CORNING

Applications Engineering Note

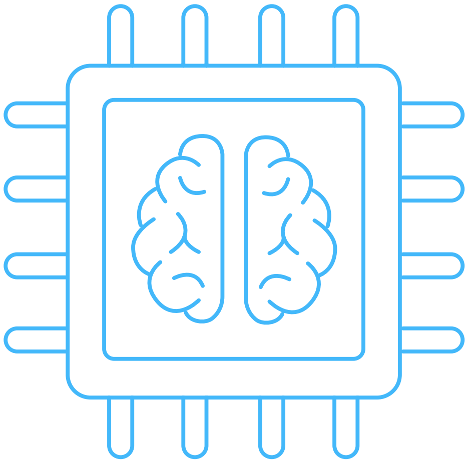
NVIDIA 800G InfiniBand and Ethernet Connectivity Solutions Utilizing Structured Cabling.

AEN185

Introduction

As we swiftly advance into the age of Generative AI, having a tailored connectivity solution becomes paramount. We're here to provide custom solutions that keep you ahead of the curve, with pricing options and free design assistance readily available. In this rapidly evolving landscape, timing is everything.

The world of AI will demand tailored connectivity solutions, future-proofed and supplied at a speed never before seen in the Data Center Industry.



Ready to purchase?
Looking for FREE design assistance?
Need a custom solution for your build?

Click here

Let us connect you to our HSDC Innovation Team.
Time is a resource you can not afford to waste.

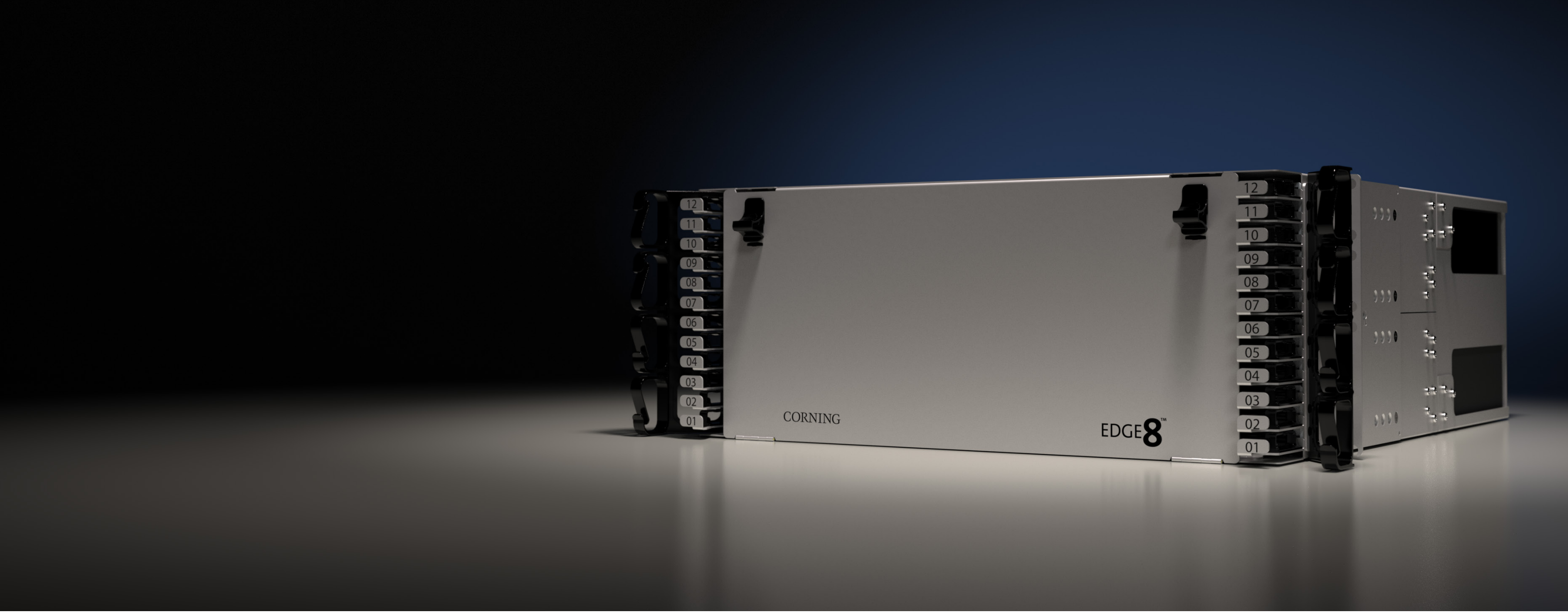


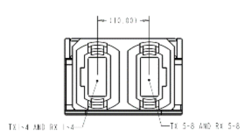
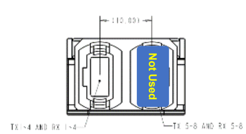
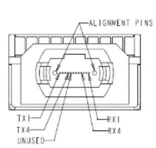
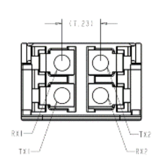
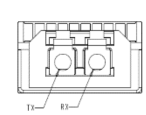
Table of Contents:

1. Understanding the transceiver types, port breakouts and cabling scenarios.

This Application Engineering Note will discuss the different available fiber optic connectivity to work with 200G, 400G and 800G transceivers, and breakout options within the same rack or row, and across the data center utilizing 400G NDR InfiniBand Quantum-2 and 400GbE Spectrum-4 Ethernet (400G IB/EN) switch capabilities.

With the introduction of 800G, NVIDIA optical transceiver modules utilize a new Twin (dual) MPO-8/12 transceiver interface utilizing 2x 8-fibers. Corning's EDGE8® solution is designed to support both single-mode and multimode optical interfaces based on the use of 2, 4, 8, and 16 fibers at the transceiver.

The following is a partial list of NVIDIA transceivers by connector type:

Twin MPO-8/12 APC Interface (OSFP)		MPO-8/12 APC Interface (QSFP112)	Twin LC Duplex UPC Interface (OSFP)	LC Duplex UPC Interface (QSFP-DD)
Single Mode [a] **800G-2xDR4 OSFP** [c] **2x 8-fiber transceiver** MMS4X00-NM MMS4X00-NS MMS4X00-NS-FLT [d] MMS4X00-NS-T	Single Mode [a] **400G-DR4 OSFP** **8-fiber transceiver** **Uses 1 out of 2 ports** MMS4X00-NS400 MMS4X00-NS400-T **200G-DR4 OSFP** [e] **4/8-fiber transceiver** **Uses 1 out of 2 ports** MMS4X00-NS400 MMS4X00-NS400-T	Single Mode [a] **400G-DR4 QSFP112** **8-fiber transceiver** MMS1X00-NS400 MMS1X00-NS400-T **200G-DR4 QSFP112** [e] **4/8-fiber transceiver** MMS1X00-NS400 MMS1X00-NS400-T	Single Mode **800G-2xFR4 OSFP** **2x 2-fiber transceiver** MMS4X50-NM MMS4X50-NM-T	Single Mode **400G-FR4 QSFP-DD** **2-fiber transceiver** MMS1V50-WM MMS1V50-WM-T
Multimode [b] **800G-2xSR4 OSFP** [c] **2x 8-fiber transceiver** MMA4Z00-NS MMA4Z00-NS-FLT [d] MMA4Z00-NS-T	Multimode [b] **400G-SR4 OSFP** **8-fiber transceiver** **Uses 1 out of 2 ports** MMA4Z00-NS400 MMA4Z00-NS400-T **200G-SR4 OSFP** [e] **4/8-fiber transceiver** **Uses 1 out of 2 ports** MMA4Z00-NS400 MMA4Z00-NS400-T	Multimode [b] **400G-SR4 QSFP112** **8-fiber transceiver** MMA1Z00-NS400 MMA1Z00-NS400-T **200G-SR4 QSFP112** [e] **4/8-fiber transceiver** MMA1Z00-NS400 MMA1Z00-NS400-T		
				

a) MPO-8/12 APC Single Mode optics are denoted by a yellow-colored pull tab and yellow-colored optical fiber. Green plastic shell on the MPO-8/12 APC connector denotes Angled Polish Connector and is not compatible with Ultra-flat Polished Connectors (UPC) used with slower line rate transceivers.

b) Multimode optics are denoted by a tan-colored pull tab and aqua-colored optical fiber. Green plastic shell on the MPO-8/12 APC connector denotes Angled Polish Connector and is not compatible with aqua colored shell for Ultra-flat Polished Connectors (UPC) for HDR.

c) Please note that in some documentation, 800G transceivers may appear as 800G-DR8 instead of 800G-2xDR4 and 800G-SR8 instead of 800G-2xSR4. However, the connectivity footprint is represented by 2x 8-fiber MPO-8/12 APC. Please follow NVIDIA's part number as the main reference.

d) The card I/Os is routed internally to four 800G Twin-port OSFP.

e) A 400G transceiver version will be able to support 200G utilizing a splitter cable (Y-Harness), activating two of the four lanes (4 out of 8-fibers) in the 400G transceiver creating a 200G device. This configuration is represented as "4/8-fiber" in the table above.

f) Transceiver part numbers ending with "-T" refer to Ethernet versions.

