

# ScaleIO IP Fabric Best Practice and Deployment Guide

## Dell EMC Networking Leaf-Spine Architecture for ScaleIO

Dell EMC Networking Solutions Engineering  
April 2018

## Revisions

| Date       | Version | Description  | Authors   |
|------------|---------|--|---|
| April 2018 | 2.1     | Updated BFD timers and added guidance on BFD settings  | Ed Blazek, Curtis Bunch, Colin King   |
| March 2017 | 2.0     | Updated best practice sections to improve document consistency. ScaleIO deployment steps added utilizing distributed switch. Leaf-Spine reference architecture was updated | Networking Solutions: Ed Blazek, Curtis Bunch , Colin King  |
| July 2016  | 1.0     | Initial release  | Networking Solutions: Curtis Bunch, Michael Matthews, Dennis Dadey, Kevin Locklear, Davis Smith<br>Networking Marketing: Drew Schulke |

©2016 - 2017 Dell Inc., All rights reserved.

Except as stated below, no part of this document may be reproduced, distributed or transmitted in any form or by any means, without express permission of Dell EMC.

You may distribute this document within your company or organization only, without alteration of its contents.

THIS DOCUMENT IS PROVIDED "AS-IS", AND WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED. IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE SPECIFICALLY DISCLAIMED. PRODUCT WARRANTIES APPLICABLE TO THE DELL PRODUCTS DESCRIBED IN THIS DOCUMENT MAY BE FOUND AT:

<http://www.dell.com/learn/us/en/vn/terms-of-sale-commercial-and-public-sector-warranties> Performance of network reference architectures discussed in this document may vary with differing deployment conditions, network loads, and the like. Third party products may be included in reference architectures for the convenience of the reader. Inclusion of such third party products does not necessarily constitute Dell EMC's recommendation of those products. Please consult your Dell EMC representative for additional information.

Trademarks used in this text:

Dell™, the Dell logo, PowerEdge™, PowerVault™, Dell Networking™, OpenManage™, and FlexAddress™ are trademarks of Dell Inc. Other Dell trademarks may be used in this document. EMC®, and EMC ScaleIO®, are registered trademarks of EMC Corporation. Intel®, Xeon®, Core® and Celeron® are registered trademarks of Intel Corporation in the U.S. and other countries. Microsoft®, Windows®, Windows Server®, Internet Explorer®, and Active Directory® are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries. VMware®, Virtual SMP®, vMotion®, vCenter® and vSphere® are registered trademarks or trademarks of VMware, Inc. in the United States or other countries. IBM® is a registered trademark of International Business Machines Corporation. Broadcom® and NetXtreme® are registered trademarks of Broadcom Corporation. QLogic® is a registered trademark of QLogic Corporation. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and/or names or their products and are the property of their respective owners. Dell disclaims proprietary interest in the marks and names of others.

# Table of contents

|   |    |
|---|----|
| Revisions.....  | 2  |
| 1 Introduction to network virtualization and hyper-converged infrastructure and networking..... | 6  |
| 1.1 ScaleIO.....  | 7  |
| 1.1.1 Benefits of ScaleIO.....  | 7  |
| 1.1.2 Components.....   | 9  |
| 1.2 Traffic model for the modern data center.....   | 9  |
| 1.3 Attachments.....  | 10 |
| 2 Building a Leaf-Spine topology.....   | 11 |
| 2.1 Management network.....   | 12 |
| 2.2 IP-based storage.....   | 13 |
| 2.3 Routing protocol selection for Leaf-Spine.....  | 13 |
| 2.3.1 BGP.....  | 13 |
| 2.3.2 OSPF.....   | 14 |
| 2.3.3 IS-IS.....  | 14 |
| 2.3.4 BFD.....  | 14 |
| 2.4 Layer 2 considerations.....   | 14 |
| 2.4.1 VLT.....  | 15 |
| 2.4.2 VRRP.....   | 15 |
| 2.4.3 Uplink Failure Detection.....   | 15 |
| 3 Configuration and Deployment.....   | 17 |
| 3.1 ScaleIO solution example.....   | 17 |
| 3.1.1 ScaleIO MDM cluster.....  | 17 |
| 3.1.2 ScaleIO SDS-SDC nodes.....  | 17 |
| 3.1.3 ScaleIO deployment options.....   | 18 |
| 3.2 Physical switch configuration.....  | 18 |
| 3.2.1 BGP ASN configuration.....  | 18 |
| 3.2.2 BGP fast fall-over.....   | 19 |
| 3.2.3 Loopback addresses.....   | 19 |
| 3.2.4 Point-to-point interfaces.....  | 20 |
| 3.2.5 Interface/IP configuration.....   | 21 |
| 3.2.6 ECMP.....   | 21 |
| 3.2.7 VRRP.....   | 22 |

|       |  |    |
|-------|--|----|
| 3.3   | Create a datacenter object and add hosts .....                       | 22 |
| 3.4   | Create clusters and add hosts .....                                  | 23 |
| 3.5   | Information on vSphere standard switches .....                       | 24 |
| 3.6   | Deploy vSphere distributed switches .....                            | 25 |
| 3.7   | Create each VDS.....   | 27 |
| 3.8   | Add distributed port groups .....                                    | 28 |
| 3.9   | Create LACP LAGs .....   | 30 |
| 3.10  | Associate hosts and assign uplinks.....                              | 32 |
| 3.11  | Configure teaming and failover on LAGs .....                         | 35 |
| 3.12  | Configure teaming and failover on MDM uplinks.....                   | 36 |
| 3.13  | Add VMkernel adapters for MDM, SDS-SDC and vMotion .....             | 37 |
| 3.14  | Add static routes for default gateways .....                         | 39 |
| 3.15  | Verify VDS configuration .....                                       | 40 |
| 3.16  | Enable LLDP.....   | 42 |
| 4     | Deploying ScaleIO.....   | 44 |
| 4.1   | Registering the ScaleIO Plug-in and uploading the OVA template ..... | 44 |
| 4.2   | Installing the SDC on a ESXi hosts .....                             | 46 |
| 4.3   | ScaleIO deployment .....   | 46 |
| 4.4   | Deployment wizard modifications .....                                | 51 |
| 4.4.1 | ScaleIO VDS.....   | 51 |
| 4.4.2 | Routed leaf-spine.....   | 52 |
| 4.4.3 | Detailed modification steps.....                                     | 52 |
| 5     | Scaling and tuning guidance .....                                    | 57 |
| 5.1   | Decisions on scaling .....   | 57 |
| 5.2   | Examples of scaling, port count and oversubscription .....           | 57 |
| 5.3   | Scaling beyond 16 racks .....  | 58 |
| 5.4   | Configure Bandwidth Allocation for System Traffic .....              | 58 |
| 5.5   | Tuning Jumbo frames.....   | 59 |
| 5.6   | Quality of Service (QoS) .....                                       | 60 |
| 5.6.1 | DSCP marking on virtual distributed switches.....                    | 60 |
| 5.6.2 | Switch QoS configuration .....                                       | 62 |
| 5.6.3 | QoS validation .....   | 63 |
| A     | Additional resources.....  | 65 |

|     |  |    |
|-----|--|----|
| A.1 | Virtualization components .....                                | 65 |
| A.2 | Dell EMC servers and switches .....                            | 65 |
| A.3 | Server and switch component details .....                      | 66 |
| A.4 | PowerEdge R730xd server .....                                  | 66 |
| A.5 | PowerEdge R630 server .....                                    | 67 |
| A.6 | S4048-ON switch .....  | 67 |
| A.7 | Z9100-ON switch .....  | 68 |
| A.8 | S3048-ON switch .....  | 68 |
| B   | Prepare your environment .....                                 | 69 |
| B.1 | Confirm CPU virtualization is enabled in BIOS .....            | 69 |
| B.2 | Confirm network adapters are at factory default settings ..... | 69 |
| B.3 | Configure the PERC H730 Controller .....                       | 70 |
| B.4 | Install ESXi .....   | 71 |
| B.5 | Configure the ESXi management network connection .....         | 71 |
| C   | Support and feedback .....                                     | 73 |
|     | About Dell EMC .....   | 73 |

# 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.

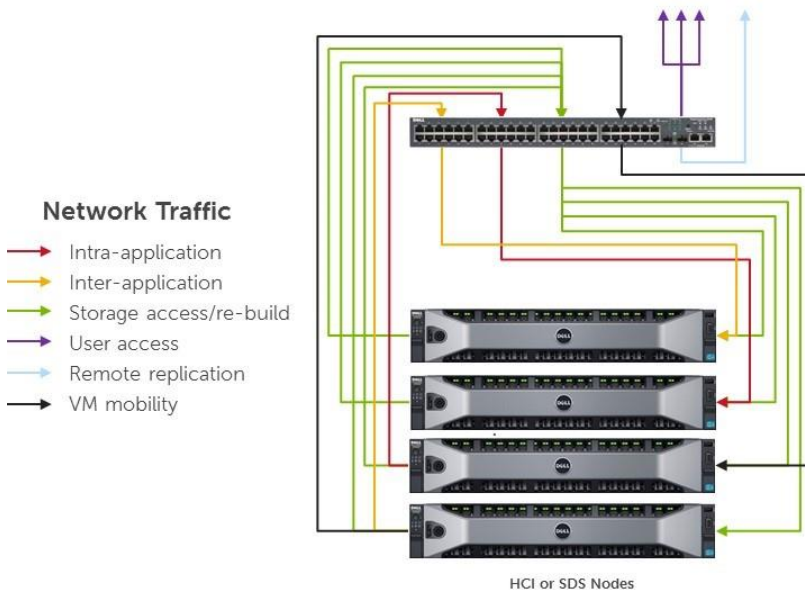


Figure 1 HCI cluster with typical network traffic types

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 from Dell EMC, ScaleIO. 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 ScaleIO 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 ScaleIO

Dell EMC ScaleIO 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. ScaleIO combines different types of storage (hard disk drives, solid-state disks and Peripheral Component Interconnect Express [PCIe] flash cards) to create shared block storage. ScaleIO is hardware-agnostic and supports physical and/or virtual application servers. By not requiring an FC fabric to connect the servers and storage, ScaleIO reduces the cost and complexity of a traditional FC SAN

**Note:** ScaleIO supports all network speeds, including 100Mb, 1Gb, 10Gb, 40Gb, 100Gb and InfiniBand (IB).

This document supplements existing ScaleIO documentation. It provides details on how to implement a ScaleIO solution using Dell EMC products. Existing ScaleIO documentation includes the following:

[EMC ScaleIO 2.0 User's Guide](#)

[EMC ScaleIO Basic Architecture](#)

[EMC ScaleIO Networking Best Practices and Design Considerations](#)

**Note:** You may need to enter EMC Community Network (ECN) login credentials or create an ECN account to access the preceding documents.

### 1.1.1 Benefits of ScaleIO

The benefits of ScaleIO include the following:

#### **Scalability**

ScaleIO 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 ScaleIO 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. ScaleIO 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**

ScaleIO 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

ScaleIO offers two flexible deployment options: two-tier (Figure 2) and hyper-converged (Figure 3). A two-tier deployment includes the application and storage installed on separate servers in the ScaleIO cluster. This provides efficient active/active paths and no single points of failure. A hyper-converged deployment includes the application and storage installed on the same server in the ScaleIO cluster. This creates a single-tier architecture with the lowest footprint and cost profile. ScaleIO provides additional flexibility by supporting mixed server brands, operating systems and storage media.

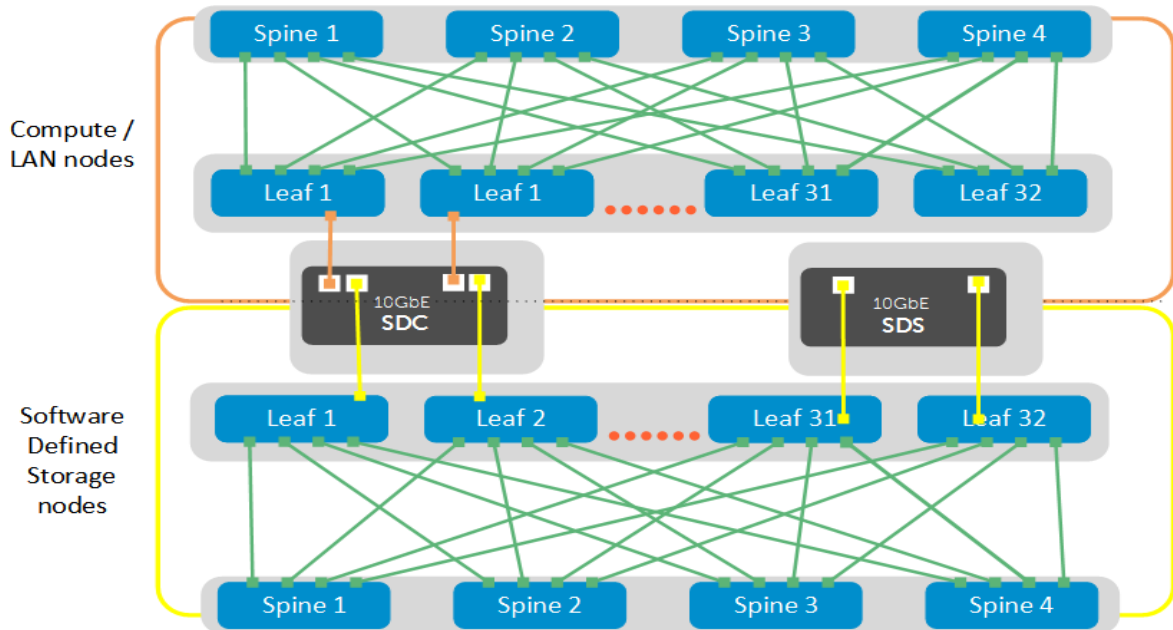


Figure 2 Traditional two-tier separated LAN and SAN topology design

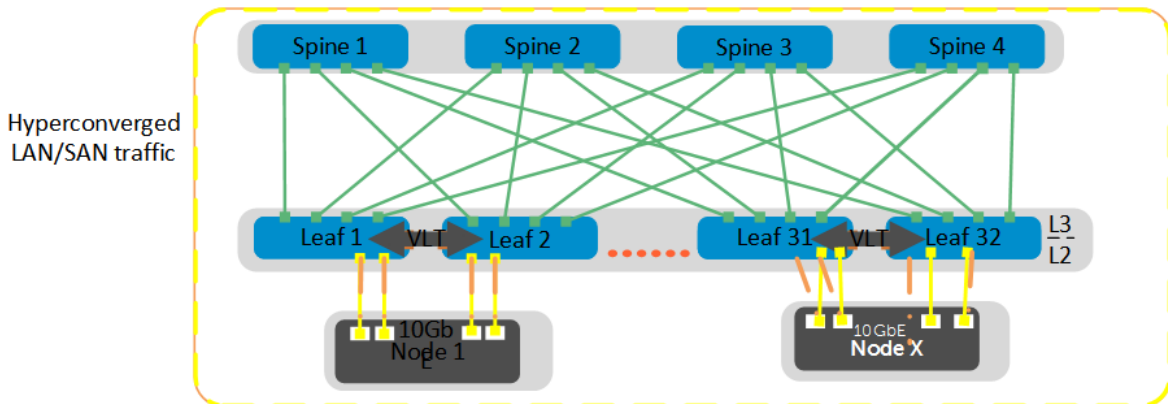


Figure 3 Hyper-converged topology design



### **Supreme Elasticity**

ScaleIO 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**

ScaleIO is comprised of three main components.

### **ScaleIO Data Client (SDC)**

The SDC is a lightweight, block device-driver that presents ScaleIO 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.

### **ScaleIO Data Server (SDS)**

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

### **Meta Data Manager (MDM)**

The MDM configures and monitors ScaleIO. It contains all the metadata required for ScaleIO 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 ScaleIO that involve hundreds of deployed nodes and require higher bandwidth and scalability.

In large data clusters, ScaleIO 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 S4048-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 4 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 connection 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 six leaf switches. Two-leaf switches per-rack provide compute node redundancy for a maximum of three racks. The first rack provides management services, ScaleIO management and VMware management. The remaining two racks provide ScaleIO 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.

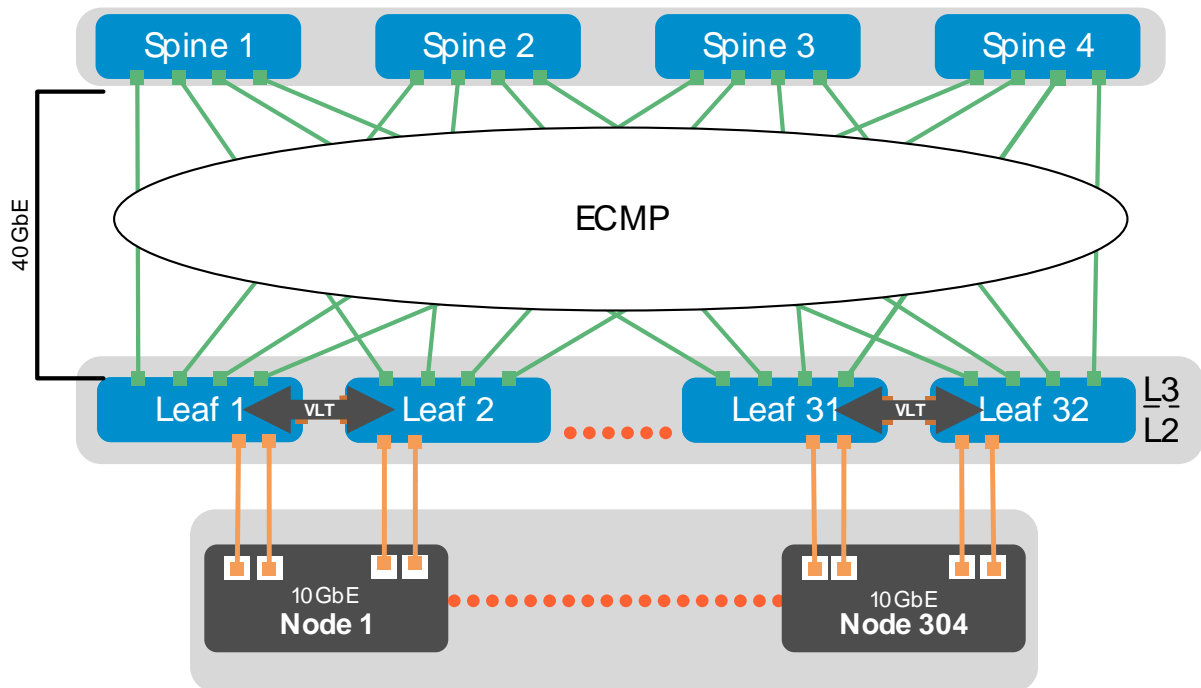


Figure 4 Leaf-Spine topology example

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

The deployment covered by this guide uses a single management-traffic network isolated from the production network. A Dell EMC S3048-ON switch in each rack provides connectivity to the management network. Out-of-band (OOB) ports on each production switch connect them to the management network. The management network is discussed early so you can use it to configure switches in the remainder of the document.

