VxFlex Network Deployment Guide using Dell EMC Networking 25GbE switches and OS10EE

A VxFlex Ready Node deployment guide using Dell EMC Networking S5200-ON switches

Abstract

The Dell EMC Networking S5248F-ON is the latest in the S-series of 25GbE switches that provide the bandwidth and low latency support for a scalable storage architecture. This document details the deployment of the Dell EMC VxFlex Ready Node solution using these Dell EMC Networking 25GbE switches.

March 2019

Revisions

Date	Description
March 2019	Initial release

Table of contents

Executive summary

Dell EMC VxFlex is an industry leading software-defined storage (SDS) solution that enables customers to extend their existing virtual infrastructure into a high performing virtual SAN. VxFlex creates a virtual server SAN using industry-standard servers with direct attached storage (DAS). You can deploy VxFlex using as few as three hosts, and up to 1024 hosts. Each host can use storage media such as flash-based SSDs, NVMe SSDs, traditional spinning disks, or a mix.

The Dell EMC Networking S5248F-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 provides two examples that detail the deployment of the Dell EMC VxFlex solution using Dell EMC Networking 25GbE switches. This document:

- Assists administrators in selecting the best hardware and topology for their Dell EMC VxFlex network
- Delivers detailed instructions and working examples for the deployment and configuration of Dell EMC Networking S-series switches using OS10 Enterprise Edition (OS10EE)
- Delivers instructions and working examples for the deployment and configuration of a sample Dell EMC VxFlex virtual SAN
- Shows conceptual, physical, and logical diagram examples for various networking topologies

1 Introduction

VxFlex 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. VxFlex uses the existing local storage devices and turns them into shared block storage. For many workloads, VxFlex storage is comparable to, or better than, external shared block storage.

The lightweight VxFlex software components are installed on the application servers and behave like a software initiator. A standard Ethernet network provides a communication channel for I/O requests sent to VxFlex block volumes. An 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.

VxFlex is designed and implemented with enterprise-grade resilience. 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 Networking offers the S5248F-ON, 25/100GbE Ethernet switch, designed to be used as a data center Leaf or Top of Rack (ToR) switch.

The example topology that is illustrated in this guide uses VxFlex Ready Nodes (formerly known as ScaleIO Ready Nodes). Dell EMC VxFlex Ready Nodes are a combination of VxFlex software-defined block storage and Dell PowerEdge servers, optimized to run the VxFlex OS, enabling the deployment of an entirely architected, software-defined, scale-out server SAN. This guide covers deploying the VxFlex solution on VMware vSphere ESXi, creating a Hyper-Converged Infrastructure (HCI) by enabling each ESXi host to present computing as well as storage resources.

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

Note: For deploying a spine-leaf architecture using Dell EMC Networking, see *Dell EMC Networking Layer 3 Leaf-Spine Deployment and Best Practices with OS10*, or use the Dell EMC Fabric Design Center discussed on page 43. For steps on deploying and configuring VMware vSphere, see the *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 that is entered in the GUI

1.2 Attachments

This document in .pdf format includes one or more file attachments. To access attachments in Adobe Acrobat Reader, click the icon in the left pane halfway down the page, then click the icon.

1.3 Dell EMC VxFlex

A VxFlex virtual SAN consists of the following software components:

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

In this environment, the VxFlex/ScaleIO Virtual Machine (SVM) hosts the MDM and SDS roles. Each VxFlex node, running ESXi, has a separate SVM. The following illustration shows the communication between these components:

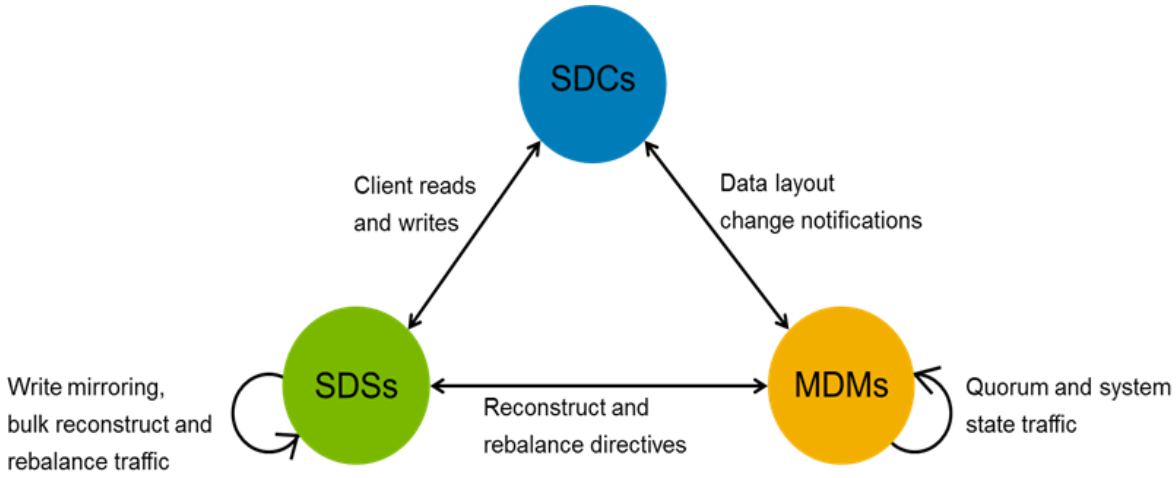

Figure 1 VxFlex component communication

2 Hardware overview

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

Steps in this document were validated using the specified Dell EMC Networking switches and OS10EE but may be used for other Dell EMC Networking switch models that use the same networking operating system or later assuming the switch has the available port numbers, speeds, and types. For example, the OS10EE commands in this document also work with 10GbE leaf switches like the Dell EMC S4128F-ON, with minor adjustments to the commands given in the examples.

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.

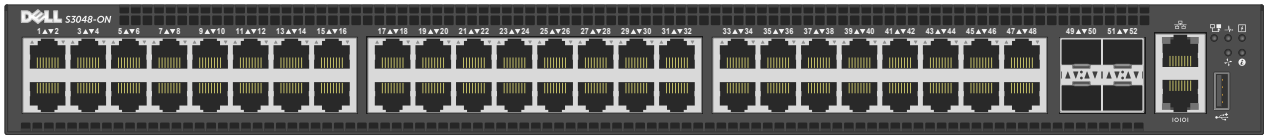

Figure 2 Dell EMC Networking S3048-ON

2.2 Dell EMC Networking S5248F-ON

The Dell EMC Networking S5248F-ON is a 1-RU switch with forty-eight 25GbE SFP28 ports, four 100GbE QSFP28 ports, and two 200GbE QSFP28-DD ports. The example in this guide deploys two of these switches as leaf switches running OS10EE, performing basic gateway functionality for attached VxFlex hosts.

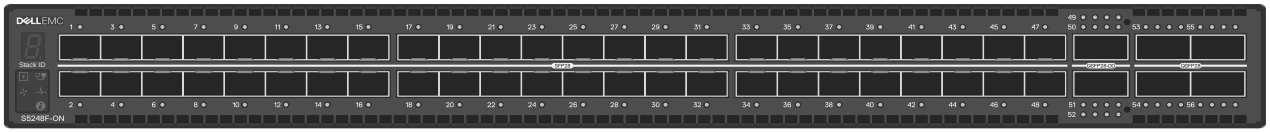

Figure 3 Dell EMC Networking S5248F-ON

2.3 Dell EMC VxFlex Ready Node (R740xd)

Dell EMC VxFlex Ready Nodes are pre-configured with certified components that simplify ordering and reduce deployment risks. The VxFlex Ready Node R740xd 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 Ready Node R740xd servers are used in the VxFlex cluster.