**For more information on NVIDIA components and design, please review the Annex 3 with the references to NVIDIA Overview White papers.*

1.1. Scenario 1 - 800G and 400G - Server to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Point-to-Point Cabling.

Application: Quantum-2 InfiniBand or Spectrum-4 Ethernet to a) Quantum-2 InfiniBand or Spectrum-4 Ethernet; b) ConnectX-7 and Bluefield-3; c) DGX H100/Cedar-7 links.

Scenario 1 is primarily used for Point-to-Point cabling applications, which involves connecting the server within a Scalable Unit to the Leaf Switches located within the same rack or row. In some cases, this type of cabling can also be used to connect different switches, such as Leaf to Spine or Spine to Core. However, it is not recommended to use Point-to-Point cabling if those switches are physically located in different areas within the data center.

Please refer to section 1.7, Table 8, to review the list of transceivers applicable for Use cases A to C in Scenarios 1 and 2. To learn more about Scalable Units and how to build a DGX SuperPOD, please refer to section 2 in this document.

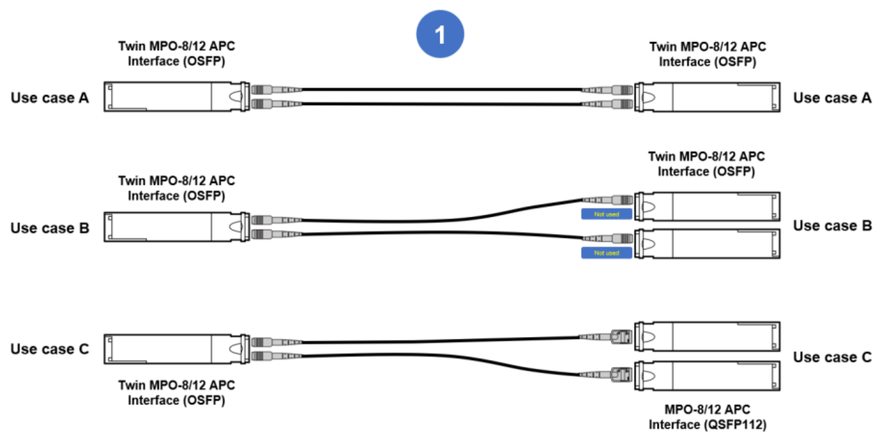


Figure 1. Use cases for Point-to-Point cabling utilizing 800G and 400G transceivers working with MPO-8/12 APC.

Item	Reference Part Number	OS2 Part Number (Americas)	OS2 Part Number (EMEA and APJ)	OM4 Part Number (Americas)	OM4 Part Number (EMEA and APJ)	Description
1	**Corning** [a]	G-BND32-E8E8G-PN000-xxxF	G-BND32-E8E8G-LZ000-xxxM	G-BND32-E9E9Q-PN000-xxxF	G-BND32-E9E9Q-LZ000-xxxM	32x 8-F Mesh Bundle, MTP® APC (non-pinned) to MTP APC (non-pinned), 78-in (2-m) legs, Type-B polarity, xxxF (feet) or xxxM (meters)
		or				
		JE8E808GE8-NBxxxF	JE8E808GEZ-NBxxxM	JE9E908QE8-NBxxxF	JE9E908QEZ-NBxxxM	EDGE8®, 8-F Jumper, MTP® APC (unpinned) to MTP APC (unpinned), Type-B polarity, xxxF (feet) or xxxM (meters)
	or					
	NVIDIA [b]	MFP7E30-Nxxx		MFP7E10-Nxxx		NVIDIA MPO-12/APC-to-MPO12/APC (8 fibers), Straight Crossover Fibers, Fiber Cable Product, Type-B polarity, xxx indicate length in meters

Notes:

a) Corning cables in the Americas use Plenum cable, while EMEA/APJ uses LSZH™/CPR rated cable. Single jumper lengths are available from 1 to 300 meters. Bundled jumpers use a meshed sleeve. Bundled lengths are available from 10 to 300 meters (furcation-to-furcation) in increments of 5 meters. Bundles are available in 16, 32 legs, and custom leg quantities with straight or staggered legs. Lengths in meters are also available for the Americas.

b) NVIDIA cables utilize a dual rated OFNR/LSZH jacket. SMF cable lengths available in 3, 5, 7, 10, 15, 20, 30, 50, 100 and 150 meters. MMF cable lengths available in 3, 5, 7, 10, 15, 20 ,25, 30, 35, 40 and 50 meters.

c) Both Corning and NVIDIA cables support InfiniBand, Ethernet and NVLink protocols.

d) Please review Corning's polarity drawings in Annex 2.

Table 2. Scenario 1 - 800G and 400G using Point-to-Point Cabling with MPO-8/12 APC. Part Number Scheme.

1.2. Scenario 2 - 800G and 400G - Switch to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Structured Cabling Across DC with Trunk.

Application: Quantum-2 InfiniBand or Spectrum-4 Ethernet to a) Quantum-2 InfiniBand or Spectrum-4 Ethernet; b) ConnectX-7 and Bluefield-3; c) DGX H100/Cedar-7 links.

In Scenario 2, the implementation of Structured Cabling utilizes an optical fiber trunk as a backbone. This application is mostly used to connect different switches, such as Leaf to Spine, and can also be implemented to connect Spine to Core. It is the recommended option to follow when two different active devices are physically located in different areas within the data center.

Please refer to section 1.7, Table 8, to review the list of transceivers applicable for Use cases A to C in Scenarios 1 and 2. To learn more about Scalable Units and how to build a DGX SuperPOD, please refer to section 2 in this document.

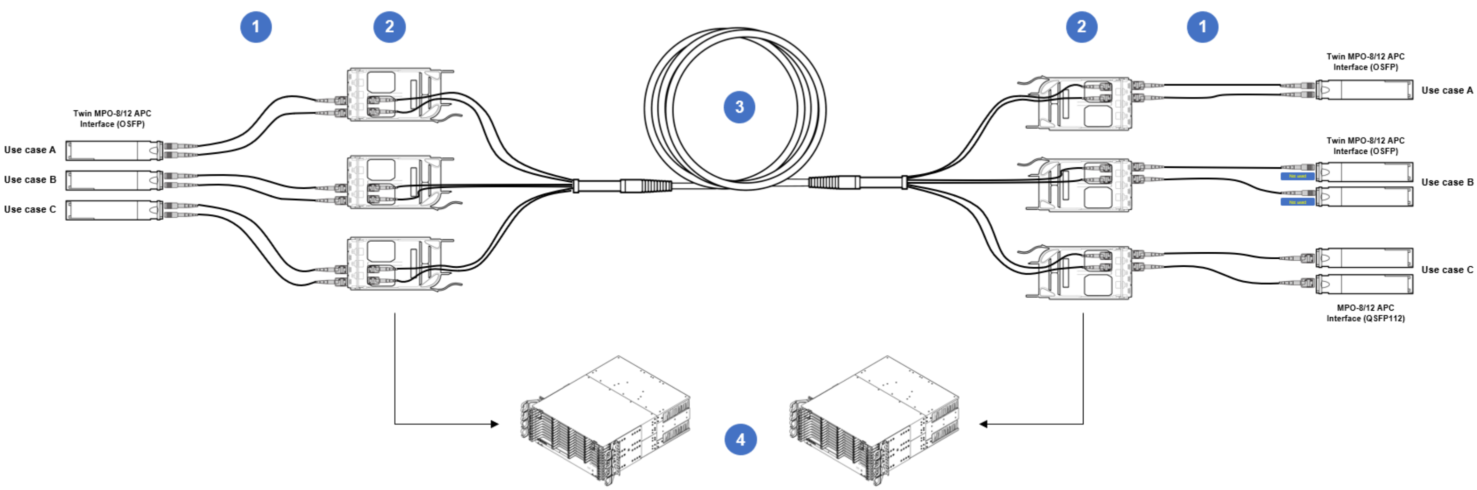


Figure 2. Use cases for Structured Cabling utilizing 800G and 400G transceivers working with MPO-8/12 APC.

