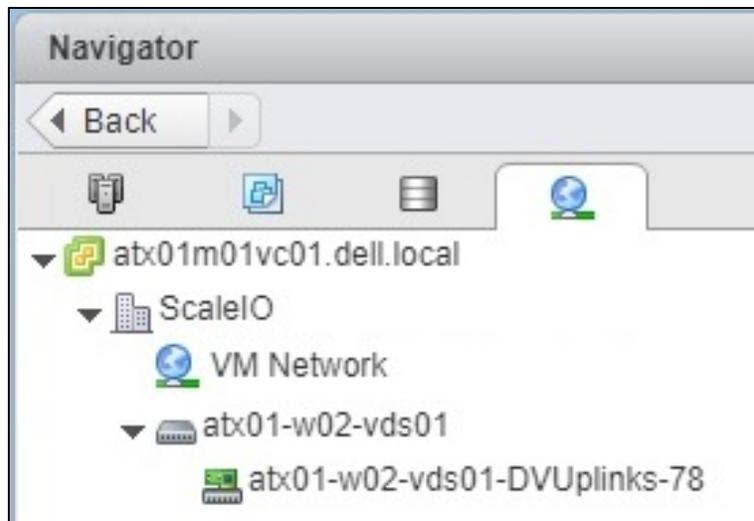


Figure 16 VDS created for compute and ScaleIO

### 3.3.5 Add distributed port groups

In this section, separate distributed port groups for management, ScaleIO-management, vMotion, and ScaleIO-data01 are added to the VDS.

To create the port group for management traffic on the VDS, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right-click the VDS. Select **Distributed Port Group > New Distributed Port Group**.
3. On the **Select name and location page**, provide a name for the distributed port group, for example, **vMotion**. Click **Next**.
4. On the **Configure settings**, next to **VLAN type**, select **VLAN**. Set the **VLAN ID** to **1731** for the **management** port group. Leave other values at their defaults as shown in Figure 17.
5. Click **Next > Finish**.

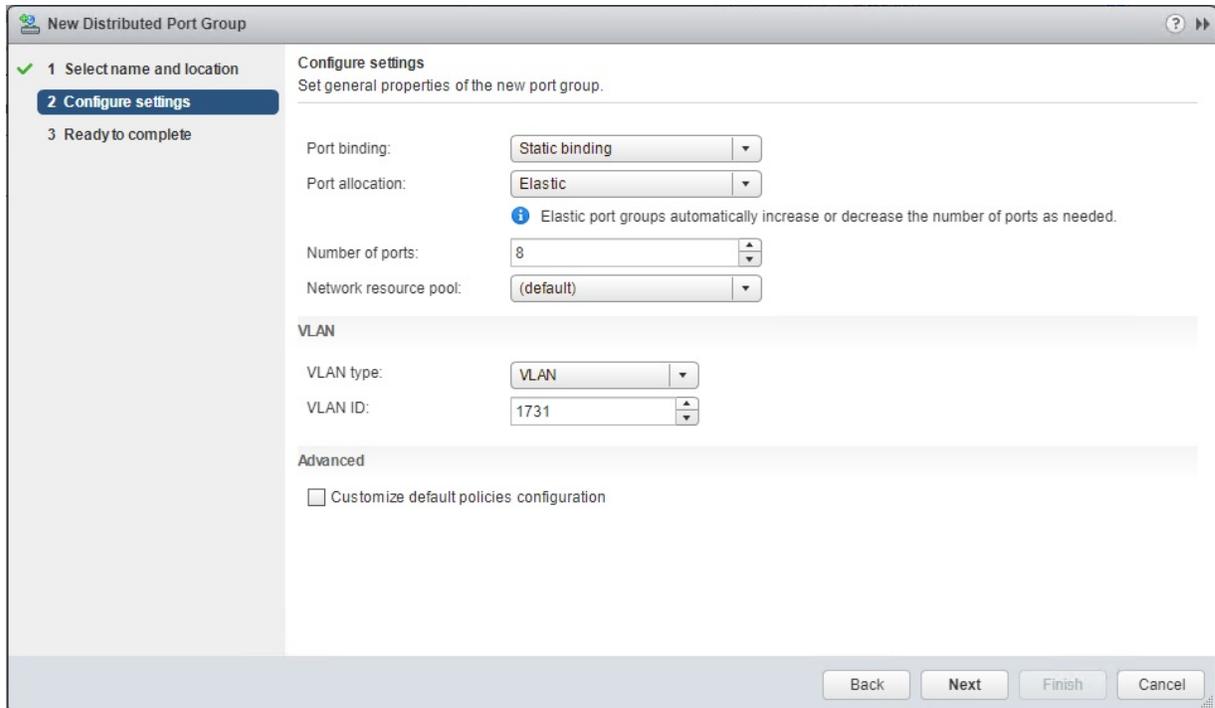


Figure 17 Distributed port group settings page – vMotion port group

Repeat steps 1-5 above to create the distributed port group for ScaleIO-management, vMotion, and ScaleIO-data01. In this example the following VLAN IDs were used:

- Management VLAN ID > 1731
- ScaleIO-management VLAN ID > 1733
- vMotion VLAN ID > 1732
- ScaleIO-data01 VLAN ID > 1734

Ensure to set each **port group name** with a unique descriptive name.

When complete, the Navigator pane appears similar to Figure 18.

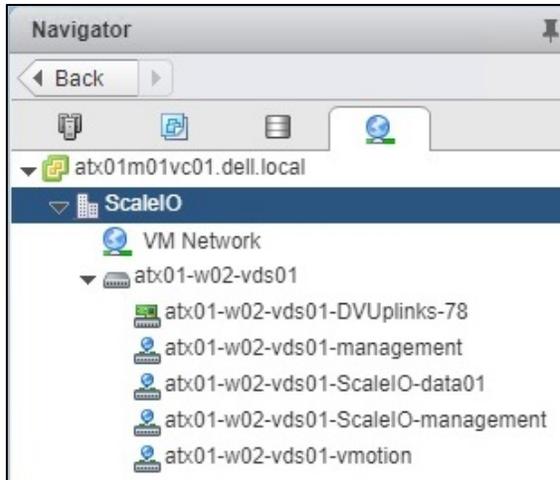


Figure 18 Distributed switches with each port group created

### 3.3.6 Create LACP LAGs

Since Link Aggregation Control Protocol (LACP) LAGs are used in the physical network between ESXi hosts and physical switches, LACP LAGs are also configured on each VDS.

To enable LACP on the VDS, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. In the Navigator pane, select the VDS.
3. In the center pane, select **Configure > Settings > LACP**.
4. Click the **+** icon. The New Link Aggregation Group dialog box opens.
5. In this example, the default **name** is kept as **lag1**. Set **number of ports** to **2**.
6. Set the **Mode** to **Active**. The remaining fields can be set to their default values as shown in Figure 19.
7. Click **OK** to close the dialog box.

**New Link Aggregation Group** ?

Name:

Number of ports:

Mode:

Load balancing mode:

**Port policies**

You can apply VLAN and NetFlow policies on individual LAGs within the same uplink port group. Unless overridden, the policies defined at uplink port group level will be applied.

VLAN type:  Override

VLAN trunk range:

NetFlow:  Override

Figure 19 LAG configuration

This creates **lag1** on the VDS. Repeat steps 1-7 to create another LACP LAG, named **lag2**.

Click the refresh icon (  ) at the top of the screen if the lag does not appear in the table as shown in Figure 20.

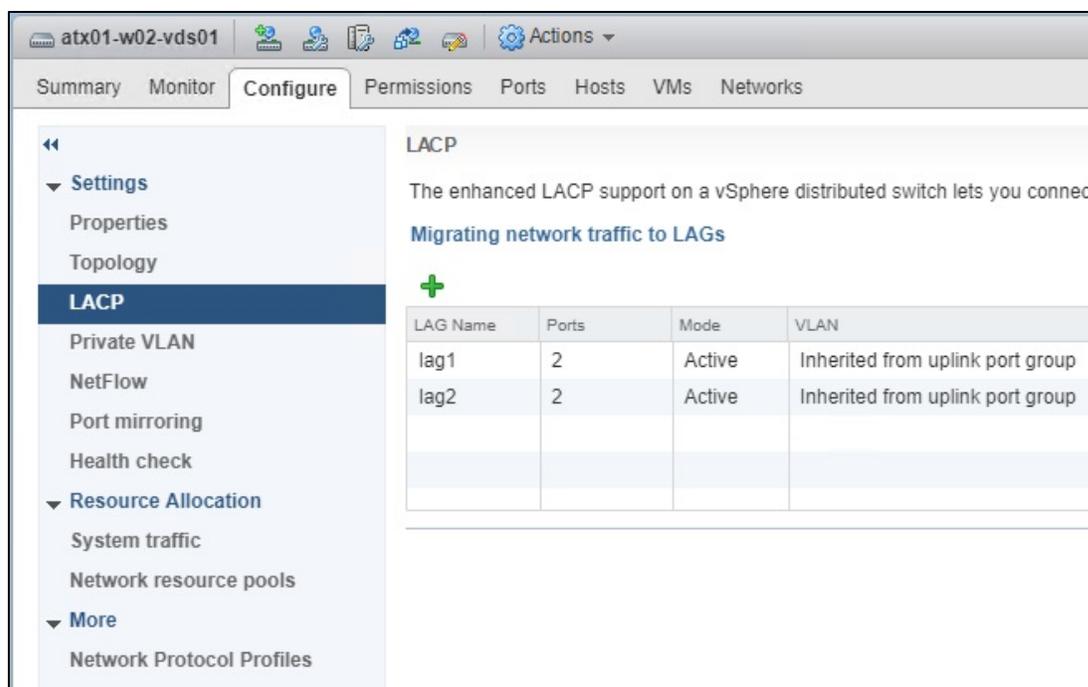


Figure 20 LAGs created on the VDS

### 3.3.7 Associate hosts and assign uplinks

Hosts and their vmnics must be associated with each vSphere-distributed switch.