Figure 4 Dell EMC VxFlex Ready Node (R740xd)

3 VxFlex networking overview

The primary purpose of this guide is to provide a step-by-step example for configuring the network for VxFlex using OS10EE. Chapter 4 provides instructions for configuring Dell EMC Networking S5248F-ON 25GbE switches, running OS10EE. Two S5248F-ON switches are used as ToR/leaf switches to connect Dell EMC VxFlex Ready Nodes (based on R740xd servers) for VxFlex installation and upstream connectivity.

In VxFlex, internode communication, which is used for managing data locations, rebuilding, 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 used, VxFlex supports VLANs. Management may be done on a separate network with access to the other VxFlex components, or on the same network.

Each VxFlex node in this guide has four 25GbE ports that are provided by two Mellanox Connect X-4 LX PCIe cards. Two of the available four ports are used to carry all traffic types (front end and back end) in this deployment and Quality of Service (QoS) is used to ensure that traffic requiring low-latency is prioritized (see Section 7.2). Since VxFlex may provide compute resources, the remaining free ports could be leveraged for compute workloads, depending on requirements.

Note: For configuring previous generation switches using version OS9 firmware, see *VxFlex Network Deployment Guide using Dell EMC Networking 25GbE switches and OS9*. Table 1 provides a breakdown and where to find each example.

Table 1 Configuration examples

DNOS	Dell EMC ToR Switch	VxFlex nodes	Example
OS10EE	S5248F-ON (25GbE)	R740xd (VxFlex Ready Nodes)	This document
OS9	S5048F-ON (25GbE)	R730xd (servers)	OS9 deployment guide

3.1 Network topologies

This section provides an overview of the production network, management network, and physical connections used in the deployment example in this guide.

3.1.1 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 VxFlex deployment and optimize the additional traffic generated in the HCI environment. Figure 5 shows the example topology using Dell EMC VxFlex Ready Nodes and OS10EE switch firmware running on Dell EMC Networking S5248F-ON leaf switches. The VxFlex components, MDM, SDS, and SDC, reside on the hosts while ESXi is managed through vCenter in the management environment. A typical leaf-spine topology is shown to demonstrate access to the existing infrastructure found in a typical data center. This configuration demonstrates a single rack deployment. Larger scale deployments may require separate storage data and VM production networks. Configuring the leaf-spine or existing production network is beyond the scope of this guide.

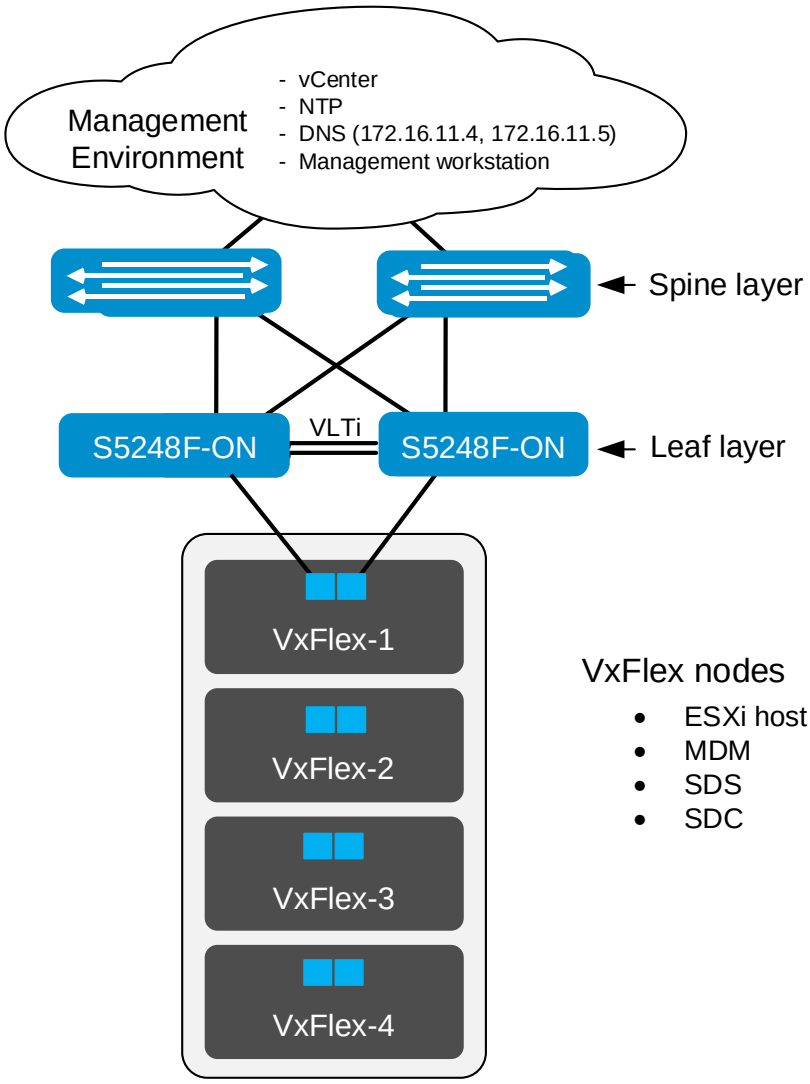

Figure 5 Production network

3.1.2 Management network

A single S3048-ON switch provides iDRAC connectivity to the VxFlex nodes. Figure 6 shows the S3048-ON is connected to the leaf switches through the OOB port on each leaf switch. In this guide, OS10EE is used on the S3048-ON management switch.

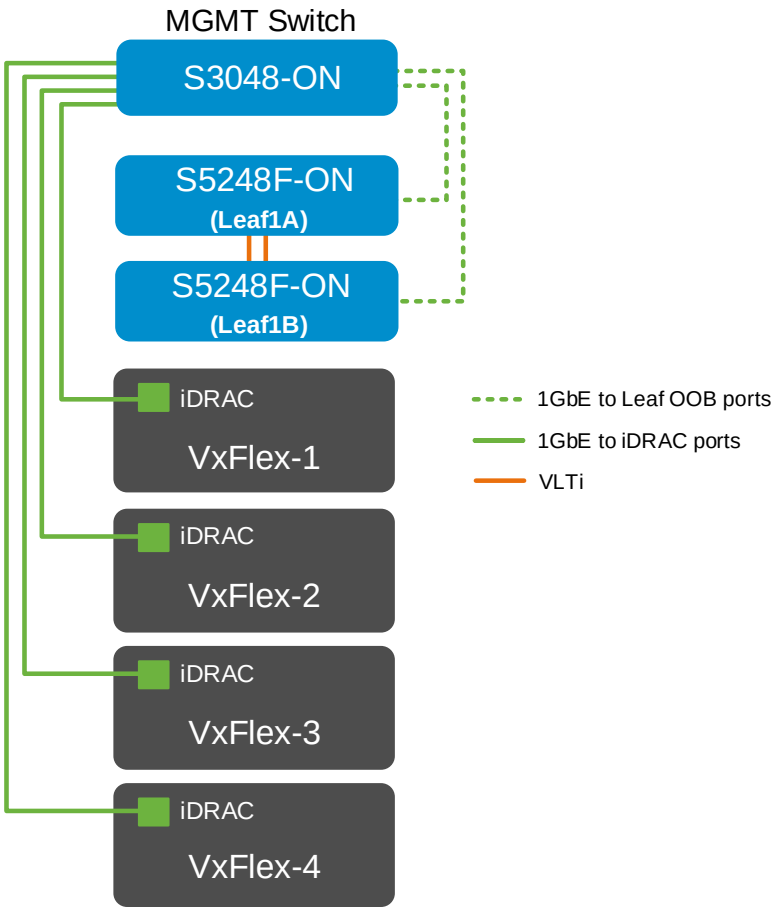

Figure 6 Management network for a single rack

3.2 Network connectivity

Figure 7 shows one VxFlex node (VxFlex-1) connected to two leaf switches using two Mellanox ConnectX-4 LX PCIe cards that are installed in PCIe slots 1 and 2. The leaf switches are Virtual Link Trunking (VLT) peers. One port from each PCIe card connects to each leaf switch. The connections for VxFlex-2 through VxFlex-4 (not shown) are done in the same manner.

