ScaleIO 2.0 Deployment Guide with Dell Networking S5048F-ON 25GbE Switches

Dell EMC Networking Infrastructure Solutions
February 2018

Revisions

Date	Rev.	Description	Authors
February 2018	1.0	Initial release	Curtis Bunch, Jordan Wilson, Andrew Waranowski

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Table of contents

Executive summary

Dell EMC ScaleIO is an industry leading software-defined storage (SDS) solution that enables customers to extend their existing virtual infrastructure into a high performing virtual SAN. ScaleIO creates a virtual server SAN using industry-standard servers with direct attached storage (DAS). You can calibrate ScaleIO to perform using as few as three hosts, to 1,024 hosts. Each host can utilize storage media such as flash-based SSDs, NVMe SSDs, traditional spinning disks, or a mix.

The Dell EMC Networking S5048F-ON is the latest in the S-series of switches that provide the bandwidth and low latency support for a scalable storage architecture. This document details the deployment of the Dell EMC ScaleIO 2.0.1.4 solution using the Dell Networking S5048F-ON 25GbE switch. This document also:

- Assists administrators in selecting the best hardware and topology for their Dell EMC ScaleIO network
- Delivers detailed instructions and working examples for the deployment and configuration of Dell EMC S-series switches
- Delivers instructions and working examples for the deployment and configuration of a sample Dell ScaleIO virtual SAN
- Shows conceptual, physical, and logical diagram examples for various networking topologies

1 Introduction

ScaleIO is a software-only solution that uses existing servers' local disks and LAN to create a virtual SAN that has all the benefits of external storage – but at a fraction of cost and complexity. ScaleIO utilizes the existing local storage devices and turns them into shared block storage. For many workloads, ScaleIO storage is comparable to, or better than, external shared block storage.

The lightweight ScaleIO software components are installed on the application servers, behaving like a software initiator, and communicate via a standard LAN to handle the application I/O requests sent to ScaleIO block volumes. An extremely efficient decentralized block I/O flow, combined with a distributed, sliced volume layout results in a massively parallel I/O system that can scale to thousands of nodes.

ScaleIO is designed and implemented with enterprise-grade resilience. Furthermore, the software features an efficient distributed self-healing process that overcomes media and server failures, without requiring administrator involvement.

In the modern data center, 100 Gbps Ethernet is now at an affordable price to respond to the increasing demands of bandwidth from storage and compute. Dell EMC offers the S5048F-ON, offering 25/100GbE Ethernet, designed to be used as a data center Top of Rack (ToR) switch.

Combining Dell PowerEdge R730xd servers, Dell EMC Networking 25GbE switching, and the flexibility of Dell EMC ScaleIO, this document covers the network deployment requirements for the ScaleIO software-only solution. This guide covers deploying the ScaleIO software-only solution on VMware vSphere ESXi, creating a Hyper-Converged Infrastructure (HCI) by enabling each ESXi host to present computing as well as storage resources.

While not used in this paper, Dell EMC offers ScaleIO Ready Nodes, which is the combination of ScaleIO software-defined block storage and Dell PowerEdge servers, optimized to run ScaleIO, enabling the deployment of an entirely architected, software-defined, scale-out server SAN. For more information on ScaleIO Ready Nodes see the *Dell EMC ScaleIO Ready Node data sheet*.

This guide does not cover physically cabling or connecting the S5048F-ON switches to existing data center infrastructure. A prerequisite for this deployment guide is access to a VMware vSphere vCenter Server capable of using virtual distributed switches.

Note: For deploying a spine-leaf architecture using Dell EMC Networking see *Dell EMC Leaf-Spine Deployment Guide* for a step-by-step guide for more information. For steps on deploying and configuring VMware vSphere, see *vSphere Networking Guide for vSphere 6.5, ESXi 6.5, and vCenter Server 6.5*.

1.1 Typographical conventions

This document uses the following typographical conventions:

`Monospaced text`	Command Line Interface (CLI) examples
`Bold monospaced text`	Commands entered at the CLI prompt
`Italic monospaced text`	Variables in CLI examples
Bold text	Graphical User Interface (GUI) fields and information entered in the GUI

1.2 Dell EMC ScaleIO

A ScaleIO virtual SAN consists of the following software components.

- Meta Data Manager (MDM) - Configures and monitors the ScaleIO system. The MDM can be configured in redundant cluster mode, with three members on three servers or five members on five servers, or in single mode on a single server. In this guide four Dell PowerEdge R730xd servers are deployed, three of them are used to create a three-member MDM configuration.
- ScaleIO Data Server (SDS) - Manages the capacity of a single server and acts as a back-end for data access. In this guide, all four R730xd servers act as an SDS.
- ScaleIO Data Client (SDC) - A lightweight device driver that exposes ScaleIO volumes as block devices. In this guide, all four R730xd servers act as an SDC.

In this environment, the ScaleIO Virtual Machine (SVM) hosts the MDM and SDS roles. Each ScaleIO node, running ESXi, has a separate SVM. The figure below illustrates the communication between these components.

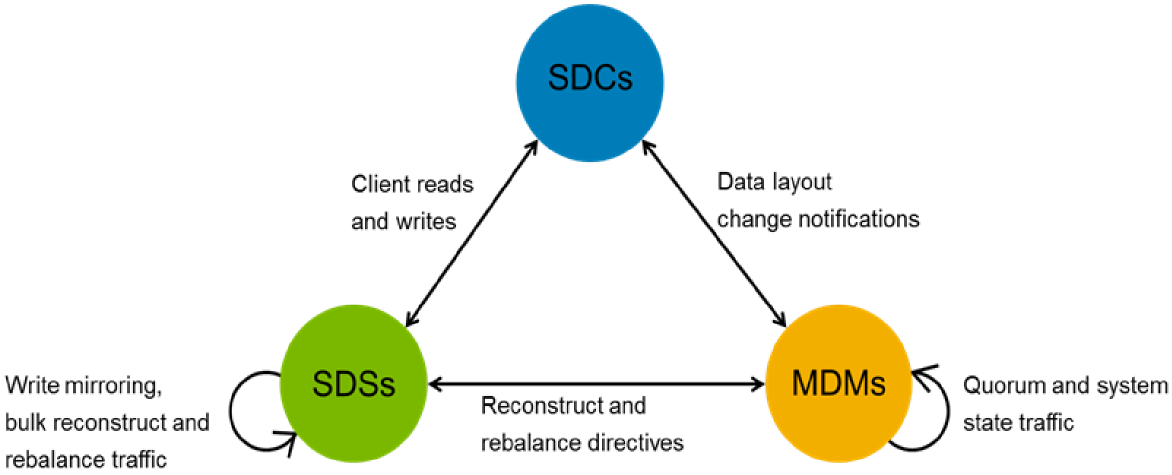

SDCs
Client reads and writes
Data layout change notifications
Write mirroring, bulk reconstruct and rebalance traffic
SDSs
Reconstruct and rebalance directives
MDMs
Quorum and system state traffic

Figure 1 ScaleIO component communication

2 Hardware overview

This section briefly describes the hardware used to validate the deployment example in this guide. Appendix C contains a complete listing of hardware and components.

2.1 Dell EMC Networking S3048-ON

The Dell EMC Networking S3048-ON is a 1-Rack Unit (RU) switch with forty-eight 1GbE Base-T ports and four 10GbE SFP+ ports. In this guide, one S3048-ON supports management traffic in each rack.

Figure 2 Dell EMC Networking S3048-ON

2.2 Dell EMC Networking S5048F-ON

The S5048F-ON is a 1RU switch with forty-eight 25GbE SFP28 ports and six ports of 100GbE. In this guide, this switch is deployed as a leaf switch performing basic gateway functionality for attached ScaleIO hosts.

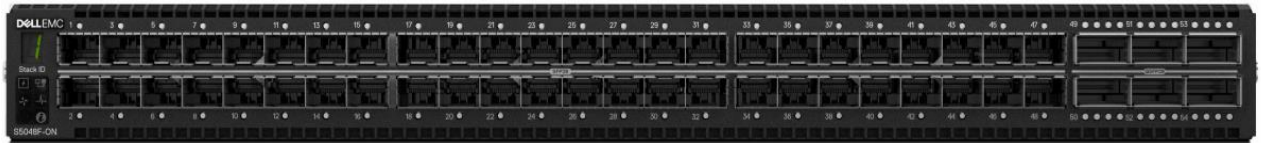


Figure 3 Dell EMC Networking S5048F-ON

2.3 Dell EMC PowerEdge R730xd

The Dell EMC PowerEdge R730xd is a 2-RU, two-socket server platform. It allows up to 32 x 2.5” SSDs or HDDs with SAS, SATA, and NVMe support. In this guide, four R730xd servers are used in the ScaleIO cluster.

Figure 4 Dell EMC PowerEdge R730xd

3 Networking

In ScaleIO, inter-node communication (for managing data locations, rebuilding, and rebalancing, and for application access to stored data) can be done on one IP network or spread across separate IP networks. Regardless of the model, ScaleIO supports VLANs. Management is done in one of two ways:

- Via a separate network with access to the other ScaleIO components
- On the same network

Each R730xd has four 25GbE ports provided by two Mellanox Connect X-4 LX PCIe cards. Two of the available four ports are used to carry all traffic types (frontend and backend) in this deployment and Quality of Service (QoS) is used to ensure that traffic requiring low-latency is prioritized (see Section 7.2).

Note: Since a node running ScaleIO can provide compute resources, the remaining two free ports can be leveraged for compute workloads depending on requirements of the environment.

3.1 Network topology

This section provides an overview of the network topology and the physical connections used in this deployment.

Production network

A non-blocking network design allows the use of all switch ports concurrently. Such a design is needed to accommodate various traffic patterns in a ScaleIO deployment and optimize the additional traffic generated in the HCI environment. Figure 5 shows a leaf-spine topology providing access to the existing infrastructure found in a typical data center. The ScaleIO components, MDM, SDS, and SDC, reside on the R730xd hosts while ESXi is managed through vCenter in the management environment.