Item	Reference Part Number	OS2 Part Number (Americas)	OS2 Part Number (EMEA and APJ)	OM4 Part Number (Americas)	OM4 Part Number (EMEA and APJ)	Description
1	Corning [a]	G-BND16-E8E8G-PN000-xxxF	G-BND16-E8E8G-LZ000-xxxM	G-BND16-E9E9Q-PN000-xxxF	G-BND16-E9E9Q-LZ000-xxxM	16x 8-F Mesh Bundle, MTP® APC (non-pinned) to MTP APC (non-pinned), 78-in (2-m) legs, Type-B polarity, xxxF (feet) or xxxM (meters)
				or		
		JE8E808GE8-NBxxxF	JE8E808GEZ-NBxxxM	JE9E908QE8-NBxxxF	JE9E908QEZ-NBxxxM	EDGE8®, MTP® APC (non-pinned) to MTP APC (non-pinned) 8-F Jumper, TIA-568 Type-B polarity, xxxF (feet) or xxxM (meters)
2		EDGE8-CP32-V1	EDGE8-CP32-V1	EDGE8-CP32-V3	EDGE8-CP32-V3	EDGE8 32-F MTP Adapter Panel, (4-port)
3		GE7E7E4GPNDDU xxxF[b]	GE7E7E4GLZDDU xxxM[b]	GE2E2E4QPNDDU xxxF[b]	GE2E2E4QLZDDU xxxM[b]	EDGE8, MTP Trunk , 144-F, MTP APC (pinned) to MTP APC (pinned), 33-in (840-mm) legs, Type-B polarity, pulling grip on first end only, xxxF (feet) or xxxM (meters)
4		EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	Please refer to Annex 1 to choose the best option for your application

Notes:

a) Corning cables in the Americas use Plenum cable, while EMEA/APJ uses LSZH™/CPR rated cable. Single jumper lengths are available from 1 to 300 meters. Bundled jumpers use a meshed sleeve. Bundled lengths are available from 10 to 300 meters (furcation-to-furcation) in increments of 5 meters. Bundles are available in 16, 32 legs, and custom leg quantities with straight or staggered legs. Lengths in meters are also available for the Americas.

b) Trunks are available in fiber counts of 8 to 288 fibers.

c) Corning cables support InfiniBand, Ethernet and NVLink protocols.

d) Please review Corning's polarity drawings in Annex 2.

Table 3. Scenario 2 - 800G and 400G using Structured Cabling with MPO-8/12 APC. Part Number Scheme.

1.3. Scenario 3 - 800G and 200G - Server to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Point-to-Point Cabling.

Application: Quantum-2 InfiniBand or Spectrum-4 Ethernet to ConnectX-7 / Bluefield-3.

Scenario 3 is used for Point-to-Point cabling applications, where a 400G transceiver version will be able to support 200G by utilizing a splitter cable (Y-Harness) that activates two of the four lanes (4 out of 8 fibers) in the 400G transceiver, effectively creating a 200G device. Point-to-Point cabling is recommended only if the active devices are within the same rack or row. However, if the active devices are physically located in different areas within the data center, then Structured Cabling is the recommended option.

Please refer to section 1.7, Table 9, to review the list of transceivers applicable for Use cases A and B in Scenarios 3 and 4. To learn more about Scalable Units and how to build a DGX SuperPOD, please refer to section 2 in this document.

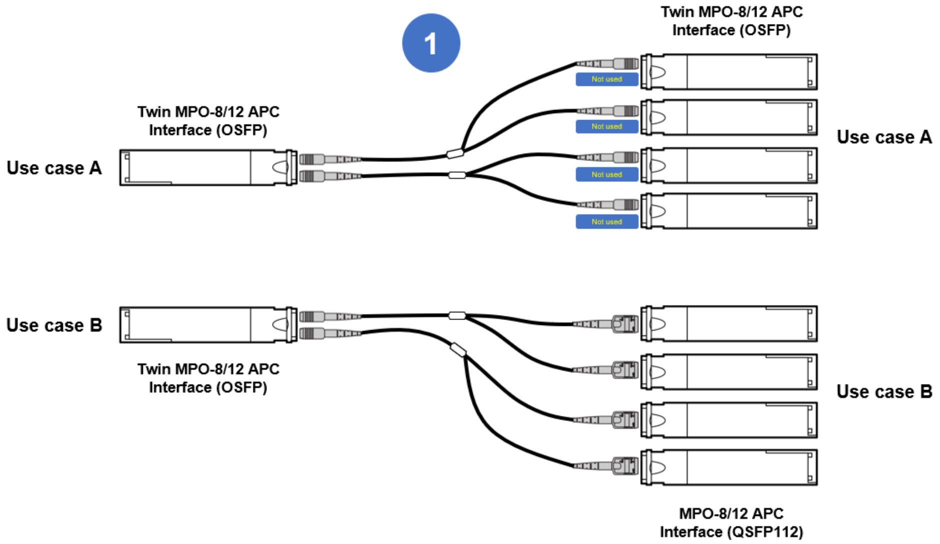


Figure 3. Use cases for Point-to-Point cabling utilizing 800G and 200G transceivers working with MPO-8/12 APC.

Item	Reference Part Number	OS2 Part Number (Americas)	OS2 Part Number (EMEA and APJ)	OM4 Part Number (Americas)	OM4 Part Number (EMEA and APJ)	Description
1	Corning [a]	HE8E808GPH-LBxxxF	HE8E808GLZ-LBxxxM	HE9E908QPH-LBxxxF	HE9E908QLZ-LBxxxM	EDGE8®, 8-F Y-Harness, MTP® APC (unpinned) to two 4-F (unpinned) MTP APC, 36-in (910-mm) breakout leg length, Type-B polarity, xxxF (feet) or xxxM (meters)
	or					
	NVIDIA [b]	MFP7E40-Nxxx		MFP7E20-Nxxx		NVIDIA MPO-12/APC-to-MPO12/APC (8 fibers), 4-channel-to-two 2-channel splitter fiber cable, Type-B polarity, xxx indicate length in meters

Notes:

a) Corning cables in the Americas use Plenum cable, while EMEA/APJ uses LSZH™/CPR rated cable.Y-Harness lengths available from 1 to 60 meters. Lengths in meters are also available for the Americas.

b) NVIDIA splitter cables utilize a dual rated OFNR/LSZH jacket. SMF and MMF splitter cable lengths available in 3, 5, 7, 10, 15, 20, 30 and 50 meters.

c) Both Corning and NVIDIA cables support InfiniBand, Ethernet and NVLink protocols.

d) A 400G transceiver version will be able to support 200G utilizing a splitter cable (Y-Harness), activating two of the four lanes (4 out of 8-fibers) in the 400G transceiver creating a 200G device.

e) Please review Corning's polarity drawings in Annex 2.

Table 4. Scenario 3 - 800G and 200G using Point-to-Point Cabling with MPO-8/12 APC. Part Number Scheme.

1.4. Scenario 4 - 800G and 200G - Switch to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Structured Cabling Across DC with Trunk.

Application: Quantum-2 InfiniBand or Spectrum-4 Ethernet to ConnectX-7 / Bluefield-3.

Scenario 4 is used with Structured Cabling components, where a 400G transceiver version will be able to support 200G by utilizing a splitter cable (Y-Harness) that activates two of the four lanes (4 out of 8 fibers) in the 400G transceiver, effectively creating a 200G device. In this scenario, as the active devices are physically located in different areas within the data center.

Please refer to section 1.7, Table 9, to review the list of transceivers applicable for Use cases A and B in Scenarios 3 and 4. To learn more about Scalable Units and how to build a DGX SuperPOD, please refer to section 2 in this document.

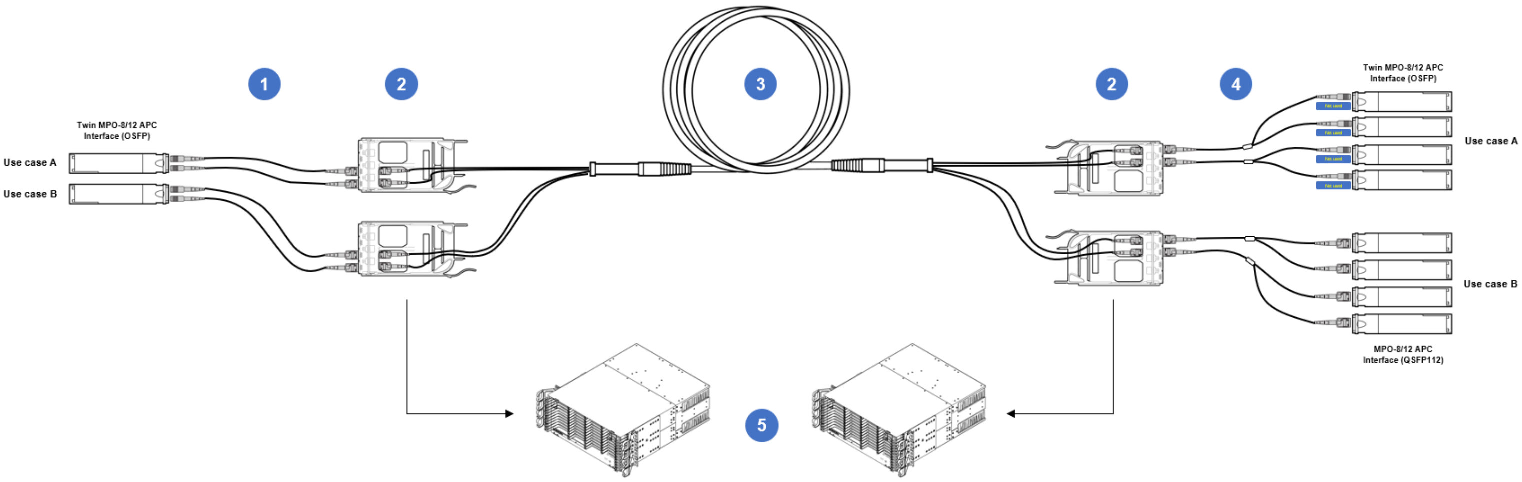


Figure 4. Use cases for Structured Cabling utilizing 800G and 200G transceivers working with MPO-8/12 APC.