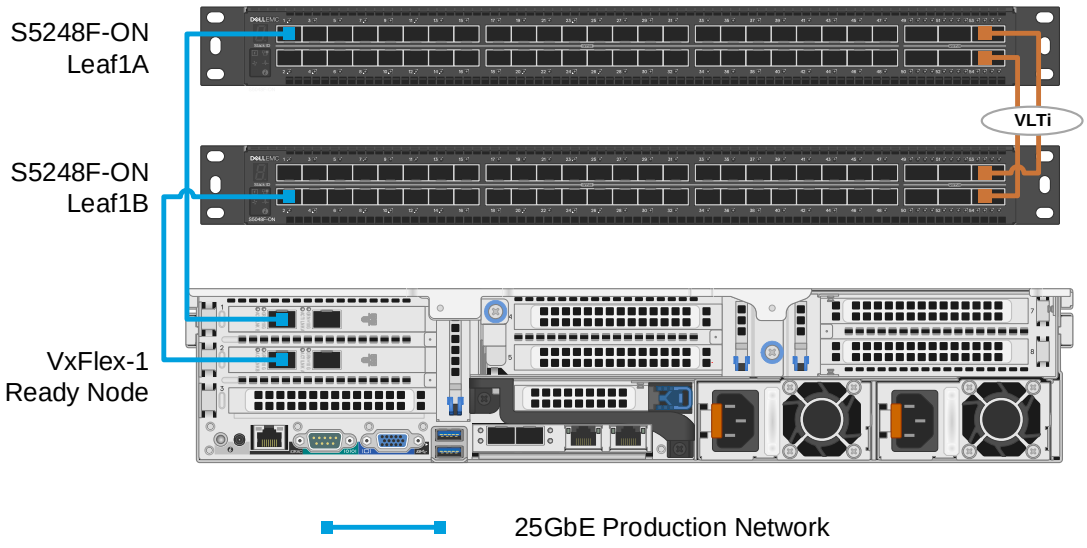

Figure 7 VxFlex-1 wiring to production network

3.1 Connections to OOB management switch

The OOB management network is an isolated network for remote management of servers, switches, and other devices. It is also used to carry heartbeat messages sent between leaf switches configured as VLT peers.

Dell EMC recommends using at least one S3048-ON switch per 42-RU rack for OOB management network connections. Each S3048-ON has forty-eight 1GbE Base-T ports for connections to server iDRACs and leaf switch management ports as shown in Figure 7. Four 10GbE SFP+ ports are available on the S3048-ON for uplinks to the OOB management network core (links not shown).

Note: When running OS10EE, the S3048-ON will function as an OOB management switch with its factory default configuration. By default, all ports are in switchport mode, in VLAN 1, administratively up, and rapid per-VLAN spanning tree plus (RPVST+) is enabled. At a minimum, Dell EMC recommends changing the admin password to a complex password during the first login.

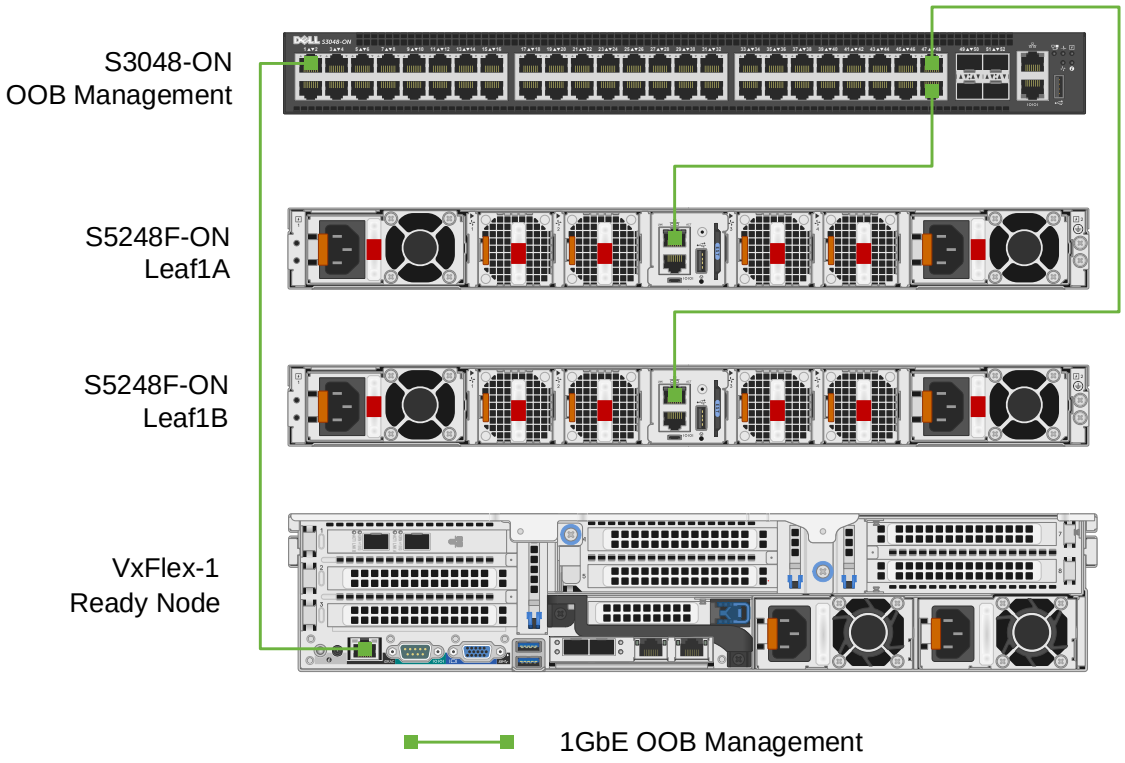

Figure 8 OOB management network connections

Figure 8 shows the first node (VxFlex-1) connected to the S3048-ON management switch using the onboard iDRAC port. The connections for VxFlex-2 through VxFlex-4 (not shown) are done in the same manner. The production switches are also connected to the S3048-ON management switch using the OOB port on the back of each switch.

The leaf 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. The out-of-band (OOB) management interface (mgmt1/1/1) is configured as a backup link to fulfill this requirement.

Note: An existing 1GbE Base-T switch may be used as an alternative to the S3048-ON. Optionally, for small deployments, the dedicated iDRAC port on each Ready Node may be connected to a port configured on the in-band management VLAN on either one of the two S5248F-ON leaf switches. This requires an SFP-1GbE Base-T adapter, Dell EMC part number XTY28, for each iDRAC connection.

3.2 IP addressing

Dell EMC VxFlex MDMs, SDSs, and SDCs can have multiple IP addresses and can reside on more than one network. Multiple IPs provide options for load balancing and redundancy. VxFlex 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.