The example in this document uses four switches for management. Each spine switch and leaf switch connects to the management network as shown in Figure 5. The management network throughout this document is on the 100.67.x.x/24 network.

**Note:** 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.

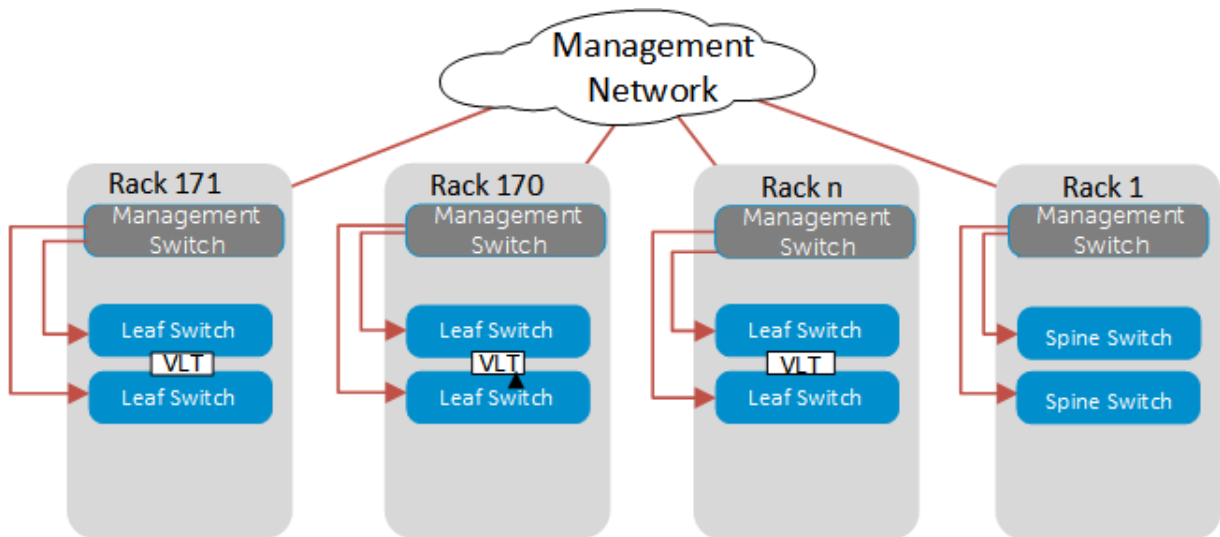


Figure 5 Management network

## 2.2 IP-based storage

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

## 2.3 Routing protocol selection for Leaf-Spine

Choose from the following routing protocols when designing a Leaf-Spine network.

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

This guide has examples for 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 fastest path among all available paths and requires no additional configurations as the protocol dynamically reacts to any changes in paths.  | BGP is a distance-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.

### 2.3.3 IS-IS

IS-IS, like OSPF, is a link-state routing protocol to compute the best path through the network. Though supported, the protocol is not widely used and is not covered in this document. To use IS-IS for routing, consult the User Guides of the switches in which the implementation is desired.

### 2.3.4 BFD

Regardless of the choice of routing protocol (OSPF or BGP), the use of Bidirectional Forwarding Detection (BFD) is required. BFD reduces the overhead associated with protocol-native hello timers, allowing link failures to be detected quickly. BFD provides faster failure detection than native protocol hello timers for a number of reasons including reduction in router CPU and bandwidth utilization. BFD is therefore strongly recommended over aggressive protocol hello timers.

For a network to converge, the event must be detected, propagated to other routers, processed by the routers, and the routing information base (RIB) or Forwarding Information Base (FIB) must be updated. All these steps must be performed for the routing protocol to converge, and they should all complete in less than 300 milliseconds.

The configuration for a 150 millisecond hold down timer consisted of 50 millisecond transmission intervals, with a 50 millisecond min\_rx and a multiplier of 3. The ScaleIO recommendation is to use a maximum hold down timer of 150 milliseconds, with the shortest achievable hold down timer is preferred. BFD should be enabled in asynchronous mode when possible.

## 2.4 Layer 2 considerations

After building the Leaf-Spine topology, the focus shifts to configuring the fabric of each participating rack. For the most part, Layer 2 configuration and tuning remains the same as any traditional Layer 2 network deployment. There are two key changes, however, as follows:

1. Across the data center, the configuration is implemented with a single set of VLANs for each rack.
2. The use of Virtual Route Redundancy Protocol (VRRP) at the leaf layer for each rack provides the default gateway.

The highlighted portions of Figure 6 each illustrate a single rack in the topology for the ScaleIO compute nodes. Each node uses a total of four physical links. Two physical links make up an LACP port channel that carries vMotion and application traffic. The remaining two physical links utilize an LACP port channel to carry the SDS-SDC traffic.

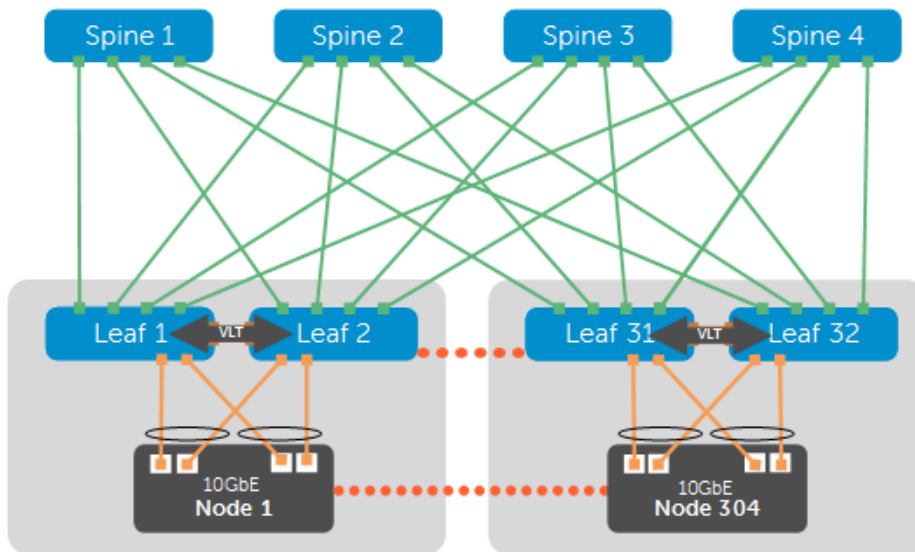


Figure 6 Server to network VLT implementation

**Note:** The MDM nodes utilize only a single LACP port channel for the vMotion traffic. The MDM traffic is carried by two individual links configured as Active-Standby uplinks.

### 2.4.1 VLT

A pair of leaf switches at the top of each rack provides redundancy. These switches' configurations include the Dell EMC Networking Virtual Link Trunking (VLT) feature. These switches are configured with multi-chassis LAG architecture, which utilizes the Dell EMC Networking OS9 VLT feature. Two 40GbE links (fo1/53 and fo1/54) are used from each leaf switch to provide the VLT interconnect (VLTi).

### 2.4.2 VRRP

Virtual router redundancy protocol (VRRP) is a protocol that provides for automatic assignment of available IP routes to participating hosts. This increases availability and reliability for each host by creating a virtual router, an abstraction of the two physical leaf switches. In the event that one of the leaf switches fails, the remaining leaf acts as the gateway until the failed unit recovers.

### 2.4.3 Uplink Failure Detection

Uplink Failure Detection (UFD) detects the loss of upstream connectivity. If a leaf switch loses connectivity to the spine layer, its attached hosts continue to send traffic without a direct path to the destination. The VLTi link handles traffic during such a network outage.

**Note:** VLTi links handling traffic is not considered a best practice.

To pre-empt this failure scenario, create an uplink-state group on each leaf switch, which creates an association between the spine uplinks and the downlink interfaces. UFD tracks the state of the uplink interfaces. In the event of a failure, UFD automatically shuts down all downstream interfaces. The hosts

attached to the leaf use a standby link or the remaining LACP port member to continue sending traffic across the fabric.



## 3 Configuration and Deployment

This section describes how to configure the physical and virtual networking environment for this hyper-converged ScaleIO 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 ScaleIO solution example

There are several options available when deploying ScaleIO. This section provides a summary of the options used in this deployment example. Options not listed here are not detailed in the setup steps beginning in section 3.

#### 3.1.1 ScaleIO MDM cluster

The ScaleIO MDM cluster in this deployment example uses the following:

**Cluster mode:** 3-node (Master, Slave and Tiebreaker)

**Hardware:** Dell PowerEdge R630 (3 qty.)

**Cluster name:** Rack 1 Management (MDM cluster name)

**Software:** ESXi 6.0

**Networking IP configuration:** MDM IPs (Control Network). IP addresses used for MDM control communications with SDSs and SDCs and to convey data migration decisions. But no user data passes through the MDM. Must be on the same network as the data network. Must be externally accessible if no MDM Management IP addresses are used.

#### 3.1.2 ScaleIO SDS-SDC nodes

The ScaleIO 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 names:** Rack 170 ScaleIO, Rack 171 ScaleIO (2 nodes per cluster)

**Software:** ESXi 6.0

**Networking IP configuration:** SDS All IPs (Rebuild and Data Path Network). IP addresses used for both SDS-SDS and SDS-SDC communications. These IP addresses are also used to communicate with the MDM.

### 3.1.3 ScaleIO deployment options

The following options were utilized during the installation of ScaleIO:

**Protection domain:** A single protection domain, PD1, used for simplicity. Protection domains are not within the scope of this document and are not contingent on the network design.

**Storage Pools:** Two storage pools defined, HDD and SSD. Best practice is to separate storage device types into capacity and performance pools, HDD and SSD respectively.

**Fault sets:** No fault sets are configured. Fault sets are not within the scope of this document and is 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 7 below shows an example for ASN numbering. The deployment steps use the ASN numbers in Table 2.

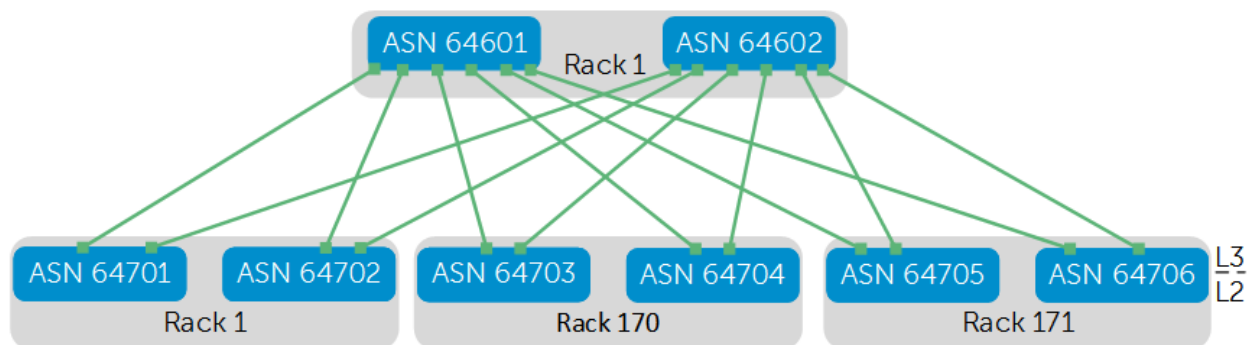


Figure 7 BGP ASN assignments

### 3.2.2 BGP fast 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 fast fall-over. Dell EMC recommends enabling fast fall-over for EBGP settings.

### 3.2.3 Loopback addresses

Figure 8 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.

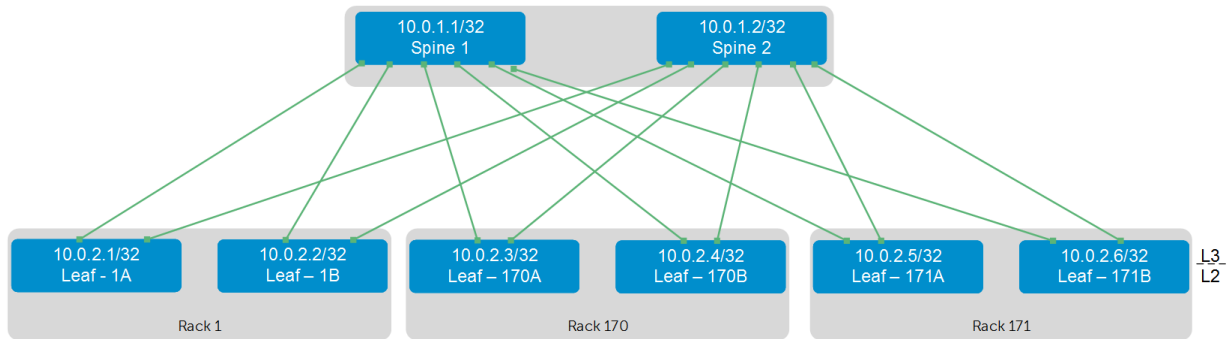


Figure 8 Loopback IP addressing

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 requires this information along with the point-to-point information in Table 3 the next section.

Table 2 Loopback interfaces and ASN associations

| Switch  | Switch Name | Loopback    | BGP ASN |
|---------|-------------|-------------|---------|
| Spine 1 | Spine 1     | 10.0.1.1/32 | 64601   |
| Spine 2 | Spine 2     | 10.0.1.2/32 | 64602   |
| Leaf 1  | Leaf 1-A    | 10.0.2.1/32 | 64701   |
| Leaf 2  | Leaf 1-B    | 10.0.2.2/32 | 64702   |
| Leaf 3  | Leaf 170-A  | 10.0.2.3/32 | 64703   |

|        |            |             |       |
|--------|------------|-------------|-------|
| Leaf 4 | Leaf 170-B | 10.0.2.4/32 | 64704 |
| Leaf 5 | Leaf 171-A | 10.0.2.5/32 | 64705 |
| Leaf 6 | Leaf 171-B | 10.0.2.6/32 | 64706 |

### 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 | Rack | Source switch | Source interface | Source IP | Network         | Destination switch | Destination interface | Destination IP |
|------|------|---------------|------------------|-----------|-----------------|--------------------|-----------------------|----------------|
| A    | 1    | Leaf 1A       | fo1/49           | .1        | 192.168.1.0/31  | Spine 1            | fo1/1/1               | .0             |
| B    | 1    | Leaf 1A       | fo1/51           | .1        | 192.168.2.0/31  | Spine 2            | fo1/1/1               | .0             |
| C    | 1    | Leaf 1B       | fo1/49           | .3        | 192.168.1.2/31  | Spine 1            | fo1/2/1               | .2             |
| D    | 1    | Leaf 1B       | fo1/51           | .3        | 192.168.2.2/31  | Spine 2            | fo1/2/1               | .2             |
| E    | 170  | Leaf 170A     | fo1/49           | .5        | 192.168.1.4/31  | Spine 1            | fo1/3/1               | .4             |
| F    | 170  | Leaf 170A     | fo1/51           | .5        | 192.168.2.4/31  | Spine 2            | fo1/3/1               | .4             |
| G    | 170  | Leaf 170B     | fo1/49           | .7        | 192.168.1.6/31  | Spine 1            | fo1/4/1               | .6             |
| H    | 170  | Leaf 170B     | fo1/51           | .7        | 192.168.2.6/31  | Spine 2            | fo1/4/1               | .6             |
| I    | 171  | Leaf 171A     | fo1/49           | .9        | 192.168.1.8/31  | Spine 1            | fo1/5/1               | .8             |
| J    | 171  | Leaf 171A     | fo1/51           | .9        | 192.168.2.8/31  | Spine 2            | fo1/5/1               | .8             |
| K    | 171  | Leaf 171B     | fo1/49           | .11       | 192.168.1.10/31 | Spine 1            | fo1/6/1               | .10            |
| L    | 171  | Leaf 171B     | fo1/51           | .11       | 192.168.2.10/31 | Spine 2            | fo1/6/1               | .10            |

Figure 9 shows the links from Table 3.