Item	Reference Part Number	OS2 Part Number (Americas)	OS2 Part Number (EMEA and APJ)	OM4 Part Number (Americas)	OM4 Part Number (EMEA and APJ)	Description
1	Corning (a)	G-BND16-E8E8G-PN000-xxxF	G-BND16-E8E8G-LZ000-xxxM	G-BND16-E9E9Q-PN000-xxxF	G-BND16-E9E9Q-LZ000-xxxM	16x 8-F Mesh Bundle, MTP® APC (non-pinned) to MTP APC (non-pinned), 78-in (2-m) legs, Type-B polarity, xxxF (feet) or xxxM (meters)
		or				
		JE8E808GE8-NBxxxF	JE8E808GEZ-NBxxxM	JE9E908QE8-NBxxxF	JE9E908QEZ-NBxxxM	EDGE8®, MTP® APC (non-pinned) to MTP APC (non-pinned) 8-F Jumper, TIA-568 Type-B polarity, xxxF (feet) or xxxM (meters)
2		EDGE8-CP32-V1	EDGE8-CP32-V1	EDGE8-CP32-V3	EDGE8-CP32-V3	EDGE8 32-F MTP Adapter Panel, (4-port)
3		GE7E7E4GPNDDU xxxF(b)	GE7E7E4GLZDDU xxxM(b)	GE2E2E4QPNDDU xxxF(b)	GE2E2E4QLZDDU xxxM(b)	EDGE8, MTP Trunk , 144-F, MTP APC (pinned) to MTP APC (pinned), 33-in (840-mm) legs, Type-B polarity, pulling grip on first end only, xxxF (feet) or xxxM (meters)
4		HE8E808GPH-LBxxxF	HE8E808GLZ-LBxxxM	HE9E908QPH-LBxxxF	HE9E908QLZ-LBxxxM	EDGE8, 8-F Y-Harness, MTP® APC (unpinned) to two 4-F (unpinned) MTP APC, 36-in (910-mm) breakout leg length, Type-B polarity, xxxF (feet) or xxxM (meters)
5		EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	EDGE8-xxU	Please refer to Annex 1 to choose the best option for your application

Notes:

a) Corning cables in the Americas use Plenum cable, while EMEA/APJ uses LSZH™/CPR rated cable. Y-Harness lengths available from 1 to 60 meters. Single jumper lengths are available from 1 to 300 meters. Bundled jumpers use a meshed sleeve. Bundled lengths are available from 10 to 300 meters (furcation-to-furcation) in increments of 5 meters. Bundles are available in 16, 32 legs, and custom leg quantities with straight or staggered legs. Lengths in meters are also available for the Americas.

b) Trunks are available in fiber counts of 8 to 288 fibers.

c) Corning cables support InfiniBand, Ethernet and NVLink protocols.

d) Please review Corning's polarity drawings in Annex 2.

Table 5. Scenario 4 - 800G and 200G using Structured Cabling with MPO-8/12 APC. Part Number Scheme.

1.5. Scenario 5 - 800G and 400G - Switch to Switch Applications.

LC-Duplex to LC-Duplex using Point-to-Point Cabling.

Application: Quantum-2 InfiniBand or Spectrum-4 Ethernet to a) Quantum-2 InfiniBand or Spectrum-4 Ethernet.

Scenario 5 is primarily used for Point-to-Point cabling applications, which involves connecting active devices located within the same rack or row. However, it is not recommended to use Point-to-Point cabling if those devices are physically located in different areas within the data center.

The main application for the Twin LC-Duplex transceiver is to link Quantum-2 or Spectrum-4 air-cooled switches together. A bright green marking on transceiver body indicates 2km maximum reach.

Please refer to section 1.7, Table 10, to review the list of transceivers applicable for Use cases A and B in Scenarios 5 and 6. To learn more about Scalable Units and how to build a DGX SuperPOD, please refer to section 2 in this document.

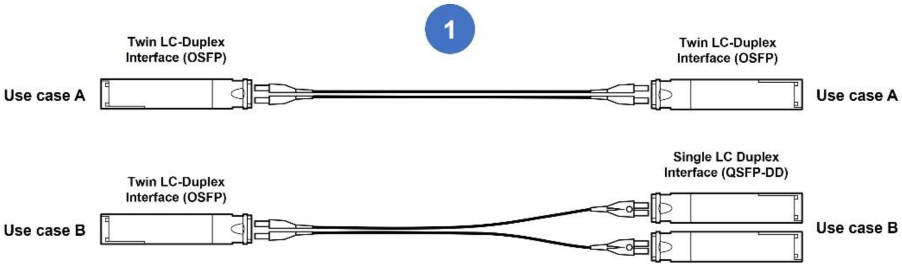


Figure 5. Use cases for Point-to-Point cabling utilizing 800G and 400G transceivers working with LC Duplex.

Item	Reference Part Number	OS2 Part Number (Americas)	OS2 Part Number (EMEA and APJ)	Description
1	**Corning** [(a)]	787802GD120xxxF	E787802GNZ20xxxM	EDGE™ LC UPC Uniboot to LC UPC Uniboot Duplex Jumper, xxxF or xxxM

Notes:

a) Corning cables in the Americas use Plenum cable, while EMEA/APJ uses LSZH™/CPR rated cable. Jumper lengths are available from 1 to 300 meters. Lengths in meters are also available for the Americas.

b) Corning cables support InfiniBand, Ethernet and NVLink protocols.

c) Please review Corning's polarity drawings in Annex 2.

Table 6. Scenario 5 - 800G and 400G using Point-to-Point Cabling with LC Duplex. Part Number Scheme.

1.6. Scenario 6 - 800G and 400G - Switch to Switch Applications.

LC-Duplex UPC to LC-Duplex UPC using Structured Cabling Across DC with Trunk.

Application: Quantum-2 InfiniBand or Spectrum-4 Ethernet to a) Quantum-2 InfiniBand or Spectrum-4 Ethernet.

In Scenario 6, the implementation of Structured Cabling utilizes an optical fiber trunk with MPO-8/12 connectivity as a backbone. This application is used to connect different switches, such as Leaf to Spine. It is the recommended option to follow when two different active devices are physically located in different areas within the data center.

The main application for the Twin LC-Duplex transceiver is to link Quantum-2 or Spectrum-4 air-cooled switches together. A bright green marking on transceiver body indicates 2km maximum reach.

Please refer to section 1.7, Table 10, to review the list of transceivers applicable for Use cases A and B in Scenarios 5 and 6. To learn more about Scalable Units and how to build a DGX SuperPOD, please refer to section 2 in this document.