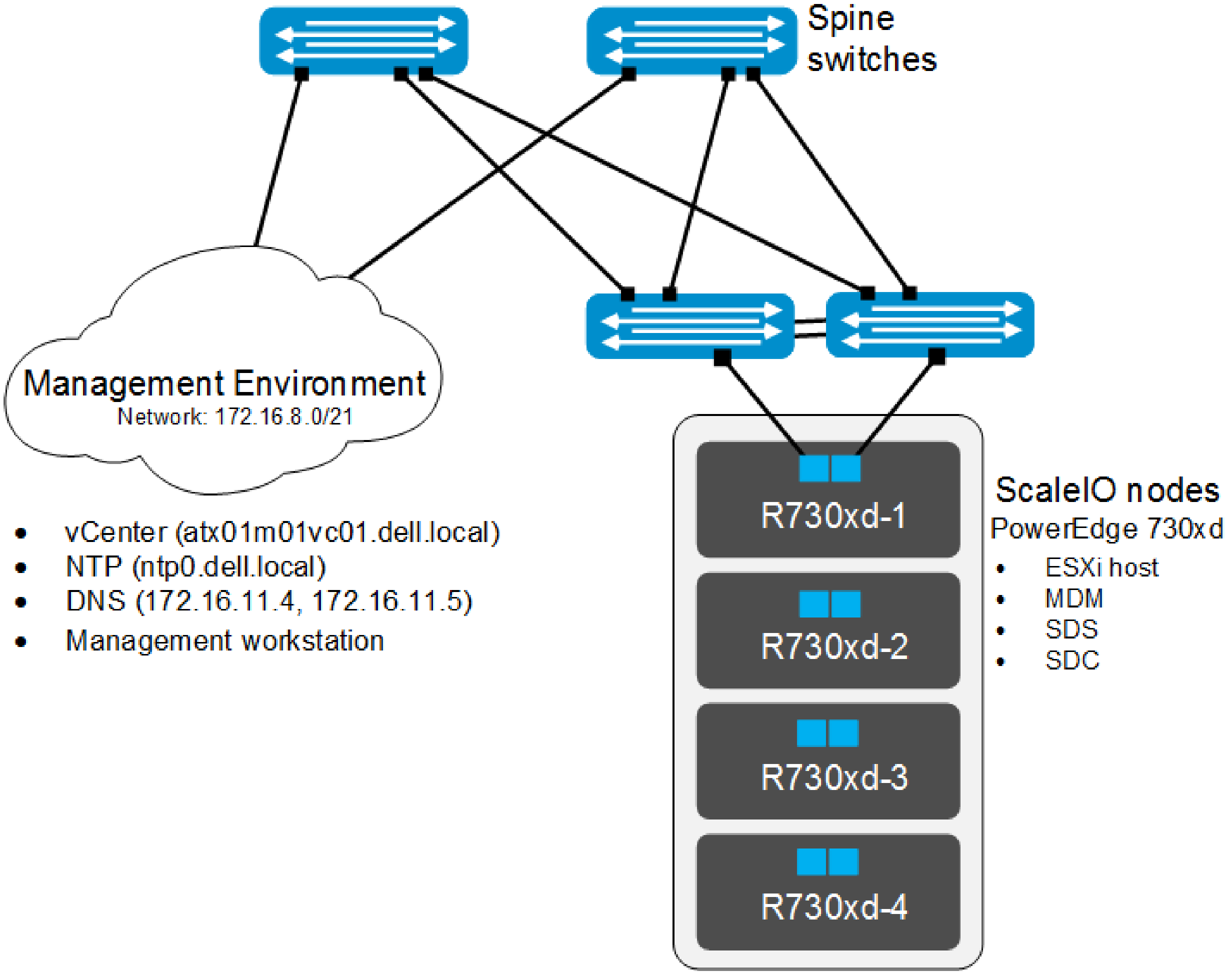


Figure 5 Production network

Management Network

A single S3048-ON switch provides iDRAC connectivity to the PowerEdge R730xd servers. Figure 6 shows the S3048-ON is connected to the S5048F-ON through a port channel. All interfaces are configured to a dedicated subnet and VLAN, and that VLAN is tagged on the upstream port channel.

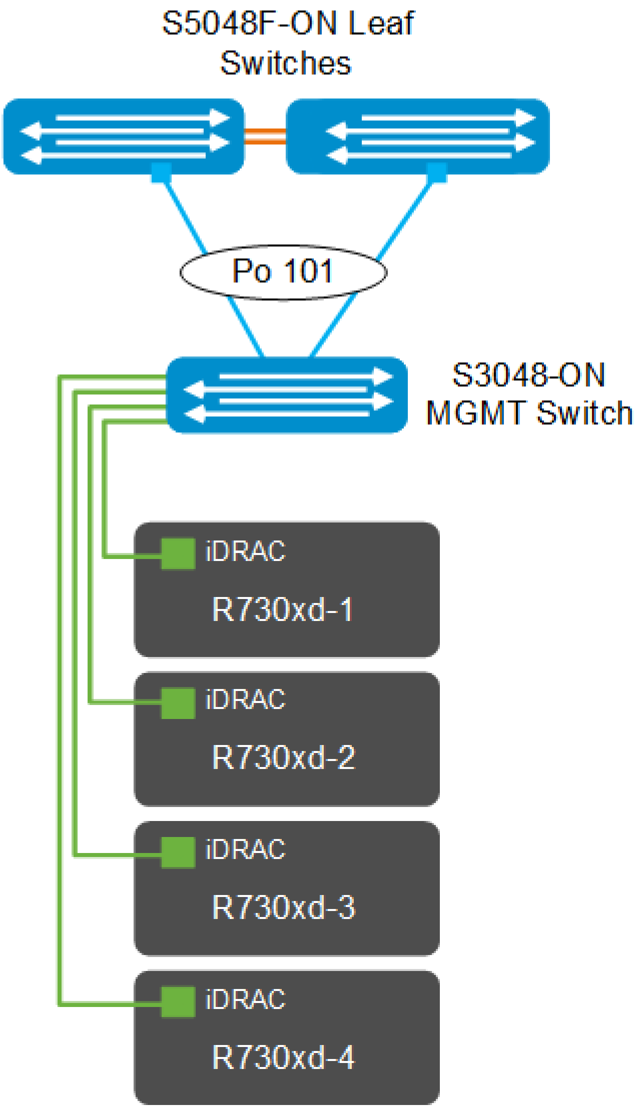


Figure 6 Management network

3.2 Network connectivity

The figure below shows one R730xd server connected to two S5048F-ON switches via two Mellanox ConnectX-4 LX PCIe cards installed in PCIe slots one and two. The leaf switches are Virtual Link Trunking (VLT) peers, and one port from each PCIe card connects to each leaf switch. The connections for R730xd-2 through R730xd-4 (not shown) are done in the same manner.

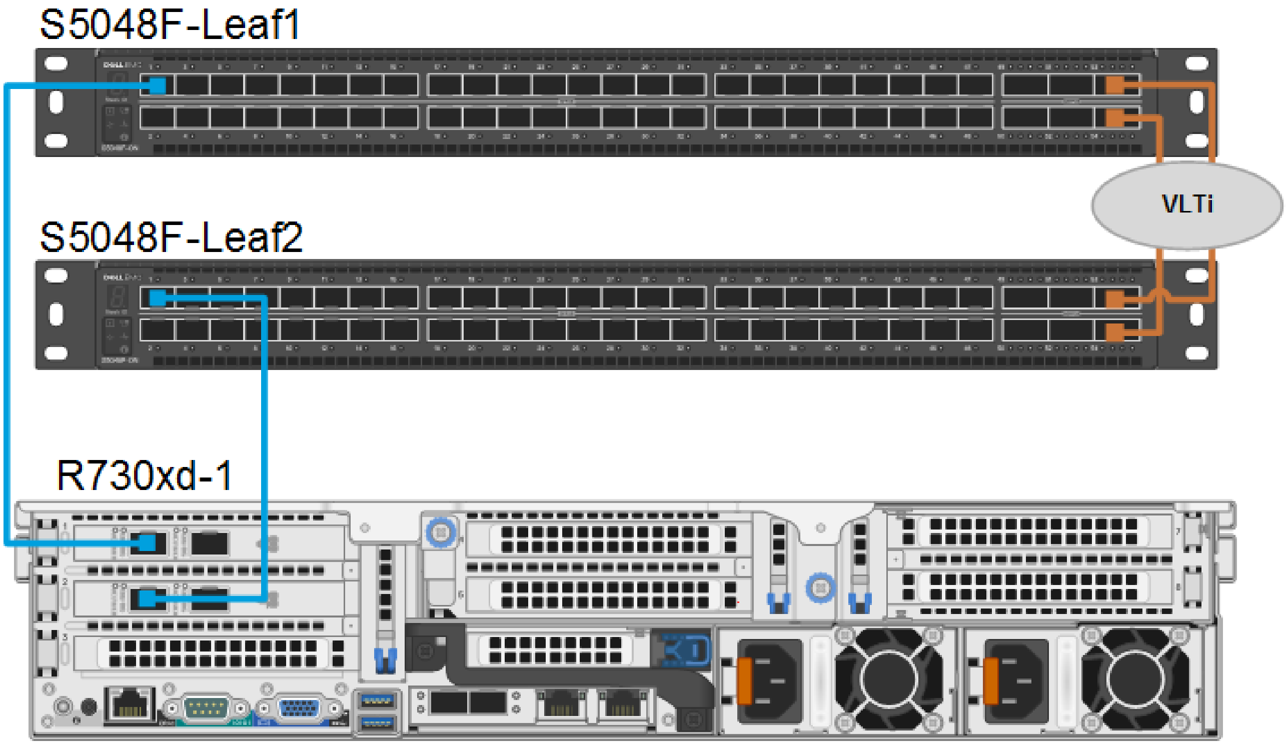


Figure 7 R730xd-1 wiring to production network

The figure below shows one R730xd server's iDRAC connected to the S3048-ON switch via the onboard iDRAC port. The S3048-Management switch is connected via a port channel to the S5048F-ON VLT pair. The connections for R730xd-2 through R730xd-4 (not shown) are done in the same manner.