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

- iDRAC Management – Used for administrative management of the VxFlex node 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

Note: Multiple IP subnets used for the VxFlex OS data network cannot be on the same subnet in a VMware setup. VxFlex OS only supports the following network configurations when deployed on VMware:

* A single data storage network
* Two or more data networks, each on separate IP subnets
* A single IP data network using several NIC-bonding configurations, or vSwitch load balancing

For more information, see the VMware limitation in the following Vmware Knowledge Base article: Multi-homing on ESXi/ESX.

Table 2 and Table 3 show how to calculate the IP address requirements. To accommodate expansion each address pool is assigned a dedicated subnet.

Table 2 VMware and management IP network calculations

Item	Description
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 HOST_AND_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 * Host and SVM_MGMT_IP + N * IDRAC_MGMT_IP	

Table 3 VxFlex 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 = VxFlex 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 = VxFlex Nodes internal (interconnect) IP addresses 2. SVM_DATA2_IP = SVM management IP addresses 3. MDM_Cluster_Virtual_IP_DATA2 = The virtual IP of the MDM cluster in the Data2 network	For clarity, the second 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 4. VMware vMotion, Data1, and Data2 do not require default gateways in this configuration. If routing is required for Data1 and Data2, see *Routing VxFlex Virtual Machine traffic* in Appendix B.

Table 4 IP address ranges

IP Pool	IP Subnet	Default gateway
ESXI_VMOTION_IP	172.16.32.0/24	n/a
HOST_AND_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 5.

Table 5 VLAN and subnet association

IP Pool	VLAN ID
ESXI_VMOTION_IP	1632
HOST_AND_SVM_MGMT_IP	1633
Node_DATA1_IP & SVM_DATA1_IP	1634
Node_DATA2_IP & SVM_DATA2_IP	1635

3.3 Switch preparation

Switches used in this guide run OS10EE version 10.4.2.2 or later. Run the **`show version`** command to check the operating system version, and then update the operating system as required for each switch.

Note: Dell EMC recommends upgrading to the latest release available on Dell Digital Locker (account required).

```
OS10# show version
Dell EMC Networking OS10-Enterprise
Copyright (c) 1999-2019 by Dell Inc. All Rights Reserved.
OS Version: 10.4.2.2
Build Version: 10.4.2.2.265
Build Time: 2019-01-14T15:15:14-0800
System Type: S5248F-ON
Architecture: x86_64
Up Time: 3 days 03:51:54
```

Run the command `show license status` to verify license installation. The `License Type:` field should indicate `PERPETUAL`. If an evaluation license is installed, licenses purchased from Dell EMC are available for download on Dell Digital Locker. Installation instructions are provided in the *OS10 Enterprise Edition User Guide Release 10.4.2.0.*

```
OS10# show license status

System Information
---------------------------------------------------------
Vendor Name     :        Dell EMC
Product Name    :        S5248F-ON
Hardware Version:        A01
Platform Name   :        x86_64-dellemc_s5248f_c3538-r0
PPID            :        CN046MRJZES0089K00016
Service Tag     :        DELLEMC
License Details
----------------
Software        :        OS10-Enterprise
Version         :        10.4.2.2
License Type    :        PERPETUAL
License Duration:        Unlimited
License Status  :        Active
License location:        /mnt/license/69100Q2.lic
```

Note: If OS10EE was factory installed, a perpetual license is already on the switch.

Set the factory defaults of your switches to remove any current configuration. The following commands set switches running OS10EE to their factory default settings:

```
OS10# delete startup-configuration
Proceed to delete startup-configuration [confirm yes/no(default)]:y

OS10# reload
System configuration has been modified. Save? [yes/no]:n
Proceed to reboot the system? [confirm yes/no]:y
```

The switch is now ready for configuration.

4 Dell EMC Networking S5248F-ON 25GbE switch configuration

This chapter provides the steps for configuring the network to connect VxFlex v2.6.1.1 on Dell EMC VxFlex Ready Nodes (based on R740xd servers) using Dell EMC Networking S5248F-ON 25GbE switches running OS10EE.

Switch configuration details with explanations are provided for one leaf switch. The remaining leaf switch uses a similar configuration.

Note: The configuration files for both leaf switches are provided as attachments in the softcopy (.pdf) of this document. To access attachments in Adobe Acrobat Reader, click the icon in the left pane halfway down the page, then click the icon.

4.1 Initializing switches

The S5248F-ON switches must be running OS10EE version 10.4.2.2 or later for the commands that are provided in this guide to work.

Note: Dell EMC recommends that the switches are using a perpetual license and are set to factory defaults. Before proceeding, use the instructions in section 3.3 to verify the operating system version of each switch. If necessary, download and install the perpetual licenses, and then set your switches to factory defaults to prepare them for the VxFlex network configuration.

4.2 S5248F-ON leaf switch configuration

The following section outlines the configuration commands that are issued to the S5248F-ON leaf switches. The switches start at their factory default settings, as indicated in section 3.3.

Note: The following configuration details are specific to S5248F, Leaf1A. The remaining leaf switch is similar. The configuration details for both leaf switches are provided in the attachments section and are named **S5248-Leaf1A.txt** and **S5248-Leaf1B.txt**.

Initial configuration requires the configuration of the hostname. The VLT backup interface is configured with an IP address.

```
OS10# configure terminal
OS10(config)# hostname S5248-Leaf1A
S5248-Leaf1A(config)# interface mgmt 1/1/1
S5248-Leaf1A(conf-if-ma-1/1/1)# no ip address dhcp
S5248-Leaf1A(conf-if-ma-1/1/1)# ip address 172.16.30.101/24
S5248-Leaf1A(conf-if-ma-1/1/1)# no shutdown
S5248-Leaf1A(conf-if-ma-1/1/1)# exit
S5248-Leaf1A(config)# management route 0.0.0.0/0 172.16.30.254
```

Enable RSTP as a precaution against loops. Configure S5248F-Leaf1A as the primary RSTP root bridge using the `spanning-tree rstp priority 0` command and S5248F-Leaf1B as the secondary RSTP root bridge using the `spanning-tree rstp priority 4096` command.

```
S5248-Leaf1A(config)# spanning-tree mode rstp
```

```
S5248-Leaf1A(config)# spanning-tree rstp priority 0
```

Configure the VLT interconnect between S5248F-Leaf1A and S5248F-Leaf1B. In this configuration, use interfaces Eth 1/1/55-56 for the VLT interconnect. The backup destination is the management IP address of the VLT peer switch, S5248F-Leaf1B. Finally, VLT peer-routing is enabled providing forwarding redundancy in the event of a switch failure.

```
S5248-Leaf1A(config)# interface range ethernet 1/1/55-1/1/56
S5248-Leaf1A(conf-range-eth1/1/55-1/1/56)# description VLTi
S5248-Leaf1A(conf-range-eth1/1/55-1/1/56)# no switchport
S5248-Leaf1A(conf-range-eth1/1/55-1/1/56)# exit

S5248-Leaf1A(config)# vlt-domain 1
S5248-Leaf1A(conf-vlt-1)# backup destination 172.16.30.102
S5248-Leaf1A(conf-vlt-1)# discovery-interface ethernet1/1/55-1/1/56
S5248-Leaf1A(conf-vlt-1)# peer-routing
S5248-Leaf1A(conf-vlt-1)# primary-priority 1
S5248-Leaf1A(conf-vlt-1)# vlt-mac de:11:de:11:de:11
S5248-Leaf1A(conf-vlt-1)# exit
```