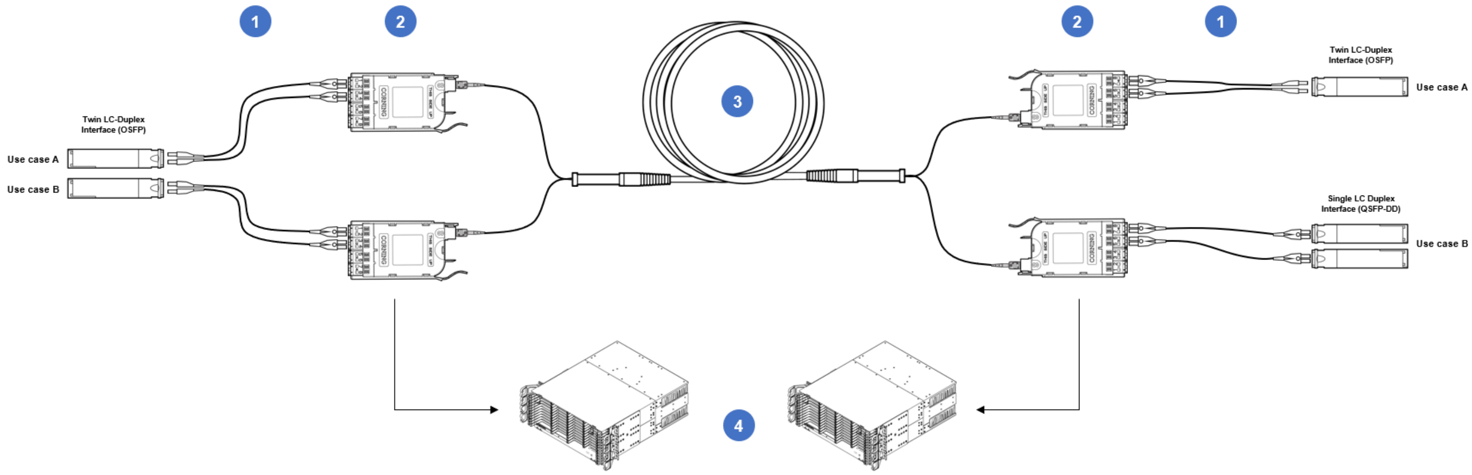


Figure 6. Use cases for Structured Cabling utilizing 800G and 400G transceivers working with LC Duplex.

Item	Reference Part Number	OS2 Part Number (Americas)	OS2 Part Number (EMEA and APJ)	Description
1	Corning [a]	787802GD120xxxF	E787802GNZ20xxxM	EDGE™ LC UPC Uniboot to LC UPC Uniboot Duplex Jumper, xxxF or xxxM
2		ECM8-UM08-04-E8G-ULL	ECM8-UM08-04-E8G-ULL	EDGE8® Ultra-Low-Loss Module, LC duplex to MTP® (unpinned), 8-F, universal polarity
3		GE7E7E4GPNDDU xxxF[b]	GE7E7E4GLZDDU xxxM[b]	EDGE8, MTP Trunk , 144-F, MTP APC (pinned) to MTP APC (pinned), 33-in (840-mm) legs, Type-B polarity, pulling grip on first end only, xxxF (feet) or xxxM (meters)
4		EDGE8-xxU	EDGE8-xxU	Please refer to Annex 1 to choose the best option for your application

Notes:

a) Corning cables in the Americas use Plenum cable, while EMEA/APJ uses LSZH™/CPR rated cable. Jumper lengths are available from 1 to 300 meters. Lengths in meters are also available for the Americas.

b) Trunks are available in fiber counts of 8 to 288 fibers.

c) Corning cables support InfiniBand, Ethernet and NVLink protocols.

d) Please review Corning's polarity drawings in Annex 2.

Table 7. Scenario 6 - 800G and 400G using Structured Cabling with LC Duplex connectivity at the active device. Part Number Scheme.

1.7. Transceiver Options and Port Breakout Connections per Scenario.

To read the following tables, please follow the examples:

Example 1: In Scenario 2, Use Case B, using Table 8, an 800G-2xDR4, MMS4X00-NS with 2x MPO-8/12 APC connections on the left side of the drawing can connect to two OSFP transceivers, MMS4X00-NS400, resulting in two individual 400G connections on the right side.

Example 2: In Scenario 4, Use Case B, using Table 9, an 800G-2xDR4, MMS4X00-NS with 2x MPO-8/12 APC connections on the left side of the drawing can connect to four 200G-DR4, QSFP112, MMS1X00-NS400 utilizing Y-Harnesses, resulting in four individual 200G connections on the right side.

Use case	Near End Optic (Left)			Far End Optic (Right)			Reach	Fiber Type
	Speed	NVIDIA Port	Footprint - Fiber/Transceiver	Speed	NVIDIA Port	Footprint - Fiber/Transceiver		
A	800G-2xDR4	MMS4X00-NM	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NM	OSFP - 2x 8F	500m	Single mode
	800G-2xDR4	MMS4X00-NM-T	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NM-T	OSFP - 2x 8F	500m	Single mode
	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	100m	Single mode
	800G-2xDR4	MMS4X00-NS-FLT	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	100m	Single mode
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	100m	Single mode
B	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	400G-DR4	MMS4X00-NS400	OSFP - 8F	100m	Single mode
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	400G-DR4	MMS4X00-NS400-T	OSFP - 8F	100m	Single mode
C	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	400G-DR4	MMS1X00-NS400	QSFP112 - 8F	100m	Single mode
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	400G-DR4	MMS1X00-NS400-T	QSFP112 - 8F	100m	Single mode
A	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	50m	Multimode
	800G-2xSR4	MMA4Z00-NS-FLT	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	50m	Multimode
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	50m	Multimode
B	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	400G-SR4	MMA4Z00-NS400	OSFP - 8F	50m	Multimode
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	400G-SR4	MMA4Z00-NS400-T	OSFP - 8F	50m	Multimode
C	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	400G-SR4	MMA1Z00-NS400	QSFP112 - 8F	50m	Multimode
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	400G-SR4	MMA1Z00-NS400-T	QSFP112 - 8F	50m	Multimode

Table 8. Transceivers Applicable for Use Cases A to C in Scenarios 1 and 2.

Use case	Near End Optic (Left)			Far End Optic (Right)			Reach	Fiber Type
	Speed	NVIDIA Port	Footprint - Fiber/Transceiver	Speed	NVIDIA Port	Footprint - Fiber/Transceiver		
A	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	200G-DR4	MMS4X00-NS400	OSFP - 4F	100m	Single mode
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	200G-DR4	MMS4X00-NS400-T	OSFP - 4F	100m	Single mode
B	800G-2xDR4	MMS4X00-NS	OSFP - 2x 8F	200G-DR4	MMS1X00-NS400	QSFP112 - 4F	100m	Single mode
	800G-2xDR4	MMS4X00-NS-T	OSFP - 2x 8F	200G-DR4	MMS1X00-NS400-T	QSFP112 - 4F	100m	Single mode
A	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	200G-SR4	MMA4Z00-NS400	OSFP - 4F	50m	Multimode
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	200G-SR4	MMA4Z00-NS400-T	OSFP - 4F	50m	Multimode
B	800G-2xSR4	MMA4Z00-NS	OSFP - 2x 8F	200G-SR4	MMA1Z00-NS400	QSFP112 - 4F	50m	Multimode
	800G-2xSR4	MMA4Z00-NS-T	OSFP - 2x 8F	200G-SR4	MMA1Z00-NS400-T	QSFP112 - 4F	50m	Multimode

Table 9. Transceivers Applicable for Use Cases A and B in Scenarios 3 and 4.

Use case	Near End Optic (Left)			Far End Optic (Right)			Reach	Fiber Type
	Speed	NVIDIA Port	Footprint - Fiber/Transceiver	Speed	NVIDIA Port	Footprint - Fiber/Transceiver		
A	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	2km	Single mode
	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	2km	Single mode
B	800G-2xFR4	MMS4X50-NM	OSFP - 2x 2F	400G-FR4	MMS1V50-WM	QSFP-DD - 2F	2km	Single mode
	800G-2xFR4	MMS4X50-NM-T	OSFP - 2x 2F	400G-FR4	MMS1V50-WM-T	QSFP-DD - 2F	2km	Single mode

Table 10. Transceivers Applicable for Use Cases A and B in Scenarios 5 and 6.

2. NVIDIA DGX SuperPOD Cabling Architecture Reference Guide.

In this section, we will review how to cable an NVIDIA DGX SuperPOD. Although this guide uses the NVIDIA DGX H100 as a reference, it's important to note that similar cabling components and infrastructure could also be used for applications involving the DGX B200 and DGX GB200. The specific considerations for each POD will depend on individual customer designs.

A DGX SuperPOD is formed by a certain number of Scalable Units (SUs), each containing a certain number of nodes (servers). This configuration determines the total number of GPUs per POD. The Scalable Unit is the building block of a SuperPOD.

PODs can also be called clusters, depending on each customer's preference. It is crucial to understand that the size of the POD or cluster will dictate the required number of InfiniBand Leaf Switches, InfiniBand Spine Switches, and InfiniBand Core Switches. Based on that configuration, we can calculate the number of cables or connections needed.

In Figure 7, if we take a medium cluster or POD with 32 Scalable Units as an example, it would total 1,024 nodes or servers, which is equivalent to 8,192 GPUs. This setup would require 256 IB Leaf Switches, 256 IB Spine Switches, and 128 IB Core Switches in the POD. We will utilize this POD size as a reference throughout the document. Please note that the classification as Small, Medium, or Large is not defined by NVIDIA; however, we will use it as a reference in this guide.

