Dell PowerScale: Ethernet Back-end Network Overview

April 2022

H16346.2

White Paper

Abstract

This white paper provides an introduction to the Ethernet back-end network for Dell PowerScale scale-out NAS.

Dell Technologies

Contents

Executive summary

Overview

This document provides design considerations for Dell PowerScale back-end (internal) networking. This back-end network, which is configured with redundant switches for high availability, acts as the backplane for the PowerScale cluster. This backplane enables each PowerScale node to act as a contributor in the cluster and provides node-to-node communication with a private, high-speed, low-latency network.

Revisions

Date	Description
June 2020	Content and template update
December 2021	Template update
April 2022	Content and template update

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Cris Banson

Note: For links to other documentation for this topic, see the PowerScale Info Hub.

Legacy Isilon back-end network

Overview

Before the introduction of the latest generation of PowerScale scale-out NAS storage platforms, inter-node communication in a Dell Isilon cluster has been performed using a proprietary, unicast (node-to-node) protocol known as Remote Block Manager (RBM). This inter-node communication uses a fast low-latency, InfiniBand (IB) network. This back-end network, which is configured with redundant switches for high availability, acts as the backplane for the Isilon cluster. This backplane enables each Isilon node to act as a contributor in the cluster and provides node-to-node communication with a private, high-speed, low-latency network. This back-end network uses Internet Protocol (IP) over IB (IPoIB) to manage the cluster. Sockets Direct Protocol (SDP) is used for all data traffic between nodes in the cluster.

Isilon platform back-end network option

Overview

PowerScale scale-out NAS storage platforms offer increased back-end networking flexibility. With PowerScale platforms, customers may choose to use either an InfiniBand or Ethernet switch on the back end. For customers electing to use an InfiniBand back-end network, the configuration and implementation will remain the same as previous generations of Isilon systems. Customers looking to add Isilon platforms (Isilon F800, H600, H5600, H500, H400, A200, and A2000) to an existing Isilon IB cluster that consisted of earlier Isilon systems, must configure the nodes with an InfiniBand back-end interface. The Ethernet back-end network option is only supported in clusters that consist

entirely of Ethernet back-end nodes. In these configurations, only Ethernet back-end switches that are provided and managed by Dell are supported.

The PowerScale Ethernet back-end connection options are detailed in the following table:

Table 1. Isilon Ethernet back-end options

Back-end option	Compute compatibility
10 GbE SFP+	Isilon H400, Isilon A200, or Isilon A2000
40 GbE QSFP+	Isilon F800/F810, Isilon H600, Isilon H5600, or Isilon H500

PowerScale platform back-end network option

Overview

The Dell PowerScale all-flash storage platforms, powered by the Dell PowerScale OneFS operating system, provide a powerful and simple scale-out storage architecture to speed up access to massive amounts of unstructured data. Powered by the new OneFS 9.x operating system, the all-flash PowerScale platforms are available in 3 product lines:

- PowerScale F200: Provides the performance of flash storage in a cost-effective form factor to address the needs of a wide variety of workloads.
- PowerScale F600: With NVMe drives, the F600 provides larger capacity with massive performance in a cost-effective compact form factor to power demanding workloads.
- PowerScale F900: Provides the maximum performance of all-NVMe drives in a cost-effective configuration to address the storage needs of demanding workloads. Each node is 2U in height and hosts 24 NVMe SSDs.

Recent additions to the hybrid storage platform include the PowerScale H700 and H7000.

- PowerScale H700: Provides optimum performance and value to support demanding file workloads. The H700 provides capacity up to 960 TB per chassis.
- PowerScale H7000: High performance, high-capacity hybrid platform with up to 1280 TB per chassis. The deep-chassis-based H7000 is an ideal to consolidate a range of file workloads on a single platform.

Recent additions to the archive storage platform include the PowerScale A300 and A3000.

- PowerScale A300: An ideal active archive storage solution that combines high performance, near-primary accessibility, value, and ease of use.
- PowerScale A3000: An ideal solution for high-performance, high-density, deep-archive storage that safeguards data efficiently for long-term retention.

