Best Practices

Dell EMC PowerScale: Leaf-Spine Network Best Practices

Abstract

This white paper provides the best practices for deploying a scalable and resilient back-end network infrastructure for Dell EMC PowerScale clusters. Dell EMC PowerScale OneFS 8.2 and later version enables the deployment of a leaf-spine back-end network switch architecture that increases the size, scale, and performance of PowerScale clusters.

August 2021

H17682

Revisions

Date	Description
August 2021	Content and format updates

Acknowledgments

Author: Abiy Mesfin

Table of contents

Executive summary

Dell EMC PowerScale OneFS 8.2 introduced support for a back-end network infrastructure that is configurable as a leaf-spine network architecture with a set of leaf switches connected to spine switches. PowerScale OneFS 9.0 now supports 252 nodes in a leaf-spine architecture. This document includes best practices for configuring this architecture while accounting for data-center space usage, rack positioning, cabling, and scale.

1 Introduction

Next-generation, multi-rack, data center solutions require performance, scale, and capacity to drive new and demanding workloads. A leaf-spine back-end network architecture supports these needs and facilitates larger deployments. The solution in this document describes a leaf-spine back-end network architecture with Dell EMC switches and a PowerScale cluster which can scale up to 252 nodes.

Dell EMC PowerScale scale-out NAS nodes use InfiniBand switches as the private network for the back-end, intracluster, node-to-node communication. The OneFS 8.1 operating system introduced the use of Ethernet switches for the back-end node-to-node communication. OneFS 8.1.1 introduced a choice of Dell EMC Ethernet switches for the back end to simplify configurations and provide a full Dell EMC solution.

2 Leaf-spine network architecture

In a leaf-spine network switch architecture, the access layer of the network is referred to as the leaf layer. The PowerScale nodes connect to leaf switches at this layer. At the next level, the aggregation and core network layers are condensed into a single spine layer. Every leaf switch connects to every spine switch to ensure that all leaf switches are no more than one hop away from one another. Also, leaf-to-spine switch connections must be evenly distributed, meaning there should be the same number of connections to each spine switch from each leaf switch. This practice minimizes latency and the likelihood of bottlenecks in the back-end network. A leaf spine network architecture is highly scalable and redundant.

Leaf spine network deployments can have a minimum of two leaf switches and one spine switch. For small to medium clusters, this back-end network consists of two redundant top-of-rack (ToR) switches. Only the Dell EMC Z9100-ON and Z9264-ON Ethernet switches (Table 1) are supported in the leaf-spine architecture.

Table 1 Dell EMC Z9100-ON and Z9264-ON switches

Dell EMC Switch	Legacy PowerScale model	Dell SKU	Back-end ports	Port type	Rack units	100 GbE and 40 GbE nodes	Mixed environment (10, 25, 40, and 100 GbE)
Z9100-ON	851-0316	210-AWOV /210-AWOU	32	All 100 GbE	1	32 or less	• Support breakout cables of 4 x 10 or 4 x 25. • Total 128 10 GbE or 25 GbE nodes as ToR back-end switch.
Z9264-ON	851- 0318	210-AWOW	64	All 100 GbE	2	64 or less	• Support breakout cables of 4 x 10 or 4 x 25. • Total 128 10 GbE or 25 GbE nodes as ToR back-end switch.

Note: Use of breakout cables on Z9264-ON switch disables the adjacent port.

While the Z9100-ON and Z9264-ON supports many features, not all capabilities of the switch are exposed or used when the switch functions as a PowerScale back-end switch (see Table 2).

Table 2 Connection considerations for solution components

Component	Description	Connection considerations
Network Spine Switch	Dell Z9100-ON 32-port switch Dell Z9264-ON 64-port switch	Back-end network with 100 GbE (uplink) connects to the leaf switch.
Network Leaf Switch	Dell Z9100-ON 32-port switch Dell Z9264-ON 64-port switch	Downlink from the leaf switch to the nodes. Supported connection type 100 GbE, 40 GbE, 25 GbE, and 10 GbE back-end nodes.
Dell EMC PowerScale nodes	F200, F600, and F900	F200: nodes support a 10 GbE or 25 GbE connection to the leaf using the same NIC. F600: nodes support a 40 GbE or 100 GbE connection to the leaf using the same NIC. F900: nodes support a 40 GbE or 100 GbE connection to the leaf using the same NIC.
Dell EMC PowerScale nodes (Gen6.5)	H700, H7000, A300, and A3000	H700 and H7000: support 40 GbE or 100 GbE connection to the leaf using the same NIC A300 and A3000: support 25 GbE or 10 GbE connection to the leaf using the same NIC
Dell EMC Isilon Performance nodes (Gen6)	F810, F800, H600, H500, and H5600	Performance nodes support a 40GbE connection to the leaf switch
Isilon Archive and Hybrid Nodes (Gen6)	A200, A2000, and H400	Archive nodes support a 10GbE connection to the leaf switch using breakout cable