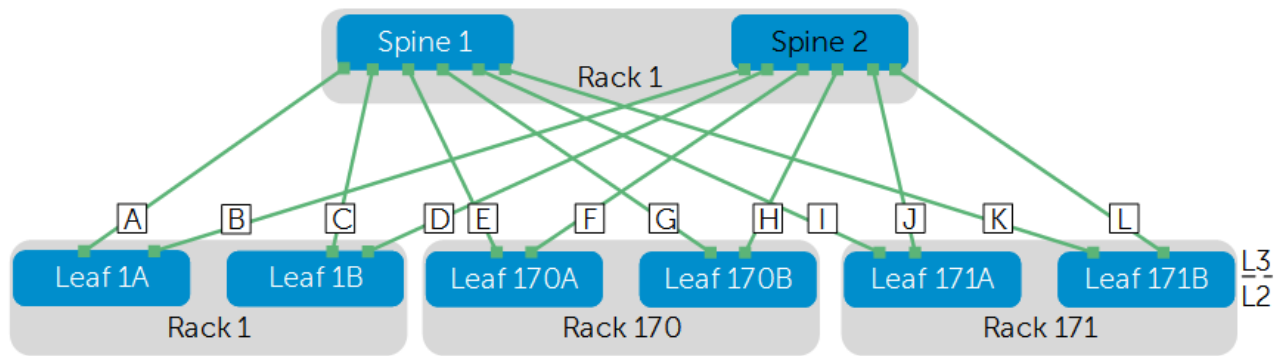


Figure 9 Point-to-point interface IP addressing

**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       | vMotion      | 10.15.1.0/24   | 15   | 10.15.1.254   |
| 1       | MDM          | 10.30.1.0/24   | 30   | 10.30.1.254   |
| 170     | vMotion      | 10.15.170.0/24 | 15   | 10.15.170.254 |
| 170     | SDS-SDC      | 10.30.170.0/24 | 30   | 10.30.170.254 |
| 171     | vMotion      | 10.15.171.0/24 | 15   | 10.15.171.254 |
| 171     | SDS-SDC      | 10.30.171.0/24 | 30   | 10.30.171.254 |

### 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.

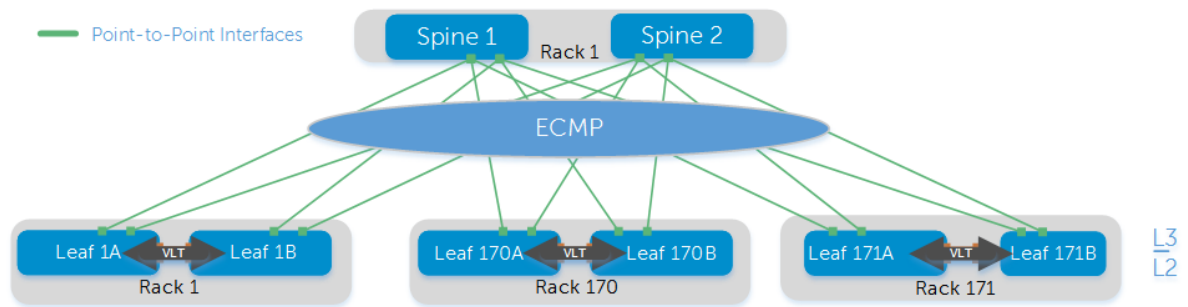


Figure 10 ECMP

### 3.2.7 VRRP

A VRRP instance is created for each VLAN/network in Table 4. As illustrated in Figure 11 below, Node 1 participates in the vMotion VLAN 15 broadcast domain in Rack 170. The host's configuration sends traffic for 10.15.170.0/24 to the Virtual IP (VIP) 10.15.170.254 provided by the VRRP instance running between leaf switches 170-A and 170-B

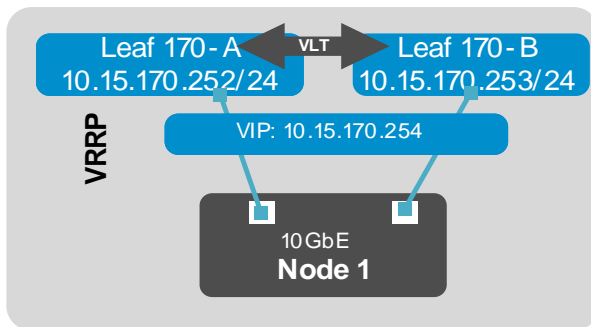


Figure 11 VRRP example

## 3.3 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.

**Prerequisite:** vCenter has been installed and all hosts to be included in the ScaleIO deployment have been added.

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 the vCenter Server object and select **New Datacenter**.
3. Provide a name (Datacenter) and click **OK** (Figure 12).

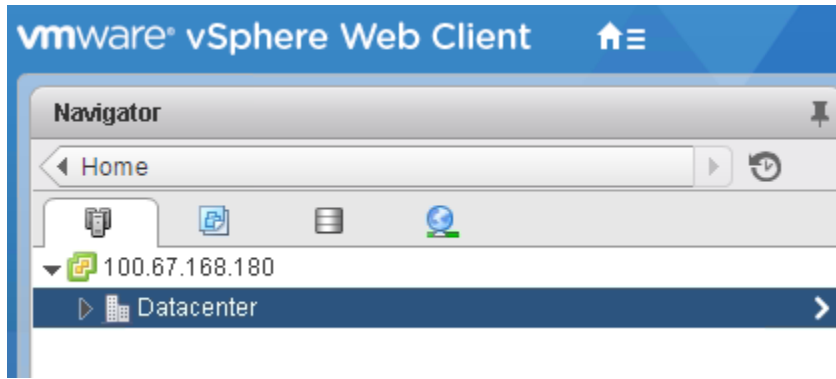


Figure 12 Datacenter created

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 **host name** 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 three R630 servers and four R730 servers for a total of seven hosts running ESXi.

## 3.4 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 guide shows how to create three clusters, with one cluster for each of the following racks:

- Rack 1 Management
- Rack 170 ScaleIO
- Rack 171 ScaleIO

All ESXi hosts are added to one of the above clusters. The rack numbers in this example are 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 first cluster is named **Rack 1 Management**.
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 6.0 Documentation](#).

Repeat for the remaining two clusters:

- Rack 170 ScaleIO
- Rack 171 ScaleIO

In the Navigator pane, drag and drop ESXi hosts into the appropriate clusters. The three ESXi hosts on R630 servers in Rack 1 are placed in the **Rack 1 Management** cluster, the four ESXi hosts on R730 servers in Rack 170 are placed in the **Rack 170 ScaleIO** cluster, and the two ESXi hosts on R730 servers in Rack 171 are placed in the **Rack 171 ScaleIO** cluster.

When complete, each cluster (📁) should contain its assigned hosts (📱) as shown in Figure 13

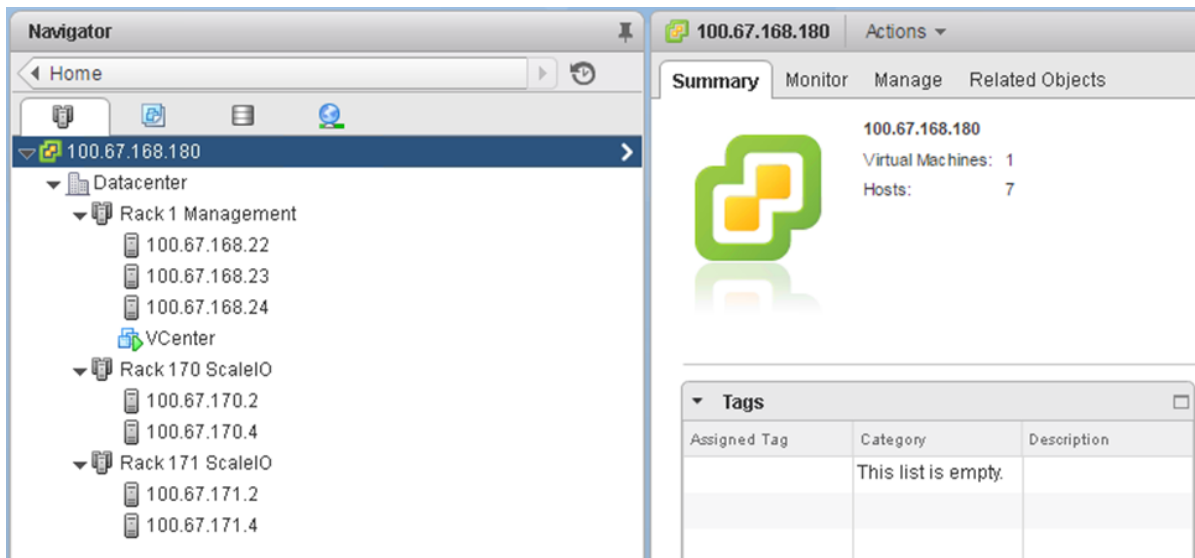


Figure 13 Clusters and hosts after initial configuration

### 3.5 Information on vSphere standard switches

A vSphere standard switch (VSS), also referred to as a standard switch, is a virtual switch that handles network traffic at the host level in a vSphere deployment. Standard switches provide network connectivity to hosts and virtual machines.



A standard switch named vSwitch0 is automatically created on each ESXi host during installation to provide connectivity to the management network.

To view a standard switch configuration, complete the following steps:

1. Go to the web client **Home** page, select **Hosts and Clusters** and select a host in the **Navigator** pane.
2. In the center pane, select **Manage > Networking > Virtual switches**.
3. Standard switch **vSwitch0** appears in the list. Click on it to view details as shown in Figure 14:

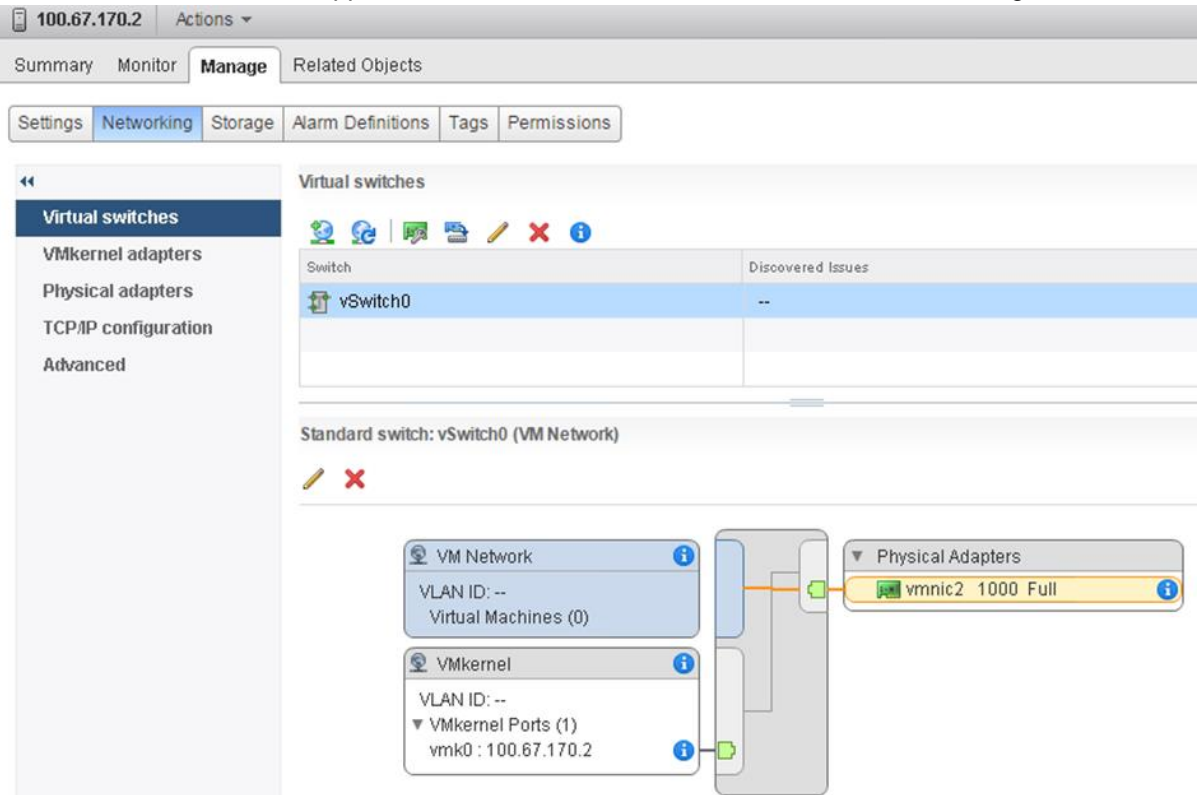


Figure 14 vSphere standard switch

**Note:** For the example in this guide, standard switches only require the default configuration. Standard switches are only used in this deployment for connectivity to the management network. Distributed switches, covered in the next section, are used for connectivity to the production network.

### 3.6 Deploy vSphere distributed switches

Two vSphere distributed switches (VDS) are created in vCenter and deployed to each host. The first VDS, called Compute VDS, has a port group called vMotion that supports all vMotion traffic. The second VDS, called ScaleIO VDS, has two port groups. The MDM port group supports MDM traffic and is connected to the ScaleIO Management hosts. The SDS-SDC port group supports storage traffic and is connected to the ScaleIO nodes.

For this guide, one VDS is created for Compute and another for ScaleIO. Each VDS is shared by all hosts within each traffic classification. The two distributed switches used in this deployment are named:

- ScaleIO VDS
- Compute VDS

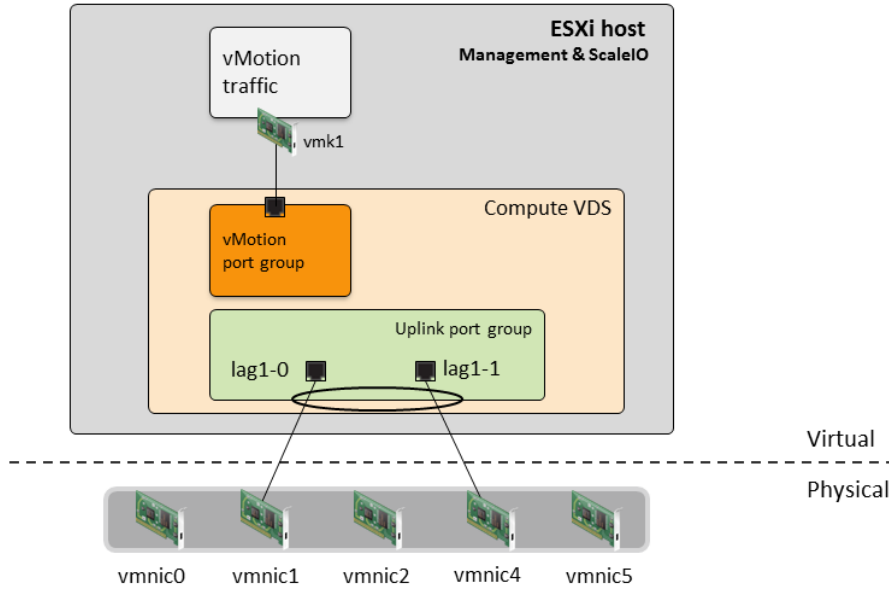


Figure 15 Distributed switch – Compute VDS

Each VDS is assigned two vmnics from each host. On the Compute VDS, each host from the Management cluster and each host from the ScaleIO node clusters use vmnic1 and vmnic4. These vmnics combine into one LACP-enabled port channel, Lag 1. This lag is mapped to the vMotion port group.

On the ScaleIO VDS, the vmnics on the Management cluster are configured differently than the vmnics on the ScaleIO node clusters. The hosts on the Management cluster use vmnic0 and vmnic5. These vmnics are configured in an Active-Standby and are mapped to the MDM port group. The hosts on the ScaleIO node clusters use vmnic0 and vmnic5. These vmnics combine into one LACP-enabled port channel, Lag 1. This lag is mapped to the SDS-SDC port group.

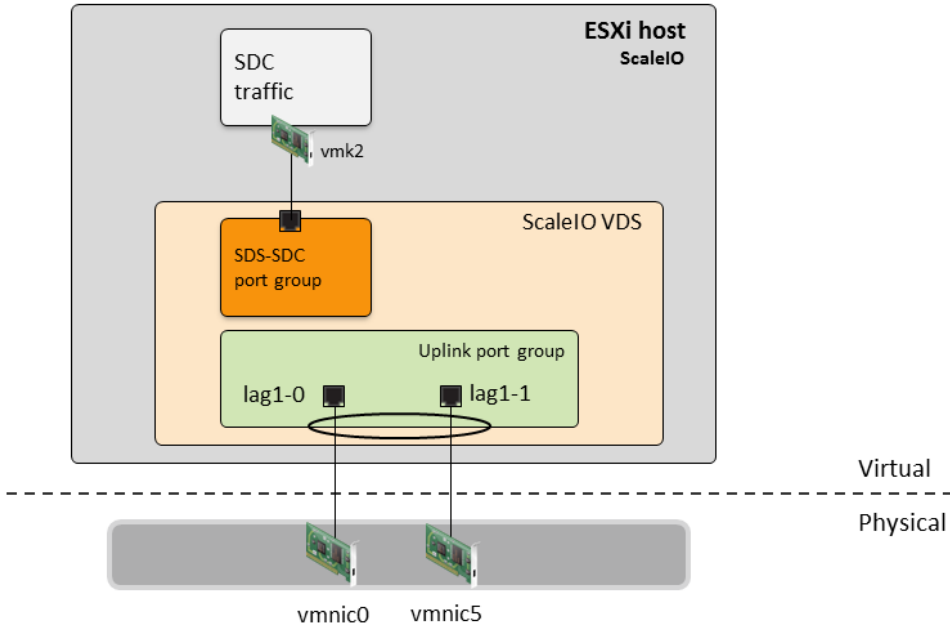


Figure 16 Distributed switch – ScaleIO VDS, SDS-SDC port group

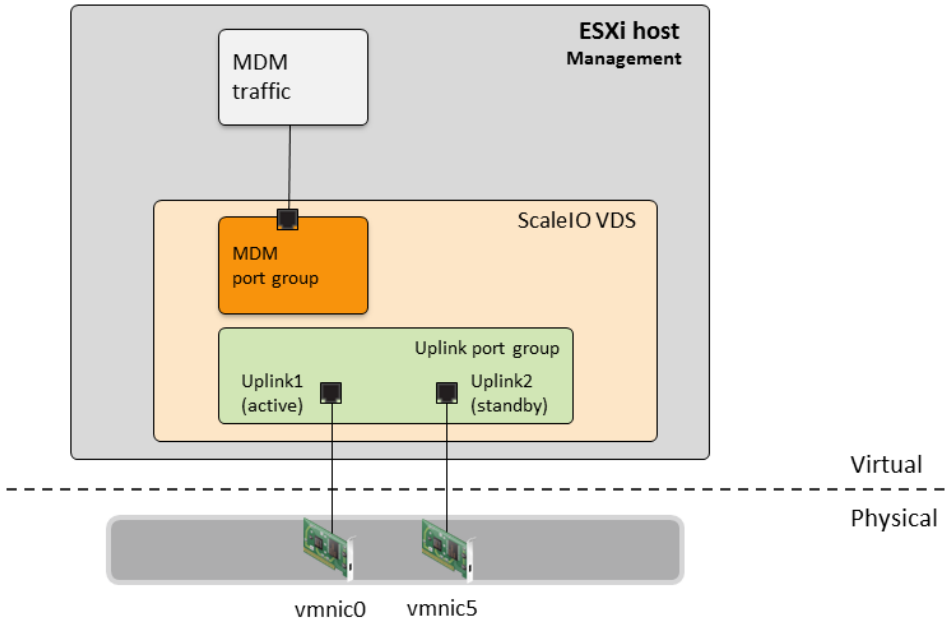


Figure 17 Distributed switch – ScaleIO VDS, MDM port group

### 3.7 Create each VDS

To create the first VDS named **ScaleIO 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 first VDS, **ScaleIO VDS**. Click **Next**.
4. On the Select version page, select **Distributed switch: 6.0.0** > **Next**.
5. On the **Edit settings** page:
  - a. Set the **Number of uplinks** set to **2**.
  - 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 uplink port group shown beneath it.