The PowerScale Ethernet back-end connection options are detailed in the following table:

Table 2. PowerScale Ethernet back-end connection options

Back-end card options	PowerScale nodes
• 25 GbE SFP28 / 10 GbE SFP+	F200 H700, H7000 A300, A3000
• 100 GbE QSFP28+ / 40 GbE QSFP+	F600, F900 H700, H7000 A300, A3000

Note: The same NIC supports both 25 GbE and 10 GbE, respectively, for the F200, H700, H7000, A300, and A3000. The same NIC supports both 100 GbE and 40 GbE respectively for the F600, F900, H700, H7000, A300, and A3000. The NIC speed change is achieved by using different transceivers or cables.

Depending on the node type, we recommend using either the 100 GbE or 25 GbE network interface.

New-generation PowerScale platforms with different back-end speeds can connect to the same switch with Isilon nodes (Isilon F800, H600, H5600, H500, H400, A200, and A2000) and not see performance issues. For example, in a mixed cluster where an archive node (such as A200 or A2000) with 10 GbE on the back end and PowerScale nodes with 40 GbE or 100 GbE on the back end, both node types can connect to a 100 GbE back-end switch without affecting the performance of other nodes on the switch. The 100 GbE back-end switch will provide 100 GbE to the ports servicing the high-performance PowerScale nodes and 10 GbE to the archive or lower performing nodes using breakout cables.

Ethernet back end

In legacy Isilon systems, back-end data traffic uses SDP and IPoIB for management. SDP has fast failover and incorporates various InfiniBand-only features that ensures optimum performance. However, because SDP only works over InfiniBand, a new method was required to get optimal performance over the Ethernet back end. For this reason, the new generation of PowerScale platforms now uses RBM over TCP on the back-end switches.

RBM now uses TCP, and the TCP stack has been enhanced to provide the performance required to support the cluster communication. All the modifications of the TCP stack have been made while conforming to the industry standard specification of the stack. The back-end and front-end networks will use the same TCP stack and modifications to the performance of the back-end TCP stack should not affect TCP traffic on the front end. RBM over Ethernet will still provide fast failover.

Dell switch support for Ethernet back end

Dell Ethernet switches to be used for the Isilon back end as a top-of-rack solution (TOR).

- S5232-ON
- Z9264-ON
- Z9100-ON
- S4112-ON
- S4148F-ON

Table 3. Dell Ethernet switches

Vendor	Model	Back-end ports	Port type	Rack units	Network ports
Dell	S5232-ON	32	All 100 GbE	1	32 x 100 GbE, 32 x 40 GbE, 124 x 10 GbE (with breakout cables), 124 x 25 GbE (with breakout cables)
Dell	Z9264-ON	64	All 100 GbE	2	64 x 100 GbE, 64 x 40 GbE, 128 x 10 GbE (with breakout cables), 128 x 25 GbE (with breakout cables)
Dell	Z9100-ON	32	All 100 GbE	1	32 x 100 GbE, 32 x 40 GbE, 128 x 10 GbE (with breakout cables), 128 x 25 GbE (with breakout cables)
Dell	S4112F-ON	15	12 port 10 GbE, 3 port 100 GbE	1	12 x 10 GbE, 3 x 100 GbE (with breakout cables, connect 12 x 10 GbE nodes using the 3 x 100 GbE ports)
Dell	S4148F-ON	48	48 port 10 GbE, 2 port 40 GbE	1	48 x 10 GbE

Note: Supported Dell Ethernet back-end switches can be found in the PowerScale OneFS Product Availability Guide and PowerScale Node Site Preparation and Planning Guide.

These Ethernet switches will be zero-touch back-end switches that are used for inter-node communication in an Isilon cluster, and those are typically what are called plug-and-play switches.

Note: The Dell supported switches are shipped with a fixed configuration and additional customer configuration is not necessary or allowed. Any switch updates (DNOS, firmware) must be performed by a Dell qualified resource.

The S5232-ON is a fixed 1U Ethernet switch is a 32 port 100 GbE multi-rate spine/leaf switch. It supports multiple interface types (32 ports of 100 GbE or 40 GbE in QSFP28, or 124 ports of 25 GbE or 10 GbE with breakout) for maximum flexibility. Port 32 does not support breakout cables.

The Z9264-ON is a fixed 2U Ethernet spine/leaf switch which provides industry-leading density of either 64 ports of 100 GbE or 40 GbE in QSFP28 or 128 ports of 25 GbE or 10 GbE by breakout. Breakout cables are only used in the odd-numbered ports and using one in odd-numbered port disables the corresponding even-numbered port.