Note: OneFS 9.0 requires the switch operating system version to be 10.5.0.6 or greater. The Dell EMC Networking OS version 10.5.0.6 is also supported in OneFS version 8.2.2. 144-node Leaf/Spine (L/S) clusters can still work with older versions of Dell EMC Networking OS version 10.4.x. Dell EMC Networking OS version 10.5.0.6 requires manual designation of leaf and spine switches through the command line from the switches. For detailed instructions for upgrading to Dell EMC Networking operating system version 10.5.0.6, see the Leaf-Spine Installation Guide and Switch OS Upgrade Guide.

3 General leaf-spine switch design considerations

Here are some general design considerations to simplify the setup and management of your PowerScale environment.

- Avoid network oversubscription between your uplink and downlink connections between the leaf and spine switches. The total throughput for all nodes connected to leaf must be less than or equal to the total uplink speed from leaf to spine.
- Mixing of Z9100 and Z9264 is supported in PowerScale leaf and spine architecture. Ensure that the Z9264 is a 2U switch and the Z9100 is a 1U switch.
- If you are replacing the Z9100 switch with the Z9264 switch, verify the Dell Networking OS version and ensure that it is supported by the Z9264 switch. Some cases may require you to upgrade your existing L/S switches before adding the Z9264 switch.
- If you have greater than one spine switch in the configuration, ensure the connections between leaf and spine switches are equally distributed among all leaf switches.
- Both Int-a and Int-b should be identical in terms of configuration and leaf spine network architectural design.
- You should strategically locate the spine switches within a data center. This practice ensures the cabling is planned, organized, and manageable when scaling out the nodes and switches within the cluster.
- Deploy a leaf-spine network topology for the expected growth of that cluster rather than the initial configuration.
- The solution supports live migration from a ToR back-end switch to an L/S back-end switch. For detailed steps, see the *Best Practices Guide for Live Migration* document.

Table 3 Z9100 switch as leaf and spine

Maximum nodes	Spines	Leaves	Cables between each pair of leaves and spines
All 40 Gb ports			
44	1	2	9
66	1	3	9
88	2	4	5
110	2	5	5
132	2	6	5
154	3	7	3
176	3	8	3
198	3	9	3
220	5	10	2
242	5	11	2
252	5	12	2

Maximum nodes	Spines	Leaves	Cables between each pair of leaves and spines
All 100 Gb ports			
32	1	2	16
64	2	4	8
112	4	7	4
128	4	8	4
135	5	9	3
150	5	10	3

Table 4 Z9264 switch as leaf and spine

Maximum nodes	Spines	Leaves	Cables between each pair of leaves and spines
All 40 Gb ports			
88	1	2	20
176	2	4	10
252	3	6	7
All 100 Gb ports			
64	1	2	32
128	2	4	16
252	4	8	8

Note: The maximum number of leaves and spines in the cluster should not exceed 17 switches per side (int-a and int-b combined 34 switches per cluster).

4 Network design examples

This section lists network design examples in your production environment.

4.1 Example 1: Network design with 252 PowerScale nodes

This example includes a proposed solution of 252 PowerScale nodes in the cluster:

Configuration: 252 performance-only nodes (40 GbE back end)

Assumptions or requirements:

- Connect not only uplink cables but also downlink cables to nodes on a different rack.
- Due to limited rack space, the customer in this example requests to minimize space in the data center.