All traffic uses the LAG uplinks to take advantage of the improved bandwidth usage and failover behavior the VLT feature provides. This section details configuration of the uplinks.

**Note:** Before starting this section, be sure you know the vmnic-to-physical adapter mapping for each host. Determine this mapping by going to **Home > Hosts and Clusters** and selecting the host in the **Navigator** pane. In the center pane, select **Configure > Networking > Physical adapters**. This example uses vmnics 0, 1, 4, and 5. Vmnic numbering varies depending on adapters installed in the host.

To add hosts to the VDS, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right click on ScaleIO VDS and select Add and Manage Hosts.
3. In the Add and Manage Hosts dialog box:
  - a. On the **Select task** page, make sure **Add hosts** is selected. Click **Next**.
  - b. On the **Select hosts** page, Click the **+ New hosts** icon. Select the check box next to each host in the ScaleIO cluster. Click **OK > Next**.
  - c. On the **Select network adapters tasks** page, be sure the **Manage physical adapters** box is checked. Be sure all other boxes are unchecked. Click **Next**.
  - d. On the **Manage physical network adapters** page, each host is listed with its vmnics beneath it.

- i. On the first host in the cluster, select the appropriate vmnic (vmnic0 in this example) and click  **Assign uplink**.
  - ii. Select **lag1-0 > OK**.
  - iii. On the same host, select the next appropriate vmnic (vmnic5 in this example) and click  **Assign uplink**.
  - iv. Select **lag1-1 > OK**.
  - v. On the same host, select the next appropriate vmnic (vmnic1 in this example) and click  **Assign uplink**.
  - vi. Select **lag2-0 > OK**.
  - vii. On the same host, select the next appropriate vmnic (vmnic4 in this example) and click  **Assign uplink**.
  - viii. Select **lag2-1 > OK**.
  - ix. Repeat steps i – viii for the remaining hosts in the cluster.
  - x. Click **Next** when done.
- e. On the Analyze impact page, **Overall impact status** should indicate  **No impact**.
  - f. Click **Next > Finish**.

When complete, the **Configure > Settings > Topology** page for the VDS should look similar to Figure 21:

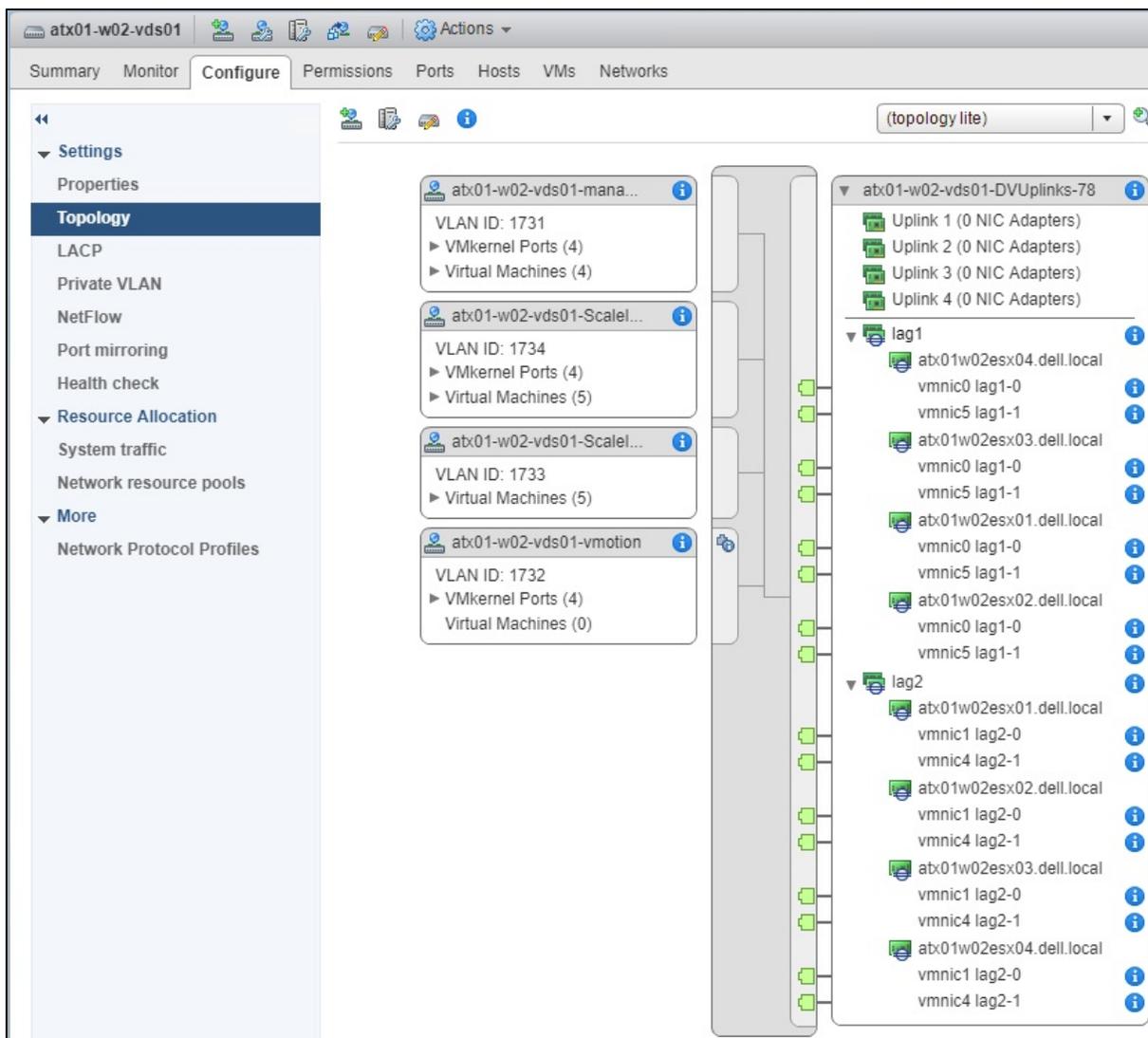


Figure 21 Uplinks configured on ScaleIO VDS

This configuration brings up the LAGs on the upstream switches. Confirm the configuration by running the **vlt-port-detail** show command on the upstream switches as shown in the example below. The **Status** column now indicates all LAGs are **up**.

```

rack1-leaf-A# show vlt 127 vlt-port-detail
vlt-port-channel ID : 1
VLT Unit ID   Port-Channel   Status   Configured ports   Active ports
-----
* 1           port-channell   up       1                   1
  2           port-channell   up       1                   1
vlt-port-channel ID : 2

```

```

VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel2     up          1                      1
  2              port-channel2     up          1                      1
vlt-port-channel ID : 3
VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel3     up          1                      1
  2              port-channel3     up          1                      1
vlt-port-channel ID : 4
VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel4     up          1                      1
  2              port-channel4     up          1                      1
vlt-port-channel ID : 5
VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel5     up          1                      1
  2              port-channel5     up          1                      1
vlt-port-channel ID : 6
VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel6     up          1                      1
  2              port-channel6     up          1                      1
vlt-port-channel ID : 7
VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel7     up          1                      1
  2              port-channel7     up          1                      1
vlt-port-channel ID : 8
VLT Unit ID      Port-Channel      Status      Configured ports      Active ports
-----
* 1              port-channel8     up          1                      1
  2              port-channel8     up          1                      1
rack1-leaf-A#

```

### 3.3.8 Configure teaming and failover on LAGs

To configure teaming and failover on LAGs, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right click on the VDS. Select **Distributed Port Group > Manage Distributed Port Groups**.
3. Select only the **Teaming and failover** checkbox. Click **Next**.
4. Click **Select distributed port groups**. Check the boxes for the management, ScaleIO-management, and vMotion port groups. Click **OK > Next**.