The Z9100-ON is a fixed 1U Ethernet spine/leaf switch which can accommodate high port density (lower and upper RUs). It also can accommodate multiple interface types (32 ports of 100 GbE or 40 GbE in QSFP28, or 128 ports of 25 GbE or 10 GbE with breakout) for maximum flexibility.

The S4148F-ON is a 10 GbE (48 ports) top-of-rack, aggregation-switch, or router product. It aggregates 10 GbE server or storage devices and provides multispeed uplinks for maximum flexibility and simple management. The switch features 48x10 GbE SFP+ ports that can be used for back-end connectivity.

The S4112F-ON supports 10/100GbE with 12 fixed SFP+ ports to implement 10 GbE and three fixed 100 GbE QSFP28 ports to implement 4x10 or 4x25 using breakout. A total of 24 10 GbE connection including the three fixed QSFP28 ports using 4x10 GbE breakout cables.

Note: These switches are qualified to be used with currently available network cables (MPO, LC, QSFP+, SFP+, and breakout cables).

Dell S5232-ON, Z9100-ON, and Z9264-ON switches have been qualified for leaf/spine deployments with PowerScale/Isilon. See the Dell PowerScale Leaf-Spine Network Best Practices and Dell PowerScale Leaf-Spine Installation Guide for more information.

Configuration and monitoring

When installing a new Isilon cluster, the Configuration Wizard has not changed. It still prompts you for int-a, int-b, and failover range. All configuration and setup steps will be the same regardless of InfiniBand or Ethernet option selected.

The following figures show the relative positioning of back-end ports provided in the Compute Assembly for each Dell PowerScale/Isilon platform node type.

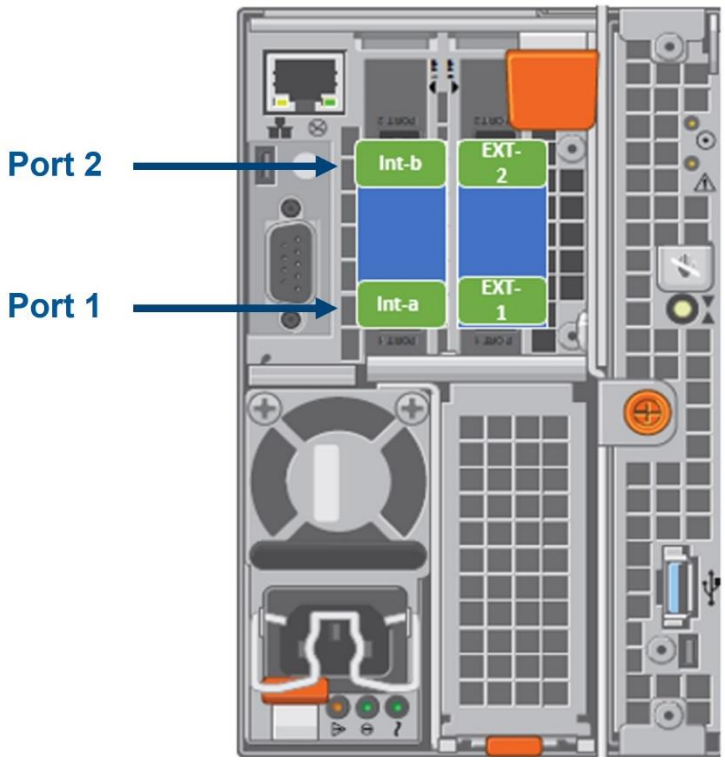


Figure 1. **F800, F810, H600, H5600, H500, H400, A200, and A2000. H700, H7000, A300, and A3000: back-end ports**