Leaf switch	Spine switch connection
L1	Port 1 and 2 → Spine switch 1, 2, 3, 4 and 5
L2	Port 3 and 4 → Spine switch 1, 2, 3, 4 and 5
L3	Port 5 and 6 → Spine switch 1, 2, 3, 4 and 5
L4	Port 7 and 8 → Spine switch 1, 2, 3, 4 and 5
L5	Port 9 and 10 → Spine switch 1, 2, 3, 4 and 5
L6	Port 11 and 12 → Spine switch 1, 2, 3, 4 and 5
L7	Port 13 and 14 → Spine switch 1, 2, 3, 4 and 5
L8	Port 15 and 16 → Spine switch 1, 2, 3, 4 and 5
L9	Port 17 and 18 → Spine switch 1, 2, 3, 4 and 5
L10	Port 19 and 20 → Spine switch 1, 2, 3, 4 and 5
L11	Port 21 and 22 → Spine switch 1, 2, 3, 4 and 5
L12	Port 23 and 24 → Spine switch 1, 2, 3, 4 and 5

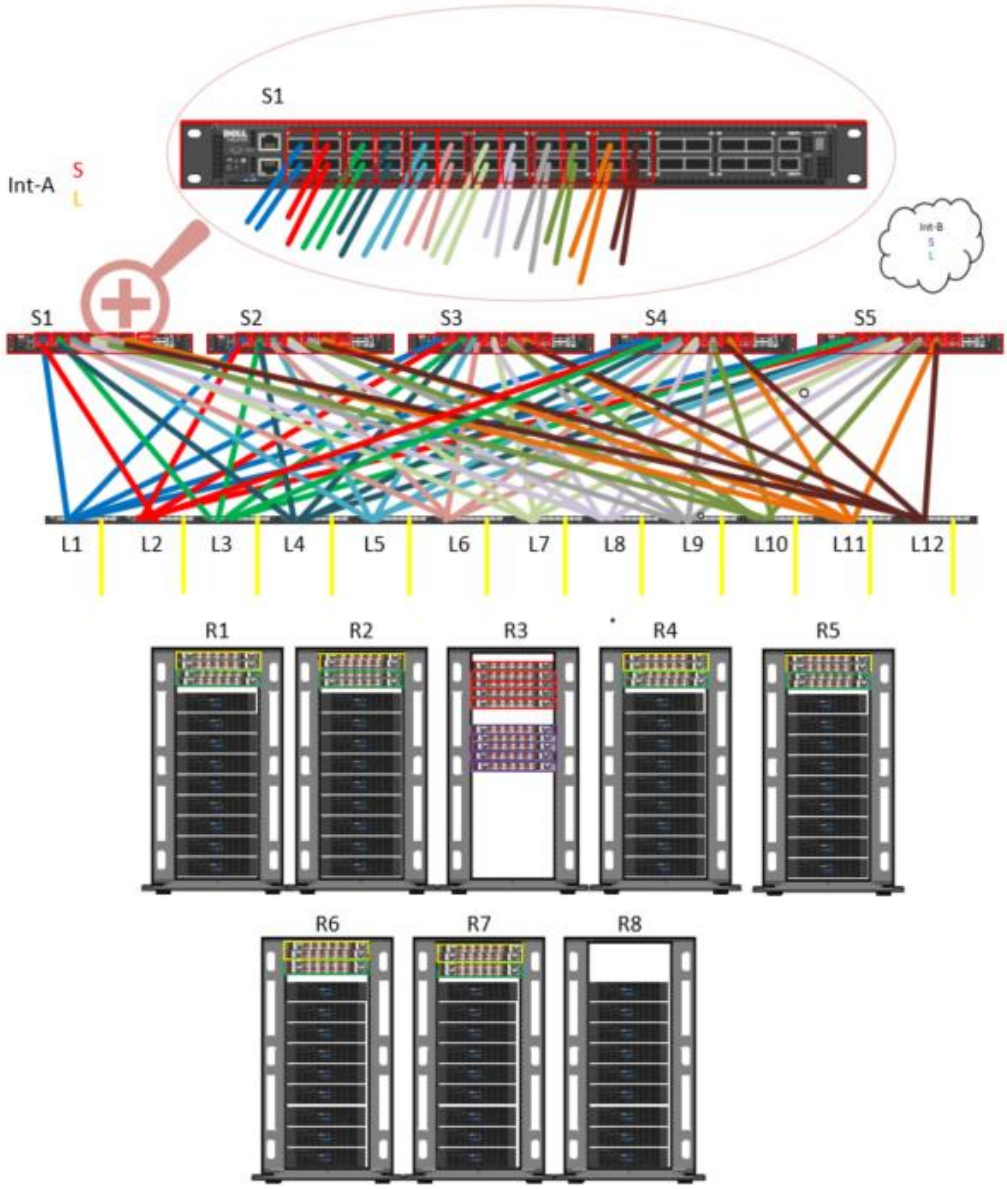

Figure 1 Sample configuration design of 252 nodes with 40 GbE back-end- network

This example configuration includes the following:

- 34 Dell EMC Networking Z9100 switches (17 per side)
  - 5 spine switches
  - 12 leaf switches
- 240 QSFP28 100 Gb uplink cables (10 uplink cables per leaf)
- 504 QSFP+ or MPO back-end cables
- 504 Optics (if MPO cables used)

Note: This example is accurate if you size all F600 systems with 40 GbE back-end connectivity.

Design considerations: To simplify and organize cabling, place your leaf switches accordingly (see Figure 1). As shown, two leaf switches from Int-a and two leaf switches from Int-b are spread across all racks except R8. No leaf switches are on R8, and the chassis on R8 must connect to a different rack.

4.2 Example 2: Network design with 252 PowerScale nodes

This example includes a proposed solution of 252 PowerScale nodes in the cluster:

Configuration: 252 performance-only nodes (100 GbE back end)

Assumptions or requirements: All nodes connect to the back-end leaf switch on the same rack, and **only** uplink cables to spine switches connect to a different rack.

Leaf switch	Spine switch connection
L1	Port 1,2,3,4,5,6,7 and 8 -> Spine switch 1, 2, 3 and 4
L2	Port 9,10,11,12,13,14,15 and 16 -> Spine switch 1, 2, 3 and 4
L3	Port 17,18,19,20,21,22,23 and 24 -> Spine switch 1, 2, 3 and 4
L4	Port 25,26,27,28,29,30,31 and 32 -> Spine switch 1, 2, 3 and 4
L5	Port 33,34,35,36,37,38,39 and 40 -> Spine switch 1, 2, 3 and 4
L6	Port 41,42,43,44,45,46,47 and 48 -> Spine switch 1, 2, 3 and 4
L7	Port 49,50,51,52,53,54,55 and 56 -> Spine switch 1, 2, 3 and 4
L8	Port 57,58,59,60,61,62,63 and 64 -> Spine switch 1, 2, 3 and 4