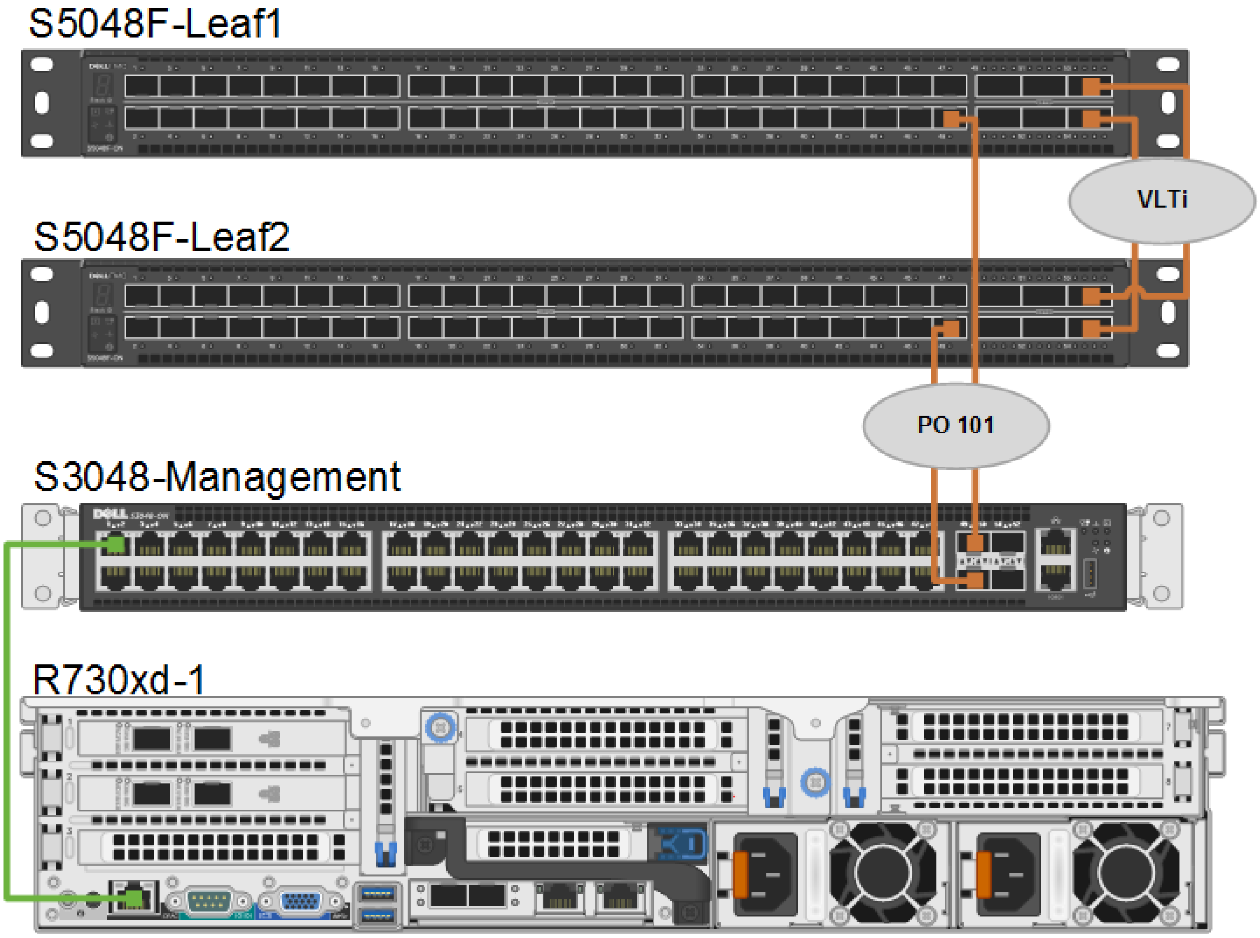


Figure 8 R730xd-1 iDRAC wiring to management switch

In this guide, the S5048F-ON switches are configured as a VLT pair. Part of a successful VLT deployment is the use of a VLT backup link. The backup link monitors the connectivity between the VLT peer switches. The backup link sends configurable, periodic keepalive messages between the VLT peer switches. Figure 9 shows that the out-of-band (OOB) management interface (ma1/1) is configured as a point-to-point link to fulfill this requirement. Administrative access is performed in-band, on a isolated VLAN, and does not require this port.

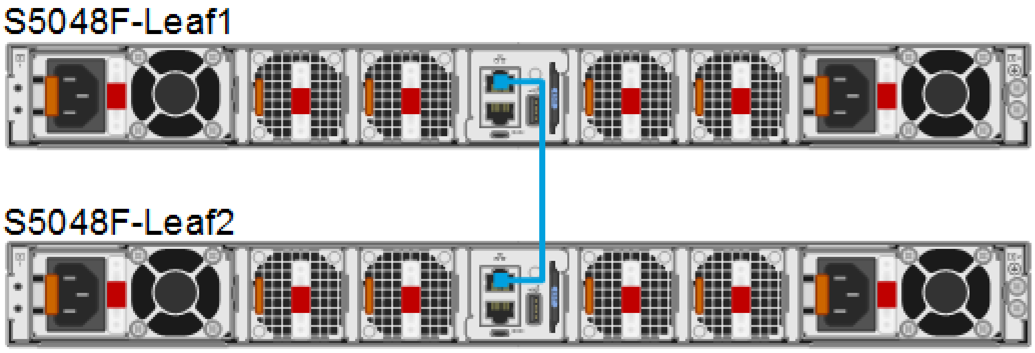


Figure 9 S5048F-ON OOB management interface used for VLT backup destination

3.3 IP addressing

Dell EMC ScaleIO MDMs, SDSs, and SDCs can have multiple IP addresses and can reside on more than one network. Multiple IPs provides options for load balancing and redundancy. ScaleIO natively provides redundancy and load balancing across physical network links when an MDM or SDS is configured to send traffic across multiple links. In this configuration, each interface available to the MDM or SDS is assigned an IP address, each in a different subnet. In this deployment guide, two networks are used for MDM and SDS traffic. When each MDM or SDS has multiple IP addresses, ScaleIO load balances more efficiently due to its awareness of the traffic pattern.

Accounting for management networks and two data networks, each ScaleIO node needs seven IPs with the VMware vMotion network being optional:

- iDRAC Management – Used for administrative management of the R730xd through the iDRAC interface.
- ESXi Management – Used to connect the host to vCenter and management.
- ESXi vMotion – Used to migrate virtual machines between ESXi hosts (optional).
- SVM Management – Used for administrative management of the SVM installed on each host.
- Data Network 1 SVM – Used by the SVM to access Data Network 1.
- Data Network 1 SDC – Used by ESXi as a VMkernel IP for data access to Data Network 1.
- Data Network 2 SVM – Used by the SVM to access Data Network 2.
- Data Network 2 SDC – Used by ESXi as a VMkernel IP for data access to Data Network 2

Table 1 and Table 2 show how to calculate the IP address requirements. To accommodate expansion each address pool is assigned a dedicated subnet.

Table 1 VMware and management IP network calculations

Item	Description	Comments
N	Number of nodes	

IP address pools	The pools of IP addresses used for the following groups:
	
	ESXI_MGMT_IP = Management IP addresses
	ESXI_VMOTION_IP = ESXi vMotion IP addresses
	SVM_MGMT_IP = SVM management IP addresses
	IDRAC_MGMT_IP = iDRAC IP address for each node	

The formula to calculate IP address subnet pool or subnet needs is: N * ESXI_MGMT_IP + N * ESXI_VMOTION_IP + N * SVM_MGMT_IP + N * IDRAC_MGMT_IP		

Table 2 ScaleIO data IP network calculations

Item	Description	Comments
N	Number of nodes	
Data network 1	The pools of IP addresses used for static allocation for the following groups: 1. Node_DATA1_IP = ScaleIO internal (interconnect) IP addresses 2. SVM_DATA1_IP = SVM management IP addresses 3. MDM_Cluster_Virtual_IP_DATA1 = The virtual IP of the MDM cluster in the Data1 network	For clarity, the first subnet is referred to as “Data1”.
	The formula to calculate the subnet pool or subnet needs is: N * Node_DATA1_IP + N * SVM_DATA1_IP + MDM_Cluster_Virtual_IP_DATA1	
Data network 2	The pools of IP addresses used for static allocation for the following groups: 1. Node_DATA2_IP = ScaleIO Nodes internal (interconnect) IP addresses 2. SVM_DATA2_IP = SVM management IP addresses MDM_Cluster_Virtual_IP_DATA2 = The virtual IP of the MDM cluster in the Data2 network	For clarity, the first subnet is referred to as “Data2”.
	The formula to calculate the subnet pool or subnet needs is: N * Node_DATA2_IP + N * SVM_DATA2_IP + MDM_Cluster_Virtual_IP_DATA2	

The IP ranges and default gateway for each IP subnet are shown in Table 3. VMware vMotion, Data1, and Data2 do not require default gateways in this configuration. If routing is required for Data1 and Data2 see, Appendix B.

Table 3 IP address ranges

IP Pool	IP Subnet	Default gateway
IDRAC_MGMT_IP	172.16.30.0/24	172.16.30.253
ESXI_MGMT_IP	172.16.31.0/24	172.16.31.253
ESXI_VMOTION_IP	172.16.32.0/24	n/a
SVM_MGMT_IP	172.16.33.0/24	172.16.33.253
Node_DATA1_IP & SVM_DATA1_IP	172.16.34.0/24	n/a
Node_DATA2_IP & SVM_DATA2_IP	172.16.35.0/24	n/a

Each data and management network can reside on a different VLAN, enabling separation at a much higher granularity. In this deployment guide, VLANs are used to separate each subnet as shown in Table 4.

Table 4 VLAN and subnet association

IP Pool	VLAN ID
IDRAC_MGMT_IP	1630
ESXI_MGMT_IP	1631
ESXI_VMOTION_IP	1632
SVM_MGMT_IP	1633
Node_DATA1_IP & SVM_DATA1_IP	1634
Node_DATA2_IP & SVM_DATA2_IP	1635

4 Configure physical switches

This section contains switch configuration details with explanations for one switch. The remaining switch uses a configuration very similar. Complete configuration files for both switches are provided as attachments.