The "Cable Counts" column shows the number of individual MPO-8/12 APC connections needed to connect the system. For example, to connect each node (server) to the IB Leaf Switch, we will require 8,192 MPO-8/12 APC connections. To connect the IB Leaf Switches to the IB Spine Switches, we will require an additional 8,192 MPO-8/12 APC connections. Lastly, to connect the IB Spine Switches to the IB Core Switches, we will need another 8,192 MPO-8/12 APC connections, resulting in a total of 24,576 MPO-8/12 APC connections across the POD.

	SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
				InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
Small Cluster or POD	4	128	1024	32	16	--	1024	1024	1024
	8	256	2048	64	32	--	2048	2048	2048
Medium Cluster or POD	16	512	4096	128	128	64	4096	4096	4096
	32	1024	8192	256	256	128	8192	8192	8192
Large Cluster or POD	64	2048	16384	512	512	256	16384	16384	16384

Figure 7. SuperPOD component counts.

2.1. Understanding a DGX-H100 Node (Server) Compute Fabric Connections.

The NVIDIA DGX-H100 Node possesses connectivity ports for an InfiniBand Compute Fabric, an InfiniBand Storage Fabric, an In-Band Management Network, and an Out-of-Band Management Network.[1]

The focus of this architecture guide will center on the InfiniBand Compute Fabric. This fabric allows for connecting 4x Dual Port OSFP on each Node, totaling 8x 400G connections. These connections are made using either Single Mode or Multimode MPO-8/12 APC connectors as shown in Figure 8.

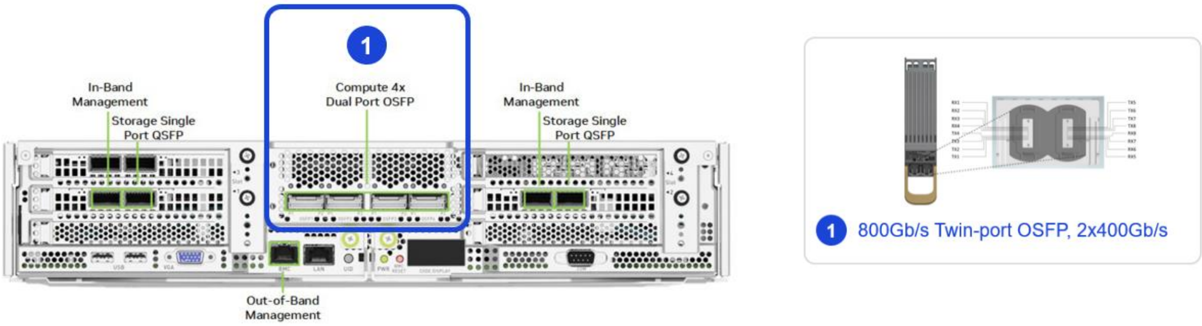


Figure 8. Back of a DGX H100 CPU tray and its connectivity.

2.2. Understanding the Quantum-2 9700 NDR IB Switches.

The Quantum-2 switches allow for 32 OSFP Twin MPO-8/12 APC ports. This means that each switch can provide up to 64 single ports of 400 Gbps.

Figure 9 illustrates an example of port mapping at the switch level, which can be used as a reference throughout the document. Note that this is just one possible method for customers to map such ports.

As Quantum-2 switches will be deployed across the network in various roles, such as Leaf, Spine, or Core, port distribution can follow the approach suggested in Figure 9 for inter-switch connectivity.

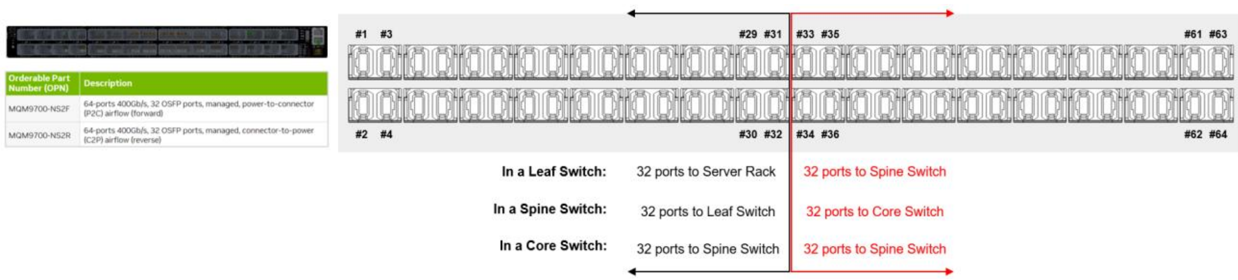


Figure 9. Example of port mapping for Quantum-2 switches.

It is important to note that the OSFP Twin transceivers located in the bottom ports of the Quantum-2 switch will be upside down, reversing the port connections, as shown in Figure 10. In this example, we can observe that Twin Port 1 provides connections X and Y, while Twin Port 2, being upside down, provides connections Y and then X. A physical way to distinguish this is by the fins on both transceivers, which will be visually opposed to each other.

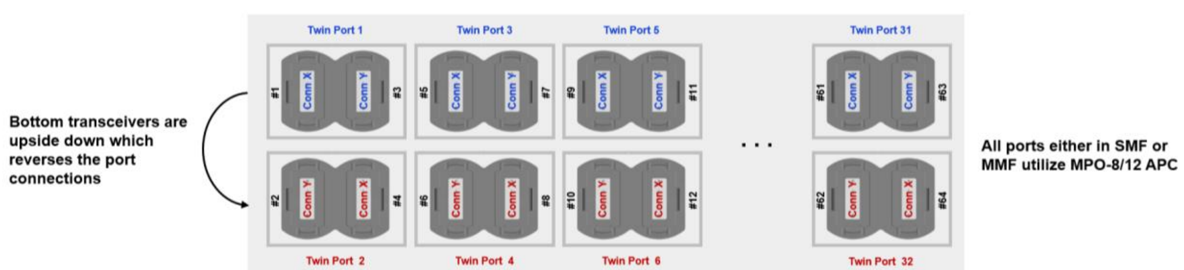


Figure 10. Twin Transceiver port consideration.

2.3. Understanding the concept of Scalable Unit, the building block of a SuperPOD.

The system is built upon building blocks of Scalable Units (SUs), each containing 32 DGX H100 systems, which allows for the rapid deployment of systems of various sizes.

Each SU contains 256 GPUs, distributed across 32 DGX H100 systems in 8 racks. There are 8 Leaf Switches for each Scalable Unit (SU).

The fabric is Rail-optimized, meaning that Host Channel Adapters (HCAs) with the same Port number from each node are connected to the same Leaf Switch. This means all Port number 1 connections from all servers will be connected to Leaf Switch number 1, Port number 2 to Leaf Switch number 2, and so on.

The fabric is constructed using Quantum-2 9700 InfiniBand switches with 800Gbps/Twin port OSFP transceivers.

Figure 11 illustrates a Middle-of-Row layout with the relevant Compute Fabric components in a Scalable Unit, based on the previous information. It is important to note that other types of layouts can be implemented. Therefore, maintaining close contact with Corning's Engineering team during the design phase is crucial to validate the best cabling options for your design choice.

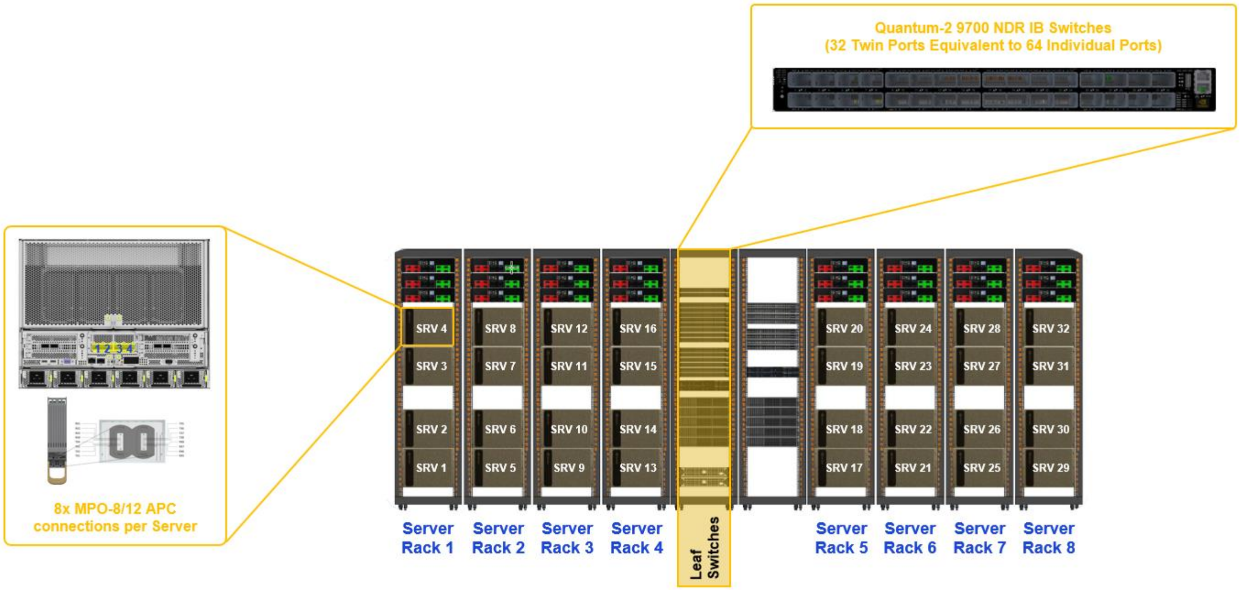


Figure 11. Relevant Compute Fabric components in a Scalable Unit.