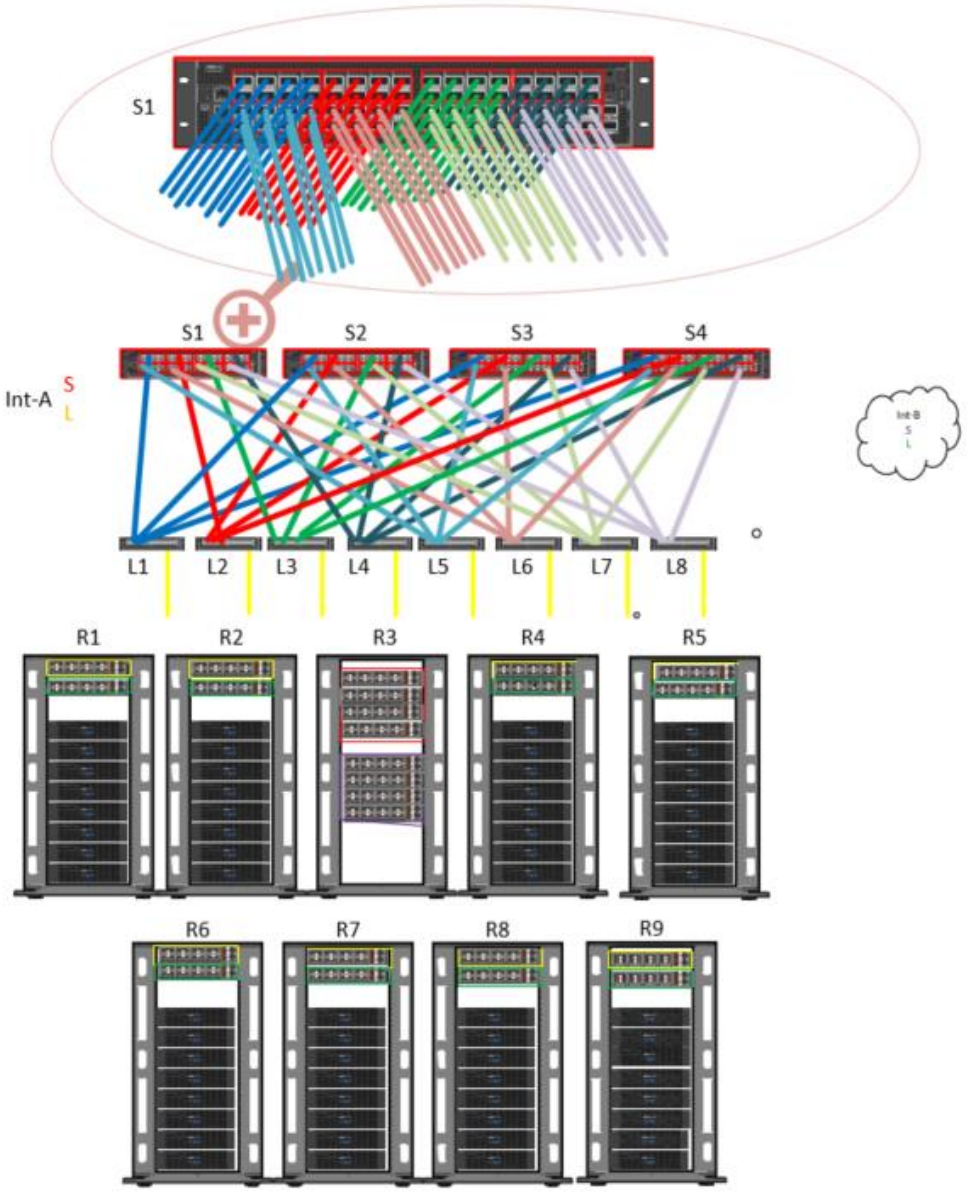

Figure 2 Sample configuration design of 252 nodes with 100 GbE back-end network

This example configuration includes the following:

- 24 Dell EMC Z9264 switches (12 per side)
  - 4 spine switches
  - 8 leaf switches
- 512 QSFP28 100 Gb uplink cables (8 uplink cables per leaf)
- 512 QSFP+ or MPO back-end cables
- 512 Optics (if MPO cables used)

Note: Since these nodes are 100 GbE back end, they require more uplinks than a typical 40 GbE back-end nodes. This example has 8 x 100 GbE uplinks per leaves.

Table 5 Legacy Dell EMC Isilon 100 GbE uplink cable options for Z9100

Cable type	Legacy Isilon model	Dell SKU number	Connector	Length
Pass Copper	851-0320	470-AEFW	QSFP28	1 m
Pass Copper	851-0321	470-AEGI	QSFP28	3 m
Pass Copper	851-0322	470-AEGO	QSFP28	5 m
Active Optical	851-0323	470-AEGJ	QSFP28	3 m
Active Optical	851-0324	470-AEGP	QSFP28	7 m
Active Optical	851-0325	470-AEFY	QSFP28	10 m
Active Optical	851-0326	470-AEGE	QSFP28	30 m

Note: Optics are added automatically when MPO cables are quoted.

Table 6 Legacy Isilon downlink 40 GbE cable options for performance nodes (F810, F800, H600, and H500)

Cable type (passive)	Legacy Isilon (model)	Dell SKU number	Connector	Length	EMC P/N
Copper	851-0253	470-AEGB	QSFP+	1 m	038-002-064-01
Copper	851-0254	470-AEGG	QSFP+	3 m	038-002-066-01
Copper	851-0255	470-AEGM	QSFP+	5 m	038-002-139-01
Optical	851-0274	407-BCIV	MPO	1 m	038-004-214
Optical	851-0275	407-BCIW	MPO	3 m	038-004-216
Optical	851-0276	407-BCJD	MPO	5 m	038-004-227
Optical	851-0224	407-BCIY	MPO	10 m	038-004-218
Optical	851-0225	407-BCJB	MPO	30 m	038-004-219
Optical	851-0226	407-BCJC	MPO	50 m	038-004-220
Optical	851-0227	407-BCIZ	MPO	100 m	038-004-221
Optical	851-0277	407-BCIX	MPO	150 m	038-000-139

Note: QSFP+ cables for Ethernet use do not require optics. MPO cables for Ethernet use require passive optics. The model is 851-0285 (019-078-046). MPO optics are added automatically when MPO cables are quoted and appear as a separate line item. Legacy Isilon downlink 10 GbE breakout cable options for archive nodes (A2000, A200, and H400)