4.1 Factory Default Settings

The configuration commands in the sections below assume switches are at their factory default settings. All switches in this guide can be reset to factory defaults as follows:

```
switch# restore factory-defaults stack-unit unit# clear-all
Proceed with factory settings? Confirm [yes/no]:yes
```

Factory settings are restored, and the switch reloads. After reload, enter **A** at the [A/C/L/S] prompt as shown below to exit Bare Metal Provisioning mode.

```
This device is in Bare Metal Provisioning (BMP) mode.
To continue with the standard manual interactive mode, it is necessary to abort
BMP.

Press A to abort BMP now.
Press C to continue with BMP.
Press L to toggle BMP syslog and console messages.
Press S to display the BMP status.
[A/C/L/S]:A

% Warning: The bmp process will stop ...

Dell>
```

The switch is now ready for configuration.

4.2 S5048F-ON leaf switch configuration

The following section outlines the configuration commands issued to S5048F-ON leaf switches. The switches start at their factory default settings per Section 4.1.

Note: The following configuration details are specific to S5048F, Leaf 1. The remaining leaf switch is similar. Complete configuration details for both leaf switches are provided in the attachments named S5048-Leaf1.txt and S5048-Leaf2.txt.

Initial configuration involves setting the hostname and enabling LLDP. LLDP is useful for troubleshooting. The VLT backup interface is configured with an IP address. Next, a policy map is created to trust all incoming DiffServ markings (see Section 7.2).

```
enable
configure
hostname S5048F-Leaf1
protocol lldp
advertise management-tlv management-address system-description system-name
advertise interface-port-desc

interface ManagementEthernet 1/1
ip address 192.168.255.1/30
no shutdown

policy-map-input TrustDSCPin
trust diffserv
```

Next, the VLT interfaces between S5048F-Leaf1 and S5048F-Leaf2 are configured. In this configuration, add interfaces hundredGigE 1/53-54 to static port channel 127 for the VLT interconnect. The backup destination is the management IP address of the VLT peer switch, S5048F-Leaf2. Finally, VLT peer-routing is enabled providing forwarding redundancy in the event of a switch failure.

Note: See Section 7.1 for Maximum Transmission Unit (MTU) guidelines used in this guide.

```
interface port-channel 127
description VLTi
mtu 9216
channel-member hundredGigE 1/53 - 1/54
no shutdown

interface range hundredGigE 1/53 - 1/54
description VLTi
mtu 9216
service-policy input TrustDSCPin
no shutdown

vlt domain 127
peer-link port-channel 127
back-up destination 192.168.255.2/30
unit-id 0
peer-routing
```

The downstream interfaces, to the ScaleIO cluster R730xd servers, in this case, are configured in the next set of commands. The downstream interface to the S3048-ON management switch is also configured.

```
interface twentygigabitethernet 1/1
description To R730xd-1
switchport
spanning-tree rstp edge-port
mtu 9216
service-policy input TrustDSCPin
```

```
no shutdown

interface twentygigabitethernet 1/3
description To R730xd-2
switchport
spanning-tree rstp edge-port
mtu 9216
service-policy input TrustDSCPin
no shutdown

interface twentygigabitethernet 1/5
description To R730xd-3
switchport
spanning-tree rstp edge-port
mtu 9216
service-policy input TrustDSCPin
no shutdown

interface twentygigabitethernet 1/7
description To R730xd-4
switchport
spanning-tree rstp edge-port
mtu 9216
service-policy input TrustDSCPin
no shutdown

interface twentygigabitethernet 1/48
description uplink from S3048-Management
port-channel-protocol LACP
port-channel 101 mode active
no shutdown

interface Port-channel 101
description uplink from S3048-Management
portmode hybrid
switchport
vlt-peer-lag port-channel 101
no shutdown
```

Five VLAN interfaces are created. Next, all server facing downstream interfaces are tagged for each VLAN with VLAN 1630 being tagged only for port-channel 101. Each interface, except the data networks, are assigned to a VRRP group, and a virtual address is assigned. VRRP priority is set to 254 to make this switch the master. (On the VRRP peer switch, priority is set to 1). Only Data 1 is created on this leaf switch, and the IP address is optionally assigned, while Data 2 is created on the other leaf switch.

```
interface Vlan 1630
description IDRAC_MGMT_IP & SW_MGMT
ip address 172.16.30.251/24
tagged Port-Channel 101
vrrp-group 1630
description IDRAC_MGMT_IP & SW_MGMT
priority 254
virtual-address 172.16.30.254
no shutdown

interface Vlan 1631
description ESXI_MGMT_IP
ip address 172.16.31.251
tagged twentyFiveGigE 1/1,1/3,1/5,1/7
vrrp-group 1631
description ESXI_MGMT_IP
priority 254
virtual-address 172.16.31.254
no shutdown

interface Vlan 1632
description ESXI_VMOTION_IP
mtu 9216
tagged twentyFiveGigE 1/1,1/3,1/5,1/7
no shutdown

interface Vlan 1633
ip address 172.16.33.251
description SVM_MGMT_IP
tagged twentyFiveGigE 1/1,1/3,1/5,1/7
vrrp-group 1633
description SVM_MGMT_IP
priority 254
virtual-address 172.16.33.254
no shutdown

interface Vlan 1634
description Node_DATA1_IP & SVM_DATA1_IP
mtu 9216
tagged twentyFiveGigE 1/1,1/3,1/5,1/7
no shutdown
```

Next, a user (admin) with privilege level 15 is created, and the SSH daemon is enabled allowing remote login. Finally, an ACL is created to allow SSH login only from the management subnet (see Section 3.1) and the configuration is saved.

```
ip ssh server enable
```

```
username admin secret 5 xxxxx privilege 15

ip access-list standard ALLOW-SSH
description Allow SSH from the management network
seq 5 permit 172.16.11.0/24
seq 20 deny any log

line vty 0 9
access-class ALLOW-SSH ipv4
```

Save the configuration.

```
end
write
```

Note: IP configuration for VLAN IDs 1634 and 1635 is optional. A gateway is not required for the two ScaleIO data networks. If multiple ScaleIO data network subnets are required, the SVMs must be modified (see Appendix B).

4.3 S3048-ON management switch configuration

The following section outlines the configuration commands issued to S3048-ON management switch. The switches start at their factory default settings per Section 5.1.

Initial configuration involves setting the hostname and enabling LLDP.

```
enable
configure
hostname S3048-Management
protocol lldp
advertise management-tlv management-address system-description system-name
advertise interface-port-desc
```

The upstream interfaces, to the S5048F-ON leaf switches, are configured in the next set of commands. Each interface is added to a port channel.

```
interface TenGigabitEthernet 1/49
description uplink to S5048F_Leaf1
port-channel-protocol LACP
port-channel 101 mode active
no shutdown

interface TenGigabitEthernet 1/50
description uplink to S5048F_Leaf2
port-channel-protocol LACP
port-channel 101 mode active
no shutdown
```

```
interface Port-channel 101
description uplink to S5048F_Leaf_Pair
no ip address
portmode hybrid
switchport
no shutdown
```

The downstream interfaces, to the R730xd server's iDRAC interfaces, are configured in the next set of commands. Each interface is set as a switch port interface.

```
interface GigabitEthernet 1/1
description R730xd-1_IDRAC
switchport
no shutdown

interface GigabitEthernet 1/3
description R730xd-2_IDRAC
switchport
no shutdown

interface GigabitEthernet 1/5
description R730xd-3_IDRAC
switchport
no shutdown

interface GigabitEthernet 1/7
description R730xd-4_IDRAC
switchport
no shutdown
```

VLAN ID 1630, used for iDRAC and switch management, is created and an IP address is assigned. The downstream interfaces, to the ScaleIO R730xd servers' iDRAC interfaces, are configured in the next set of commands. The upstream port channel to the S5048-ON is tagged.

```
interface vlan 1630
description IDRAC_MGMT_IP & SW_MGMT
ip address 172.16.30.250/24
no shutdown
untagged GigabitEthernet 1/1,1/3,1/5,1/7
tagged port-channel 101
```

A gateway of last resort is configured pointing at the virtual IP address configured for the VLAN ID 1630 on the S5048F-ON leaf pair (see Section 4.2). Next, a user account with privilege level 15 is created, and the SSH daemon is enabled allowing remote login. Finally, an ACL is created to allow SSH login only from the management subnet (see Section 3.1) and the configuration is saved.

```
ip route 0.0.0.0 0.0.0.0 172.16.30.254
ip ssh server enable
username admin secret 5 xxxxx privilege 15

ip access-list standard ALLOW-SSH
description Allow SSH from the management network
seq 5 permit 172.16.11.0/24
seq 20 deny any log

line vty 0 9
access-class ALLOW-SSH ipv4
```

Save the configuration.

```
end
write
```

4.4 Verify switch configuration

The following sections show commands and output to verify switches are configured and connected correctly. Except where there are fundamental differences, the only output shown is the S5048F-Leaf1 switch. The output from remaining devices is similar.