Repeat steps 1-7 above, substituting **Compute VDS** for the switch name in step 3 to create the remaining distributed switch.

When complete, the Navigator pane should look similar to Figure 18.

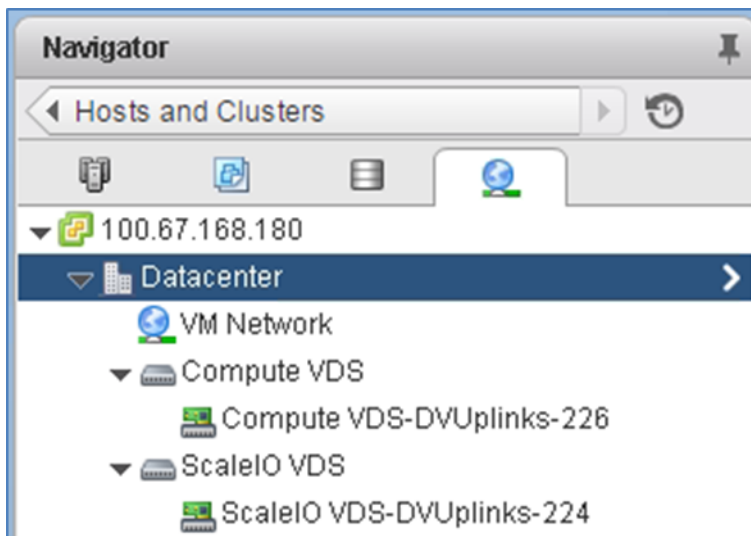


Figure 18 VDS created for compute and ScaleIO

## 3.8 Add distributed port groups

In this section, separate distributed port groups for vMotion, MDM and SDS-SDC traffic are added to the appropriate VDS.

To create the port group for vMotion traffic on the **Compute VDS**, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right click Compute 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 **15** for the vMotion port group. Leave other values at their defaults as shown in Figure 19.

5. Click **Next > Finish**.

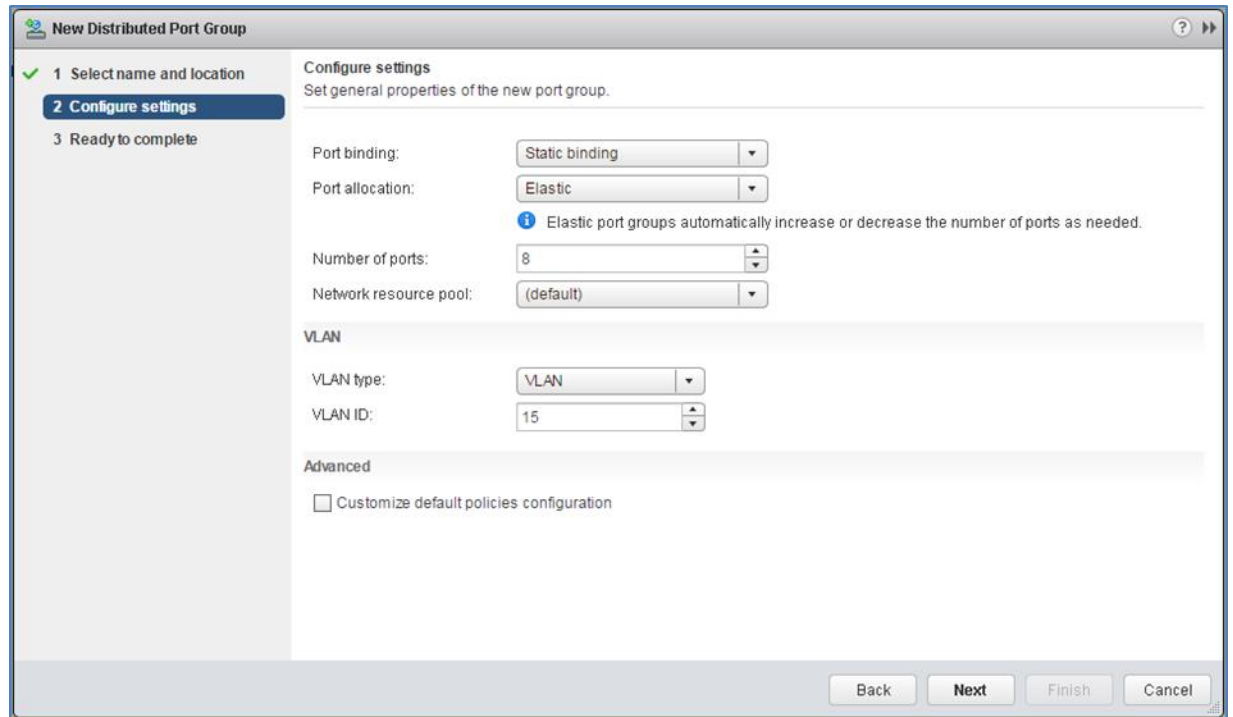


Figure 19 Distributed port group settings page – vMotion port group

To create the port group for MDM traffic on the **ScaleIO VDS**, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. Right click on ScaleIO 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, **MDM**. Click **Next**.
4. On the **Configure settings** page, next to **VLAN type**, select **VLAN**. Set the **VLAN ID** to **30** for the MDM port group. Leave other values at their defaults.
5. Click **Next > Finish**.

Repeat steps 6-10 above to create the distributed port group for SDS-SDC traffic, except replace "MDM" with "SDS-SDC" in the **port group name**.

When complete, the Navigator pane appears similar to Figure 20.

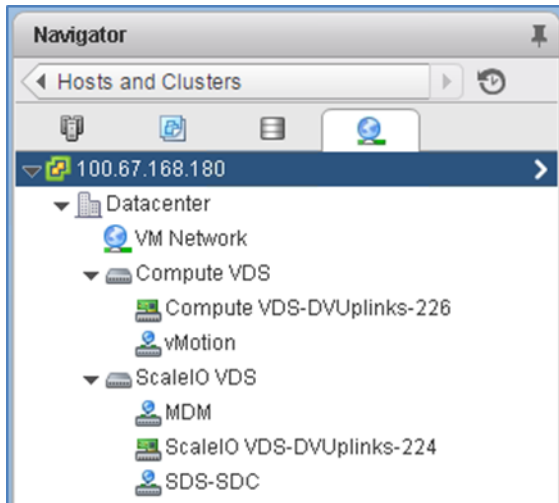


Figure 20 Distributed switches with vMotion, MDM and SDS-SDC port groups created

## 3.9 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 **Compute VDS**, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. In the Navigator pane, select **Compute VDS**.
3. In the center pane, select **Manage > Settings > LACP**.
4. Click the **+** icon. The New Link Aggregation Group dialog box opens.
5. Set the number of ports equal to the number of physical uplinks on each ESXi host. In this deployment, R630 and R730 hosts use two ports in a LAG to connect to the upstream switches so this number is set to 2.
6. Set the **Mode** to **Active**. The remaining fields can be set to their default values as shown in Figure 21
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

OK Cancel

Figure 21 LAG configuration

This creates **lag1** on the VDS. Click the refresh icon (🔄) at the top of the screen if the lag does not appear in the table as shown in Figure 22.

Compute VDS Actions

Getting Started Summary Monitor **Manage** Related Objects

Settings Alarm Definitions Tags Permissions Network Protocol Profiles Ports Resource Allocation

← Properties Topology **LACP** Private VLAN NetFlow Port mirroring Health check

**LACP**

The enhanced LACP support on a vSphere distributed switch lets you connect ESXi hosts

**Migrating network traffic to LAGs**

+

| LAG Name | Ports | Mode   | VLAN                             |
|----------|-------|--------|----------------------------------|
| lag1     | 2     | Active | Inherited from uplink port group |

Figure 22 LAG1 created on Compute VDS

Repeat steps 1-7 above to enable LACP on **ScaleIO VDS**.

## 3.10 Associate hosts and assign uplinks





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

MDM traffic requires failover times that prevent the use of LAG uplinks. For MDM traffic, the uplinks use individual vmnics configured to use active/standby.

All other traffic uses the LAG uplinks to take advantage of the improved bandwidth usage the VLT feature provides. This section details configuration of both uplink types at the same time.

**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 **Manage > Networking > Physical adapters**. This example uses vmnics 1 and 3. Vmnic numbering varies depending on adapters installed in the host.

To add hosts to ScaleIO 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 **Rack 1 Management, Rack 170 ScaleIO and Rack 171 ScaleIO** clusters. 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 within the Rack 1 Management cluster, select the appropriate vmnic (vmnic0 in this example) and click  **Assign uplink**.
    - ii. Select **Uplink 1 > OK**.
    - iii. On the same host, select the next appropriate vmnic (vmnic5 in this example) and click  **Assign uplink**.
    - iv. Select **Uplink 2 > OK**.
    - v. Repeat steps i – iv for the remaining two hosts in the Rack 1 Management cluster.
    - vi. On the first host within the Rack 170 ScaleIO cluster, select the appropriate vmnic (vmnic0 in this example) and click  **Assign uplink**.
    - vii. Select **lag1-0 > OK**.
    - viii. On the same host, select the next appropriate vmnic (vmnic5 in this example) and click  **Assign uplink**.
    - ix. Select **lag1-1 > OK**.



- x. Repeat steps vi – ix for the remaining hosts in the Rack 170 ScaleIO and Rack 171 ScaleIO clusters. Click **Next** when done.
- e. On the Analyze impact page, **Overall impact status** should indicate ✔ **No impact**.
- f. Click **Next > Finish**.

When complete, the **Manage > Settings > Topology** page for ScaleIO VDS should look similar to Figure 23:

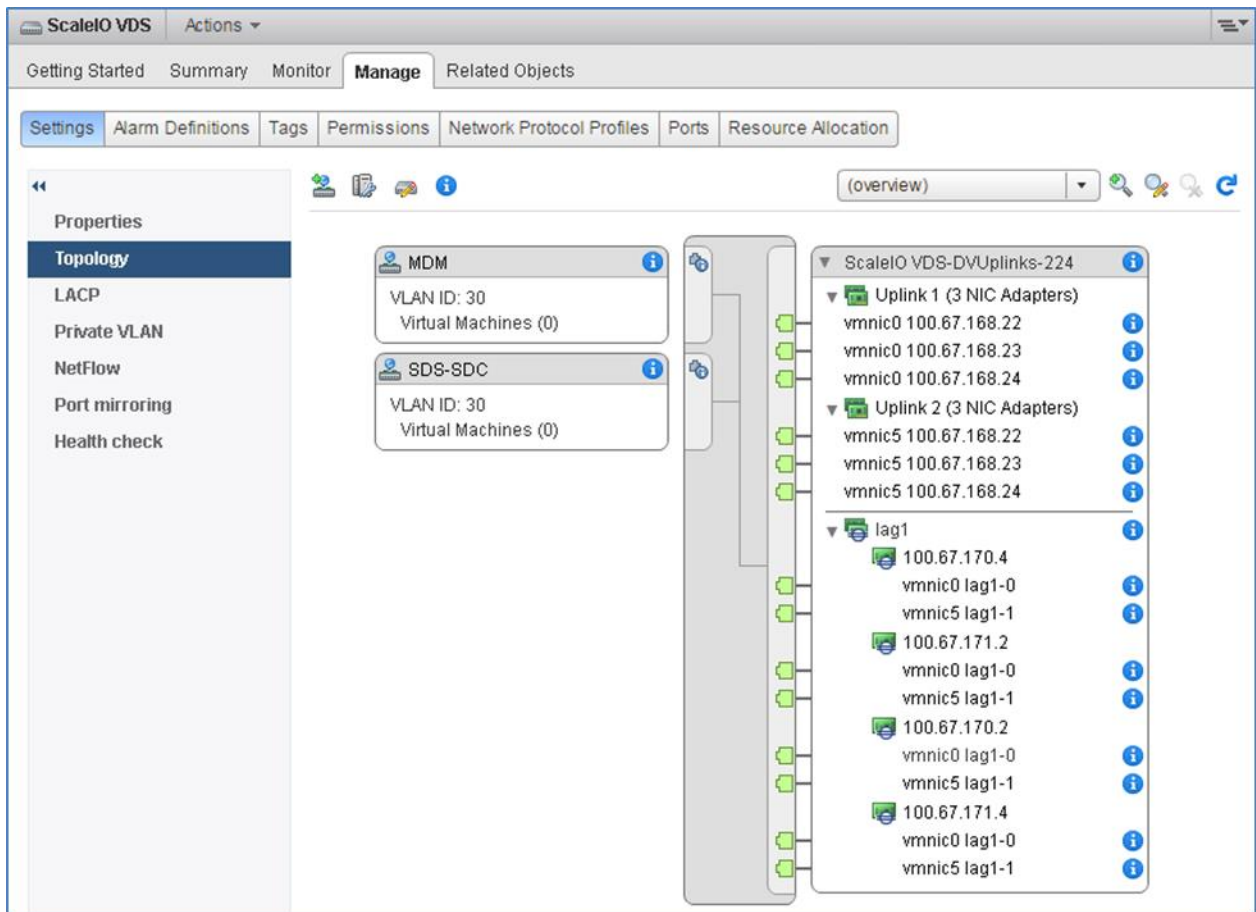





Figure 23 Uplinks configured on ScaleIO VDS

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

1. On the web client **Home** screen, select **Networking**.
2. Right click Compute 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 **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 **Rack 1 Management, Rack 170 ScaleIO and Rack 171 ScaleIO** clusters. 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. Select the appropriate vmnic (vmnic1 in this example) on the first host and click  **Assign uplink**.
  - ii. Select **lag1-0 > OK**.
  - iii. On the same host, select the next appropriate vmnic (vmnic4 in this example) and click  **Assign uplink**.
  - iv. Select **lag1-1 > OK**.
  - v. Repeat steps i – iv for all the remaining hosts in the Rack 1 Management, Rack 170 ScaleIO and Rack 171 ScaleIO clusters. Click **Next** when done.
- e. On the Analyze impact page, **Overall impact status** should indicate  **No impact**.
- f. Click **Next > Finish**.

The completed **Manage > Settings > Topology** page for Compute VDS should look similar to Figure 24:

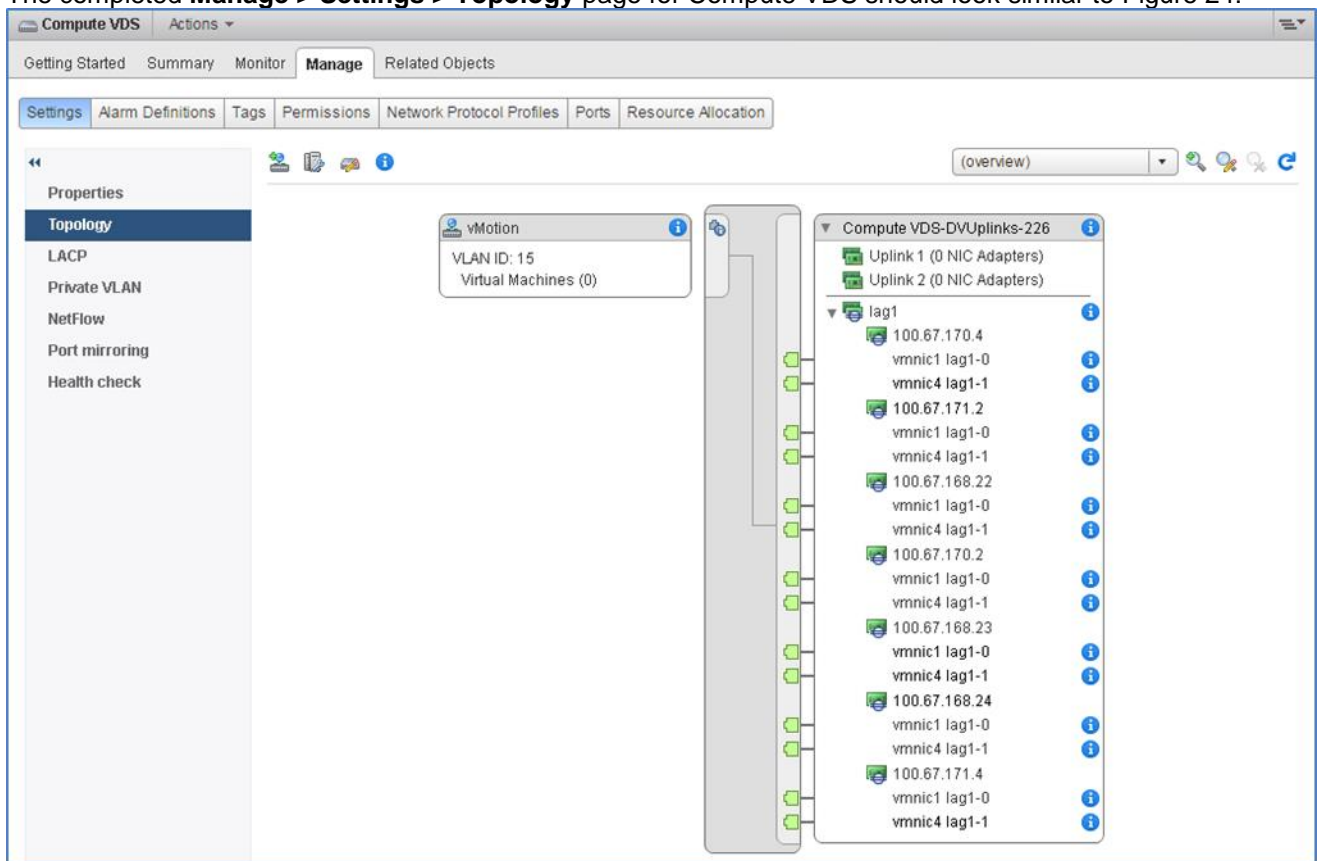


Figure 24 Uplinks configured on Compute VDS

This configuration brings up the LAGs on the upstream switches. Confirm the configuration by running the show vlt detail command on the upstream switches as shown in the examples below. The Local and Peer Status columns now indicate all LAGs are UP.

```
rack170-leaf-A#show vlt detail
```

| Local LAG Id | Peer LAG Id | Local Status | Peer Status | Active VLANs |
|--------------|-------------|--------------|-------------|--------------|
| 1            | 1           | UP           | UP          | 1, 15        |
| 2            | 2           | UP           | UP          | 1, 30        |
| 3            | 3           | UP           | UP          | 1, 15        |
| 4            | 4           | UP           | UP          | 1, 30        |

### 3.11 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 ScaleIO 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 top box to select only the **SDS-SDC** port group. Click **OK > Next**.
5. On the **Teaming and failover page**, click **lag1** and move it up to the **Active uplinks** section by clicking the up arrow. Move **Uplinks 1-2** 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 25 when complete.