Cable type	Legacy Isilon model	Length	Dell SKU number	Connector	EMC P/N	Reason
Copper	851-0278	1 m	470-AEGC	(1) QSFP to (4) SFP+	038-004-506-03	Breakout: 40Ge/10Ge (4)
Copper	851-0279	3 m	470-AEGH	(1) QSFP to (4) SFP+	038-004-507-03	Breakout: 40Ge/10Ge (4)
Copper	851-0280	5 m	470-AEGN	(1) QSFP to (4) SFP+	038-004-508-03	Breakout: 40Ge/10Ge (4)

Table 7 Dell EMC PowerScale supported cables and optics

PowerScale 10 GbE breakout cables	SKU
Dell Networking Cable, 40 GbE, QSFP+ to 4x10GbE SFP+, Passive Copper Breakout Cable, 1 Meter Customer Kit	470-AAVO
Dell Networking Cable, 40 40 GbE (QSFP+) to 4 x 10 GbE SFP+ Passive Copper Breakout Cable, 2 Meters, Customer Kit	470-ABXO
Dell Networking Cable 40 40 GbE (QSFP+) to 4 x 10 GbE SFP+ Passive Copper Breakout Cable, 3 Meters, Customer Install	470-AAXG
Dell Networking 40 40 GbE (QSFP+) to 4x10 GbE SFP+ Passive Copper Breakout Cable, 5 Meters, Customer Kit	470-AAXH
Dell Networking Cable, 40 40 GbE, QSFP+ to 4x10 GbE SFP+, Passive Copper Breakout, 7 Meters, Customer Kit	470-AAWU
PowerScale 25 GbE breakout cables	
Dell Networking Cable, 100 GbE QSFP28 to 4xSFP28 Passive DirectAttachBreakout Cable, 1 Meter, Customer Kit	470-ABPR
Dell Networking Cable,100 GbE QSFP28 to 4xSFP28 Passive DirectAttachBreakout Cable, 2 Meters, Customer Kit	470-ABQF
Dell Networking Cable,100 GbE QSFP28 to 4xSFP28 Passive DirectAttachBreakout Cable, 3 Meters, Customer Kit	470-ABQB
Dell Networking Cable QSFP28-4XSFP28, 25G, Passive Copper DAC, Breakout, 5 Meters, Customer Kit	470-AECY

PowerScale 100 GbE cables	
Copper cables	
Dell Networking Cable 100 GbE, QSFP28 to QSFP28, Passive Copper Direct Attach Cable, 1 Meter, Customer Kit	470-ABPY
Dell Networking Cable, 100 GbE QSFP28 to QSFP28, Passive Copper Direct Attach Cable, 2 MetersMeters, Customer Kit	470-ADDP
Dell Networking Cable, 100 GbE QSFP28 to QSFP28, Passive Copper Direct Attach Cable, 3 Meters, Customer Kit	470-ABQE
Dell Networking Cable, 100 GbE QSFP28 to QSFP28, Passive Copper Direct Attach Cable, 5 MetersMeters, Customer Kit	470-ABPU
Active optical cables	
Dell Networking Cable, QSFP28 to QSFP28, 100 GbE, Active Optical (Optics included),3 Meter, Customer Kit	470-ACLU
Dell Networking Cable, QSFP28 to QSFP28, 100 GbE, Active Optical (Optics included) Cable, 7 Meters, Customer Kit	470-ABPI
Dell Networking Cable, QSFP28 to QSFP28, 100 GbE, Active Optical (Optics included) Cable,10 MetersMeters, Customer Kit	470-ABPM
Dell Networking Cable, QSFP28 to QSFP28, 100 GbE, Active Optical (Optics included), 30 MetersMeters, Customer Kit	470-ABPJ
Optics	
Node Side: Dell EMC PowerEdge QSFP28 SR4 100 GbE 85C optic Customer Install	407-BCEX
Switch Side: Dell Networking, Transceiver, 100 GbE QSFP28 SR4, No FEC Capable, MPO, MMF, Customer Kit	407-BBWV
MPO/MPT passive optical cables	
Dell Networking MPO Type B Crossover Cable, Multi Mode Fiber OM4, 1 Meter, Customer kit	470-ABPO
Dell Networking MPO Type B Crossover Cable, Multi Mode Fiber OM4, 3 Meters, Customer kit	470-ABPN
Dell Networking MPO Type B Crossover Cable, Multi Mode Fiber OM4, 5 Meters, Customer kit	470-ABPQ
Dell Networking MPO Type B Crossover Cable, Multi Mode Fiber OM4, 7 Meters, Customer kit	470-ABPP