2.4. Implementing the Cabling Scenarios within the NVIDIA DGX SuperPOD.

To facilitate the identification of the different cabling components used in building an AI/ML cluster, Corning will utilize three levels of connectivity in this guide. These levels and the number of switches are based on the example of a 32 Scalable Unit Cluster or POD:

Level A - Server to Leaf cabling.

Level B - Leaf to Spine cabling.

Level C - Spine to Core cabling.

2.4.1. Level A – Server to Leaf cabling.

A Scalable Unit can be cabled using point-to-point connections between the node (server) and the Leaf Switch (see Figure 1), with at least two cabling product options available (See Figure 12). In some specific custom designs, implementing structured cabling at the Scalable Unit level is also possible (see Figure 2).

The first approach is to use 256 individual 8-fiber jumpers to create the connections from each twin transceiver. The second approach is to use 8 bundled jumpers instead. A bundled jumper consists of a group of 8-fiber MPO-8/12 APC patch cords grouped as a single 256-fiber unit with a flame-retardant wrap-around mesh. The choice will depend on the specific requirements of the customer's design. Regardless of the choice, we will bring a total of 256 MPO-8/12 APC connections to our Leaf Switches, which, in this example, are in a Middle-of-Row layout (see Figure 12).

Previously, we described the compute fabric as Rail-optimized, with a total of 8 Rails. These Rails correspond to the 8 MPO-8/12 APC connections from each server. In the following examples, we will use color references, as indicated in Figures 12 and 13, to indicate each of these Rail connections.

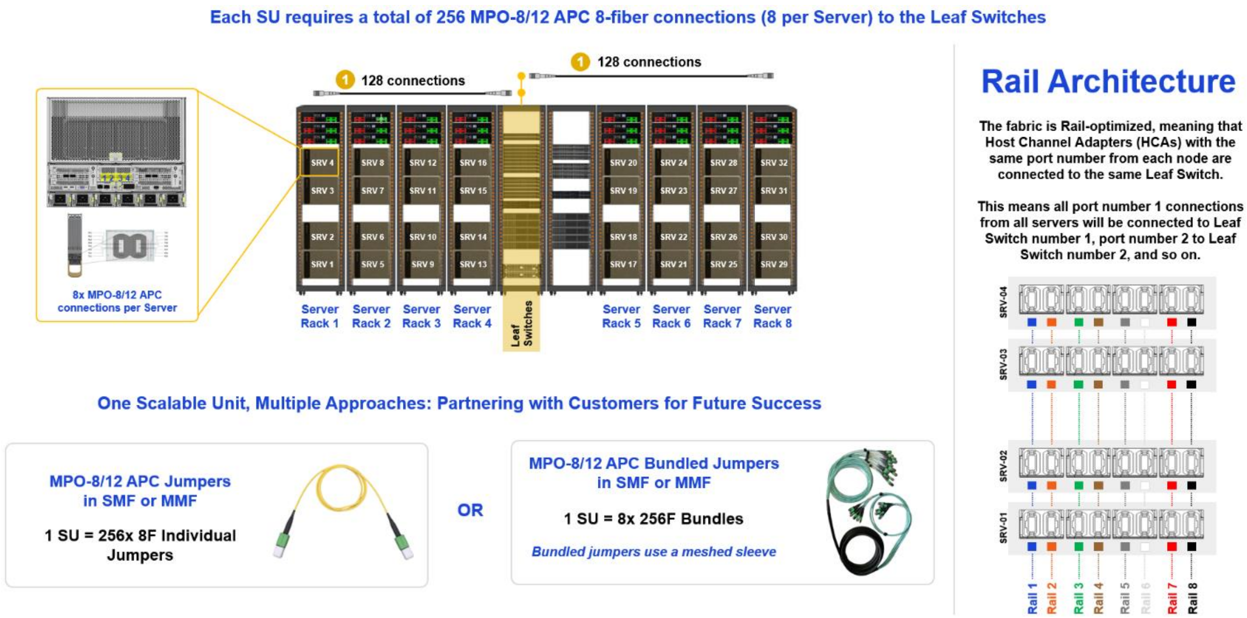


Figure 12. Level A - Server to Leaf cabling describing number of connections and two product approaches.

For Level A, we will illustrate the connections from the Server Rack to the InfiniBand Leaf Switches located in the Middle of the Row (MOR) of our Scalable Unit using a bundled jumper. If individual jumpers are chosen, the Rail mapping can be followed accordingly.

In Figure 13, on the left, we can visualize a Server Rack with 4 servers, each depicting their respective Rails. On the right, we find the Leaf Switch Rack with 8 Leaf Switches, one for each Rail. As we know, these Quantum-2 switches allow for 32 Twin MPO-8/12 APC ports, meaning each switch can support up to 64 individual connections.

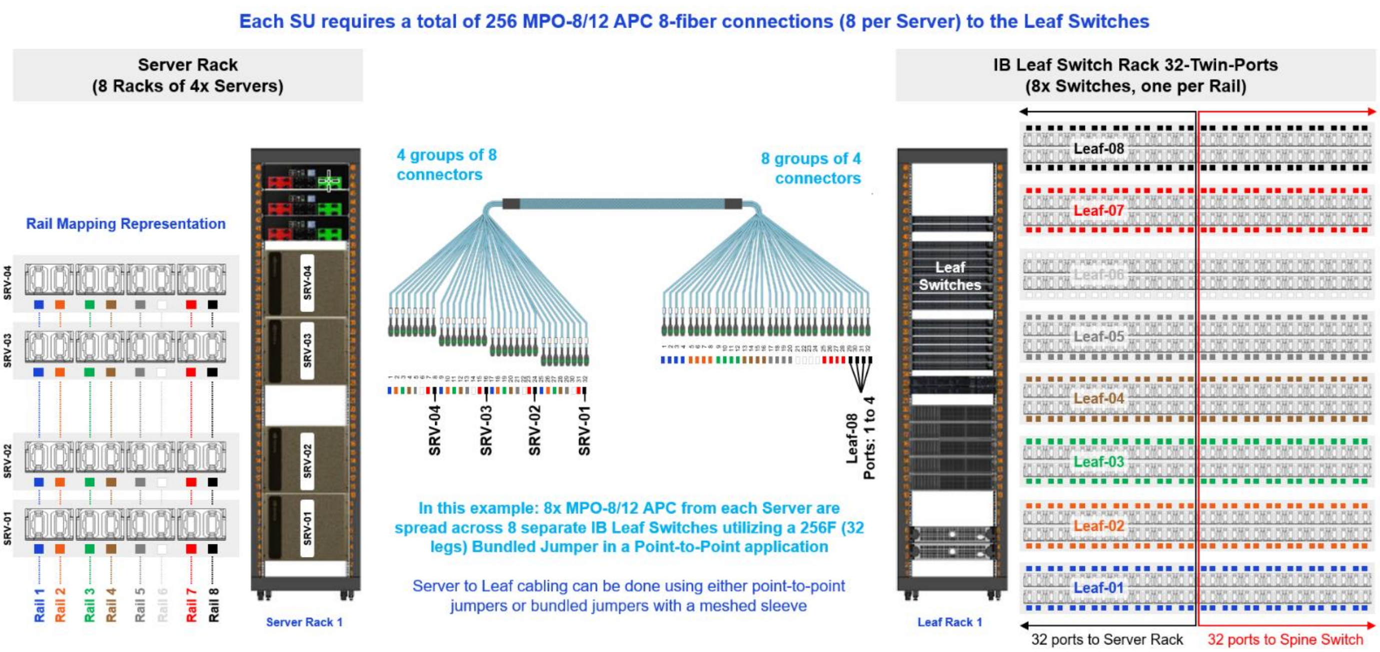


Figure 13. Level A - Server to Leaf cabling utilizing Point-to-Point connection with a bundled jumper - Example based on 32 SU Cluster/POD.