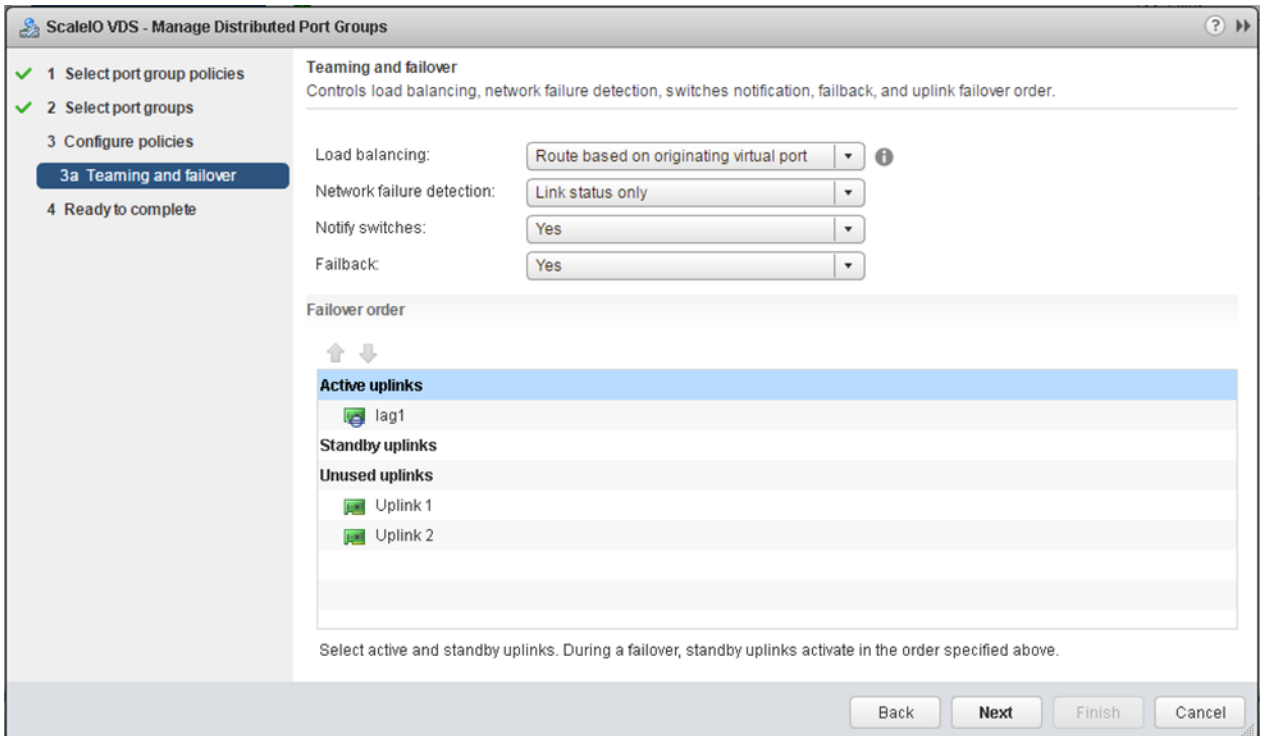


Figure 25 Teaming and failover settings for LAGs

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

Repeat steps 1-6 above for the remaining Compute VDS, except select the “vMotion” port group.

## 3.12 Configure teaming and failover on MDM uplinks

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

1. On the web client **Home** screen, select **Networking**.
2. Right click ScaleIO 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 top box to select only the **MDM** port group. Click **OK > Next**.
5. On the **Teaming and failover** page, click **Uplink 2** and move it down to the **Standby uplinks** section by clicking the down arrow.
6. Change the **Failback** option to **No**. Leave other settings at their defaults. The **Teaming and failover** page should look similar to Figure 26 when complete.

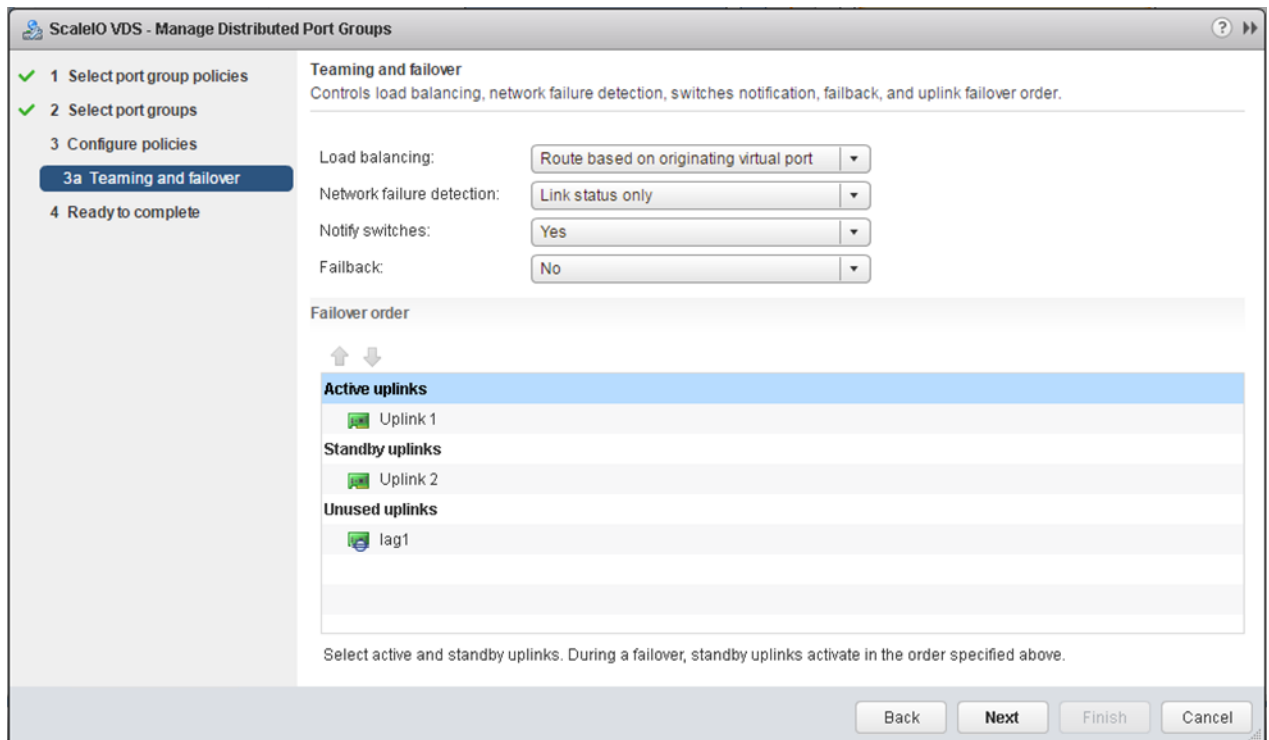


Figure 26 Teaming and failover for MDM uplinks

7. Click **Next** followed by **Finish** to apply the settings.

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

To allow virtual machines to be moved between compute nodes in each rack a vMotion-enabled VMkernel is created. This VMkernel is configured to use the vMotion port group on the default TCP/IP stack. An additional VMkernel interface is created to handle SDC to SDS/MDM traffic. This VMkernel is associated with the SDS-SDC port group and also uses the default TCP/IP stack. No VMware specific services are enabled on this second VMkernel

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

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

This deployment uses the following addressing scheme for the MDM, Storage and vMotion networks, where "xx" represents the rack number and "yy" represents the node number:


Table 5 VLAN and network examples

| PortGroup Name | VLAN ID | IP Address  | Subnet Mask   |
|----------------|---------|-------------|---------------|
| SDS-SDC_VMK    | 30      | 10.30.xx.yy | 255.255.255.0 |
| vMotion_VMK    | 15      | 10.15.xx.yy | 255.255.255.0 |

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

1. On the web client **Home** screen, select **Networking**.
2. Right click Compute 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 vMotion adapter.

    - 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 traffic** box. Click **Next**.
    - v. On the **IPv4 settings page**, if DHCP is not used, select **Use static IPv4 settings**. Set the IP address, for example 10.15.1.22, and subnet mask for the host on the vMotion network, 255.255.255.0. Click **Next > Finish**.
  - d. Repeat steps i-v for the remaining hosts, and then click **Next**.
  - e. On the **Analyze impact** page, **Overall impact status** should indicate  **No impact**.
  - f. Click **Next > Finish**.

When complete, the VMkernel adapters' page for each ESXi host in the vSphere datacenter should look similar to Figure 27. View this page by going to **Hosts and Clusters**, selecting a host in the **Navigator** pane, then selecting **Manage > Networking > VMkernel adapters** in the center pane.

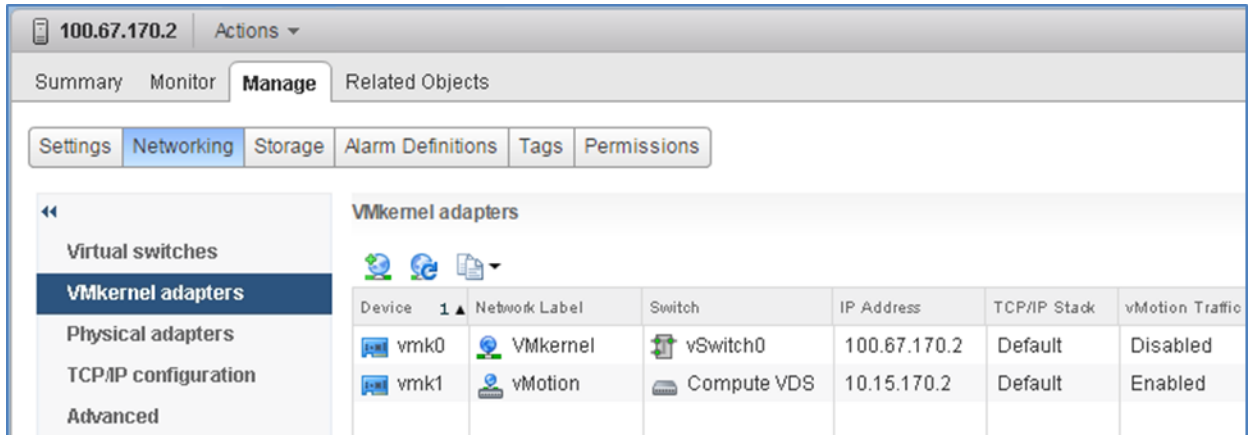


Figure 27 Host VMkernel adapters page

Adapter vmk0 was installed by default for host management. Adapter vmk1 was created in this section.

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.

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

1. On the web client **Home** screen, select **Networking**.
2. Right click 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 **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**.
  - d. The Manage VMkernel network adapters page opens.

#### SDS-SDC adapter

- i. To add the SDS-SDC 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 **SDS-SDC**, click **OK**. Click **Next**.

- iv. On the **Port properties** page, leave **IPv4** selected. Leave all **Available services** unchecked. Click **Next**.
- v. On the **IPv4 settings page**, if DHCP is not used, select **Use static IPv4 settings**. Set the IP address, for example 10.30.1.22, and subnet mask for the host on the SDS-SDC network, 255.255.255.0. Click **Next > Finish**.
- e. Repeat steps i-v for the remaining hosts, and then click **Next**.
- f. On the **Analyze impact** page, **Overall impact status** should indicate ✔ **No impact**.
- g. Click **Next > Finish**.

When complete, the VMkernel adapters' page for each ESXi host in the vSphere datacenter should look similar to Figure 28. View this page by going to **Hosts and Clusters**, selecting a host in the **Navigator** pane, then selecting **Manage > Networking > VMkernel adapters** in the center pane.

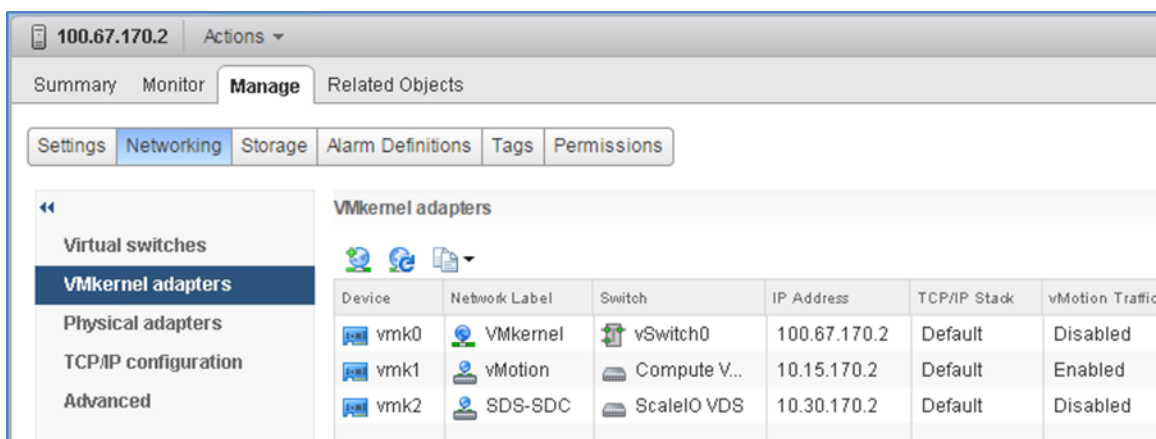


Figure 28 Host VMkernel adapters page

Adapter vmk0 was installed by default for host management. Adapter vmk2 was created in this section.

Verify the information on other hosts as needed.

### 3.14 Add static routes for default gateways

The default gateway for each network configured on the VMkernals cannot be modified during VMkernel creation. To set the default gateway requires configuring a static route at the command line of each ESXi host.

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 VLAN 15 and 30:

```
[root@MGMT2:~] esxcli network ip route ipv4 add -n=10.15.0.0/16 -g=10.15.1.254
```

```
[root@MGMT2:~] esxcli network ip route ipv4 add -n=10.30.0.0/16 -g=10.30.1.254
```

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

```
[root@MGMT2:~] esxcli network ip route ipv4 list
Network          Netmask          Gateway          Interface        Source
-----          -
default          0.0.0.0          100.67.168.254  vmk0             MANUAL
10.15.0.0         255.255.0.0      10.15.1.254     vmk1             MANUAL
10.15.1.0         255.255.255.0    0.0.0.0         vmk1             MANUAL
10.30.0.0         255.255.0.0      10.30.1.254     vmk2             MANUAL
10.30.1.0         255.255.255.0    0.0.0.0         vmk2             MANUAL
100.67.168.0     255.255.255.0    0.0.0.0         vmk0             MANUAL
```

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

The example commands above show the default gateway IP for a host located in Rack 1 Management cluster. The default gateway used in each host should be modified for the appropriate network. For example, systems in the Rack 170 ScaleIO cluster would use the gateways 10.15.**170**.254 and 10.30.**170**.254. This follows the IP structure outlined in Table 4.