Dell Networking MPO Type B Crossover Cable, Multi Mode Fiber OM4, 10 Meters, Customer kit	470-ABPV
Dell Networking MPO Type B Crossover Cable, Multi Mode Fiber OM4, 25 Meters, Customer kit	470-ABPT
40 GbE for back-end compatibility with existing Isilon Gen 6 clusters	
Dell Networking Cable QSFP+ to QSFP+ 40 GbE Passive Copper Direct Attach Cable, 1 Meter, Customer Kit	470-AAVR
Dell Networking Cable, QSFP+ to QSFP+, 40 GbE Passive Copper Direct Attach Cable, 2 MetersMeters, Customer Kit	470-ACIW
Dell Networking Cable QSFP+ to QSFP+ 40 GbE Passive Copper Direct Attach Cable, 3 Meters, CK	470-AAWN
Dell Networking Cable QSFP+ to QSFP+ 40 GbE Passive Copper Direct Attach Cable, 5 Meters, CK	470-AAWE
Dell Networking Cable, QSFP+, 40 GbE Active Optical (no optics required), 3 Meters, , Customer Kit	470-ACOR
Dell Networking, Cable, QSFP+, 40 GbE, Active Fiber Optical, 10 Meters (No optics required), Customer Kit	470-AAZM
Optics	
Node Side: Mellanox, Transceiver, QSFP, 40 GbE, Short-Range, for use in Mellanox network adapter only, Customer Kit	407-BBOI
Switch Side: Dell Networking, Transceiver, 40 GbE QSFP+ SR4 Optics, 850nmWavelength, 100-150m Reach on OM3/OM4, CK	407-BBOZ

5 Back-end Ethernet switch monitoring

The PowerScale OneFS operating system provides an inbound back-end Ethernet switch monitoring through the CELOG system. CELOG monitors, logs, and reports important events and error conditions on the node and cluster including the back-end Ethernet switches. CELOG provides a single point from which notifications are generated, including sending alert emails and SNMP traps.

Table 8 Available CELOG alerts on the back-end Ethernet switches

Monitoring event	Alert	Applies to	Frequency check	Description
NODE_BE_SWITCH_MGMT_SERVICE	Failed to talk with back-end switch management {protocol} service on {ifname}	All Ethernet switches	Every 3 minutes	Determines the responsiveness of the back-end Ethernet switch management service.
NODE_BE_SWITCH_FAN	Fan failure detected in switch {switch_serial} on {int} fabric	All Ethernet switches	Every 11 minutes	Determines the state of the back-end Ethernet switch's FRUs.
NODE_BE_SWITCH_PWR	Power supply failure detected in switch {switch_serial} on {int} fabric	All Ethernet switches	Every 7 minutes	Determines the state of the back-end Ethernet switch FRUs.
NODE_BE_SWITCH_FABRIC	{int} fabric error detected: {desc} {id} {switch_serial}	Only Dell switches	Every 5 minutes	Determines if there is back-end Ethernet miscabling and other validation errors reported by the fabric.
NODE_BE_SWITCH_BANDWIDTH_DISCREPANCY	{int} fabric bandwidth discrepancy detected: {desc} {id} {switch_serial}	Leaf-Spine Only	Every 59 minutes	Determines if there is a discrepancy between total uplink and total downlink bandwidth on a leaf switch.
NODE_BE_FABRIC_BANDWIDTH_INCONGRUENCE	Fabric bandwidth incongruence detected: {desc}	Leaf-Spine Only	Every 61 minutes	Determines if there is incongruity between total uplink (trunk) bandwidth between int-a and int-b fabrics.

Besides these listed CELOG inbound alerts, users can also monitor the switches through SNMP MIBs. Dell Networking OS supports standard and private SNMP MIBs. For a list of MIBs supported in the OS10 version running on a switch, see the *OS10 Release Notes and Guide*.

A Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Storage technical documents and videos provide expertise that helps to ensure customer success on Dell EMC storage platforms.

A.1 Related resources

See the following related resources:

- Dell EMC PowerScale ToR network best practices
- Dell Switch OS Upgrade Guide
- PowerScale Leaf-Spine Installation Guide