4.4.1 show vlt brief

The Inter-chassis link (ICL) Link Status, Heart Beat Status, and VLT Peer Status must all be up. The role of one switch in the VLT pair is Primary, and its peer switch (not shown) is assigned the Secondary role. Peer routing is also enabled leaving timers at system defaults.

```
S5048F-Leaf-1#show vlt brief
 VLT Domain Brief
------------------
 Domain ID:                      127
 Role:                           Primary
 Role Priority:                  32768
 ICL Link Status:                Up
 HeartBeat Status:               Up
 VLT Peer Status:                Up
 Local Unit Id:                  0
 Version:                        6(7)
 Local System MAC address:       34:17:eb:37:21:00
 Remote System MAC address:      34:17:eb:37:1c:00
 Remote system version:          6(7)
 Delay-Restore timer:            90 seconds
 Delay-Restore Abort Threshold:  60 seconds
 Peer-Routing :                  Enabled
 Peer-Routing-Timeout timer:     0 seconds
 Multicast peer-routing timeout: 150 seconds
```

4.4.2 show vlt detail

Port channel 101 will be up with active VLAN 1630.

```
S5048F-Leaf-1#show vlt detail
Local LAG Id  Peer LAG Id  Local Status  Peer Status  Active VLANs
------------  -----------  ------------  -----------  -------------
101           101          UP            UP            1, 1630
```

4.4.3 show vlt mismatch

Show VLT mismatch lists VLANs configured on a single switch in the VLT domain. The output shows the two VLANs associated with Data 1 and Data 2 are listed.

```
S5048F-Leaf-1#sh vlt mismatch
Domain
------
Parameters        Local              Peer
----------        -----              ----

Vlan-config
------------
Vlan-ID   Local Mode    Peer Mode
---------   ------------  -----------

    1634             L3            --

    1635             --            L3
```

4.4.4 show vrrp brief

The output from the show vrrp brief command should be similar to that shown below. The priority (Pri column) of the master router in the pair is 254, and the backup router (not shown) is assigned priority 1. Note that some of the descriptions have been truncated to fit the document.

```
S5048F-Leaf1#show vrrp brief
Interface Group   Pri Pre State  Master addr   Virtual addr(s) Description
--------------------------------------------------------------------------------
Vl 1630   IPv4 30 254 Y   Master 172.16.30.251 172.16.30.254  IDRAC_MGMT_IP…
Vl 1631   IPv4 31 254 Y   Master 172.16.31.251 172.16.31.253  ESXI_MGMT_IP
Vl 1633   IPv4 33 254 Y   Master 172.16.33.251 172.16.33.253  SVM_MGMT_IP
```

5 VMware virtual network design

In this section, tables are provided that outline the virtual network design used in this deployment. Specific steps to create the distributed switches, VMkernels, and setting NIC teaming policies are not covered in this document. See *vSphere Networking Guide for vSphere 6.5, ESXi 6.5, and vCenter Server 6.5*.

5.1 ESXi management

The default VMkernel, vmk0 is used for ESXi management and is migrated from the default standard switch to the VDS created in this section. See: *How to migrate service console / VMkernel port from standard switches to VMware vSphere Distributed Switch*.

5.2 Load balancing

In the deployment, two different load balancing algorithms are used. ScaleIO data networks (ScaleIO Data 1 and ScaleIO Data 2) use the route based on originating virtual port. Each port group is assigned a single interface as active while the other interface is unused. This creates a traditional storage topology where each host has two separate networks both logically and physically.

The remaining port groups use Route based on Physical NIC Load. Both uplinks are set as active, and I/O is automatically balanced across both interfaces. The Virtual Distributed Switches (VDS) tests the associated uplinks every 30 seconds, and if their load exceeds 75 percent of usage, the port ID of the virtual machine with the highest I/O is moved to a different uplink.

5.3 Configuration details

The following tables contain the pre and post-installation configuration details for the VDS used for the ScaleIO cluster.

Table 5 Virtual switch details

VDS switch name	Function	Physical NIC port count	MTU
atx01-w01-vds01	• ESXI_MGMT_IP • ESXI_VMOTION_IP • SVM_MGMT_IP • Node_DATA1_IP & SVM_DATA1_IP • Node_DATA2_IP & SVM_DATA2_IP	2	9000

Table 6 Port group configuration settings

Parameter	Settings
Failover Detection	Link status only
Notify switches	Enabled
Failback	Yes

Table 7 Port group settings

VDS	Port group name	Teaming policy	Teaming and Failover	VLAN ID
atx01-w01-vds01	atx01-w01-vds01-management	Route based on physical NIC load	Active: Uplink 1 and Uplink 2	1631
atx01-w01-vds01	atx01-w01-vds01-vmotion	Route based on physical NIC load	Active: Uplink 1 and Uplink 2	1632
atx01-w01-vds01	atx01-w01-vds01-ScaleIO-management	Route based on physical NIC load	Active: Uplink 1 and Uplink 2	1633
atx01-w01-vds01	atx01-w01-vds01-ScaleIO-data01	Route based on originating virtual port	Active: Uplink 1 Unused: Uplink 2	1634
atx01-w01-vds01	atx01-w01-vds01-ScaleIO-data02	Route based on originating virtual port	Active: Uplink 2 Unused: Uplink 1	1635

Table 8 Physical, virtual, and VDS uplink NIC mapping

VDS	Physical NIC	Virtual NIC	Uplink Mapping
atx01-w01-vds01	Mellanox ConnectX-4 LX	vmnic5	Uplink 1
atx01-w01-vds01	Mellanox ConnectX-4 LX	vmnic7	Uplink 2

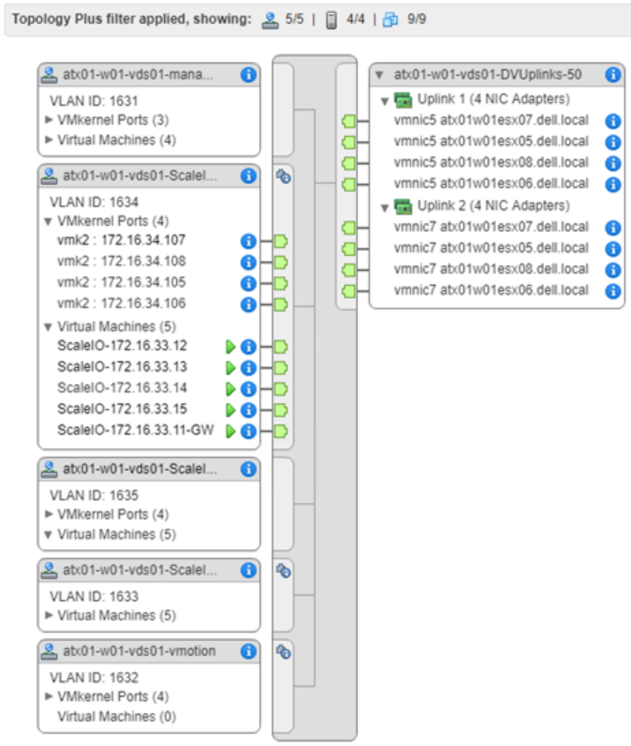
5.4 VMware vSphere VMkernel configuration

The following table contains the configuration details for the ScaleIO VDS with four VMkernel adapters assigned (see Section 6).

Table 9 VMkernel adapter settings

Network label	Connected port group	Enabled services	MTU	Comment
management	atx01-w01-vds01-management	Management traffic	1500	ESXi management
vMotion	atx01-w01-vds01-vmotion	vMotion traffic	9000	Optional for ScaleIO deployment
ScaleIO-data01	atx01-w01-vds01-ScaleIO-data01	SDC	9000	Used by SDC driver
ScaleIO-data02	atx01-w01-vds01-ScaleIO-data02	SDC	9000	Used by SDC driver

Figure 10 is taken from **Home** > **Networking** > **atx01-w01-vds** > **Configure** > **Topology** and shows the completed topology of vDS-ScaleIO showing port groups and VLAN assignments, VMkernels and IP addresses, and physical NIC uplinks. Note that some port groups, like atx01-w01vds-vmotion, have been collapsed for brevity.



Topology Plus filter applied, showing: 5/5 | 4/4 | 9/9

atx01-w01-vds01-mana...
VLAN ID: 1631
▶ VMkernel Ports (3)
▶ Virtual Machines (4)

atx01-w01-vds01-ScaleI...
VLAN ID: 1634
▼ VMkernel Ports (4)
vmk2 : 172.16.34.107
vmk2 : 172.16.34.108
vmk2 : 172.16.34.105
vmk2 : 172.16.34.106
▼ Virtual Machines (5)
ScaleIO-172.16.33.12
ScaleIO-172.16.33.13
ScaleIO-172.16.33.14
ScaleIO-172.16.33.15
ScaleIO-172.16.33.11-GW

atx01-w01-vds01-ScaleI...
VLAN ID: 1635
▶ VMkernel Ports (4)
▼ Virtual Machines (5)

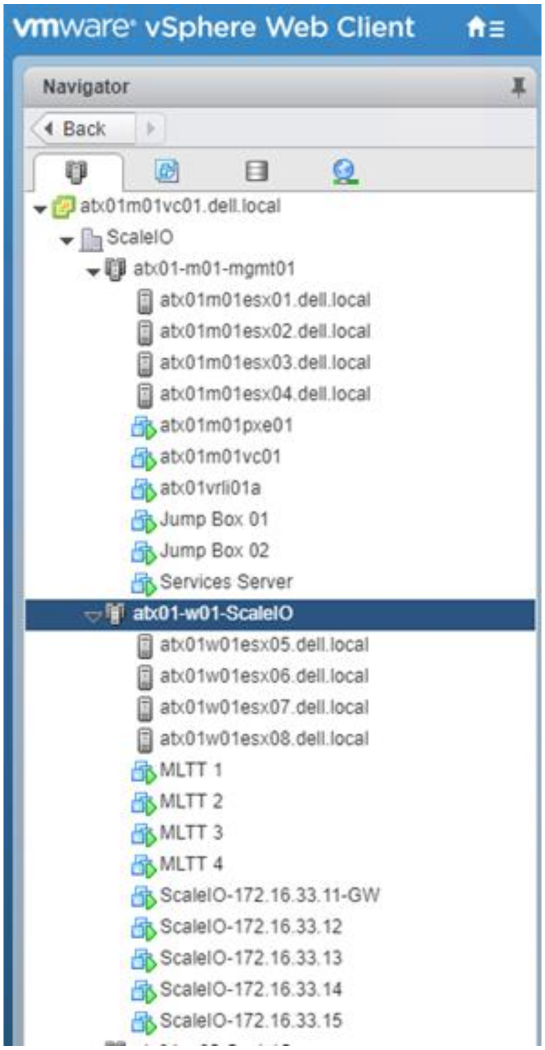
atx01-w01-vds01-ScaleI...
VLAN ID: 1633
▶ Virtual Machines (5)

atx01-w01-vds01-vmotion
VLAN ID: 1632
▶ VMkernel Ports (4)
Virtual Machines (0)

▼ atx01-w01-vds01-DVUplinks-50
▼ Uplink 1 (4 NIC Adapters)
vmnic5 atx01w01esx07.dell.local
vmnic5 atx01w01esx05.dell.local
vmnic5 atx01w01esx08.dell.local
vmnic5 atx01w01esx06.dell.local
▼ Uplink 2 (4 NIC Adapters)
vmnic7 atx01w01esx07.dell.local
vmnic7 atx01w01esx05.dell.local
vmnic7 atx01w01esx08.dell.local
vmnic7 atx01w01esx06.dell.local

Figure 10 VDS atx01-w01-vds01 topology

Figure 11 is taken from **Home** > **Hosts and Clusters** and shows a successfully deployed ScaleIO HCI platform.