Three VLAN interfaces are created. The management network is assigned to a VRRP group, and a virtual address is assigned. To designate this switch as the master, the VRRP priority is set to 254 (the VRRP peer switch priority is set to 1). Only Data1 is created on this leaf switch, and the IP address is optionally assigned, while Data 2 is created on the other leaf switch. The switch with the largest priority value becomes the master VRRP router.

```
S5248-Leaf1A(conf-vlt-1)# interface vlan 1632
S5248-Leaf1A(conf-if-vl-1632)# description "ESXI VMOTION IP"
S5248-Leaf1A(conf-if-vl-1632)# mtu 9216
S5248-Leaf1A(conf-if-vl-1632)# no shutdown
```

Note: It is considered a best practice to use jumbo frames (mtu 9216) in this configuration. See chapter 7 for the Maximum Transmission Unit (MTU) guidelines used in this guide.

```
S5248-Leaf1A(conf-if-vl-1632)# interface vlan 1633
S5248-Leaf1A(conf-if-vl-1633)# ip address 172.16.33.251/24
S5248-Leaf1A(conf-if-vl-1633)# description "HOST AND SVM MGMT IP"
S5248-Leaf1A(conf-if-vl-1633)# mtu 9216
S5248-Leaf1A(conf-if-vl-1633)# vrrp-group 33
S5248-Leaf1A(conf-vlan1633-vrid-33)# priority 254
S5248-Leaf1A(conf-vlan1633-vrid-33)# virtual-address 172.16.33.253
S5248-Leaf1A(conf-vlan1633-vrid-33)# exit
S5248-Leaf1A(conf-if-vl-1633)# no shutdown

S5248-Leaf1A(conf-if-vl-1633)# interface Vlan 1634
S5248-Leaf1A(conf-if-vl-1634)# description "Node DATA1 IP & SVM DATA1 IP"
S5248-Leaf1A(conf-if-vl-1634)# mtu 9216
S5248-Leaf1A(conf-if-vl-1634)# no shutdown
S5248-Leaf1A(conf-if-vl-1634)# exit
```

The downstream interfaces to the VxFlex cluster R740xd servers are configured in the next set of commands. All server-facing interfaces are tagged for each VLAN.

```
S5248-Leaf1A(config)# interface range ethernet 1/1/1-1/1/4
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# no ip address
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# description "LINK TO VXFLEX NODES"
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# mtu 9216
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# switchport mode trunk
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# switchport access vlan 1
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# switchport trunk allowed vlan
1632,1633,1634
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# spanning-tree port type edge
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# spanning-tree guard root
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# no shut
```

The following `mode` command converts 25GbE switch ports to 10GbE, allowing 10GbE network devices to attach to the 25GbE switch at the lower speed. The converted port's new port names end in :1.

```
S5248-Leaf1A(config)# port-group 1/1/12
S5248-Leaf1A(conf-pg-1/1/12)# mode Eth 10g-4x
S5248-Leaf1A(conf-pg-1/1/12)# exit
```

Save the configuration.

```
S5248-Leaf1A(config-line-vty)# end
S5248-Leaf1A# write memory
```

Note: IP configuration for VLAN IDs 1634 (on Leaf1A) and 1635 (on Leaf1B) is optional. A gateway is not required for the two VxFlex data networks. When connecting external SDC's it is recommended to use IP's on the same subnet. Connecting SDC's via router on a different subnet is supported and will require manual configuration of static routes on the SVM and the ESXi as described in Appendix B.

4.3 Verify switch configuration

The following sections provide the commands and outputs that are used to verify the configuration of the switches, and to ensure that they are correctly connected. Except where there are fundamental differences, the only output that is shown is the S5248F-Leaf1A switch. The output from S5248F-Leaf1B is similar.

4.3.1 show vlt domain id

The `show vlt` *`domain-id`* command validates the VLT configuration status when the VLTi Link Status is up. The role of one switch in the VLT pair is primary (shown) and its peer switch is assigned the secondary role.

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.

```
S5248-Leaf1A# show vlt 1
Domain ID                  : 1
Unit ID                    : 1
Role                       : primary
Version                    : 2.0
Local System MAC address   : 54:bf:64:e6:dd:c1
Role priority              : 1
```

```
VLT MAC address            : de:11:de:11:de:11
IP address                 : fda5:74c8:b79e:1::1
Delay-Restore timer        : 90 seconds
Peer-Routing               : Enabled
Peer-Routing-Timeout timer : 0 seconds
VLTi Link Status
    port-channel1000       : up

VLT Peer Unit ID    System MAC Address    Status    IP Address           Version
--------------------------------------------------------------------------------
  2                 3c:2c:30:10:35:00      up      fda5:74c8:b79e:1::2   2.0
```

4.3.2 show vlt domain id mismatch

The `mismatch` option lists VLANs configured on a single switch in the VLT domain. The output shows the two VLANs associated with Data1 and Data2.

```
S5248-Leaf1A# show vlt 1 mismatch
VLT-MAC mismatch:                              No mismatch
Peer-routing mismatch:                         No mismatch
VLAN mismatch:
VLT Unit ID    Mismatch VLAN List
---------------------------------------------------------
* 1              1634
  2              1635
VLT VLAN mismatch:                             No mismatch
VLT Virtual Network Mismatch:                  No mismatch
Virtual Network Name Mismatch:                 No mismatch
Virtual Network VLTi-VLAN Mismatch:            No mismatch
Virtual Network Mode Mismatch:                 No mismatch
Virtual Network Tagged Interfaces Mismatch:    No mismatch
Virtual Network Untagged Interfaces:           No mismatch
Virtual Network VNI Mismatch:                  No mismatch
Virtual Network Remote-VTEP Mismatch:          No mismatch
```

4.3.3 show vlt domain id backup-link

The `show vlt` *`domain-id`* `backup-link` command shows the heartbeat status of the vlt peer.

```
S5248-Leaf1A# show vlt 1 backup-link
VLT Backup Link
------------------------
Destination                    : 172.16.30.102
Peer Heartbeat status          : Up
Heartbeat interval             : 30
Heartbeat timeout              : 90
Destination VRF                : default
```

4.3.4 show vlt mac-inconsistency

The `show vlt mac-inconsistency` command shows the inconsistencies in dynamic MAC addresses learned between VLT peers.

```
S5248-Leaf1A# show vlt mac-inconsistency
Inconsistency check for VLAN based MAC
--------------------------------------
Fetching MACs from unit 2
Fetching MACs from unit 1
Identifying inconsistencies ..
No inconsistencies found
```

4.3.5 show spanning-tree brief

The `show spanning-tree brief` command validates that STP is enabled on the leaf switches. All connected interfaces are forwarding (`FWD`), as shown in the **Sts** column.

```
S5248-Leaf1A# show spanning-tree brief
Spanning tree enabled protocol rstp with force-version rstp
Executing IEEE compatible Spanning Tree Protocol
Root ID    Priority 4096, Address 54bf.64e6.ddc1
Root Bridge hello time 2, max age 20, forward delay 15
Bridge ID    Priority 8192, Address 3c2c.3010.3500
Configured hello time 2, max age 20, forward delay 15
Flush Interval 200 centi-sec, Flush Invocations 59
Flush Indication threshold 65535

Interface                                          Designated
Name            PortID    Prio   Cost  Sts   Cost   Bridge ID   PortID
-------------------------------------------------------------------------
ethernet1/1/1   128.516   128    800   FWD     1      8192     3c2c.3010.3500
ethernet1/1/2   128.520   128    800   FWD     1      8192     3c2c.3010.3500
ethernet1/1/3   128.524   128    800   FWD     1      8192     3c2c.3010.3500
ethernet1/1/4   128.528   128    800   FWD     1      8192     3c2c.3010.3500
ethernet1/1/5   128.532   128    20000 BLK     0      0        0000.0000.0000
ethernet1/1/6   128.536   128    20000 BLK     0      0        0000.0000.0000
ethernet1/1/7   128.540   128    20000 BLK     0      0        0000.0000.0000
                                    :
                            <output truncated>
```

4.3.6 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.

```
S5248F-Leaf1A#show vrrp brief
Interface Group  Priority Prempt   State     Version  Master addr   Virtual addr
-------------------------------------------------------------------------------
vlan1633  IPv4 33   1      true   master-state  2   172.16.33.251 172.16.33.253
```

5 VMware virtual network design

Tables are provided in this section 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* for details on configuring ESXi and the virtual network environment.