- On the **Teaming and failover page**, click **lag1** and move it up to the **Active uplinks** section by clicking the up arrow. Move **Uplinks 1-4** down to the **Unused uplinks** section by clicking the down arrow. Leave other settings at their defaults. The **Teaming and failover** page should look similar to Figure 22 when complete.

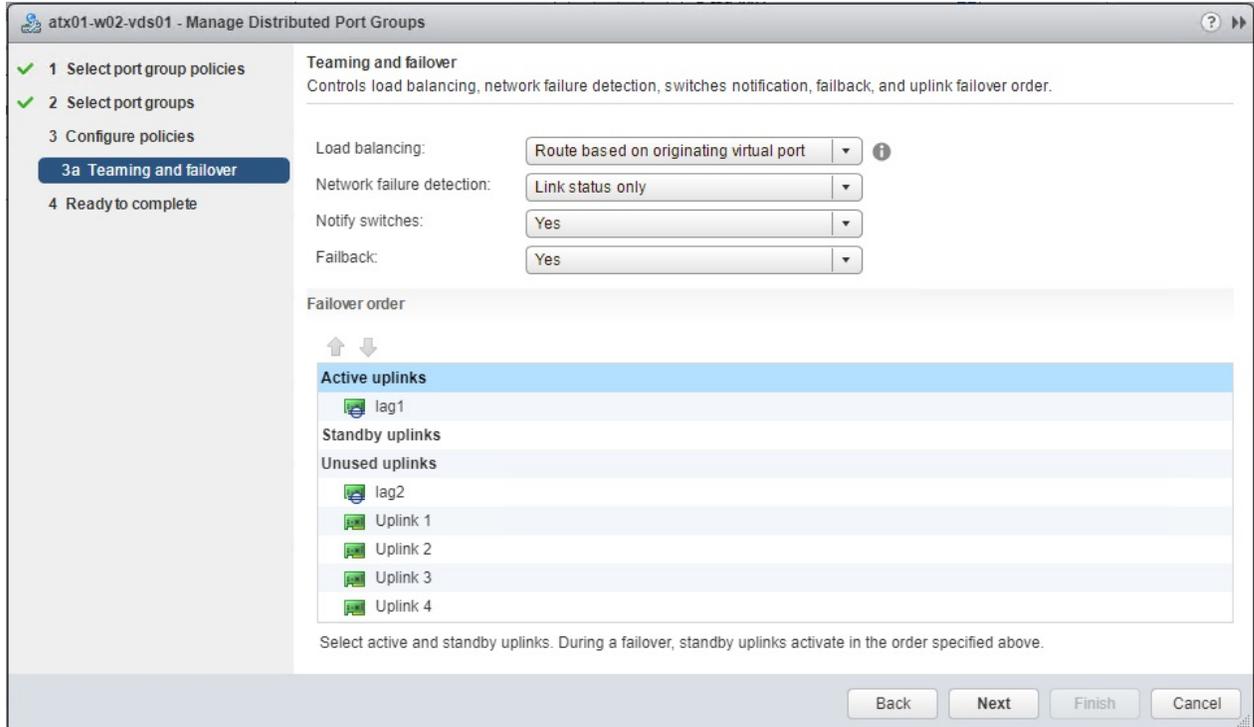


Figure 22 Teaming and failover settings for LAGs

- Click **Next** followed by **Finish** to apply settings.

Repeat steps 1-6 above for the **ScaleIO-data01** port group, except select **lag2** to move to the **Active Uplinks**.

### 3.3.9 Add VMkernel adapters for MDM, SDS-SDC, and vMotion

To allow virtual machines to be moved between compute nodes a vMotion-enabled VMkernel is created. This VMkernel is configured to use the vMotion port group on the default TCP/IP stack. A VMkernel interface is also created to handle VxFlex OS storage traffic. The VMkernel is associated with the ScaleIO-data01 and uses the default TCP/IP stack. No VMware specific services are enabled for storage traffic.

The procedure in this section, adds ScaleIO-data01, management, and vMotion VMkernel adapters (referred to as VMkernel ports) to each ESXi host to allow for ScaleIO, management, and vMotion traffic.

Either assign IP addresses statically to VMkernel adapters upon creation or use DHCP. This guide uses Static IP addresses.

Table 5 VLAN and network examples

VMkernel Name	VLAN ID	IP Address	Subnet Mask
Management > vmk0	1731	172.17.31.yyy	255.255.255.0
vMotion > vmk1	1732	172.17.32.yyy	255.255.255.0
ScaleIO-data01 > vmk2	1734	172.17.34.yyy	255.255.255.0

**Note:** VMkernel adapter vmk0 is installed by default for host management at the time of ESXi installation.

To add the vMotion VMkernel adapters to all hosts connected to the VDS, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right-click on the VDS and select **Add and Manage Hosts**.
3. In the Add and Manage Hosts dialog box complete the following steps:
  - a. On the **Select task** page, make sure **Manage host networking** is selected. Click **Next**.
  - b. On the **Select hosts** page, click **+ Attached hosts**. Select all hosts. Click **OK > Next**.
  - c. On the **Select network adapter tasks** page, make sure the **Manage VMkernel adapters** box is checked, and all other boxes are unchecked. Click **Next**.  
The Manage VMkernel network adapters page opens.
    - i. To add the vMotion adapter, select the first host and click **+ New Adapter**.
    - ii. On the **Select target device** page, click the radio button next to **Select an existing network** and click **Browse**.
    - iii. Select the port group created for vMotion > **OK**. Click **Next**.
    - iv. On the **Port properties** page, leave **IPv4** selected and check only the **vMotion** box for **Enabled services**. Click **Next**.
    - v. On the **IPv4 settings page**, if DHCP is not used, select **Use static IPv4 settings**. Set the IP address, for example, 172.17.32.101, and subnet mask for the host on the vMotion network,. Click **Next > Finish**.
  - d. Repeat steps i-v to add ScaleIO-data01 VMkernels for the remaining hosts, choosing a different IP setting for each host.
  - e. Click **Next**.
  - f. On the **Analyze impact** page, **Overall impact status** should indicate **✔ No impact**.
  - g. Click **Next > Finish**.

To add ScaleIO VMkernel adapters to all hosts connected to the VDS, repeat the above steps except step 3.c.iv. On the **Port Properties** page, uncheck all **Enabled services** boxes.

When complete, the VMkernel adapters' page for each ESXi host in the vSphere datacenter should look similar to Figure 23. View this page by going to **Hosts and Clusters**, selecting a host in the **Navigator** pane, then selecting **Configure > Networking > VMkernel adapters** in the center pane.

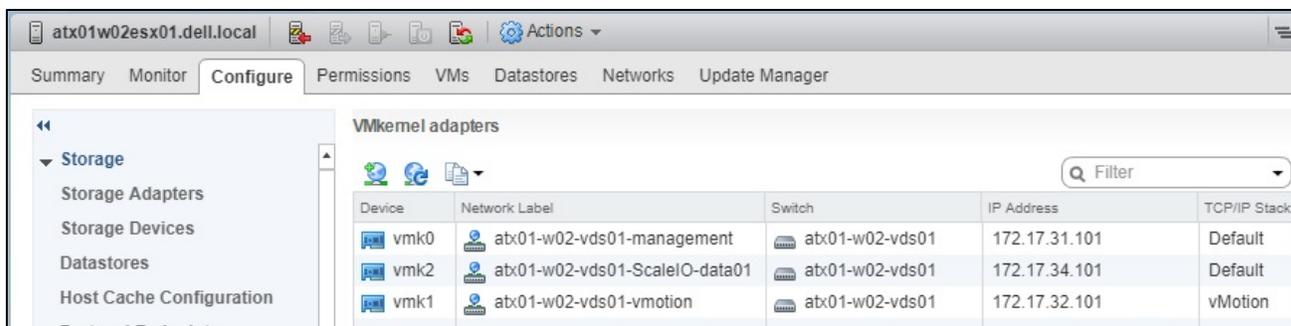


Figure 23 Host VMkernel adapters page

To verify the configuration, ensure the vMotion adapter, **vmk1** in this example is shown as **Enabled** in the **vMotion Traffic** column. Verify the VMkernel adapter IP addresses are correct.

Verify the information is correct on other hosts, as needed.

**Note:** The example used in this guide only uses four hosts on a single leaf pair. For large scale designs with multiple routed leaf pairs, additional configuration steps are required. The default gateway for each network configured on the VMkernels cannot be modified during VMkernel creation. Setting the default gateway requires configuring a static route at the command line of each ESXi host. For instructions on how to set the default gateway see Appendix C.

**Note:** A separate vMotion TCP/IP stack can be used to isolate vMotion traffic according to the topology of the network. For more information and advantages of using a separate vMotion TCP/IP stack, see the following VMware documentation: <https://docs.vmware.com/en/VMware-vSphere/6.5/com.vmware.vsphere.vcenterhost.doc/GUID-5211FD4B-256B-4D9F-B5A2-F697A814BF64.html>

### 3.3.10 Verify VDS configuration

To verify the distributed switches have been configured correctly, the **Topology** page for each VDS provides a summary.