- atx01m01vc01.dell.local
  - ScaleIO
    - atx01-m01-mgmt01
      - atx01m01esx01.dell.local
      - atx01m01esx02.dell.local
      - atx01m01esx03.dell.local
      - atx01m01esx04.dell.local
      - atx01m01pxe01
      - atx01m01vc01
      - atx01vrli01a
      - Jump Box 01
      - Jump Box 02
      - Services Server
    - atx01-w01-ScaleIO
      - atx01w01esx05.dell.local
      - atx01w01esx06.dell.local
      - atx01w01esx07.dell.local
      - atx01w01esx08.dell.local
      - MLTT 1
      - MLTT 2
      - MLTT 3
      - MLTT 4
      - ScaleIO-172.16.33.11-GW
      - ScaleIO-172.16.33.12
      - ScaleIO-172.16.33.13
      - ScaleIO-172.16.33.14
      - ScaleIO-172.16.33.15

Figure 11 vCenter host and clusters

6 Deploying Dell EMC ScaleIO

Deploying ScaleIO in this environment consists of the following topics:

- Register the ScaleIO plug-in
- Upload the ScaleIO Open Virtual Appliance (OVA) template
- Deploy ScaleIO

This section does not contain step-by-step instructions for deploying ScaleIO. For a detailed step-by-step guide, see the *ScaleIO IP Fabric Best Practice and Deployment Guide*.

6.1 Deploy the ScaleIO plug-in

All physical connectivity and configuration of the VMware vSphere distributed switches are complete. The ScaleIO plug-in for vSphere simplifies the installation and management of the ScaleIO system in a vSphere environment.

Use the following parameters during the installation:

Table 10 Dell EMC ScaleIO VMware vSphere plug-in parameters

Parameter	**Setting**
vCenter Server	atx01m01vc01.dell.local
Registration mode	Standard

6.2 Upload the ScaleIO OVA template

Once the VMware vCenter Dell EMC ScaleIO plug-in installation is completed, upload the ScaleIO virtual machine OVA to an R730xd local data store. The OVA serves as a virtual machine template for deploying all the software components of ScaleIO. From the VMware PowerCLI program, select Create SVM Template. While it is possible to specify separate datastores to correspond to each ScaleIO host, this deployment uses a single datastore for all hosts. VMware vSphere vMotion is used to copy the template to each remaining host during the ScaleIO implementation.

During the installation process, use the parameters shown in the following table:

Table 11 ScaleIO Vmware vSphere plug-in parameters

Parameter	**Setting**
vCenter Server	atx01m01vc01.dell.local
Data center name	ScaleIO
Path to OVA	local OVA path including file name
Datastore name	atx01w01esx05-lds01

6.3 Deploy ScaleIO

This section describes how the deployment wizard is used in the deployment example provided in this guide.

ScaleIO deployment has four steps:

- SDC deployment and configuration
- ScaleIO advanced configuration settings
- Deploy the ScaleIO environment
- Install the ScaleIO GUI (optional)

Before an ESXi host can consume the virtual SAN, the SDC kernel driver must be installed on each ESXi host, regardless of the role that host is playing. The process outlined below installs the SDC driver on the target host.

To start the installation wizard, perform the following steps:

1. From the **Basic tasks** section of the **EMC ScaleIO** screen, click **Install SDC on ESX**.
2. Select all hosts under the ScaleIO data center as targets for the installation.
3. Once complete, reboot all hosts before continuing with the deployment.

To deploy ScaleIO, perform the following steps:

1. From the Basic tasks section of the EMC ScaleIO screen, click **Deploy ScaleIO environment**.
2. Using the following table, assign the settings listed to the parameters provided.

Note: The parameters and settings provided in the table address the selections necessary through step 4 of the installation wizard. A setting that is not listed indicates that the default setting has been applied.

Table 12 ScaleIO Wizard deployment settings

Parameter	Setting
Select installation	Create a new system
System name	SIO01
Admin password	ScaleIO Admin Password
vCenter server	atx01m01vc01.dell.local
Host selection	atx01w01esx05, atx01w01esx06, atx01w01esx07, atx01w01esx08
ScaleIO components	3-node mode
Initial Master MDM	atx01w01esx05
Manager MDM	atx01w01esx06
TieBreaker MDM	atx01w01esx07
DNS Server 1	172.16.11.4
DNS Server 2	172.16.11.5

3. Using the following table, select the ScaleIO wizard parameter settings for steps 5 through 7.

Table 13 ScaleIO Wizard deployment settings

Parameter	**Setting**
Protection domain name	PD01
RAM read cache size per SDS	1,024 MB
Storage Pools	SSD01
Enable zero padding	True
SDS host selection	atx01w01esx05, atx01w01esx06, atx01w01esx07, atx01w01esx08
Selected devices	All empty device categorized into the appropriate storage pool.
SDC host selection	atx01w01esx05, atx01w01esx06, atx01w01esx07, atx01w01esx08
Enable/Disable SCSI LUN	Enable

4. Using the following tables, use the appropriate values to complete the wizard setup.

Table 14 ScaleIO Wizard deployment settings

Parameter	Setting
Host for ScaleIO gateway	atx01w01esx08
Gateway admin password	ScaleIO Admin Password
Gateway LIA password	ScaleIO Admin Password
Select OVA template	EMC ScaleIO SVM Template (v2.0.140000.228) 1
OVA root password	ScaleIO Admin Password
OVA LIA password	ScaleIO Admin Password
Management network label	atx01-w01-vds01-ScaleIO-management
Data network label	atx01-w01-vds01-ScaleIO-data01
2nd data network label	atx01-w01-vds01-ScaleIO-data02

Table 15 ScaleIO networking addressing

ESXi name	Management IP	Default gateway	Data 1 IP	Data 2 IP
atx01w01esx08 (ScaleIO Gateway)	172.16.33.11/24	172.16.33.253	172.16.34.11/24	172.16.35.11/24
atx01w01esx05 (Master MDM)	172.16.33.12/24	172.16.33.253	172.16.34.12/24	172.16.35.12/24
atx01w01esx06 (Slave 1 MDM)	172.16.33.13/24	172.16.33.253	172.16.34.13/24	172.16.35.13/24
atx01w01esx07 (TieBreaker 1)	172.16.33.14/24	172.16.33.253	172.16.34.14/24	172.16.35.14/24
atx01w01esx08	172.16.33.15/24	172.16.33.253	172.16.34.15/24	172.16.35.15/24

A virtual IP is assigned which is used for communications between the MDM cluster and the SDCs. Only one virtual IP address is mapped to each NIC, with a maximum of four virtual IPs per system. This virtual IP is mapped to the manager MDM dynamically and is moved if the primary MDM is under maintenance.

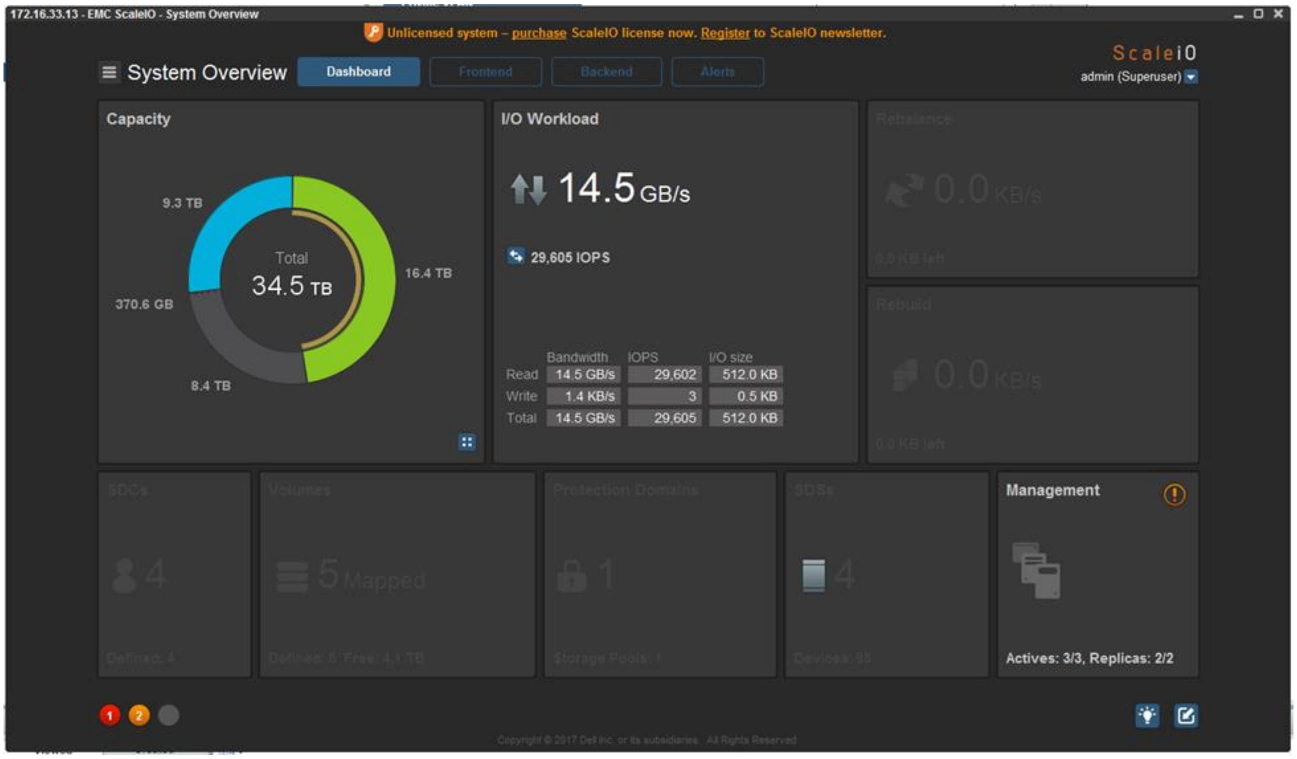
Table 16 ScaleIO networking virtual IP addresses

Parameter	Setting
Data (atx01-w01-vds01-ScaleIO-data01)	172.16.34.4
2nd Data (atx01-w01-vds01-ScaleIO-data02)	172.16.35.4

Once the summary screen displays, the deployment begins. At this point, the installation wizard stops on any errors allowing issues to be resolved and the deployment can be continued.