5.1 ESXi management

The default VMkernel, vmk0 is used for 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. VxFlex data networks (Data1 and Data2) use *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 Switch (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 VxFlex cluster.

Table 6 Virtual switch details

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

Table 7 Port group configuration settings

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

Table 8 Port group settings

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

Table 9 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 VxFlex VDS with four VMkernel adapters assigned (see chapter 6).

Table 10 VMkernel adapter settings

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

Figure 9 is taken from **vCenter > Home** > **Networking** > **atx01-w01-vds** > **Configure** > **Topology** and shows the completed topology of vDS-VxFlex 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.

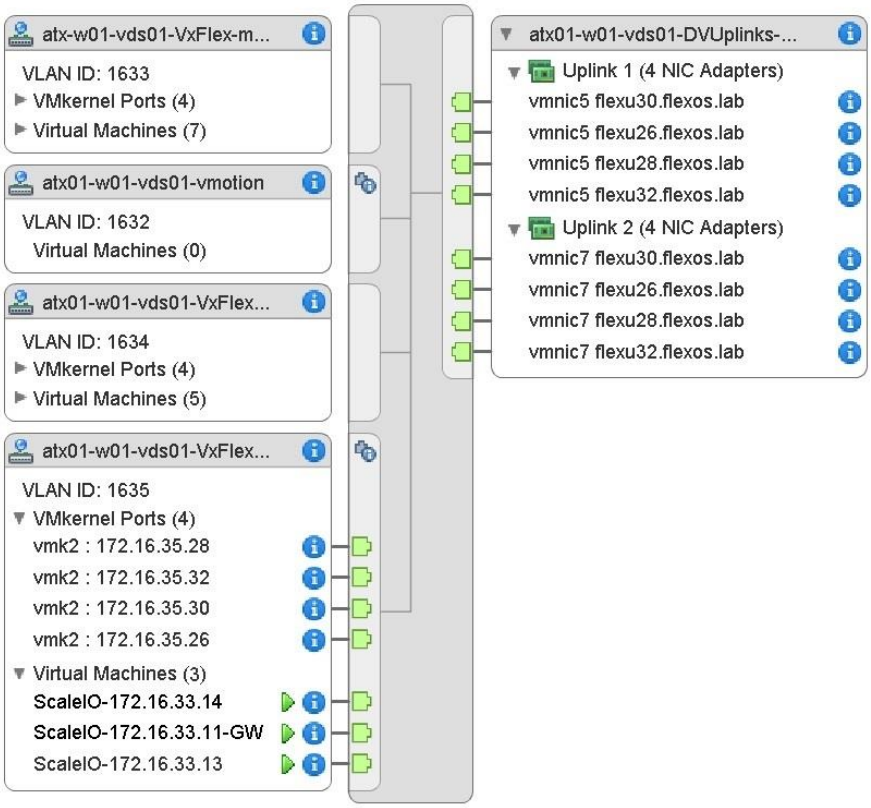

Figure 9 VDS atx01-w01-vds01 topology shown in vCenter

Physical connectivity and configuration of the VMware vSphere distributed switches are now complete.

Figure 10 represents a successfully deployed VxFlex HCI platform which is created in the next chapter. The screen is taken from **Home** > **Hosts and Clusters.**

Continue to chapter 6 for instructions to deploy the VxFlex HCI platform.

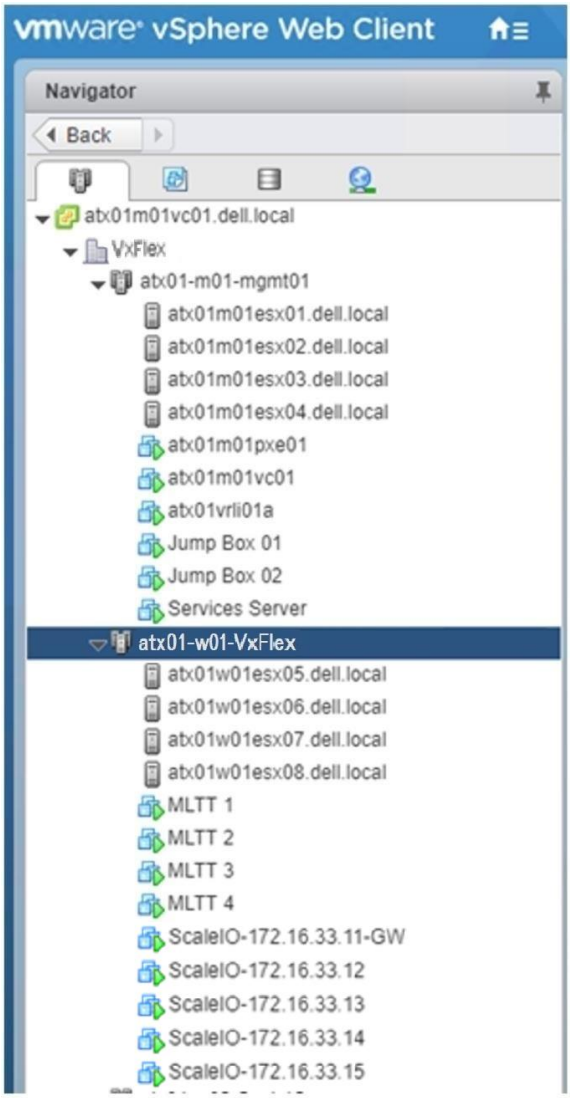

Figure 10 vCenter host and clusters

6 Deploy Dell EMC VxFlex

Deploying VxFlex in this environment consists of the following steps:

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

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

6.1 Deploy the VxFlex plug-in

The VxFlex plug-in for vSphere simplifies the installation and management of the VxFlex system in a vSphere environment. Use the following parameters during the installation:

Table 11 Dell EMC VxFlex VMware vSphere plug-in parameters

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

6.2 Upload the VxFlex OVA template

Once the VMware vCenter Dell EMC VxFlex plug-in installation is completed, upload the VxFlex virtual machine OVA to an R740xd local datastore. The OVA file has a virtual machine template for deploying all the software components of VxFlex. From the VMware PowerCLI program, select Create SVM Template. While it is possible to specify separate datastores to correspond to each VxFlex 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 VxFlex implementation.

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

Table 12 VxFlex VMware vSphere plug-in parameters

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

6.3 Deploy VxFlex using the deployment wizard

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

VxFlex deployment has four steps:

1. SDC deployment and configuration
2. VxFlex advanced configuration settings
3. Deploy the VxFlex environment
4. Install the VxFlex 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 that is 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 VxFlex** screen, click **Install SDC on ESX**.
2. Select all hosts under the VxFlex data center as targets for the installation.
3. Once complete, reboot all hosts before continuing with the deployment.