Figure 2. **F200: back-end ports**

Figure 3. F600: back-end ports

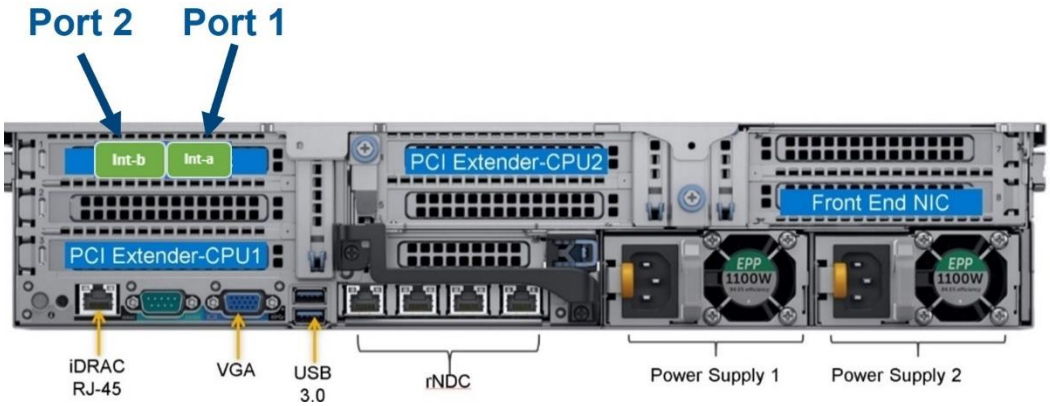


Figure 4. F900: back-end ports

The following table provides configuration information for the back-end ports in PowerScale platforms:

Table 4. Configuration for int-a, int-b, and failover

Setting	Description
Int-a network setting	The network settings used by the int-a network. The int-a network is used for communication between nodes.
Netmask	The int-a network must be configured with IPv4.
IP range	The int-a network must be on a separate or distinct subnet from an int-b/failover network.
Int-b and failover network setting	The network settings used by the optional int-b/failover network.
Netmask	The int-b network is used for communication between nodes and provides redundancy with the int-a network.
IP range	The int-b network must be configured with IPv4.
Failover IP range	The int-a, int-b, and failover networks must be on separate or distinct subnets.

The monitoring capabilities on PowerScale/Isilon back-end switches correspond to the field replaceable unit (FRU) components such as power supply, the fan, or others. Protocol and performance monitoring capability is not provided.

Note: Customers should not attempt to alter the back-end network configurations provided by Dell. Any attempt to do so can result in a cluster-wide outage.

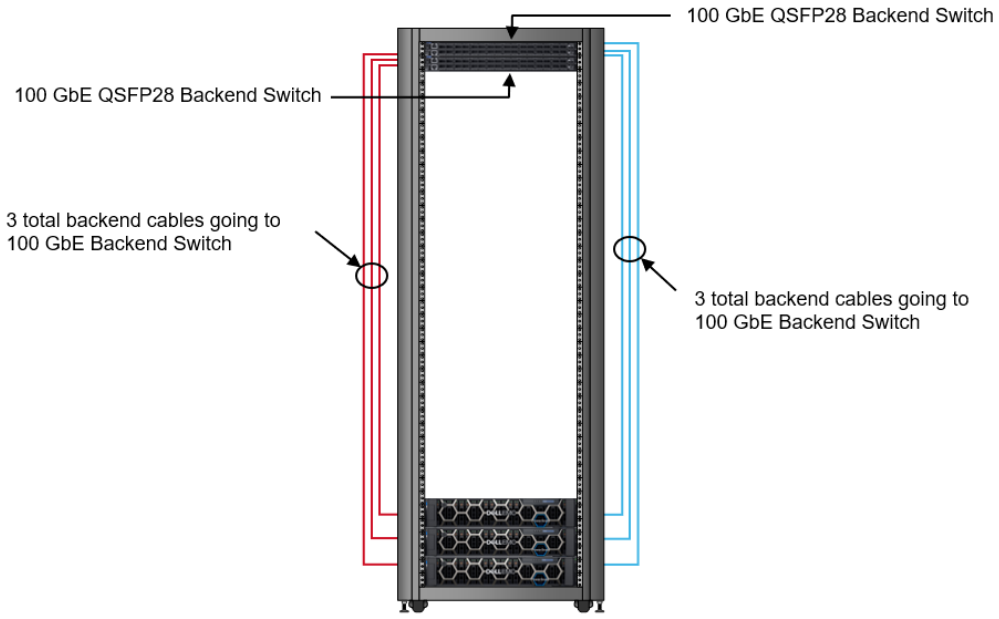
For SNMP capabilities, customer may send an SNMP alert through the CELOG system. In today's back-end Ethernet world, we no longer have **opensm** topology files to view all connected devices on the back-end network. If you want to know what is connected to the fabric of back-end Ethernet (int-a or int-b) you may use the **isi_dump_fabric int-a** (or int-b) command.