6.4 ScaleIO GUI

The ScaleIO graphical user interface (GUI) can be installed on a management workstation to provide an easy way to monitor and configure the ScaleIO system. Once installed, the virtual IP assigned to Data1 (172.16.34.4) can be used to access the ScaleIO GUI. The installation file is part of the ScaleIO for Windows download. The ScaleIO cluster created is shown below in Figure 12.



172.16.33.13 - EMC ScaleIO - System Overview

Unlicensed system – purchase ScaleIO license now. Register to ScaleIO newsletter.

	Bandwidth	IOPS	I/O size
Read	14.5 GB/s	29,602	512.0 KB
Write	1.4 KB/s	3	0.5 KB
Total	14.5 GB/s	29,605	512.0 KB

Figure 12 ScaleIO under load showing throughput

7 Best practices

The post-installation information provided in this section consists of the following:

- Increase the Maximum Transmission Unit (MTU) for VMware vSphere, and Dell EMC ScaleIO
- Configure Quality of Service using Differentiated Services (DiffServ)

For additional performance tuning, including ESXi hosts and ScaleIO VMs, see the *ScaleIO v2.0.x Performance Fine-Tuning Technical Notes Guide*.

7.1 Maximum Transmission Unit size

In this environment, the distributed switch is assigned an MTU value of 9000. Also, any storage-related interface/port group has an MTU value of 9000. The following table summarizes the port groups that have an MTU value of 9000:

Table 17 VDS port groups with modified MTU values

VDS switch name	Network label	Connected port groups	MTU
atx01-w01-vds01	vMotion	atx01-w01-vds01-vMotion	9000
atx01-w01-vds01	ScaleIO-data01	atx01-w01-vds01- ScaleIO-data01	9000
atx01-w01-vds01	ScaleIO-data02	atx01-w01-vds01- ScaleIO-data02	9000

To verify that jumbo frames are working the ESXi CLI tool **vmkping** is used. After establishing an SSH connection with atx01w01esx01, a non-defragment capable ping with an MTU value of 8972 is sent from the host using the data01 VMkernel adapter to atx01w01esx02.dell.local.

Note: The maximum frame size that vmkping can send is 8972 due to IP (20 bytes) and ICMP (8 bytes) overhead.

```
[root@atx01w01esx01:~] vmkping -d -s 8972 –I vmk2 172.16.34.106
PING 172.16.34.102 (172.16.34.102): 8972 data bytes
8980 bytes from 172.16.34.106: icmp_seq=0 ttl=64 time=0.360 ms
8980 bytes from 172.16.34.106: icmp_seq=1 ttl=64 time=0.373 ms
8980 bytes from 172.16.34.106: icmp_seq=2 ttl=64 time=0.451 ms
```

To enable jumbo frames for the SVM, perform the following steps:

1. Run the `ifconfig` command to get the NIC information. The following is an example from an SVM deployed in this solution, ScaleIO-172.16.33.12:

```
ScaleIO-172-16-33-12:~ # ifconfig
eth0   Link encap:Ethernet  HWaddr 00:50:56:B7:81:28
       inet addr:172.16.33.12  Bcast:172.16.33.255  Mask:255.255.255.0
       inet6 addr: fe80::250:56ff:feb7:8128/64 Scope:Link
       UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
       RX packets:438 errors:0 dropped:0 overruns:0 frame:0
       TX packets:118 errors:0 dropped:0 overruns:0 carrier:0
       collisions:0 txqueuelen:1000
       RX bytes:32011 (31.2 Kb)  TX bytes:14726 (14.3 Kb)

eth1   Link encap:Ethernet  HWaddr 00:50:56:B7:A1:31
       inet addr:172.16.34.12  Bcast:172.16.34.255  Mask:255.255.255.0
       inet6 addr: fe80::250:56ff:feb7:a131/64 Scope:Link
       UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
       RX packets:44348476 errors:0 dropped:1 overruns:0 frame:0
       TX packets:20392679 errors:0 dropped:0 overruns:0 carrier:0
       collisions:0 txqueuelen:1000
       RX bytes:150867588100 (143878.5 Mb)  TX bytes:187404794725 (178723.1 Mb)

eth2   Link encap:Ethernet  HWaddr 00:50:56:B7:C4:C8
       inet addr:172.16.35.12  Bcast:172.16.35.255  Mask:255.255.255.0
       inet6 addr: fe80::250:56ff:feb7:c4c8/64 Scope:Link
       UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
       RX packets:44461343 errors:0 dropped:1 overruns:0 frame:0
       TX packets:20318888 errors:0 dropped:0 overruns:0 carrier:0
       collisions:0 txqueuelen:1000
       RX bytes:159199012195 (151824.0 Mb)  TX bytes:181980663115 (173550.2 Mb)
```

In this example, eth1 and eth2 correspond to ScaleIO Data Network 1 (Subnet 172.16.34.0/24, VLAN 1634) and ScaleIO Data Network 2 (Subnet 172.16.35.0/24, VLAN 1635). Administrative access uses Eth0.

Using the interface name, edit the appropriate network configuration files, and append MTU value of 9000 to the end of the configuration. The following is an example for interface eth1:

```
ScaleIO-172-16-33-12:~ # vi /etc/sysconfig/network/ifcfg-eth1
DEVICE=eth1
STARTMODE=onboot
USERCONTROL=no
BOOTPROTO=static
NETMASK=255.255.255.0
IPADDR=172.16.34.12
MTU=9000
```

2. Save the file (**:qw** [ENTER]) then enter the following command to restart the network services for the virtual machine:

```
ScaleIO-172-16-33-12:~ # service network restart
Shutting down network interfaces:
    eth0      device: VMware VMXNET3 Ethernet Controller             done
    eth1      device: VMware VMXNET3 Ethernet Controller             done
    Eth2      device: VMware VMXNET3 Ethernet Controller             done
Shutting down service network  .  .  .  .  .  .  .  .  .             done
Hint: you may set mandatory devices in /etc/sysconfig/network/config
Setting up network interfaces:
    eth0      device: VMware VMXNET3 Ethernet Controller
    eth0      IP address: 172.16.33.12/24                            done
    eth1      device: VMware VMXNET3 Ethernet Controller
    eth1      IP address: 172.16.34.12/24                            done
    eth2      device: VMware VMXNET3 Ethernet Controller
    eth2      IP address: 172.16.35.12/24                            done
Setting up service network  .  .  .  .  .  .  .  .  .  .             done
```

3. Use the `ping` command to validate jumbo frames connectivity to another, already-configured, SVM:

```
ScaleIO-172-16-33-12:~ # ping -M do -s 8972 172.16.31.13
PING 172.16.34.13 (172.16.34.13) 8972(9000) bytes of data.
8980 bytes from 172.16.34.13: icmp_seq=1 ttl=64 time=0.393 ms
8980 bytes from 172.16.34.13: icmp_seq=2 ttl=64 time=0.398 ms
8980 bytes from 172.16.34.13: icmp_seq=3 ttl=64 time=0.366 ms
```

7.2 Configure Quality of Service using Differentiated Services (DiffServ)

The basis of Quality of Service (QoS) is traffic differentiation. In this guide, different types of traffic are classified accordingly and then are given priorities throughout the network using DiffServ. The priorities are marked using Differentiated Services Code Point (DSCP). Once marking is complete, QoS tools are applied to each specific traffic type to affect traffic behavior. DiffServ operates independently on each router in a standalone model, meaning that each router can use the classification to affect traffic behavior differently.

Traffic is separated into two broad categories and marked using DSCP markings based on what the network needs to provide for that type of traffic:

1. Traffic with higher priority requirements is given a DSCP mark of 46
2. Traffic with lower priority requirements are not marked

ScaleIO has native QoS capabilities that are not demonstrated in this paper. For instance, the amount of traffic the SDC generates can be limited per volume with a high level of granular control. For more information, see the *EMC ScaleIO 2.0.X Deployment Guide*.

A DSCP value of 46 corresponds to expedited forwarding and maps to queue 5 on the S5048F-ON switch. These actions ensure that the switch prioritizes this type of traffic above unmarked traffic that uses the default queue 0. In this deployment guide, both the physical and virtual networks are configured to classify traffic. The VDS is configured to initially insert the DSCP value while the physical switches are configured to trust the default DSCP value mapping.

In the switch configuration section, a policy map is created and instructs both switches to trust the DSCP value mapping. The configuration below shows the commands set to trust DSCP value mapping for the S5048F-ON switch. Configuration for the second leaf switch is identical.

```
policy-map-input TrustDSCPin
trust diffserv

interface range tw1/1,tw1/3,tw1/5,tw1/7
description ScaleIO nodes
service-policy input TrustDSCPin
```

DSCP values are inserted on a port-group basis. In the table below two port groups are enabled to filter traffic, atx01-w01-vds01-management, and atx01-w01-vds01-ScaleIO-management. To create marking navigate to **Home** > **Networking** > **port group** > **Edit** > **Traffic filtering and marking**. Table 18 shows the values used.