To view the **Topology** page for the VDS, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. In the Navigator pane, select the VDS.
3. In the center pane, select **Configure > Settings > Topology** and click the ► icon next to **VMkernel Ports** to expand. The screen should look similar to Figure 24:

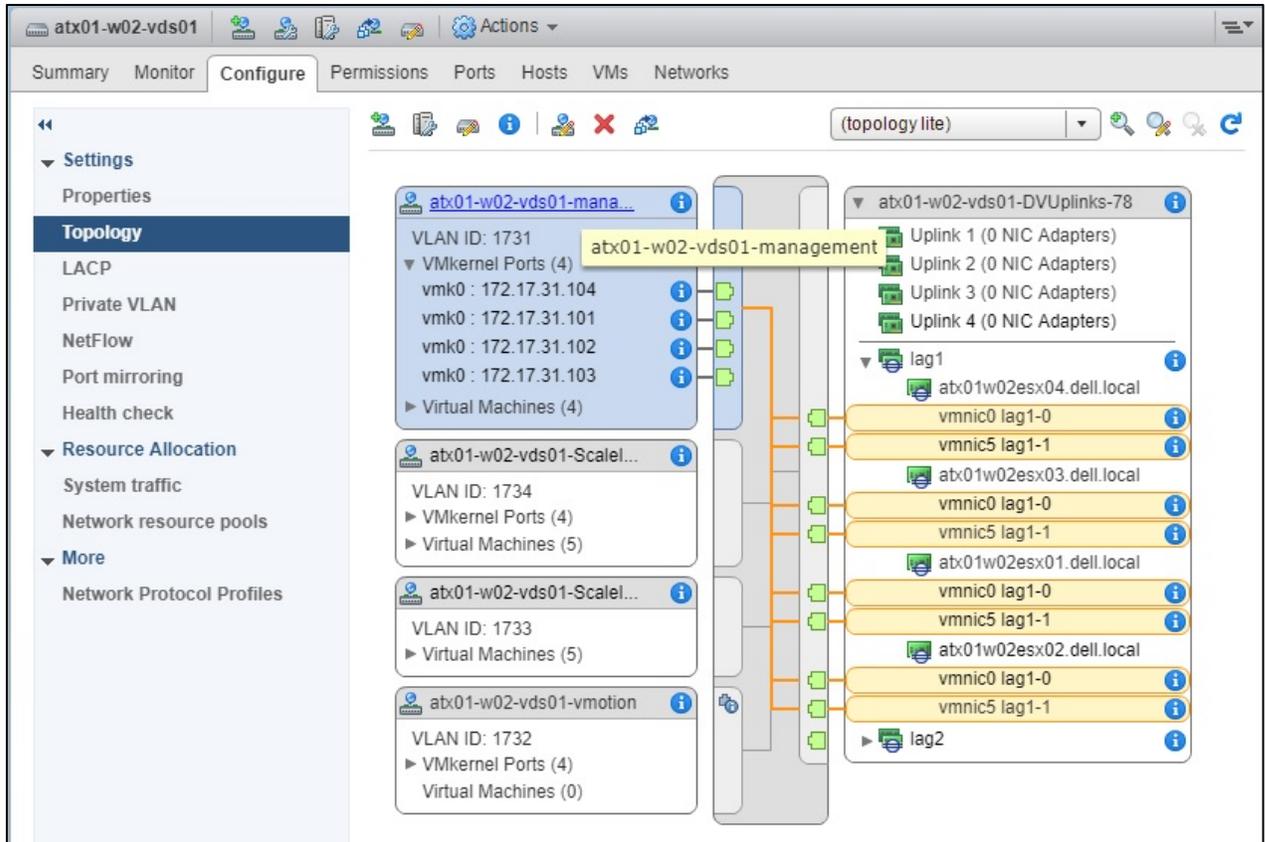


Figure 24 VDS VMkernel ports, VLANs, and IP addresses for management port group

Repeat steps 1-3 above for all port groups within the VDS.

### 3.3.11 Enable LLDP

Enabling Link Layer Discovery Protocol (LLDP) on vSphere-distributed switches is optional but can be helpful for link identification and troubleshooting.

**Note:** LLDP functionality may vary with adapter type. LLDP must also be configured on the physical switches per the switch configuration instructions provided earlier in this guide.

Enabling LLDP on vSphere-distributed switches enables them to send information such as vmnic numbers and MAC addresses to the physical switch connected to the ESXi host.

To enable LLDP on each VDS, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right-click on the VDS, and select **Settings > Edit Settings**.
3. In the left pane of the **Edit Settings** page, click **Advanced**.
4. Under Discovery protocol, set **Type** to Link Layer Discovery Protocol, set **Operation** to Both.
5. Click **OK**.

To view LLDP information sent from the ESXi host adapters, run the following command from the CLI of a directly connected switch:

```
Dell#show lldp neighbors
Loc PortID          Rem Host Name          Rem Port Id          Rem Chassis Id
-----
ethernet1/1/1      atx01w02esx01.del...  00:50:56:50:f6:cf    vmnic1
ethernet1/1/2      atx01w02esx01.del...  00:50:56:50:da:f0    vmnic0
ethernet1/1/3      atx01w02esx02.del...  00:50:56:5e:2d:0a    vmnic1
ethernet1/1/4      atx01w02esx02.del...  00:50:56:56:8e:85    vmnic0
ethernet1/1/5      atx01w02esx03          00:50:56:58:df:59    vmnic1
ethernet1/1/6      atx01w02esx03          00:50:56:58:45:78    vmnic0
ethernet1/1/7      atx01w02esx04          00:50:56:53:50:d3    vmnic1
ethernet1/1/8      atx01w02esx04          00:50:56:5d:8d:93    vmnic0
ethernet1/1/43     rack1-leaf-B           ethernet1/1/43       34:17:eb:0c:89:00
ethernet1/1/44     rack1-leaf-B           ethernet1/1/44       34:17:eb:0c:89:00
ethernet1/1/46     Spine2                 ethernet1/1/1        34:17:eb:19:e2:80
mgmt1/1/1          Rack173-N2048          Gi1/0/14             f4:8e:38:3f:c9:a9
```

## 4 Scaling and tuning guidance

### 4.1 Decisions on scaling

Dell EMC Networking provides a resilient, high-performance architecture that improves availability and meets Service Level Agreements (SLAs) more effectively. This example uses the Dell EMC Networking S4248FB-ON switch because of its ability to provide a low latency with deep buffers for optimum performance. The forty 10GbE, two 40GbE, and six 100GbE ports in a single rack unit provide flexibility in the data center for 1/10/40/100GbE uplink compatibility.

The S4248FB-ON switches connect to Z9100-ON switches as the spine of the networking topology. The Z9100-ON switch adds multiline rate capability supporting 10GbE, 25GbE, 40GbE, 50GbE and 100GbE. It provides for substantial growth with this initial configuration using 40GbE links but able to move to 25GbE downlinks and 100GbE uplinks.

### 4.2 Examples of scaling, port count, and oversubscription

An example of scaling this solution is a rack pod configuration of 16 racks. This pod consists of one rack that contains WAN-edge connectivity and one rack for management servers. There are 14 racks of the SDS storage/compute nodes each holding 19 PowerEdge R730xd's.

In this example, each R730xd has four 10Gb uplinks with 19 servers/nodes per rack. Additionally, the example architecture has four spine switches.

Table 6 Connections for 16 racks with four spine switches

	PowerEdge R730xd	Server connections to leaf switches	Leaf connections to spine switches per rack	Total connections for leaf switches to four spine switches
Connections	4 NIC ports	19 X 4 = 76	4 per leaf switch, 2 leaf switches per rack = 8 links	16 racks * 8 = 128
Speed of ports	10Gb	10Gb	40Gb/ 100Gb	40Gb/ 100Gb
Total theoretical available bandwidth	4 x 10 = 40Gb	76 X 4 = 760Gb	8 * 40 per rack = 320Gb 8 * 100Gb per rack = 800Gb	16 * 320Gbs = 5120 Gb 16 * 800 = 12,800Gb

This example provides for an oversubscription rate of 2.375:1 for 40Gb or 1:1 oversubscription for 100Gb connectivity.

## 4.3 Scaling beyond 16 racks

The proof-of-concept scaling that Figure 25 shows allows four 16-rack pods connected using an additional spine layer to scale in excess of 1,000 nodes with the same oversubscription ratio. This scenario reduces the number of racks available per pod to accommodate the uplinks required to connect to the super spine layer.

It is important to understand the port-density of switches used and their feature sets' impact on the number of available ports. This directly influences the number of switches necessary for proper scaling.