Sample configurations

Following are examples of cluster configurations with varying node types and the corresponding back-end connectivity infrastructure.

Example 1: All performance Dell PowerScale 100 GbE back end

When using performance nodes (for example, F600 and F900), the back end must be 100/40 GbE (10 GbE is not supported).

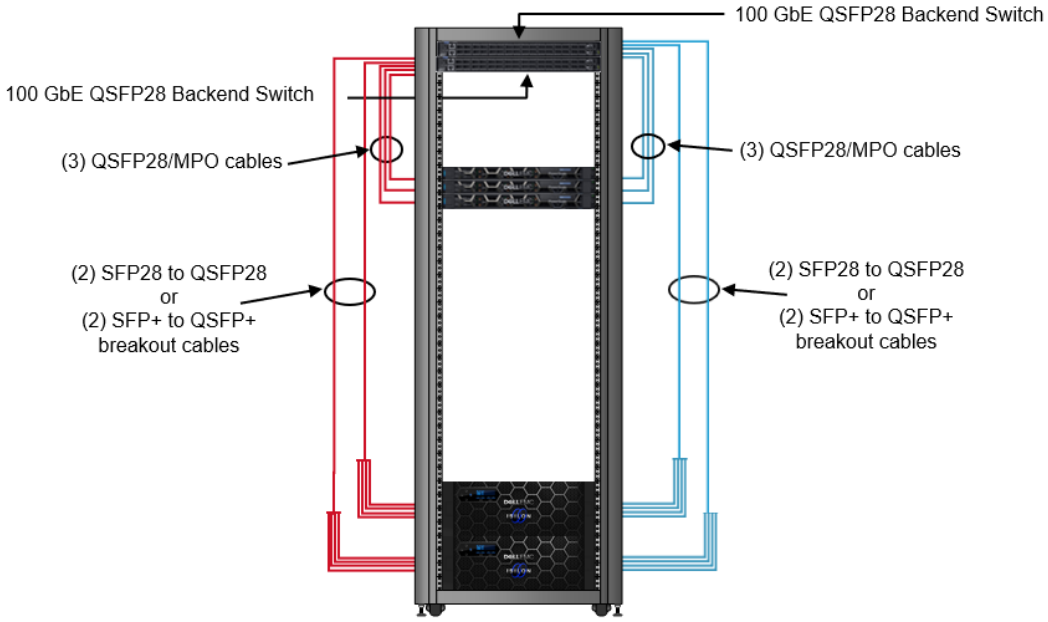


In this example, your configuration will include:

- (2) 100 GbE back-end switches
- (6) QSFP+/MPO back-end cables
- (6) Optics (If MPO cables used)

Example 2: Mixed environment of PowerScale 100 GbE and 25/10 GbE back end

When mixing performance and archive nodes, use a 100 GbE infrastructure with 100 GbE connections to the performance nodes. Also, use 8 x 10 GbE or 4 x 25 GbE breakout cables (depending on switch support) to the archive nodes.