## 3.15 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 **Compute VDS**, complete the following steps:

1. On the web client **Home** screen, select **Networking**.
2. In the Navigator pane, select **Compute VDS**.
3. In the center pane, select **Manage > Settings > Topology** and click the ► icon next to **VMkernel Ports** to expand. The screen should look similar to Figure 29:



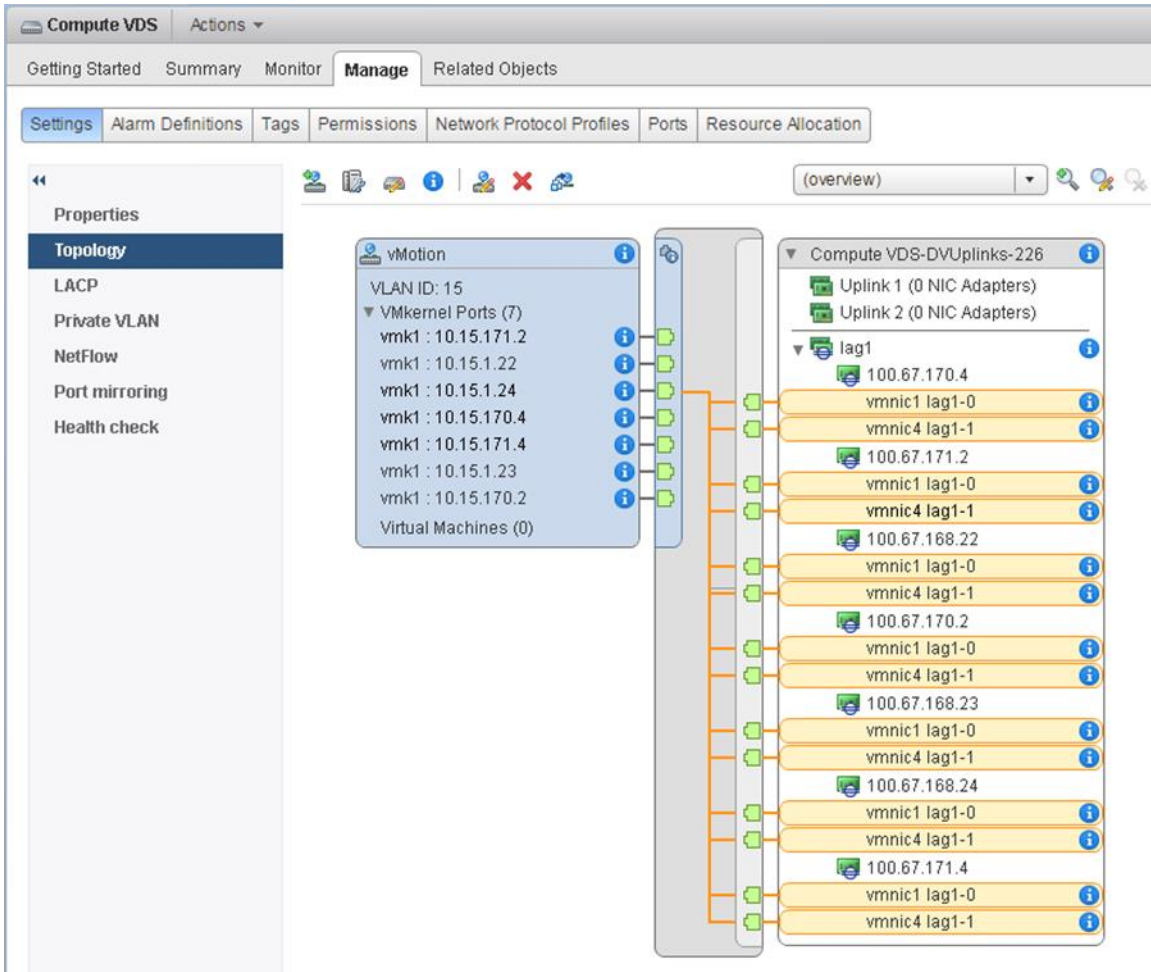


Figure 29 Compute VDS VMkernel ports, VLANs, and IP addresses

Repeat steps 1-3 above for the **SDS-SDC VDS**.

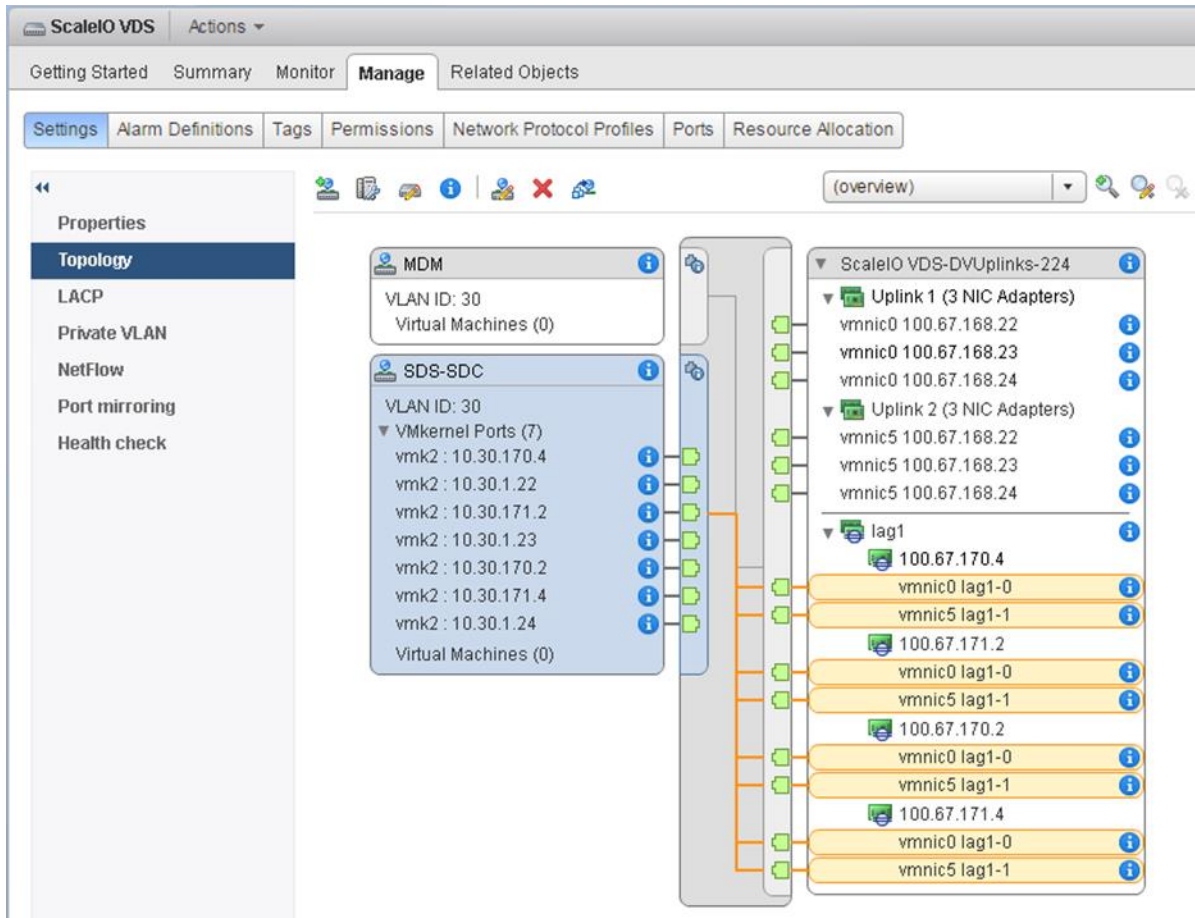


Figure 30 SDS-SDC VDS VMkernel ports, VLANs, and IP addresses

### 3.16 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 a **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**.

Repeat for remaining distributed switches.

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

```
rack170-leaf-A#show lldp neighbors
Loc PortID      Rem Host Name      Rem Port Id      Rem Chassis Id
-----
Te 1/1          R170U2-ESXI        00:50:56:55:81:79      vmnic4
Te 1/2          R170U2-ESXI        00:50:56:5e:c1:ac      vmnic0
Te 1/3          R170U4-ESXI        00:50:56:5a:ba:c3      vmnic4
Te 1/4          R170U4-ESXI        00:50:56:56:ba:be      vmnic0
Fo 1/49         Spine-1             fortyGigE 1/5/1        4c:76:25:e8:d6:40
Fo 1/51         Spine-2             fortyGigE 1/5/1        4c:76:25:e5:86:c0
Fo 1/53         rack170-leaf-B     fortyGigE 1/53         f4:8e:38:2e:76:f7
Fo 1/54         rack170-leaf-B     fortyGigE 1/54         f4:8e:38:2e:76:f7
Ma 1/1         -                   GigabitEthernet 1/34    74:e6:e2:f5:c8:80
```

## 4 Deploying ScaleIO

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

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

See the following documents for detailed installation instructions:

[ScaleIO 2.0.x Deployment on VMware-Quick Start Guide](#)

[EMC ScaleIO 2.0.x Deployment Guide](#)

**Note:** You may need to enter EMC Community Network (ECN) login credentials or create an ECN account to access the preceding documents

### 4.1 Registering the ScaleIO Plug-in and uploading the OVA template

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

To register the ScaleIO plugin follow the procedure in the **EMC ScaleIO 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 6.3.0 R1 Patch1
  - filename: VMware-PowerCLI-6.3.0-3737840.exe
- ScaleIO v2.0.1.1 Complete VMware Software
  - filename: ScaleIO\_2.0.1.1\_Complete\_VMware\_SW\_Download.zip
    - > ScaleIOVM\_2nics\_2.0.11000.174.ova
    - > EMC-ScaleIO-vSphere-plugin-installer-2.0-11000.174.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.

After installation of the EMC ScaleIO plugin, the Home window should look similar to Figure 31:

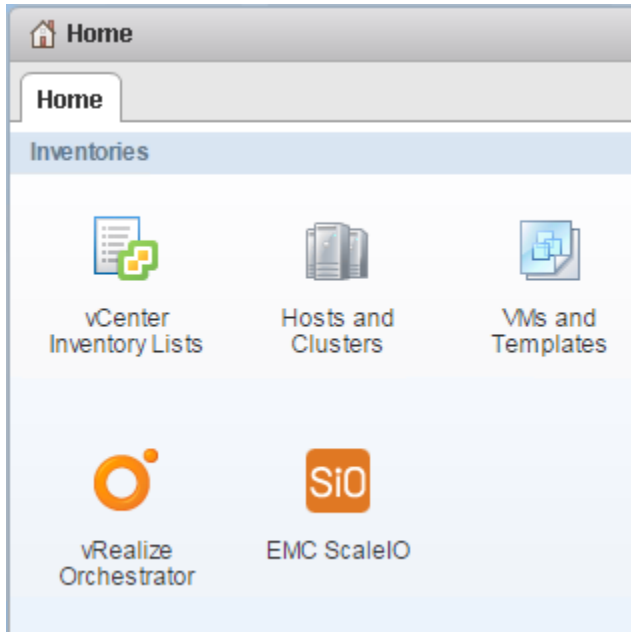


Figure 31 EMC ScaleIO plugin icon

The example in this guide uses seven hosts. For faster, parallel deployment, SVMs were created on seven data stores.

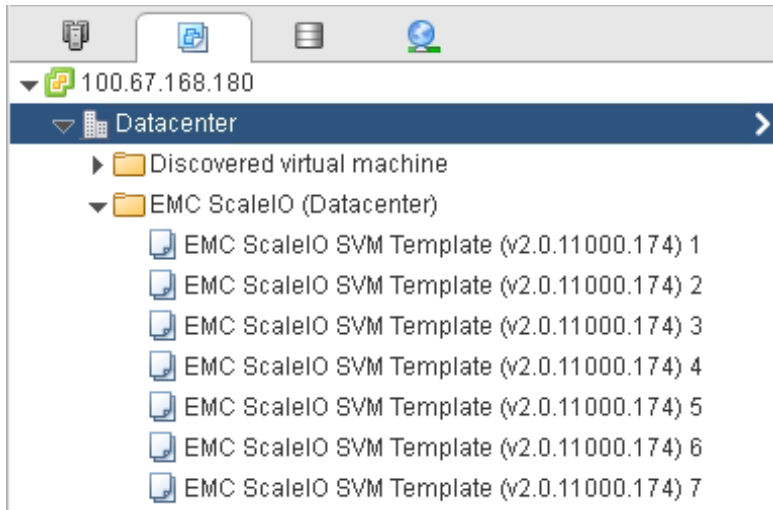


Figure 32 vSphere SVMs

## 4.2 Installing the SDC on a ESXi hosts

The SDC component must be applied to every ESX host within the ScaleIO system. This includes all hosts within the ScaleIO management and ScaleIO node clusters.

Complete the procedure summarized below within the EMC ScaleIO plugin application:

- Execute Basic tasks > Install SDC on ESX.
- Select all the ESX hosts, provide root passwords.
- Complete install and restart the ESX hosts.

## 4.3 ScaleIO 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 2.0 Deployment Guide](#).

Complete the procedures summarized after the following note from within the EMC ScaleIO plugin application.

Prior to 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.

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

- Execute 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 > 100.67.168.22
  - Manager MDM > 100.67.168.23
  - TieBreaker MDM > 100.67.168.24
  - Hosts for this example are chosen from the Rack 1 Management cluster.
  - Leave optional settings as default.
- Performance profile, Sizing, Syslog and DNS servers setting are left as their default.
- Enter a **Protection Domain** name (example: PD1).
- Enter **Storage Pool** names (example: HDD and SSD).
  - HDD, check Enable zero padding.
  - SSD, check Enable zero padding.
- **Create new Fault Sets** is left as its default; none are created in this example.
- Assign ESX host devices to ScaleIO SDS components.
  - Select all ScaleIO node hosts (not MDM hosts)
  - Select devices tab > Select all empty devices (ensure only SCSI devices checked)
  - Ensure all SSD devices are placed into the SSD storage pool. (By default, devices are placed into the first configured storage pool from the previous step.)
- Select ESXs to add as SDCs to ScaleIO system.
  - Select all ScaleIO node hosts (not MDM hosts), enter root passwords.
  - Disable SCSI LIN number comparison for hosts.
- Select OVA Template. Select an existing template for the ScaleIO virtual machines.
  - Select all the templates that were uploaded to the datastores, enter root passwords.
- Configure networks.
  - Management network label, select **IPv4** and **VM Network**.
  - Data network label, select **IPv4** and **MDM (ScaleIO VDS)**.
    - > See Section 3.6 for a discussion of the networks used in this deployment step.
- Configure SVM. Configure ScaleIO Virtual Machine (SVM) IP addresses and hosting Datastore.
  - Enter IP information for all SVMs.
  - See Table 6 for IP information for this deployment example.
  - This example uses **10.30.1.1** is used for the Cluster Virtual IP address, **Data (MDM(ScaleIO VDS))**.

Table 6 ScaleIO Wizard – Configure SVM

| ESX Name                          | Mgmt IP & Subnet Mask             | Default Gateway | Data IP & Subnet Mask            |
|-----------------------------------|-----------------------------------|-----------------|----------------------------------|
| 100.67.170.2<br>(ScaleIO Gateway) | 100.67.170.52<br>/ 255.255.255.0  | 100.67.170.254  | 10.30.170.52<br>/ 255.255.255.0  |
| 100.67.168.22<br>(Master MDM)     | 100.67.168.52<br>/ 255.255.255.0  | 100.67.168.254  | 10.30.1.52<br>/ 255.255.255.0    |
| 100.67.168.23<br>(Slave 1 MDM)    | 100.67.168.53<br>/ 255.255.255.0  | 100.67.168.254  | 10.30.1.53<br>/ 255.255.255.0    |
| 100.67.168.24<br>(TieBreaker 1)   | 100.67.168.54<br>/ 255.255.255.0  | 100.67.168.254  | 10.30.1.54<br>/ 255.255.255.0    |
| 100.67.170.2                      | 100.67.170.202<br>/ 255.255.255.0 | 100.67.170.254  | 10.30.170.202<br>/ 255.255.255.0 |
| 100.67.170.4                      | 100.67.170.204<br>/ 255.255.255.0 | 100.67.170.254  | 10.30.170.204<br>/ 255.255.255.0 |
| 100.67.171.2                      | 100.67.171.202<br>/ 255.255.255.0 | 100.67.171.254  | 10.30.171.202<br>/ 255.255.255.0 |
| 100.67.171.4                      | 100.67.171.204<br>/ 255.255.255.0 | 100.67.171.254  | 10.30.171.204<br>/ 255.255.255.0 |