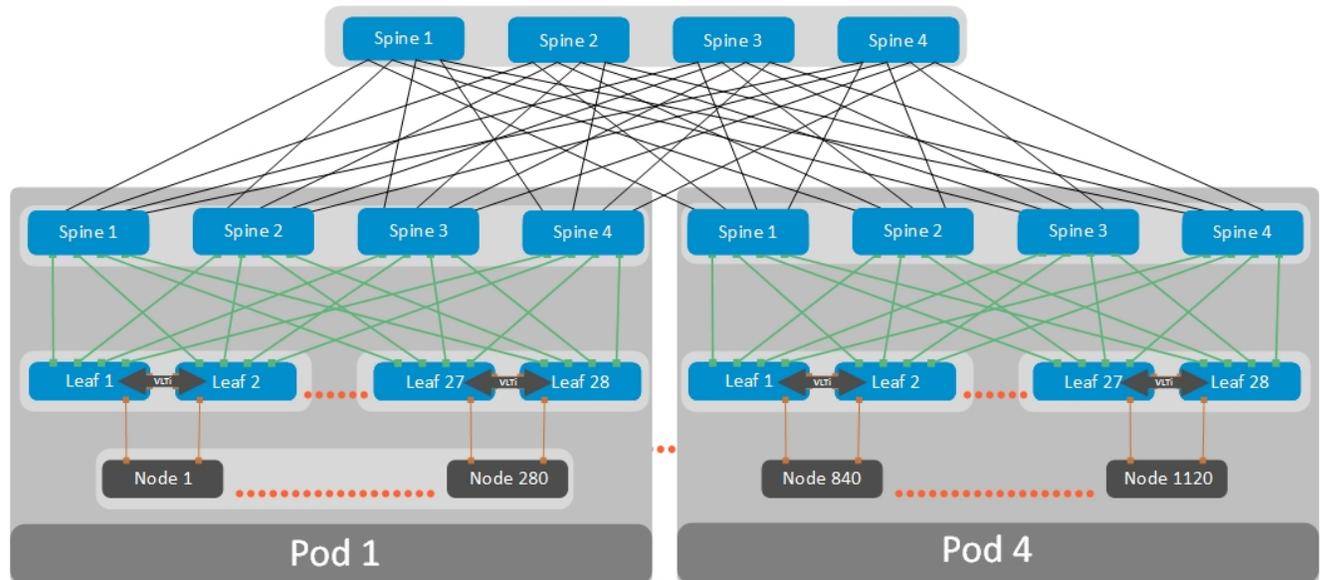


Figure 25 Scaling out the existing networking topology

## 4.4 Configure Bandwidth Allocation for System Traffic

Assign bandwidth for host management, virtual machines, iSCSI storage, NFS storage, vSphere vMotion, vSphere Fault Tolerance, Virtual SAN and vSphere Replication on the physical adapters that are connected to a vSphere Distributed Switch.

To enable bandwidth allocation for virtual machines by using Network I/O Control, configure the virtual machine system traffic. The bandwidth reservation for virtual machine traffic is also used in admission control. When you power on a virtual machine, admission control verifies the availability of sufficient bandwidth.

1. In the vSphere Web Client, navigate to the distributed switch.
2. On the Configure tab > Resource Allocation > System Traffic.
3. Select the traffic type according to the vSphere feature that you want to provision and click **Edit**.
  - a. The network resource settings for the traffic type appear.
4. From the **Shares** drop-down menu, edit the share of the traffic in the overall flow through a physical adapter. Network I/O Control applies the configured shares when a physical adapter is saturated.
  - a. You can select an option to set a pre-defined value, or select **Custom** and enter a number from 1 to 100 to set another share.
5. In the **Reservation** text box, enter a value for the minimum required bandwidth for the traffic type.

- a. The total reservation for system traffic must not exceed 75% of the bandwidth supported by the physical adapter with the lowest capacity of all adapters connected to the distributed switch.
6. In the **Limit** text box, enter the maximum bandwidth that system traffic of the selected type can use.
7. Click **OK** to apply the allocation settings.

## 4.5 Tuning Jumbo frames

In the initial system solution that was put together for this set of VxFlex OS validation scenarios, jumbo frames was not enabled. After stability of the environment was confirmed, jumbo frames were enabled on all networking components to improve the overall performance of the converged environment. This is a typical industry method for improving the performance of any networking infrastructure, especially storage and converged networks. This scenario was implemented to match the Dell EMC recommendation of initially starting with jumbo frames disabled based on the complexity of the configuration. In summary, Dell EMC recommends starting without jumbo frames for ease of configuration and initial setup. Only enable jumbo frames after evaluating the complexity and benefits to the environment. Figure 26 below shows the interfaces that require increased MTU values for a single path from SDC 1 to SDS 1.

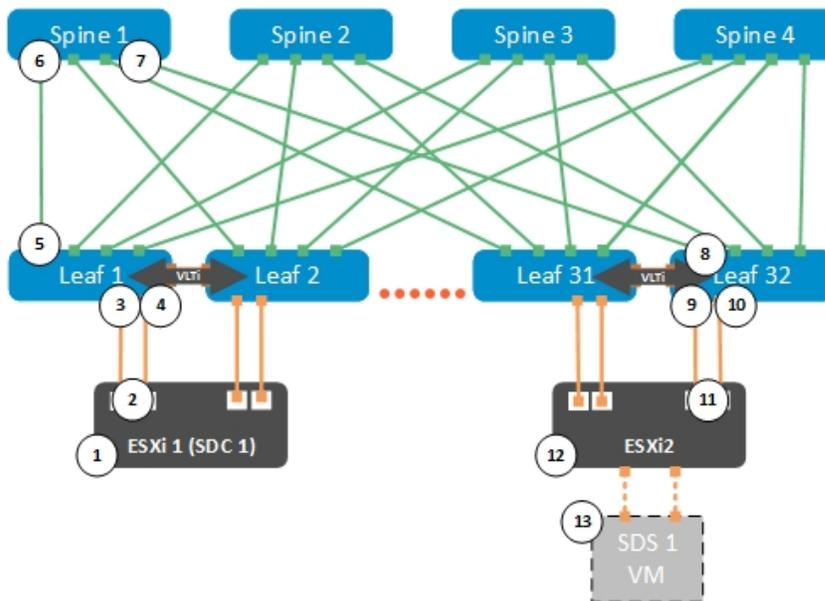


Figure 26 Jumbo frame interface configuration points

Table 7 Jumbo frames checklist

Step Number	Device	Checked?
1	VMkernel interface	
2	Rack designated VDS	
3	VLAN interface	
4	Port Channel interface and members	

5	Leaf switch interface leading to spine switch	
6	Spine switch interface from originating leaf switch	
7	Spine switch interface to designated leaf switch	
8	Leaf switch interface from originating spine switch	
9	VLAN interface	
10	Port channel interface and members	
11	Rack designated VDS	
12	VMkernel interface	
13	SDS VM interface	

## 4.6 Quality of Service (QoS)

The hyper-converged solution includes application and storage traffic distributed across the entire leaf-spine network architecture. To improve critical VxFlex OS and storage-related traffic, QoS features on the leaf and spine switches can be utilized. This section describes some basic QoS configurations and settings to enable traffic policing through the use of Differentiated Services Code Point (DSCP) marking and priority queuing.

### 4.6.1 DSCP marking on virtual distributed switches

Differentiated services provide a means to classify traffic using DSCP values. The DSCP values give storage-related traffic a higher priority than application traffic.

The deployment example uses separate physical NICs, virtual distributed switches and distributed port groups for storage and application traffic. This configuration allows the use of the traffic filtering and marking feature on distributed port groups to easily mark different DSCP values on storage and application traffic.

To mark traffic at the VM, use the settings shown below:

1. In vCenter, navigate to **Home > Networking**.
2. On the VDS, select the **ScaleIO-data01** distributed port group.
3. Right-click on the **ScaleIO-data01** port group name in the left-hand column and select **Edit Settings**.
4. Select Traffic filtering and marking.
  - a. Change the **Status** to **Enabled**.
  - b. Click **+** to add a traffic rule.
  - c. Enter a descriptive name into the **Name** field.
  - d. Keep the default **Action** as **Tag**.
  - e. Check the **DSCP value** checkbox, and enter **46** as the value.
  - f. Keep the Traffic direction as Ingress/Egress.
  - g. Click **+** and add a **New IP Qualifier**.
  - h. Change the Protocol to **any**.

- i. Leave the **Source** and **Destination Address** settings as default.
- j. Click **OK**.