To deploy VxFlex, perform the following steps:

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

Note: The parameters and settings that are 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 13 VxFlex Wizard deployment settings

Parameter	Setting
Select installation	Create a new system
System name	SIO01
Admin password	VxFlex Admin Password
vCenter server	atx01m01vc01.dell.local
Host selection	atx01w01esx05, atx01w01esx06, atx01w01esx07, atx01w01esx08
VxFlex 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 VxFlex wizard parameter settings for steps 5 through 7.

Table 14 VxFlex Wizard deployment settings

Parameter	Setting
Protection domain name	PD01
RAM read cache size per SDS	128 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 15 VxFlex Wizard deployment settings

Parameter	Setting
Host for VxFlex gateway	atx01w01esx08
Gateway admin password	VxFlex Admin Password
Gateway LIA password	VxFlex Admin Password
Select OVA template	EMC VxFlex SVM Template (v2.0.140000.228) 1
OVA root password	VxFlex Admin Password
OVA LIA password	VxFlex Admin Password
Management network label	atx01-w01-vds01-VxFlex-management
Data01 network label	atx01-w01-vds01-VxFlex-data01
Data02 network label	atx01-w01-vds01-VxFlex-data02

Table 16 VxFlex networking addressing

ESXi name	Management IP	Default gateway	Data 1 IP	Data 2 IP
atx01w01esx08 (VxFlex 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 (Standby MDM)	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 17 VxFlex networking virtual IP addresses

Parameter	Setting
Data01 (atx01-w01-vds01-VxFlex-data01)	172.16.34.4
Data02 (atx01-w01-vds01-VxFlex-data02)	172.16.35.4

Once the summary screen displays, the deployment begins. If errors are encountered, the installation wizard will halt and resume once the issue is resolved.

6.3.1 MDM Virtual IP Addresses

Up to four virtual IP addresses can be defined for the MDM cluster and are highly recommended. SDCs are then mapped to the MDM cluster's virtual IP addresses, instead of to static MDM IP addresses. MDMs are sometimes switched during normal operation of the cluster, and the virtual IP address will always be mapped to the active MDM.

The use of virtual IP addresses simplifies maintenance procedures on the MDM cluster, because system components communicate via the virtual IP addresses. Therefore, SDCs do not need to be reconfigured when a server hosting an MDM is replaced.

In new installations in Linux environments, the MDM cluster's virtual IP address can be added and mapped using the VxFlex OS Installer CSV file. In VMware environments, virtual IP addresses are mandatory, and configuration is performed using the VxFlex OS VMware Installation Wizard, in the Configure SVM stage. The

REST API can be used to add virtual IP addresses to the cluster. In all cases, a virtual IP NIC placeholder must be mapped to each virtual IP address. Ensure that there are NICs available for this purpose. Existing systems may be extended to include additional MDMs to a cluster. The new MDMs should be mapped to the existing virtual IP addresses. If virtual IP addresses need to be modified, you must use the CLI or the REST API (not the VxFlex OS Installer or the vSphere plug-in), and it must be done with extreme caution. All SDCs will require reconfiguration, to reflect the changes made to the MDM cluster. Otherwise, the SDCs will not be able to communicate with the MDM cluster, and volumes will not be accessible.

6.4 VxFlex GUI

Installing the VxFlex graphical user interface (GUI) on a management workstation provides a way to monitor and configure the VxFlex system. Once installed, the virtual IP assigned to Data1 (172.16.34.4) can be used to access the VxFlex GUI. The installation file is part of the VxFlex for Windows download. The VxFlex cluster that is created is shown below in Figure 11.

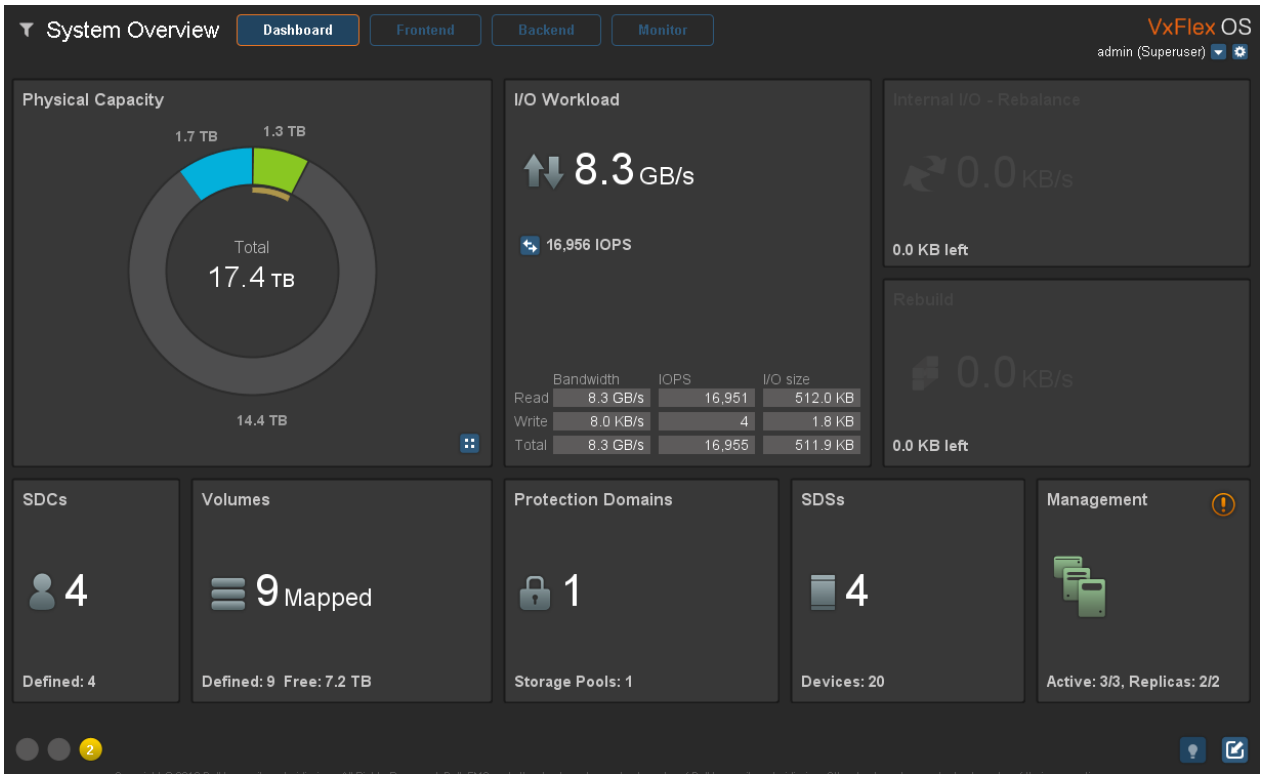

Figure 11 VxFlex under load showing throughput

7 Best practices

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

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

For more information on performance tuning, including ESXi hosts and VxFlex VMs, see the *VxFlex v2.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 18 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	VxFlex-data01	atx01-w01-vds01- VxFlex-data01	9000
atx01-w01-vds01	VxFlex-data02	atx01-w01-vds01- VxFlex-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, VxFlex-172.16.33.12:

```
VxFlex-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 VxFlex Data Network 1 (subnet 172.16.34.0/24, VLAN 1634) and VxFlex Data Network 2 (subnet 172.16.35.0/24, VLAN 1635). Administrative access uses Eth0.

2. 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:

```
VxFlex-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
```

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

```
VxFlex-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
```

4. Use the `ping` command to validate jumbo frames connectivity to another configured SVM:

```
VxFlex-172-16-33-12:~ # ping -M do -s 8972 172.16.33.13
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 DiffServ

The basis of Quality of Service (QoS) is traffic differentiation. In this guide, different types of traffic are classified accordingly and are thereby given priorities throughout the network using differentiated services (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 stand-alone 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:

- Traffic with higher priority requirements is given a DSCP mark of 46
- Traffic with lower priority requirements is not marked

VxFlex 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 *Dell EMC VxFlex v2.x Deployment Guide*.