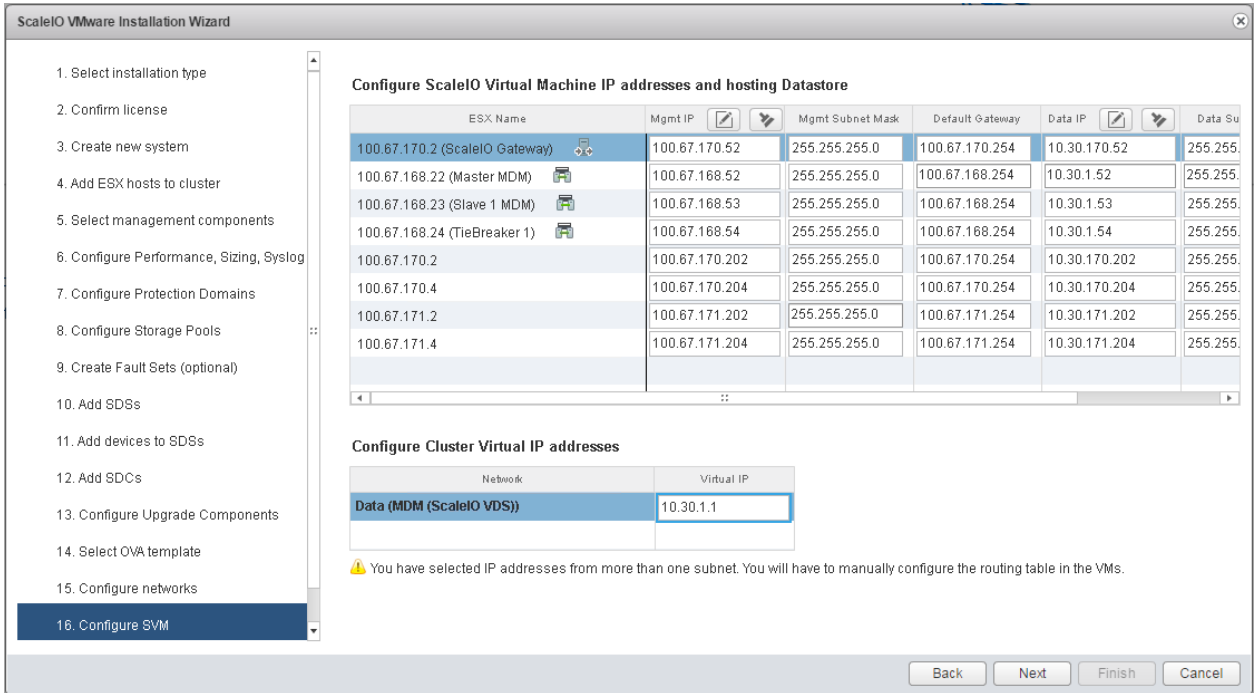


Figure 33 Configure SVM deployment step



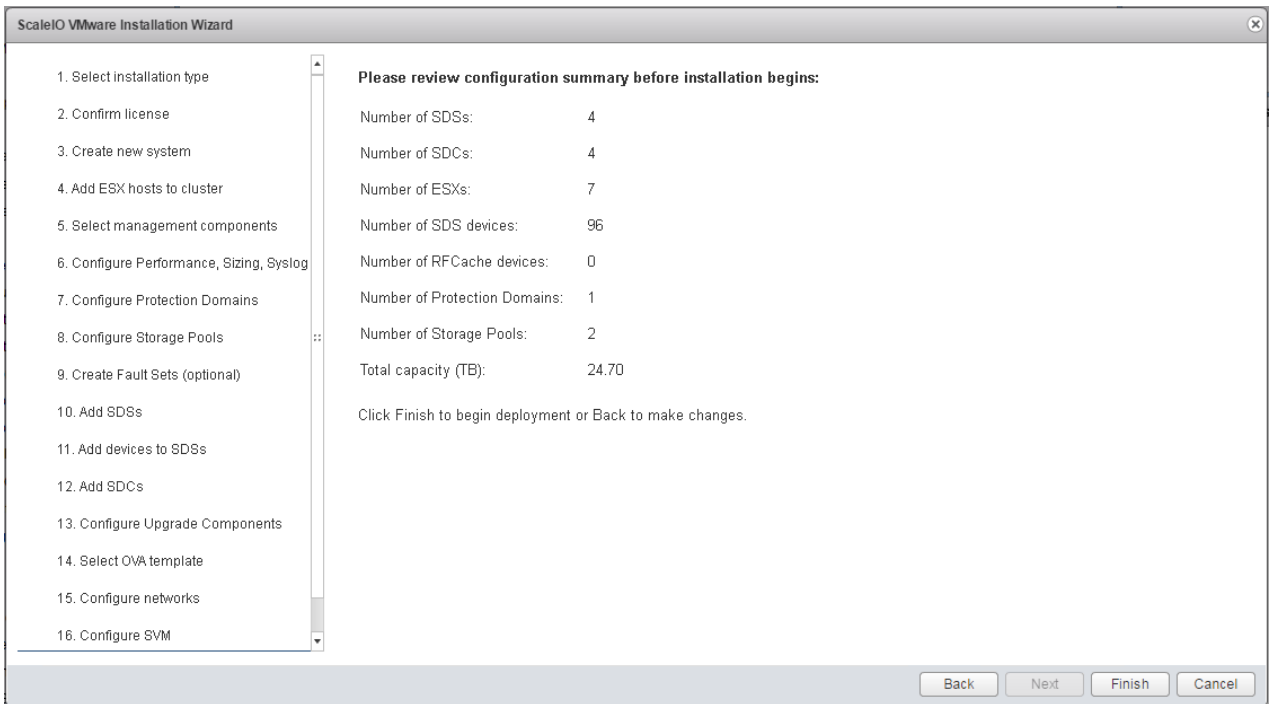


Figure 34 ScaleIO deployment wizard summary screen

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

Simple network designs typically require no more changes or interventions. The network design in this guide employs some advanced features that the deployment wizard cannot account for at the time of publication.

The ScaleIO deployment wizard is expected to fail tasks throughout the deployment process. Figure 35 shows the errors expected. The next section discusses the errors.

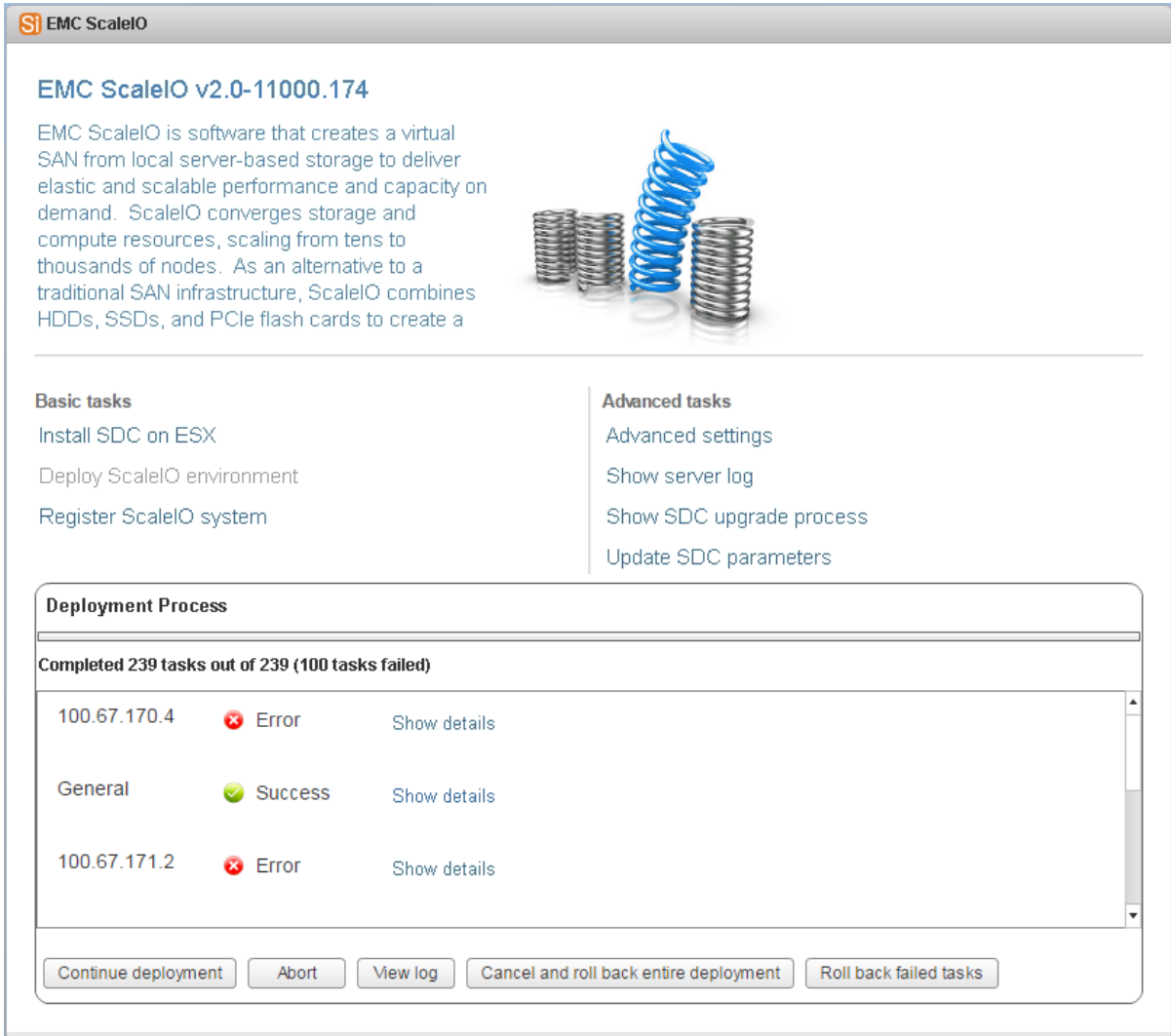


Figure 35 Expected errors during deployment process




| Deployment Process   |   |                              |
|--|---|------------------------------|
| Completed 239 tasks out of 239 (100 tasks failed)  |   |                              |
| 100.67.170.4   |  Error   | <a href="#">Hide details</a> |
| Completed: Deploy VM ScaleIO-100.67.170.204<br>Completed: Perform initial configuration for VM: 100.67.170.204<br>Completed: Install ScaleIO LIA module<br>Completed: Install ScaleIO SDS module<br>Completed: Install ScaleIO RFCACHE module<br>Failed: Configure SDC driver on ESX (The SDC cannot communicate with the MDM cluster. Verify vmkernel and/or portgroup settings.)<br>Completed: Create RDM ScaleIO-RDM-499076573 on datastore LDS 170.4<br>Completed: Create RDM ScaleIO-RDM-499076574 on datastore LDS 170.4 |   |                              |
| General  |  Success | <a href="#">Show details</a> |
| 100.67.171.2   |  Error   | <a href="#">Show details</a> |

Figure 36 SDC to MDM failure message

**Important:** Before taking any action on an error message, read the next section to identify expected errors and the steps to move forward.

## 4.4 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.

ScaleIO Virtual Machines were created during the deployment of ScaleIO. The SVMs can be categorized into the following three groups:

- MDM SVMs (3)
- SDS-SDC SVMs (4)
- ScaleIO Gateway SVM (1)

### 4.4.1 ScaleIO VDS

The ScaleIO VDS configured in Section 3.8 has two port groups. The MDM port group handles MDM traffic between the three MDM hosts. In the deployment wizard, the MDM network was chosen during the Configure networks step. When the deployment started, the tasks relating to MDM SVMs successfully completed.

The failed tasks related to SDS-SDC SVMs are expected due to the configuration of the SDC-SDS port group within the ScaleIO VDS. The SDS-SDC port group is configured as an active LAG and carries the SDS-SDC traffic between the SVMs. The LAG is not active and does not pass traffic until the virtual network adapter within the SDS-SDC SVMs is changed to the SDS-SDC network. The network wizard does not allow for separation of the MDM and SDS-SDC SVM network settings.

Section 4.4.3 shows the steps to change the virtual network adapter.

## 4.4.2 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.

Section 4.4.3 shows the steps for adding the routes to each SVM.

## 4.4.3 Detailed 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 ScaleIO plugin screen looks similar to Figure 36.
2. In vCenter, navigate to **Home > Hosts and Clusters**.
3. Right click on the first SDS-SDC SVM in one of the ScaleIO clusters, select **Edit Settings**.

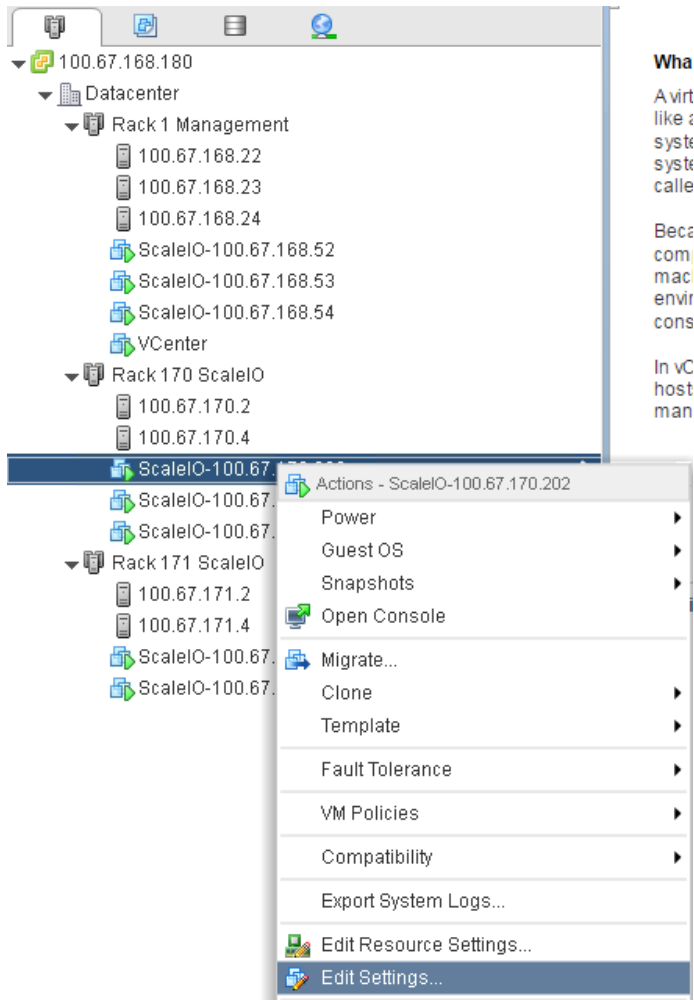


Figure 37 Edit Settings on ScaleIO SDS-SDC SVMs

4. Select the **SDS-SDC (ScaleIO VDS)** network from the dropdown menu of **Network adapter 2**

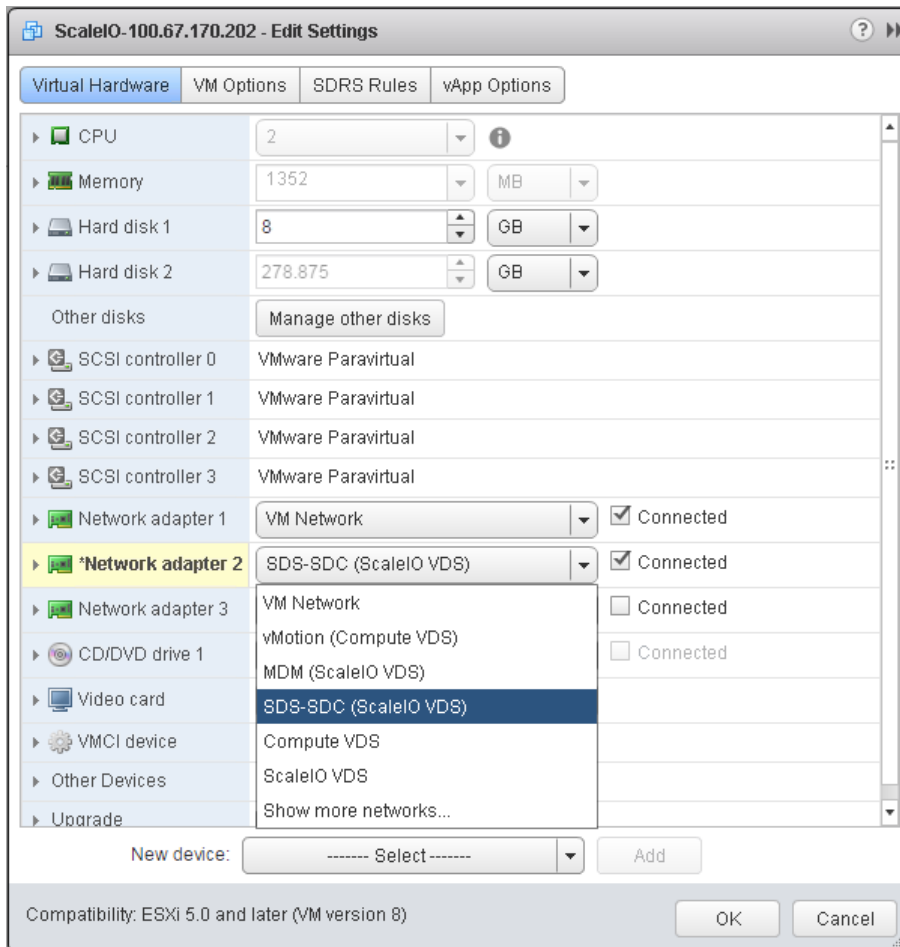


Figure 38 Changing SDS-SDC SVM data network

5. Click **OK**.
6. Repeat steps 3-5 for the remaining SDS-SDC SVMs in all the ScaleIO clusters. For this deployment example, there are a total of five SDS-SDC SVMs. Do not modify the MDM SVMs.
7. Navigate to the console of any SVM (MDM or SDC). Use the **Open Console** option or a SSH client to open a session to the **Mgmt IP** listed in Table 6.
8. 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-100-67-168-52:~ #echo "10.30.0.0/16 10.30.1.254 255.255.255.0 eth1" >> /etc/sysconfig/network/routes
```

```
ScaleIO-100-67-168-52:~ #service network restart
```

9. Repeat steps 7-8 for each SVM, changing the number for the third octet to correspond with the correct rack IP scheme. The command below shows the generic command for all SVMs. Replace the "xx" with the correct number that correlates with the Data IP in Table 6.