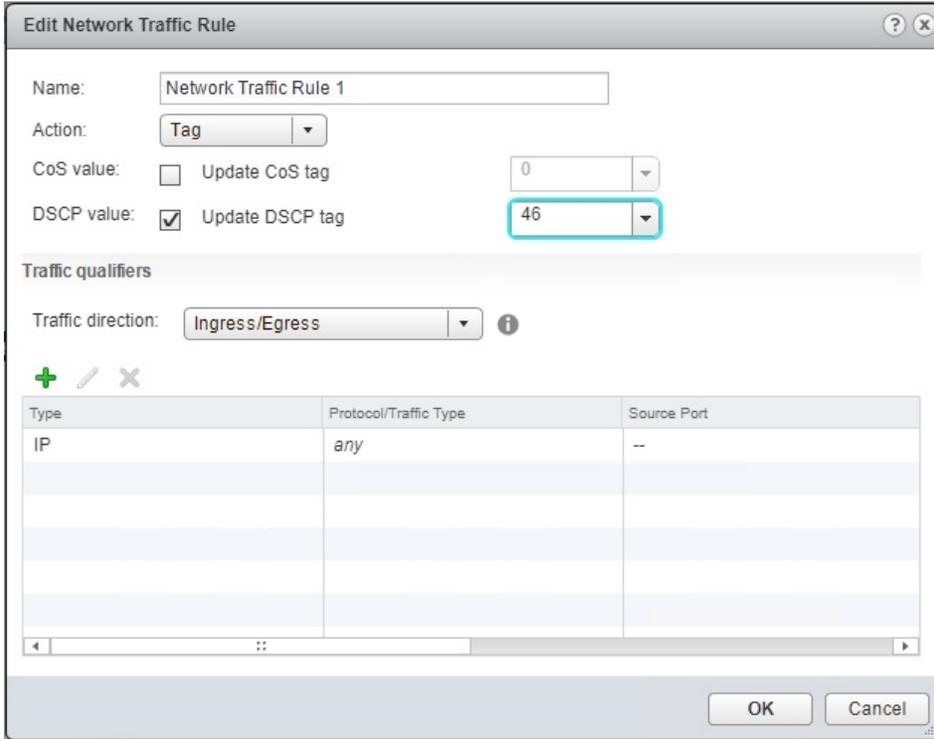


Figure 27 Traffic rule for ScaleIO-data01 distributed port group

- 5. Click **OK** to save the MDM port group settings.

This simplified example for QoS only demonstrates the use of a single class of traffic, VxFlex OS storage traffic. For more complex traffic priority requirements, administrators can assign multiple unique DSCP values to the appropriate port group. Administrators may then modify the switch configuration shown in the following sections to support additional DSCP values and queues. The DSCP value used in this example is only for demonstrating the configuration process.

## 4.6.2 Switch QoS configuration

The Dell EMC Networking switches use DSCP marking to place the appropriate traffic into queues for prioritization. This section details a simple configuration to place a higher priority on storage traffic than all other traffic.

There is not a generic, one-size-fits-all approach to QoS and Dell EMC Networking switches provide some customization. The strict queuing used in this example could easily be substituted with explicit bandwidth assignments. Administrators can use this example as a starting point for their QoS strategy.

The configuration example below accomplishes the following:

- Uses the DSCP values as configured on the ScaleIO-data01 port group
- Maps DSCP value to a specified queue
- Prioritizes egress traffic on uplinks through strict queuing

**Note:** The example below uses a DSCP value of 46 and a queue value of 3. The choice of values is arbitrary and is used to show any DSCP value can be manually mapped to any queue.

Leaf switch configuration procedure:

1. Access the command line and enter configuration mode.
2. Create a class map to match traffic for the DSCP value.

```
class-map type qos ClassQoS_Storage
match ip dscp 46
```

3. Create a policy map to map the class of traffic to a specific queue.

```
policy-map type qos PolicyQoS_Storage
class ClassQoS_Storage
set qos-group 3
```

4. Apply the policy map to the ingress interfaces using an input service policy.

```
interface ethernet 1/1/1
service-policy input type qos PolicyQoS_Storage
```

- a. Repeat step 4 for all ingress interfaces that have ScaleIO-data01 traffic.

5. Create a class map to match traffic for the queue.

```
class-map type queuing ClassQueue_Storage
match queue 3
```

6. Create a policy map to map the class of traffic to a strict priority.

```
policy-map type queuing PolicyQueue_Strict
class ClassQueue_Storage
priority
```

7. Apply the policy map to an uplink interface using an output service policy.

```
interface ethernet 1/1/42
service-policy output type queuing PolicyQueue_Strict
```

- a. Repeat step 7 for all leaf uplink interfaces.

The configuration in the above steps can be applied to any input and uplink interface throughout the leaf-spine topology. This example only implements priority queuing at the uplink interfaces. Tagging or marking occurs at the distributed port groups, and mapping occurs at the switch interfaces to the nodes.

### 4.6.3 QoS validation

Monitor QoS marking and performance on the switches through show commands. This section details the show command that can be used to evaluate if the QoS configuration is functioning. The statistics below are from application and storage traffic generated by a test traffic generator.

Following is from OS10 with QoS policy applied above. Shows no dropped packets exiting queue 3.

```
OS10# show queuing statistics interface eth 1/1/21
```

```
Interface ethernet1/1/21
```

Queue	Packets	Bytes	Dropped-Packets	Dropped-Bytes
0	5748974	8623461000	592542	888813000
1	0	0	0	0
2	0	0	0	0
3	634152	951228000	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0

## A Additional resources

This section tells you where to find documentation and other support resources for components as used in the examples this document describes.

### A.1 Virtualization components

The table below lists the software components used by this document:

Table 8 Software Components

Software	Version	Link to Documentation
VMware vSphere ESXi	6.5.0 U1 5969303	<a href="http://www.vmware.com/products/vsphere/">http://www.vmware.com/products/vsphere/</a>
VMware vCenter Server Appliance (vCSA)	6.5.0 U1 5973321	<a href="http://www.vmware.com/products/vcenter-server/">http://www.vmware.com/products/vcenter-server/</a>
EMC ScaleIO/VxFlex OS	2.0-14000.228	<a href="http://www.emc.com/storage/scaleio/index.htm">http://www.emc.com/storage/scaleio/index.htm</a>

### A.2 Dell EMC servers and switches

This section lists the servers and switches used in the examples shown in this document.

Table 9 Servers and Switches

Product	Description	Link to product information
PowerEdge R730xd	Dell EMC rack server that provides core compute infrastructure for EMC Converged Infrastructure.	<a href="http://www.dell.com/us/business/p/powered-ge-r730xd/pd">http://www.dell.com/us/business/p/powered-ge-r730xd/pd</a>
S4248FB-ON	The Dell EMC Networking S-Series S4248FB-ON is an ultralow-latency 10/40GbE top-of-rack (ToR) switch with deep buffers, built for applications in high-performance data center and computing environments.	<a href="http://www.dell.com/en-us/work/shop/povw/networking-s-series-10gbe">http://www.dell.com/en-us/work/shop/povw/networking-s-series-10gbe</a>
Z9100-ON	The Dell EMC Networking Z9100-ON is a 10/25/40/50/100GbE top-of-rack (ToR) fixed switch purpose-built for applications in high-performance data center and computing environments.	<a href="http://www.dell.com/en-us/work/shop/povw/networking-z-series">http://www.dell.com/en-us/work/shop/povw/networking-z-series</a>

## A.3 Server and switch component details

The table below lists the BIOS, firmware and driver components used in the examples shown in this document:

Table 10 BIOS, Firmware, and Switch OS Components

Component	Version	Notes
PowerEdge R730xd Server BIOS	2.4.3	BIOS facilitates the hardware initialization process and transitions control to the operating system.
Integrated Remote Access Controller (iDRAC)	2.41.40.40	iDRAC is a systems management hardware and software solution that provides remote management capabilities, crashed-system recovery and power control functions for PowerEdge systems.
PERC H730 RAID Controller	Firmware: 25.5.2.0001 ESXi lsi_mr3 Driver: 7.700.50.00	12 Gbps RAID controller supporting SAS or SATA hard disk or solid-state drives, provides unsurpassed performance and enterprise-class reliability.
Intel X520 1/10Gb Ethernet Network Adapter	Firmware: 18.0.6 ESXi net-ixgbe Driver: 4.5.1	Intel's X520 Converged Network Ethernet Adapters are flexible and scalable for today's demanding data center and cloud environments by providing unmatched features for virtualization, SAN networking and proven reliable performance.
DNOS Z9100-ON	10.4.0E(R3)	The Dell EMC Networking Z9100-ON switch adds multiline rate capability supporting 10GbE, 25GbE, 40GbE, 50GbE and 100GbE. This switch provides for substantial growth with this initial configuration using 40GbE links but able to move to 25GbE downlinks and 100GbE uplinks.
DNOS S4248FB-ON	10.4.0E(R3)	The Dell EMC Networking S4048-ON switch can provide low latency, non-blocking layer-2 network architecture. The forty-eight 10GbE and six 40GbE ports in a single rack unit provide flexibility in the data center for 1/10/40GbE uplink compatibility.

## A.4 PowerEdge R730xd server

PowerEdge R730xd is a 2-socket CPU, 2U, multi-purpose server, offering an excellent balance of ultra-dense internal storage, redundancy and value in a compact form factor. It is a hardware building block for any mid-size or large business that provides scalability in memory density and storage capacity and IOPS performance in a dense 2U form-factor.