A DSCP value of 46 corresponds to expedited forwarding and maps to queue 5 on the S5248F-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 S5248F-ON switch. Configuration for the second leaf switch is identical.

```
S5248-Leaf1A(config)# interface range ethernet1/1/1-1/1/4
S5248-Leaf1A(conf-range-eth1/1/1-1/1/4)# trust-map dscp default
```

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-VxFlex-management. To create markings navigate to **Home** > **Networking** > **port group** > **Edit** > **Traffic filtering and markings**. Table 19 shows the values that are used.

Table 19 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-VxFlex-management	Enable	46	Virtual Machines
atx01-w01-vds01	atx01-w01-vds01- VxFlex-data01	Disable	n/a	n/a
atx01-w01-vds01	atx01-w01-vds01- VxFlex-data02	Disable	n/a	n/a

The `show qos interface ethernet` *`interface_number`* command is used to confirm that DSCP is enabled on the interface.

```
S5248-Leaf1A# show qos interface ethernet 1/1/1
  Interface
  unknown-unicast-storm-control : Disabled
  multicast-storm-control : Disabled
  broadcast-storm-control : Disabled
  flow-control-rx : Disabled
  flow-control-tx : Disabled
  ets mode : Enabled
  pfc mode : Enabled
  Dscp-tc-mapping : default
```

A Troubleshoot SDS connectivity

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

The following commands test all SDS nodes with a payload of 10 GB, using eight parallel threads:

```
VxFlex-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

VxFlex-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 VxFlex 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 VxFlex management
- Eth1 for VxFlex Data01
- Eth2 for VxFlex Data02

The SVM uses a single TCP/IP stack, and any unknown networks are limited to this single default gateway. If VxFlex Data1 or Data2 need 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 use policy-based routing (PBR).

1. To show the routing table of an SVM, enter the `ip route` command:

```
VxFlex-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.

```
VxFlex-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 issue, create two routing tables for eth1 and eth2 and verify that they have been created.

```
VxFlex-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.

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

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

```
VxFlex-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

VxFlex-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. Repeat the ping test to validate that the solution is working.

```
VxFlex-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
8 packets captured
8 packets received by filter
0 packets dropped by kernel
```

Note: This example solution is not persistent across reboots but can be automated by the administrator.

C Validated hardware and components

The tables in this section list the hardware and components that are used to configure and validate the examples in this guide. Table 20 shows the Dell EMC Networking switches and associated versions used. It is recommended to update your switches to the latest operating systems available on Dell Digital Locker (account required).

Table 20 Dell EMC Networking switches

Qty	Item	OS/Firmware version
1	S3048-ON - Management switch	OS10EE 10.4.2.2
		System CPLD: 9
		Module CPLD: 7
2	S5248F-ON - Leaf switch	OS10EE 10.4.2.2
		System CPLD: 0.4
		Slave CPLD: 1.0

Table 21 lists details of the VxFlex Ready Nodes used to validate the examples in this guide. It is recommended that you update your hardware and firmware with the latest VxFlex Ready Node 14G Hardware Update .

Table 21 VxFlex hardware and firmware

Qty per server	Item	Firmware version
2	Intel® Xeon® Gold 6126 CPU @ 2.60GHz, 12 cores	-
192	GB RAM	-
5	960GB SATA SSD	-
1	Dell HBA330	15.17.09.06
2	MLNX 25GbE 2P ConnectX-4 LX	14.23.10.20
1	Integrated Intel® 2P X710/2P I350 rNDC	18.8.9
-	BIOS	1.6.12
-	iDRAC	3.21.26.22

D Validated software

Table 22 lists the software components used to validate the example configurations in this guide. It is important to update to the latest versions located at VxFlex Ready Nodes driver and firmware version requirements.

Table 22 Software versions

Item	Version
Dell EMC VxFlex	2.6.1.1_113
VMware vSphere Power CLI	11.0
VxFlex vSphere Plug-in Installer	2.6.11000_113
SVM OVA	2.6.11000_113.ova
VMware ESXi	6.5 U1 - Dell EMC customized image, A07 build 7388607
VMware vCenter Server Appliance	6.5.0 - 9451637

E Supported switches

Steps in this document were validated using specific Dell EMC Networking switches and OS10EE but may be used for other Dell EMC Networking switch models that use the same networking operating system or later. Depending on a switch's available port numbers, speeds, and types, minor adjustments may be required to the example commands provided in this guide to achieve the same results.

The table below shows all switches that can be configured using the directions provided in this guide. For the leaf switches, the switches must support OS10EE version 10.4.2.2 or later and must also support 25GbE connections.

Table 23 Supported switches for this deployment guide

Dell EMC Networking switches supported for VxFlex in this guide*

Dell EMC Networking Switch	Leaf Switches Using OS10EE v10.4.2.2 and later	Management Switch Using OS10EE v10.4.2.2 and later
S5148F-ON	Supported	
S5248F-ON	Supported	
S5232F-ON	Supported	
S5296F-ON	Supported	
S6100-ON	Supported	
S3048-ON		Supported
S3124 / S3124F / S3124P / S3148 / S3148P		Supported

*DELL EMC PROVIDES THIS SUPPORTED SWITCH LIST AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND. THIS LIST IS FOR INFORMATIONAL PURPOSES ONLY AND MAY CONTAIN TYPOGRAPHICAL AND TECHNICAL INACCURACIES. DELLEMC SHALL NOT BE LIABLE FOR ANY DAMAGES ARISING OUT OF OR IN CONNECTION WITH THE USE OF THIS LIST.

For the latest list of 25GbE switches from Dell EMC, visit Dell Networking S-Series 25-100GbE switches.

F Product manuals and technical guides

Dell EMC

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

Dell EMC Networking Guides

VxFlex Ready Nodes solution overview

VxFlex Ready Nodes driver and firmware version requirements

Manuals and documentation for Dell Networking S5248F-ON

Manuals and documentation for Dell Networking S3048-ON

Manuals and documentation for PowerEdge R740xd

Dell EMC VxFlex OS only: Documentation Library

Dell EMC VxFlex OS Documentation Library (Guides and Tech Notes)

VxFlex Ready Node comparisons

Dell EMC VxFlex OS Networking Best Practices and Design Considerations

VMware

VMware vSphere Documentation

VMWare 6.5 Installation and Setup: vSphere, ESXi, and vCenter

VMware Compatibility Guide

Dell EMC VxFlex SDS with ESXi 5.5, ESXi 6.0, and ESXi 6.5 (2146203) (account required)

G Fabric Design Center

The Dell EMC Fabric Design Center (FDC) is a cloud-based application that automates the planning, design and deployment of network fabrics that power Dell EMC compute, storage and hyper-converged infrastructure solutions, including VxFlex. The FDC is ideal for turnkey solutions and automation based on validated deployment guides like this one.

FDC allows design customization and flexibility to go beyond validated deployment guides. For additional information, visit the Dell EMC Fabric Design Center.

H 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