```
echo "10.30.0.0/16 10.30.xx.254 255.255.255.0 eth1" >>
/etc/sysconfig/network/routes
```

```
service network restart
```

10. Verify that the ScaleIO SVMs can send traffic across the spines by pinging the Cluster Virtual IP address, 10.30.1.1, from the console of each SDS-SDC SVM.
11. Verify the host routing table is correct by pinging the Cluster Virtual IP address, 10.30.1.1, from the console of each ESXi host.
12. Run additional ping tests as follows:
  - a. Master MDM SVM to:
    - i. ESXi hosts (e.g. 100.67.170.2)
    - ii. SDS-SDC VMkernel Ports (e.g. 10.30.170.2)
    - iii. Data IP SDS-SDC SVMs (e.g. 10.30.170.202)
  - b. SDS-SDC SVM to:
    - i. Master MDM SVM (100.67.168.52)
    - ii. Cluster Virtual IP (10.30.1.1)
  - c. ESXi hosts to Master MDM SVM (100.67.168.52)
13. Return to the ScaleIO deployment wizard:
  - a. Click **Continue deployment**.
  - b. Choose **No** when the message appears:

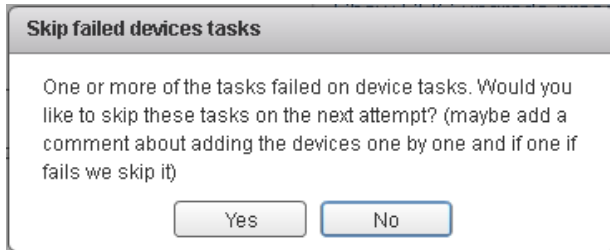


Figure 39 Continue deployment failed task message


The deployment wizard should now complete without failing any tasks.

Click **Finish** after the deployment process completes.

**EMC ScaleIO**

### EMC ScaleIO v2.0-11000.174

EMC ScaleIO is software that creates a virtual SAN from local server-based storage to deliver elastic and scalable performance and capacity on demand. ScaleIO converges storage and compute resources, scaling from tens to thousands of nodes. As an alternative to a traditional SAN infrastructure, ScaleIO combines HDDs,



**Basic tasks**

- Install SDC on ESX
- Deploy ScaleIO environment
- Register ScaleIO system

**Advanced tasks**

- Advanced settings
- Show server log
- Show SDC upgrade process
- Update SDC parameters

**Deployment Process**

**Completed 239 tasks out of 239 (0 tasks failed)**

|               |           |                              |
|---------------|-----------|------------------------------|
| 100.67.171.2  | ✔ Success | <a href="#">Show details</a> |
| 100.67.171.4  | ✔ Success | <a href="#">Show details</a> |
| 100.67.168.24 | ✔ Success | <a href="#">Show details</a> |

Figure 40 Deployment process completed



## 5 Scaling and tuning guidance

### 5.1 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 S4048-ON switch because of its ability to provide a 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.

The S4048-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.

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

The first example of scaling this solution is a rack pod configuration of 16 racks. This pod consists of one rack that contains WAN-edge connectivity. Secondly, it contains one rack of management servers including the management servers for the ScaleIO solution. Lastly, there are 14 racks of the SDS storage/compute nodes. Each of these 14 racks holds 19 PowerEdge R730xd's.

Because this particular example with each R730xd has four 10Gb uplinks, there would be 19 servers/nodes per rack. Additionally, the example architecture has four spine switches

Table 7 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 10 = 760Gb                     | 8 * 40 per rack = 320Gb<br>8 * 100Gb per rack = 800Gb | 16 * 320Gbs = 5120 Gb<br>16 * 800 = 12,800Gb               |

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

As an additional example, if redundancy is not important, a single, higher-density switch can be used as the leaf switch, for example a Z9100-ON switch with 10Gb breakout connections.

## 5.3 Scaling beyond 16 racks

The proof-of-concept scaling that Figure 41 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.

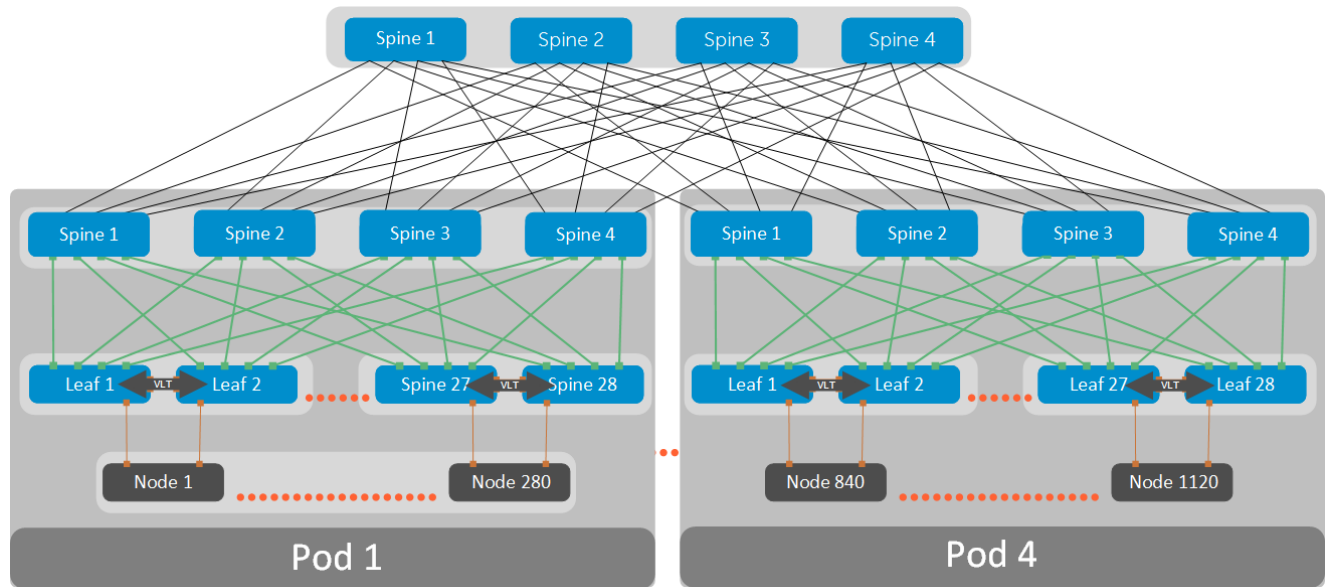


Figure 41 Scaling out the existing networking topology

## 5.4 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 Manage tab, click Resource Allocation.
3. Click System Traffic.  
You see the bandwidth allocation for the types of system traffic.
4. Select the traffic type according to the vSphere feature that you want to provision and click **Edit**.  
The network resource settings for the traffic type appear.
5. 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.

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.

6. In the **Reservation** text box, enter a value for the minimum required bandwidth for the traffic type. 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.
7. In the **Limit** text box, enter the maximum bandwidth that system traffic of the selected type can use.
8. Click **OK** to apply the allocation settings.

## 5.5 Tuning Jumbo frames

In the initial system solution that was put together for this set of ScaleIO validation scenarios, jumbo frames was not enabled. After stability of the environment was confirmed, the MTU size was increased to 9,000 bytes (jumbo frames) 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 42 below shows the interfaces that require an MTU value of 9,000 bytes for a single path from SDC 1 to SDS 1.

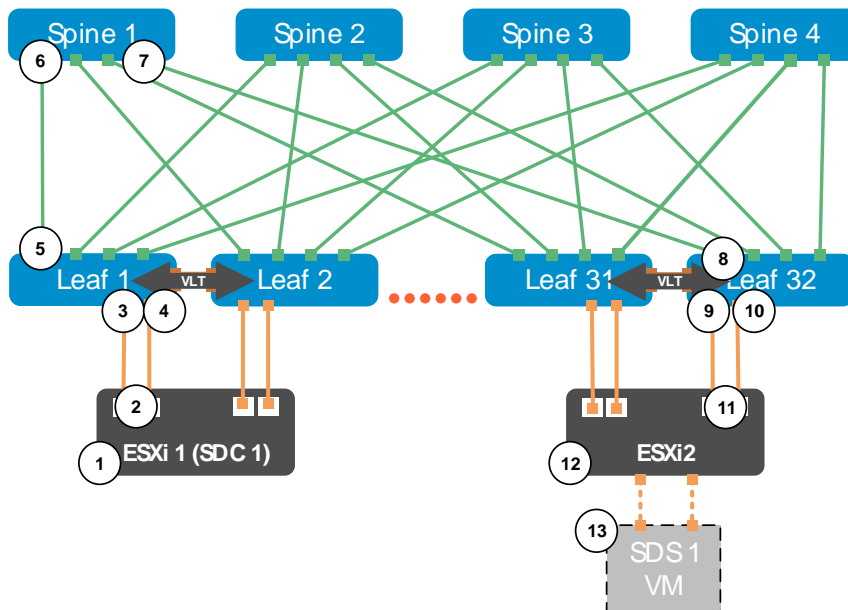


Figure 42 Jumbo frame interface configuration points

Table 8 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                                    |  |

## 5.6 Quality of Service (QoS)

The hyper-converged solution includes application and storage traffic distributed across the entire leaf-spine network architecture. To improve critical ScaleIO 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.

### 5.6.1 DSCP marking on virtual distributed switches

Differentiated services provides 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 **ScaleIO VDS**, select the **MDM** distributed port group.
3. Right click on the **MDM** 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**.

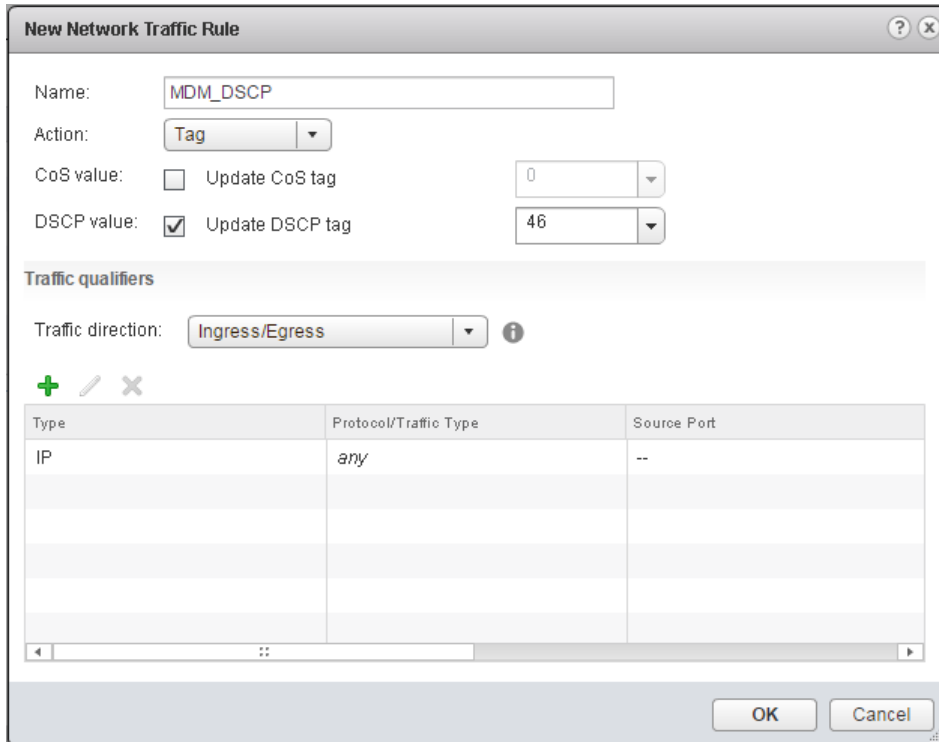


Figure 43 Traffic rule for MDM distributed port group

5. Click **OK** to save the MDM port group settings.
6. Repeat steps 2-5 for each port group in the deployment. For the **SDS-SDC** port group, use the same DSCP value, 46. For the port groups contained within the **Compute VDS**, use a DSCP value of 14.

The example configuration in section 3 does not show application VMs and their associated port groups. After installation of ScaleIO, the administrator can add application VMs and use the Compute VDS for any application-related traffic.

This simplified example for QoS only demonstrates the use of two classes of traffic, application and storage. 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. DSCP values used in this example are simply for demonstrating the configuration process.

## 5.6.2 Switch QoS configuration

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

Dell EMC Networking switches provide a high degree of customization for QoS. Some options available include bandwidth limitations, Weighted Random Early Detection (WRED) and Explicit Congestion Notification (ECN). There is not a generic, one-size-fits-all approach to QoS. 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 distributed port groups
- Maps DSCP input traffic to specified queues and DSCP color
- Prioritizes egress traffic on uplinks through strict queuing and WRED

Leaf switch configuration procedure:

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

```
class-map match-any class_compute
  match ip dscp 14
```

```
class-map match-any class_storage
  match ip dscp 46
```

3. Create an input policy map to map each class of traffic to a specific queue.

```
policy-map-input pmap_ingress
  service-queue 1 class-map class_compute
  service-queue 3 class-map class_storage
```

4. Create a DSCP color map profile for the application traffic.

```
qos dscp-color-map Colormap_DSCP
  dscp yellow 14
```

5. Apply to each input interface the input policy map with an input service policy and the DSCP color map profile with a qos dscp color map policy.

```
interface TenGigabitEthernet 1/1
  service-policy input pmap_ingress
  qos dscp-color-policy Colormap_DSCP
```

Repeat step 4 for each input interface. (leaf interfaces to each node)

6. Create a qos output policy for the storage traffic

```
qos-policy-output qpol_egress
```

```
scheduler strict
```

7. Create a WRED profile for the color used in Step 4.

```
wred-profile Yellow_profile  
  threshold min X max X max-drop-rate X
```

Replace the “**X**” in the command above with appropriate values for your deployment

8. Create a qos output policy for the application traffic.

```
qos-policy-output qpol_egress2  
  wred yellow Yellow_profile
```

9. Create an output policy map for each qos policy

```
policy-map-output pmap_egress  
  service-queue 1 qos-policy qpol_egress2  
  service-queue 3 qos-policy qpol_egress
```

10. Apply the output policy with an output service policy to each switch uplink interface.

```
interface fortyGigE 1/49  
  service-policy output pmap_egress
```

Repeat step 10 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.

### 5.6.3 QoS validation

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

To verify the traffic is being marked by the virtual distributed port group and the switch is properly mapping those values to a specific queue, perform the following command at the switch CLI:

```
Leaf1#show qos statistics  
Interface Te 1/1  
Queue#  Matched Pkts  
0       0  
1       9653  
2       0  
3       0  
4       0  
5       0  
6       0  
7       0
```

```

Interface Te 1/2
Queue#  Matched Pkts
  0      0
  1      0
  2      0
  3    2117934561
  4      0
  5      0
  6      0
  7      0

```

The above-abbreviated output shows that queues 1 and 3 contain traffic. Our example maps the DSCP values to those specific queues. If the DSCP values were not marked on the traffic, the class map would not place traffic in any queue. Ensure that each interface shows the proper traffic type as configured for your deployment.

To verify if the strict queuing and WRED output policies are functioning:

```
Leaf1#show qos statistics egress-queue fortyGigE 1/51
```

```

Interface Fo 1/51
Unicast/Multicast Egress Queue Statistics
Queue# Q# Type      TxPkts      TxPkts/s    DroppedPkts  DroppedPkts/s
-----
  0      UCAST          0            0             0             0
  1      UCAST    6056811      221628        96486         3551
  2      UCAST          0            0             0             0
  3      UCAST    5033473      220002         0             0
  4      UCAST          0            0             0             0
  5      UCAST          0            0             0             0
  6      UCAST          0            0             0             0
  7      UCAST          0            0             0             0

```

**Note:** The above output of the show command has been truncated to show only the columns with packet statistics.

The above-abbreviated output shows that queue #3 is has no dropped packets. In strict-priority queuing, the system de-queues all packets from the assigned queue before servicing any other queues. The queue assigned the WRED policy, queue #1, shows dropped packets as expected. The test traffic used in this example included application traffic above the total available bandwidth while simultaneously generating storage traffic.



## 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 9 Software Components

| Software                               | Version                        | Link to Documentation   |
|--|--------------------------------|---|
| VMware vSphere ESXi                    | 6.0.0 Update 2 A03             | <a href="http://www.vmware.com/products/vsphere/">http://www.vmware.com/products/vsphere/</a>               |
| VMware vCenter Server Appliance (vCSA) | 6.0.0 Update 2 - build 3634788 | <a href="http://www.vmware.com/products/vcenter-server/">http://www.vmware.com/products/vcenter-server/</a> |
| EMC ScaleIO                            | R2_0.5014.0                    | <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 by this document.

Table 10 Servers and Switches

| Product          | Description   | Link to documentation   |
|------------------|---|---|
| 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>   |
| PowerEdge R630   | Dell EMC rack server that provides management infrastructure for the EMC Converged Infrastructure   | <a href="http://www.dell.com/us/business/p/powered-ge-r630/pd">http://www.dell.com/us/business/p/powered-ge-r630/pd</a>   |
| S4048-ON         | The Dell EMC Networking S-Series S4048-ON is an ultralow-latency 10/40GbE top-of-rack (ToR) switch built for applications in high-performance data center and computing environments. | <a href="http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/Dell-Networking-S4048-ON-Spec-Sheet.pdf">http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/Dell-Networking-S4048-ON-Spec-Sheet.pdf</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/learn/us/en/12/shared-content~data-sheets~en/documents~dell-networking-z9100-spec-sheet.pdf">http://www.dell.com/learn/us/en/12/shared-content~data-sheets~en/documents~dell-networking-z9100-spec-sheet.pdf</a>       |

## A.3 Server and switch component details

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

Table 11 BIOS, Firmware and Switch OS Components

| Component                                   | Version   | Notes  |
|---|---|--|
| PowerEdge R730xd Server BIOS                | 2.3.4   | 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: 25.5.0.0018<br>ESXi lsi_mr3 Driver: 6.903.85.00 | 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: 17.5.10<br>ESXi net-ixgbe Driver: 4.1.1.1       | 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                               | 9.11.0  | 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 S4048-ON                               | 9.11.0  | The Dell EMC Networking S4048-ON switch has the ability to provide a 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 44 R730xd front view with bezel



Figure 45 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 46 R730xd back view

## A.5 PowerEdge R630 server

The PowerEdge R630 server is a 2-socket, 1-RU server. This server functions in monitoring and configuration agent (MDM) in the ScaleIO system. The MDM is mainly used for management, which consists of migration, rebuilds and all system-related functions.



Figure 47 R630 front view without bezel

## A.6 S4048-ON switch

The S4048-ON is a 1RU Layer 2/3 switch with 48, 10GbE SFP+ ports and 6, 40GbE QSFP+ ports. Six S4048-ON switches are used as leaf switches in the Leaf-Spine topology covered in this guide.



Figure 48 S4048-ON

## A.7 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.

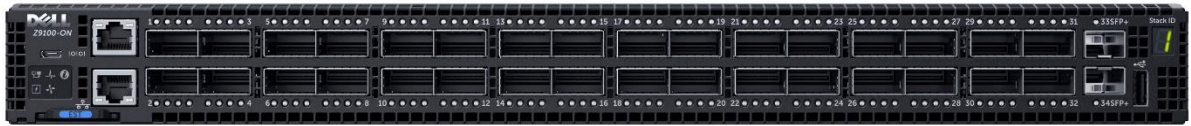


Figure 49 Z9100-ON

## A.8 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 management traffic in this guide.



Figure 50 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 ScaleIO 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. For each HDD disk attached to the PERC controller needs to be placed in a separate RAID-0 container:

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

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

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

20. Connect to the iDRAC in a web browser and launch the virtual console.
21. In the virtual console, from the Next Boot menu, select BIOS Setup.
22. Reboot the server.

23. From the System Setup Main Menu, select Device Settings.
24. From the list of devices, select the PERC controller. This opens the Modular RAID Controller Configuration Utility Main Menu.
25. Select Virtual Disk Management
26. 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 [support.dell.com](https://support.dell.com). 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, the Broadcom QLogic 5700 1GbE on-board adapter provides this connection for R630 servers and 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 Support and feedback

### Contacting Technical Support

Support Contact Information

Web: <http://Support.Dell.com/>

Telephone: USA: 1-800-945-3355

### Feedback for this document

We encourage readers of this publication to provide feedback on the quality and usefulness of this best practices guide by sending an email to [Dell\\_Networking\\_Solutions@Dell.com](mailto:Dell_Networking_Solutions@Dell.com).

## About Dell EMC

Dell EMC is a worldwide leader in data center and campus solutions, which includes the manufacturing and distribution of servers, network switches, storage devices, personal computers and related hardware and software. For more information on these and other products, please visit the Dell EMC website at <http://www.dell.com>.