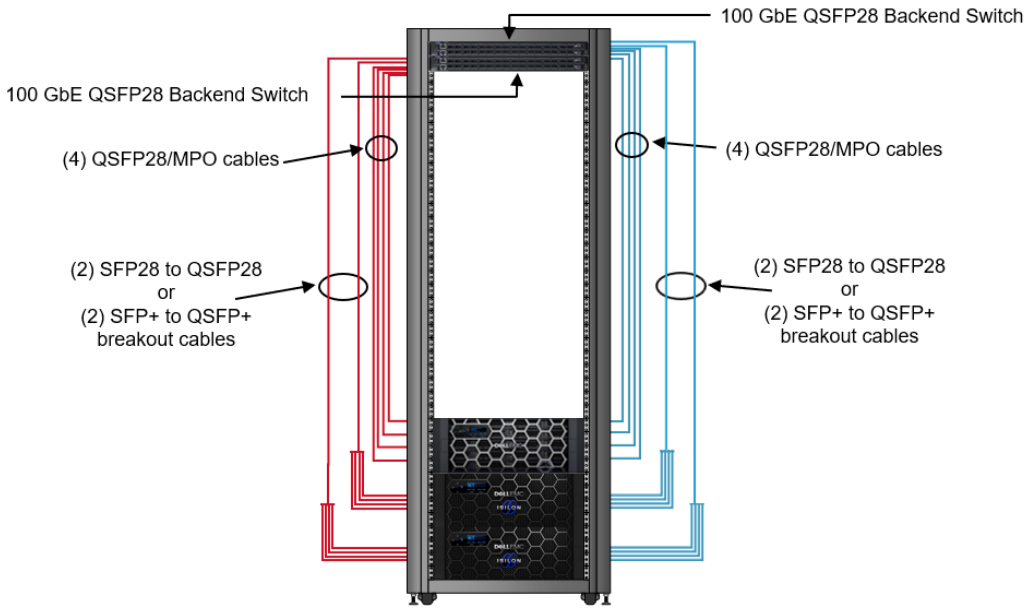


In this example, your configuration will include:

- Two 100/40 GbE back-end switches
- Six QSFP+/MPO back-end cables
- Six optics (If MPO cables used)
- Four SFP28 to QSFP28 or 4 SFP+ to QSFP+ breakout cables

Example 3: Mixed environment of PowerScale 100 GbE and 25/10 GbE back end

When mixing hybrid and archive nodes, use a 100 GbE infrastructure with 100 GbE connections to the hybrid nodes and 4 x 25 GbE or 4 x 10 GbE breakout cables (depending on node type) to the archive nodes.

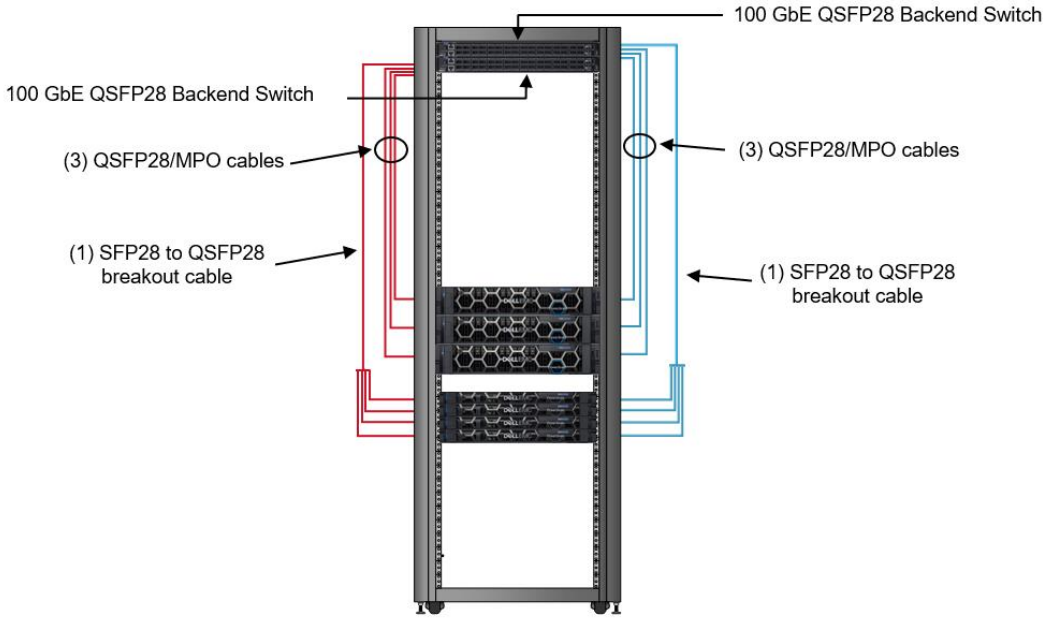


In this example, your configuration will include:

- Two 100/40 GbE back-end switches
- Eight QSFP28/MPO back-end cables
- Eight optics (If MPO cables used)
- Four SFP28 to QSFP28 or four SFP+ to QSFP+ breakout cables

Example 4: All PowerScale F200 and F600/F900

When mixing F900/F600 and F200 nodes, use a 100 GbE infrastructure with 100 GbE connections to the F900/F600 nodes and 4 x 25 GbE breakout cables to the F200 nodes.



In this example, your configuration will include:

- Two 100/40 GbE back-end switches
- Six QSFP28/MPO back-end cables
- Six optics (If MPO cables used)
- Two SFP28 to QSFP28 breakout cables

Supported Ethernet back-end switches

The following switches are supported for PowerScale back-end connectivity.