Figure 28 R730xd front view with bezel



Figure 29 R730xd front view without bezel

In addition to the R730xd back-panel features, the R730xd includes two optional 2.5" hot-plug drives in the back of the system.



Figure 30 R730xd back view

## A.5 S4248FB-ON switch

The S4248FB-ON is the latest generation S-Series multi-purpose 10/40/100GbE switch with deep buffers for optimum performance and connectivity, featuring 40 x 10GbE SFP+ ports + 2 x 40GbE QFSP+ ports + 6 x 100GbE QSFP28 ports. S4248FB-ON switches are used as leaf switches in the Leaf-Spine topology covered in this guide.



Figure 31 S4248FB-ON

## A.6 Z9100-ON switch

The Z9100-ON is a 1RU layer 2/3 switch with 32 ports supporting 10/25/40/50/100GbE. Two Z9100-ON switches are used as spine switches in the leaf-spine topology covered in this guide.

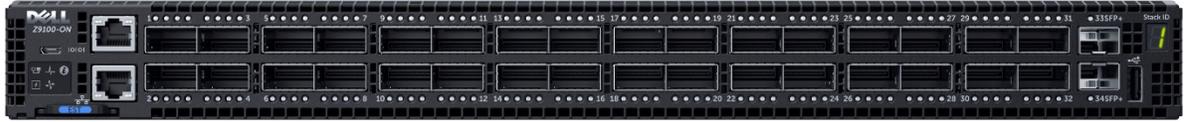


Figure 32 Z9100-ON

## A.7 S3048-ON switch

The S3048-ON is a 1RU layer 2/3 switch with 48, 1GbE Base-T ports. One S3048-ON switch is used for OOB management traffic in this guide.



Figure 33 S3048-ON

## B Prepare your environment

This section covers basic PowerEdge server preparation and ESXi hypervisor installation. Installation of guest operating systems (Microsoft Windows Server, Red Hat Linux, etc.) is outside the scope of this document.

**Note:** Exact iDRAC console steps in this section may vary slightly depending on hardware, software and browser versions used. See your PowerEdge server documentation for steps to connect to the iDRAC virtual console.

### B.1 Confirm CPU virtualization is enabled in BIOS

**Note:** CPU virtualization is typically enabled by default in PowerEdge server BIOS. These steps are provided for reference in case this required feature has been disabled.

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the Next Boot menu, select **BIOS Setup**.
3. Reboot the server.
4. From the System Setup Main Menu, select **System BIOS**, and then select **Processor Settings**.
5. Verify Virtualization Technology is set to Enabled.
6. To save the settings, click **Back**, **Finish**, and **Yes** if prompted to save changes.
7. If resetting network adapters to defaults, proceed to step 4, System Setup Main Menu, in the next section. Otherwise, reboot the server.

### B.2 Confirm network adapters are at factory default settings

Complete the following steps:

**Note:** These steps are only necessary if installed network adapters have been modified from their factory default settings.

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the Next Boot menu, select **BIOS Setup**.
3. Reboot the server.
4. From the System Setup Main Menu, select **Device Settings**.
5. From the Device Settings page, select the first port of the first NIC in the list.
6. From the Main Configuration Page, click the **Default button** followed by **Yes** to load the default settings. Click **OK**.
7. To save the settings, click **Finish** then **Yes** to save changes. Click **OK**.
8. Repeat for each NIC and port listed on the Device Settings page.
9. Reboot the server.

## B.3 Configure the PERC H730 Controller

As a best practice, Dell EMC recommends using the PERC H730 controller in RAID mode and create a RAID-0 container for each disk attached to the controller. This allows RDM (Raw Device Mapping) for all hard disk to be mapped directly to the SVM.

Storage controllers used in an EMC VxFlex OS deployment should be set to RAID mode. For the deployment used in this guide, this applies to all PERC H730 controllers in each of the R730XD servers.

To verify storage controllers are in RAID mode, complete the following steps:

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the Next Boot menu, select **BIOS Setup**.
3. Reboot the server.
4. From the System Setup Main Menu, select **Device Settings**.
5. From the list of devices, select the **PERC controller**. This opens the Modular RAID Controller Configuration Utility Main Menu.
6. Select **Controller Management**. Scroll down to Controller Mode and verify it is set to RAID. If set to HBA, select **Advanced Controller Management > Switch to RAID Mode > OK**.

The H730 controller can handle both RAID and non-RAID disks. Each HDD disk attached to the PERC controller needs to be placed in a separate RAID-0 container:

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the Next Boot menu, select **BIOS Setup**.
3. Reboot the server.
4. From the System Setup Main Menu, select **Device Settings**.
5. From the list of devices, select the **PERC controller**. This opens the Modular RAID Controller Configuration Utility Main Menu.
6. Select **Configuration Management > Create Virtual Disk**.
7. Choose RAID0 for the RAID level.
8. Click **Select Physical Disks**.
9. Set the Media Type to HDD.
10. Choose the first available disk > Select **Apply Changes > OK**.
11. Scroll to the bottom of the Create Virtual Disk window and select **Create Virtual Disk**.
12. On the Virtual Disk warning window, check the Confirm box and choose **Yes**.
13. Choose **OK**.

Repeat for all remaining disks that are part of the VxFlex OS environment. This deployment example uses four R730XD servers each using 24 disks for 96 RAID-0 containers.

To verify that 24 virtual disks have been created, complete the following steps:

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, from the Next Boot menu, select **BIOS Setup**.
3. Reboot the server.

4. From the System Setup Main Menu, select **Device Settings**.
5. From the list of devices, select the **PERC controller**. This opens the Modular RAID Controller Configuration Utility Main Menu.
6. Select **Virtual Disk Management**.
7. The total number of virtual disks should equal 24 (Virtual Disk 0 through Virtual Disk 23).

## B.4 Install ESXi

Dell EMC recommends using the latest Dell EMC customized ESXi .iso image available on [www.dell.com/support](http://www.dell.com/support). The correct drivers for your PowerEdge hardware are built into this image.

Install ESXi on all servers that will be part of your deployment. For the example in this guide, ESXi is installed to redundant internal SD cards in the PowerEdge servers. This includes three R630 servers and four R730xd servers.

A simple way to install ESXi on a PowerEdge server remotely is by using the iDRAC to boot the server directly to the ESXi .iso image. To do this, complete the following steps:

1. Connect to the iDRAC in a web browser and launch the virtual console.
2. In the virtual console, select **Virtual Media > Connect Virtual Media**.
3. Select **Virtual Media > Map CD/DVD >** browse to the Dell EMC customized ESXi .iso image **> Open > Map Device**.
4. Select **Next Boot > Virtual CD/DVD/ISO > OK**.
5. Select **Power > Reset System (warm boot)**. Answer **Yes** to reboot the server.
6. The server reboots to the ESXi .iso image and installation starts.
7. Follow the prompts to install ESXi. Select the server's Internal Dual SD Module (IDSMD) when prompted for a location.
8. After installation is complete, click **Virtual Media > Disconnect Virtual Media > Yes**.
9. Reboot the system when prompted.

## B.5 Configure the ESXi management network connection

Be sure the host is physically connected to the management network. For this deployment, Intel 10G 2P X520 adapters provide this connection for R730xd servers.

1. Log in to the ESXi console and select **Configure Management Network > Network Adapters**.
2. Select the correct vmnic for the management network connection. Follow the prompts on the screen to make the selection.
3. Go to **Configure Management Network > IPv4 Configuration**. If DHCP is not used, specify a static IP address, mask, and default gateway for the management interface.
4. Optionally, configure DNS settings from the Configure Management Network menu if DNS is used on your network.
5. Press **Esc** to exit and answer **Y** to apply the changes.
6. From the ESXi main menu, select **Test Management Network**. Verify pings are successful. If there is an error, be sure you have configured the correct vmnic.

7. Optionally, under Troubleshooting Options, enable the ESXi shell and SSH to enable remote access to the CLI.
8. Log out of the ESXi console.

## C VxFlex OS with routed leaf - multiple rack considerations

The section provides information relating to scaling the VxFlex OS example described throughout this document. When using a layer 3 design to include the leaf and spine, additional steps are needed to ensure VxFlex OS and VMware network settings are complete.

### C.1 VMkernel configuration - add static routes for default gateways

The default gateway for each network configured on the VMkernels cannot be modified during VMkernel creation. To set the default gateway requires configuring a static route at the command line of each ESXi host. These steps are normally taken immediately after configuring the VMkernels, which are detailed in section [3.3.9](#).

To add the static routes on each ESXi host, complete the following steps:

1. Open a SSH session to an ESXi host and login as the root user.
2. Use the following `esxcli` commands to add the default gateway routes for VLANs 1732 and 1734:

```
[root@ atx01w02esx01:~] esxcli network ip route ipv4 add -n=172.17.32.0/24 -g=172.17.32.253
```

```
[root@ atx01w02esx01:~] esxcli network ip route ipv4 add -n=172.17.34.0/24 -g=172.17.34.253
```

3. To verify the routes are configured, use the following `esxcli` command:

```
[root@ atx01w02esx01:~] esxcli network ip route ipv4 list
```

Repeat steps 1-3 for each ESXi host in each cluster.