Table 18 Atx01-w01-vds01 port group DSCP values

VDS switch name	Port group names	Traffic filtering and markings	DSCP Values	Protocol/Traffic Type
atx01-w01-vds01	atx01-w01-vds01-management	Enable	46	Management
atx01-w01-vds01	atx01-w01-vds01-vMotion	Disable	n/a	n/a
atx01-w01-vds01	atx01-w01-vds01-ScaleIO-management	Enable	46	Virtual Machines
atx01-w01-vds01	atx01-w01-vds01- ScaleIO-data01	Disable	n/a	n/a
atx01-w01-vds01	atx01-w01-vds01- ScaleIO-data02	Disable	n/a	n/a

A Troubleshooting SDS connectivity

SDS connectivity problems affect ScaleIO performance. ScaleIO has a built-in tool to verify all SDS nodes in a given protection domain have connectivity. From the ScaleIO Command Line Interface (SCLI), run the ScaleIO internal network test to verify the network speed between all the SDS nodes in the Protection Domain.

The command below tests all SDS nodes with a payload of 10 GB, using 8 parallel threads.

```
ScaleIO-172-16-33-12:~ # scli --mdm_ip 172.16.33.13 --start_all_sds_network_test
--protection_domain_name PD01 --parallel_messages 8 --network_test_size 10
Started test on atx01w01esx08.dell.local-ESX 172.16.34.15 at 11:58:59..........
Test finished.
Started test on atx01w01esx05.dell.local-ESX 172.16.34.12 at 11:59:10.........
Test finished.
Started test on atx01w01esx07.dell.local-ESX 172.16.34.14 at 11:59:20.........
Test finished.
Started test on atx01w01esx06.dell.local-ESX 172.16.34.13 at 11:59:30.........
Test finished.
Protection Domain PD01  contains 4 SDSs

ScaleIO-172-16-33-12:~ # scli --mdm_ip 172.16.33.13 --
query_all_sds_network_test_results --protection_domain_name PD01
Protection Domain PD01
Connection: atx01w01esx08.dell.local-ESX 172.16.34.15 <--->
atx01w01esx07.dell.local-ESX 172.16.34.14   --> 2.9 GB (2934 MB) per-second <--
3.1 GB (3160 MB) per-second
Connection: atx01w01esx08.dell.local-ESX 172.16.34.15 <--->
atx01w01esx05.dell.local-ESX 172.16.34.12   --> 2.9 GB (2968 MB) per-second <--
3.1 GB (3190 MB) per-second
Connection: atx01w01esx08.dell.local-ESX 172.16.34.15 <--->
atx01w01esx06.dell.local-ESX 172.16.34.13   --> 3.1 GB (3220 MB) per-second <--
3.1 GB (3150 MB) per-second
Connection: atx01w01esx05.dell.local-ESX 172.16.34.12 <--->
atx01w01esx06.dell.local-ESX 172.16.34.13   --> 3.3 GB (3346 MB) per-second <--
3.1 GB (3160 MB) per-second
Connection: atx01w01esx07.dell.local-ESX 172.16.34.14 <--->
atx01w01esx06.dell.local-ESX 172.16.34.13   --> 3.1 GB (3180 MB) per-second <--
3.2 GB (3282 MB) per-second
Connection: atx01w01esx05.dell.local-ESX 172.16.34.12 <--->
atx01w01esx07.dell.local-ESX 172.16.34.14   --> 3.3 GB (3424 MB) per-second <--
3.2 GB (3250 MB) per-second
```

B Routing ScaleIO Virtual Machine traffic

In this section a possible solution to solve routing between SVMs in separate subnets is outlined. Each SVM contains three virtual NICs:

- Eth0 for ScaleIO management
- Eth1 for ScaleIO Data01
- Eth2 for ScaleIO Data02

The SVM uses a single TCP/IP stack, and any unknown networks are limited to this single default gateway. If ScaleIO Data01 or ScaleIO Data02 needs to reach an SVM in another subnet, for instance in another rack in the data center, this traffic fails. Each SVM in the environment can be updated to leverage policy-based routing (PBR).

1. To show the routing table of an SVM issue the command `ip route`.

```
ScaleIO-172-16-33-12:~ # ip route
default via 172.16.33.253 dev eth0
127.0.0.0/8 dev lo  scope link
169.254.0.0/16 dev eth0  scope link
172.16.33.0/24 dev eth0  proto kernel  scope link  src 172.16.33.12
172.16.34.0/24 dev eth1  proto kernel  scope link  src 172.16.34.12
172.16.35.0/24 dev eth2  proto kernel  scope link  src 172.16.35.12
```

2. When attempting to ping the SVM from another network, the ping fails due to asymmetric routing. The ping reaches the host as shown by TCP dump.

```
ScaleIO-172-16-33-12:~ # tcpdump -n -i eth1 icmp
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on eth1, link-type EN10MB (Ethernet), capture size 96 bytes
22:43:40.067441 IP 172.18.31.101 > 172.16.34.12: ICMP echo request, id 56600,
seq 0, length 64
22:43:41.067630 IP 172.18.31.101 > 172.16.34.12: ICMP echo request, id 56600,
seq 1, length 64
22:43:42.069783 IP 172.18.31.101 > 172.16.34.12: ICMP echo request, id 56600,
seq 2, length 64
```

3. However, due to the routing table, the response is sent back out eth0 resulting in a failed connection. To address this problem, create two routing tables for eth1 and eth2 and verify that they have been created.

```
ScaleIO-172-16-33-12:~ # echo '100 eth1' >> /etc/iproute2/rt_tables && echo '200
eth2' >> /etc/iproute2/rt_tables && cat /etc/iproute2/rt_tables

# reserved values
#
255     local
```

```
254     main
253     default
0       unspec
#
# local
#
#1      inr.ruhep
100    eth1
200    eth2
```

4. For each interface enable PBR. Below is an example for eth1 and needs to be repeated for eth2.

```
ScaleIO-172-16-33-12:~ # ip route flush table eth1
ScaleIO-172-16-33-12:~ # ip route add 172.16.34.0/24 dev eth1 proto kernel scope
link table eth1
ScaleIO-172-16-33-12:~ # ip route add default via 172.16.34.253 dev eth1 table
eth1
ScaleIO-172-16-33-12:~ # ip rule add from 172.16.34.0/24 lookup eth1
```

5. Verify the route tables and ip rules to ensure a new default route for each interface has been created.

```
ScaleIO-172-16-33-12:~ # ip route list table eth1
default via 172.16.34.253 dev eth1
172.16.34.0/24 dev eth1  proto kernel  scope link

ScaleIO-172-16-33-12:~ # ip rule
0:      from all lookup local
32763:  from 172.16.35.0/24 lookup eth2
32764:  from 172.16.34.0/24 lookup eth1
32765:  from all lookup main
32766:  from all lookup default
```

6. Finally, validate the solution is working by repeating the ping test.

```
ScaleIO-172-16-33-12:~ # tcpdump -n -i eth2 icmp

tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on eth2, link-type EN10MB (Ethernet), capture size 96 bytes
15:18:25.730704 IP 172.18.35.101 > 172.16.35.12: ICMP echo request, id 36916,
seq 674, length 64
15:18:25.730709 IP 172.16.35.12 > 172.18.35.101: ICMP echo reply, id 36916, seq
674, length 64
15:18:26.733704 IP 172.18.35.101 > 172.16.35.12: ICMP echo request, id 36916,
seq 675, length 64
15:18:26.733709 IP 172.16.35.12 > 172.18.35.101: ICMP echo reply, id 36916, seq
675, length 64
15:18:27.736662 IP 172.18.35.101 > 172.16.35.12: ICMP echo request, id 36916,
seq 676, length 64
```

```
15:18:27.736668 IP 172.16.35.12 > 172.18.35.101: ICMP echo reply, id 36916, seq
676, length 64
15:18:28.739532 IP 172.18.35.101 > 172.16.35.12: ICMP echo request, id 36916,
seq 677, length 64
15:18:28.739537 IP 172.16.35.12 > 172.18.35.101: ICMP echo reply, id 36916, seq
677, length 64
^C
8 packets captured
8 packets received by filter
0 packets dropped by kernel
```

Note: This solution is not persistent across reboots and needs to be automated. That effort is left to the reader.

C Validated hardware and components

The following tables list the hardware and components used to configure and validate the example configurations in this guide.

C.1 Dell EMC Networking switches

Qty	Item	OS/Firmware version
1	S3048-ON - Management switch	OS: DNOS 9.13.0.0
		System CPLD: 9
		Module CPLD: 7
2	S5048F-ON - Leaf switch	OS: DNOS 9.12(1.0)
		CPLD: 1.0

C.2 Dell EMC PowerEdge 730xd servers

Qty per server	Item	Firmware version
2	Intel(R) Xeon(R) CPU E5-2698 v4@2.2.20GHz, 20 cores	-
512	GB RAM	-
24	400GB SAS SSD	-
1	Dell PERC H730 Mini	25.5.2.0001
1	Mellanox ConnectX-4 LX 25GbE SFP Adapters	14.17.20.52
2	Mellanox ConnectX-4 LX 25GbE DP	14.17.20.52
-	BIOS	2.4.3
-	iDRAC with Lifecycle Controller	2.41.41.4.1

D Validated software

The Software table lists the software components used to validate the example configurations in this guide.

Item	Version
Dell EMC ScaleIO	2.0.-14000.228
VMware vSphere Power CLI	6.5.0
ScaleIO vSphere Plug-in Installer	2.0.1.4
SVM OVA	2.0.1.4000.288.ova
VMware ESXi	6.5 U1 - Dell EMC customized image build 6765664
VMware vCenter Server Appliance	6.5 build 5973321

E Product manuals and technical guides

E.1 Dell EMC

Dell EMC TechCenter - An online technical community where IT professionals have access to numerous resources for Dell EMC software, hardware, and services.

Dell EMC TechCenter Networking Guides

Manuals and documentation for Dell Networking S5048F-ON

Manuals and documentation for Dell Networking S3048-ON

Manuals and documentation for PowerEdge R730xd

Dell EMC ScaleIO Software Only: Documentation Library – This link requires a username and password to access. Please contact your sales representative for login credentials.

E.2 VMware

VMware vSphere Documentation

vSphere Installation and Setup – This document includes ESXi 6.5 and vCenter Server 6.5

VMware Compatibility Guide

Dell EMC ScaleIO Software Defined Storage with ESXi 5.5, ESXi 6.0, and ESXi 6.5 (2146203)

F 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 to provide feedback on the quality and usefulness of this publication by sending an email to Dell_Networking_Solutions@Dell.com.