Table 5. Supported back-end switches

Vendor	Model	Switch speed
Dell	S5232-ON	100 GbE
Dell	Z9264-ON	100 GbE
Dell	Z9100-ON	100 GbE
Dell	S4112F-ON	10/100 GbE
Dell	S4148F-ON	10 GbE
Celestica	D4040	40 GbE
Arista	DCS-7308	40 GbE
Celestica	D2024	10 GbE

Vendor	Model	Switch speed
Celestica	D2060	10 GbE
Arista	DCS-7304	10 GbE

Note: Supported Ethernet back-end switches are listed in the PowerScale OneFS Product Availability Guide.

Cable options

The following table lists the back-end cable options for the PowerScale/Isilon platform nodes.

Table 6. Supported cable options

Cable type	Connector	Length	Description	Supported back-end switch	Node types		
					All-Flash	Hybrid	Archive
Copper	QSFP+	1, 2, 3, 5 m	40 GbE cable; QSFP+ to QSFP+	40/100 GbE switch	F900, F600, F800, F810	H600, H500, H5600 H700, H7000	A300, A3000 (with 100 Gb / 40 Gb Ethernet back-end card)
Optical	MPO	1, 3, 5, 7, 10, 25, 30, 50, 100, 150 m	100 GbE/40 GbE Ethernet MPO to MPO Optical (optics required)	40/100 GbE switch	F900, F600, F800, F810	H600, H500, H5600 H700, H7000	A300, A3000 (with 100 Gb / 40 Gb Ethernet back-end card)
Optical	QSFP+	3, 10 m	40 GbE Ethernet Cable QSFP+ to QSFP+	40/100 GbE switch	F900, F600	H700, H7000	A300, A3000 (with 100 Gb / 40 Gb Ethernet back-end card)
Optical	QSFP28	3, 7, 10, 30 m	100 GbE Ethernet Cable QSFP28 to QSFP28	100 GbE switch	F900, F600	H700, H7000	A300, A3000 (with 100 Gb / 40 Gb Ethernet back-end card)
Copper	QSFP28	1, 3, 5 m	100 GbE Ethernet Cable QSFP28 to QSFP28	100 GbE switch	F900, F600	H700, H7000	A300, A3000 (with 100 Gb / 40 Gb Ethernet back-end card)
Copper	(1) QSFP+ to (4) SFP+	1, 3, 5, 7 m	Breakout: 40 GbE /10 GbE (4)	Switch(es) that support breakout cables	F200	H400	A200, A2000 A300, A3000

Cable type	Connector	Length	Description	Supported back-end switch	Node types		
					All-Flash	Hybrid	Archive
Optical	(1) QSFP+ to (4) SFP+	10, 30 m	Breakout: 40 Gbe /10 GbE (4)	Switch(es) that support breakout cables	F200	H400	A200, A2000 A300, A3000
Copper	(1) QSFP28 to (2) SFP28	1, 2, 3, 5 m	Breakout: 100 GbE / 25 GbE (4)	Switch(es) that support breakout cables	F200		A300, A3000
Copper	SFP+	1, 2, 3, 5, 7 m	10 GbE Ethernet Cable SFP+ to SFP+	10 GbE Switch	F200	H400 H700, H7000	A200, A2000 A300, A3000
Optical	LC	1, 2, 3, 5, 10, 30, 50, 100, 150 m	10/25 GbE Ethernet Cable LC to LC (optics required)	10 GbE Switch	F200	H400 H700, H7000	A200, A2000 A300, A3000

Note: Breakout cables do not require optics. QSFP+ cables for Ethernet use do not require optics. MPO cables for Ethernet use require passive optics. The optics for the LC-LC cables are bundled with the cable bill of materials (BOM).

Appendix: Technical support and resources

Technical support

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

The Dell Technologies Info Hub provides expertise that helps to ensure customer success on Dell storage platforms.

Related resources

Related resources include:

- Dell PowerScale Leaf-Spine Network Best Practices:

  https://www.dellemc.com/resources/en-us/asset/white-papers/products/storage/h17682-dell-emc-powerscale-leaf-spine-network-best-practices.pdf

- Dell PowerScale Leaf-Spine Installation Guide:

  https://dl.dell.com/content/manual55426266--leaf-spine-cluster-installation-guide.pdf