### C.2 Deploying VxFlex OS

This section summarizes how to deploy VxFlex OS in the VMware environment. Deployment entails the following tasks:

- Registering the VxFlex OS plug-in
- Uploading the OVA template to vCenter
- Accessing the plug-in
- Installing the SDC on ESXi hosts
- Deploying VxFlex OS using the VxFlex OS VMware Deployment Wizard

**Note:** Navigate to [Dell EMC ScaleIO/VxFlex OS Software Only: Documentation Library](#) to download the full documentation package and [Dell EMC ScaleIO/VxFlex OS Documentation Hub](#) for additional VxFlex OS documentation. You may need to enter EMC Community Network (ECN) login credentials or create an ECN account to access the documentation package.

## C.2.1 Registering the VxFlex OS Plug-in and uploading the OVA template

The VxFlex OS plugin for vSphere simplifies the installation and management of the VxFlex OS system in an ESXi environment. The initial VxFlex OS system configuration plugin utilizes the plugin as well as for adding additional SDS nodes.

To register the VxFlex OS plugin follow the procedure in the ***EMC ScaleIO/VxFlex OS 2.0.x Deployment guide, Chapter 4: Registering the ScaleIO plug-in and uploading the OVA template.***

Software versions used in this document:

- VMware vSphere PowerCLI 6.5.4.7155375
- ScaleIO v2.0-14000.228
  - filename: ScaleIO\_2.0.1.4\_Complete\_VMware\_SW\_Download.zip
    - > ScaleIOVM\_2nics\_2.0.14000.231.ova
    - > EMC-ScaleIO-vSphere-plugin-installer-2.0-14000.231.zip

Procedure (summary):

- Copy the .ova and plugin-installer zip files into the vCenter host.
- Extract the zip file.
- Use PowerCLI for VMware to Run plugin setup script (.ps1).
- Verify that the EMC ScaleIO icon is visible in the vCenter GUI within **Home > Inventories**.
- Use PowerCLI for VMware and plugin script to upload the .ova template to the datastore(s).
  - Create SVM template
- Exit the plug-in script.



Figure 34 EMC ScaleIO plugin icon

## C.2.2 Installing the SDC on ESXi hosts

The SDC component must be applied to every ESXi host within the VxFlex OS system.

Complete the Pre-Deployment procedure summarized below within the EMC VxFlex OS plugin application:

- **Basic tasks > Pre-Deployment Actions.**
- Select all the ESXi hosts, check Install SDC, and provide root passwords.
- Complete install and restart the ESXi hosts.

## C.2.3 VxFlex OS deployment

This section provides information on using the deployment wizard for the deployment example in this guide. For detailed information on each deployment step and how to apply changes based on a specific topology or hardware configuration, see the *ScaleIO/VxFlex OS v2.0.x Deployment Guide* located at [Dell EMC ScaleIO/VxFlex OS Software Only: Documentation Library](#).

Complete the procedures summarized in the following note from within the EMC VxFlex OS plugin application.

Before using the deployment wizard, use the **Advanced Settings** link to **Enable RDMs on nonparallel SCSI controllers** (check the box). This enables non-SCSI controller devices to be added as RDM (Raw Device Mapping).

**Note:** Do not enable this option if the device does not support SCSI Inquiry Vital Data Product (VPD) page code 0x83.

- **Basic Tasks > Deploy ScaleIO environment.**

- **Create a new ScaleIO system;** agree to license terms.
- Enter a **System Name** and **password**.
- Select the vCenter server and all hosts in all clusters.
- Select a **3-node cluster**.
  - Initial Master MDM > 172.17.33.12
  - Manager MDM > 172.17.33.13
  - TieBreaker MDM > 172.17.33.14
  - Leave optional settings as default.
- **Configure Performance, Sizing, and Syslog;** leave the setting as default.
- Enter a **Protection Domain** name.
- Enter **Storage Pool** name(s), enable zero padding.
- **Create new Fault Sets** is left as its default; none are created in this example.
- **Add SDSs**.
  - Select **SDS** to assign an SDS role. Assign a Protection Domain.
- **Add devices to SDSs**.
  - **Information tab**, select an ESXi host, click **Assign devices**.
  - For each SCSI device, select **Storage**, select the appropriate storage pool, then click **Assign**.
  - Repeat or replicate for other hosts.
- **Add SDCs**.
  - For each ESXi host to be added as an SDC, select the **SDC** check box.
  - Enter root passwords.
  - Disable SCSI LUN number comparison for hosts.
- **Configure Upgrade Components**.
  - Confirm an ESXi host selected to be the Gateway VM.
  - Provide a Gateway administrative password.
  - Provide an LIA password.
- **Select OVA Template**.
  - Select the template to use to create the SVM. (Uploading templates to multiple datastores is recommended for faster deployment, see the *ScaleIO/VxFlex OS Deployment Guide* located at [Dell EMC ScaleIO/VxFlex OS Software Only: Documentation Library](#)).
  - Provide a root password for the SVMs.
- **Configure networks**.
  - Management network label, select **IPv4** and **management**.
  - Data network label, select **IPv4** and **ScaleIO-data01**.
    - > Network labels used are the management and storage networks configured in section 3.

- **Configure SVM.**
  - Configure ScaleIO Virtual Machine (SVM) IP addresses.
  - Enter IP information for all SVMs.
  - See [Table 11](#) for IP information for this deployment example.
  - This example uses **172.17.34.4** is used for the Cluster Virtual IP address, network **ScaleIO-data01**.

Table 11 ScaleIO Wizard – Configure SVM

ESXi Name (function)	Mgmt IP & Subnet Mask	Default Gateway	Data IP & Subnet Mask
ScaleIO-172.17.33.11-GW (ScaleIO Gateway)	172.17.33.11 / 255.255.255.0	172.17.33.253	172.17.34.11 / 255.255.255.0
ScaleIO-172.17.33.12 (Master MDM)	172.17.33.12 / 255.255.255.0	172.17.33.253	172.17.34.12 / 255.255.255.0
ScaleIO-172.17.33.13 (Slave 1 MDM)	172.17.33.13 / 255.255.255.0	172.17.33.253	172.17.34.13 / 255.255.255.0
ScaleIO-172.17.33.14 (TieBreaker 1)	172.17.33.14 / 255.255.255.0	172.17.33.253	172.17.34.14 / 255.255.255.0
ScaleIO-172.17.33.15	172.17.33.15 / 255.255.255.0	172.17.33.253	172.17.34.15 / 255.255.255.0

Press **Finish** to begin the ScaleIO deployment process.

## C.2.4 Deployment wizard modifications

The wizard does not support all deployment and network topology scenarios. This section describes two specific design details used in this deployment example that need to be addressed midway through the deployment. The deployment wizard is expected to fail device tasks and pauses the deployment at those points. After changes are made to the configuration, the wizard resumes the deployment.

### Routed leaf-spine

The deployment wizard is designed for networks that do not deploy routing at the leaf layer. The SVM .ova supplied with the ScaleIO package includes a single IP stack which has a single default gateway. That gateway supports the out of band management network.

The network architecture for routed leaf-spine requires a route to add a next hop to the storage network. Each rack in a routed leaf-spine is its own L2 domain.

## Modification steps

This section provides step-by-step instructions for modifying the SVMs to continue with the deployment wizard.

Procedure:

1. Let the deployment process complete with failed tasks. The EMC VxFlex OS plugin screen will show errors.
2. Navigate to the console of any SVM. Use the **Open Console** option or an SSH client to open a session to the **Mgmt IP** listed in [Table 11](#).
3. On the console, enter the following to add a route to the storage network gateway and restart the network services: (the command below is for the Master MDM SVM in this deployment example.)

```
ScaleIO-172-17-33-12:~ #echo "172.18.34.0/24 172.17.34.253 255.255.255.0 eth1" >> /etc/sysconfig/network/routes
```

```
ScaleIO-172-17-33-12:~ #service network restart
```

**Note:** IP information for multiple racks are not included in the main body of this document. The IP information above is provided as an example only and not part of any configuration. Apply the appropriate settings for your network.

4. Verify that the VxFlex OS SVMs can send traffic across the spines by pinging the Cluster Virtual IP address from the console of each SVM.
5. Verify the host routing table is correct by pinging the Cluster Virtual IP address, from the console of each ESXi host.
6. Run additional ping tests as follows:
  - a. Master MDM SVM to:
    - i. ESXi hosts
    - ii. ScaleIO-data01 VMkernel Ports
  - b. SDS SVM to:
    - i. Master MDM SVM
    - ii. Cluster Virtual IP
  - c. ESXi hosts to Master MDM SVM