Cabling is executed using one bundled jumper per Server Rack to route all Rails of the same color from each server to their corresponding Leaf Switch color. In this example, Rail 8 (Black Rail) from each server is distributed to Leaf-08 Switch ports 1 to 4. With a total of 8 Server Racks in a Scalable Unit, Rail 8 will complete the first 32 port connections to Leaf-08, following the mapping shown in Figure 9, while the remaining 32 port connections are routed to the Spine Switches. This process repeats for each Server Rack and Rail until all connections are completed within the Scalable Unit.

2.4.2. Level B – Leaf to Spine cabling.

To connect the Leaf Switches to the Spine Switches, structured cabling is the recommended approach. However, in some designs, connections can also be made using point-to-point cabling. It is important to note that these connections can extend up to 500 meters (based on 800G-2xDR4, MMS4X00-NM transceivers), which can present significant challenges in managing the cabling for point-to-point configurations.

As mentioned at the beginning of the architecture section, we are using a 32 Scalable Unit Cluster or POD as an example. By referring to the logic diagram presented in Figure 14 and the component count mentioned in Figure 7, we can observe that this design comprises a total of 256 Spine Switches, which are distributed across 16 Spine Racks. This indicates that each rack contains 16 Spine Switches.

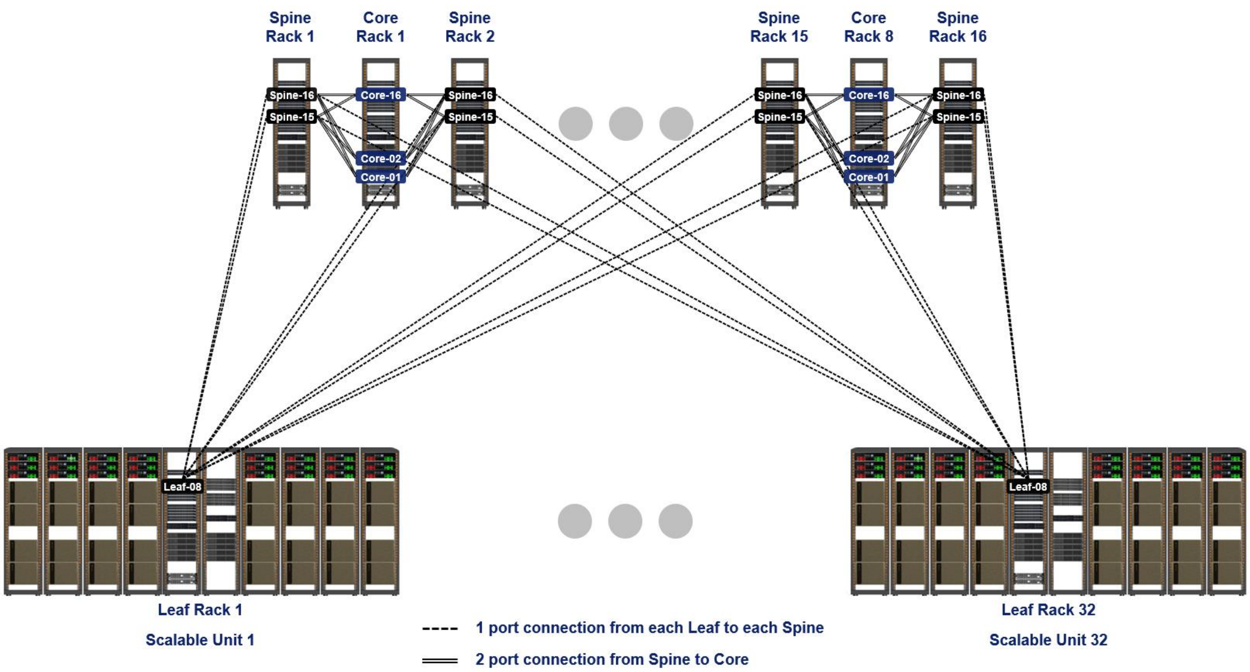


Figure 14. Logic Diagram for Rail 8 (Black Rail) in a 32 SU DGX H100 SuperPOD.

Even though it is not the recommended option, let’s start by reviewing our cabling example using a point-to-point connection with a bundled jumper to simplify the port mapping explanation. In Figure 15, on the left, we can see a Leaf Rack from one of our Scalable Units, where each switch is depicted with its respective Rail colors. On the right, we find one Spine Rack (out of 16) with 16 Spine Switches, two for each Rail color.

For this example, let's use the Black Rail (Rail 8) as a reference. If we consider the connections from Leaf-08 in Leaf Rack 1 (which is considered as the Leaf Rack in Scalable Unit 1) and use the Leaf Switch ports 33 and 34 (as shown in the port mapping in Figure 9), we can route them to Spine Rack 1, connecting to Spine-15, Port 1 and Spine-16, Port 1. Since every Leaf Switch on the same Rail needs to be connected to every Spine Switch on the same Rail (two per Spine Rack), ports 33 and 34 from a different Leaf Switch in a different Scalable Unit will also connect to Spine Rack 1, connecting to Spine-16, Port 2 and Spine-15, Port 2, and so on. The objective is to establish connections between the Leaf Switches in all Scalable Units and the two Spine Switches in all the Spine Racks, as shown in Figure 15. Taking Rail-8 as a reference, this would require using 16 bundled jumpers from a single Leaf-08 Switch in Scalable Unit 1 to connect to 16 different Spine Racks, resulting in connections to a total of 32x Rail 8 (Black Rail) Spine Switches, as shown in Figure 14.

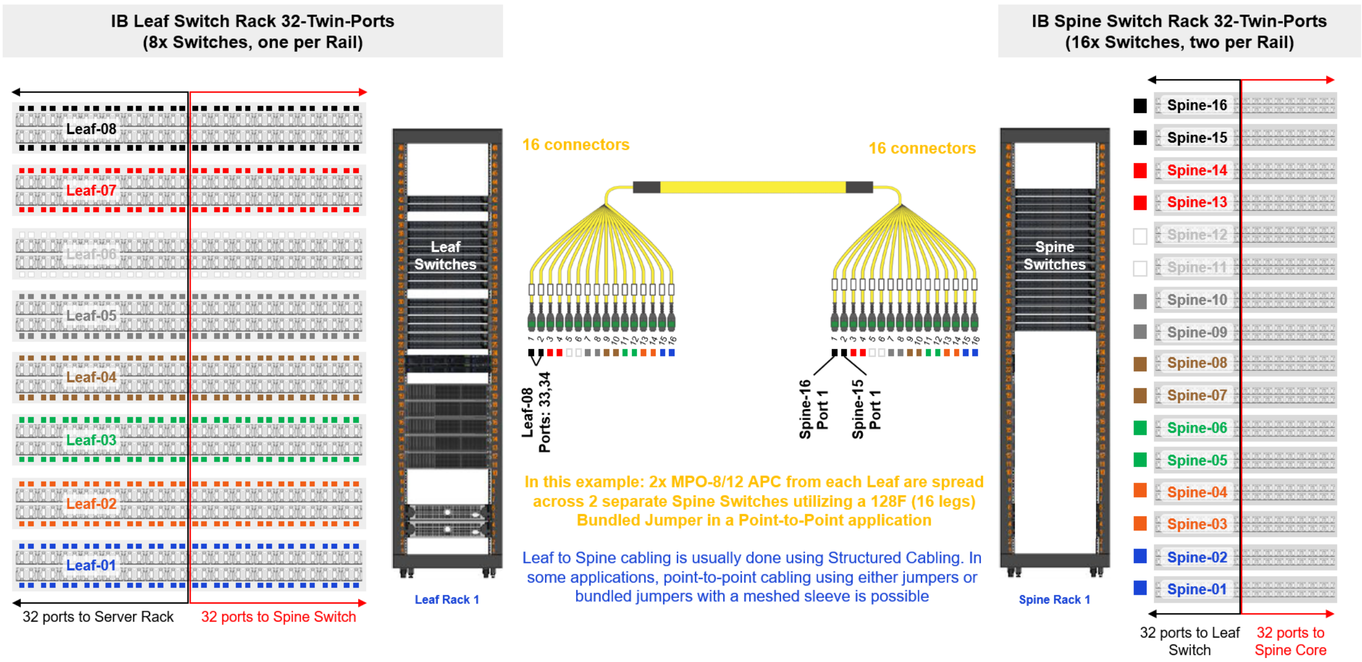


Figure 15. Level B – Leaf to Spine cabling using Point-to-Point connection with a bundled jumper - Example based on 32 SU Cluster/POD.

Once we have understood how the mapping works for our Leaf to Spine Switch connection using a point-to-p oint connection with a bundled jumper, let's apply the same logic to the Structured Cabling Option.

In Figure 16, on the left, we can see our Leaf Rack from one of our Scalable Units, where each switch is depicted with its respective Rail colors. On the right, we find the Spine Rack 1 (out of 16) with 16 Spine Switches, two for each Rail color.

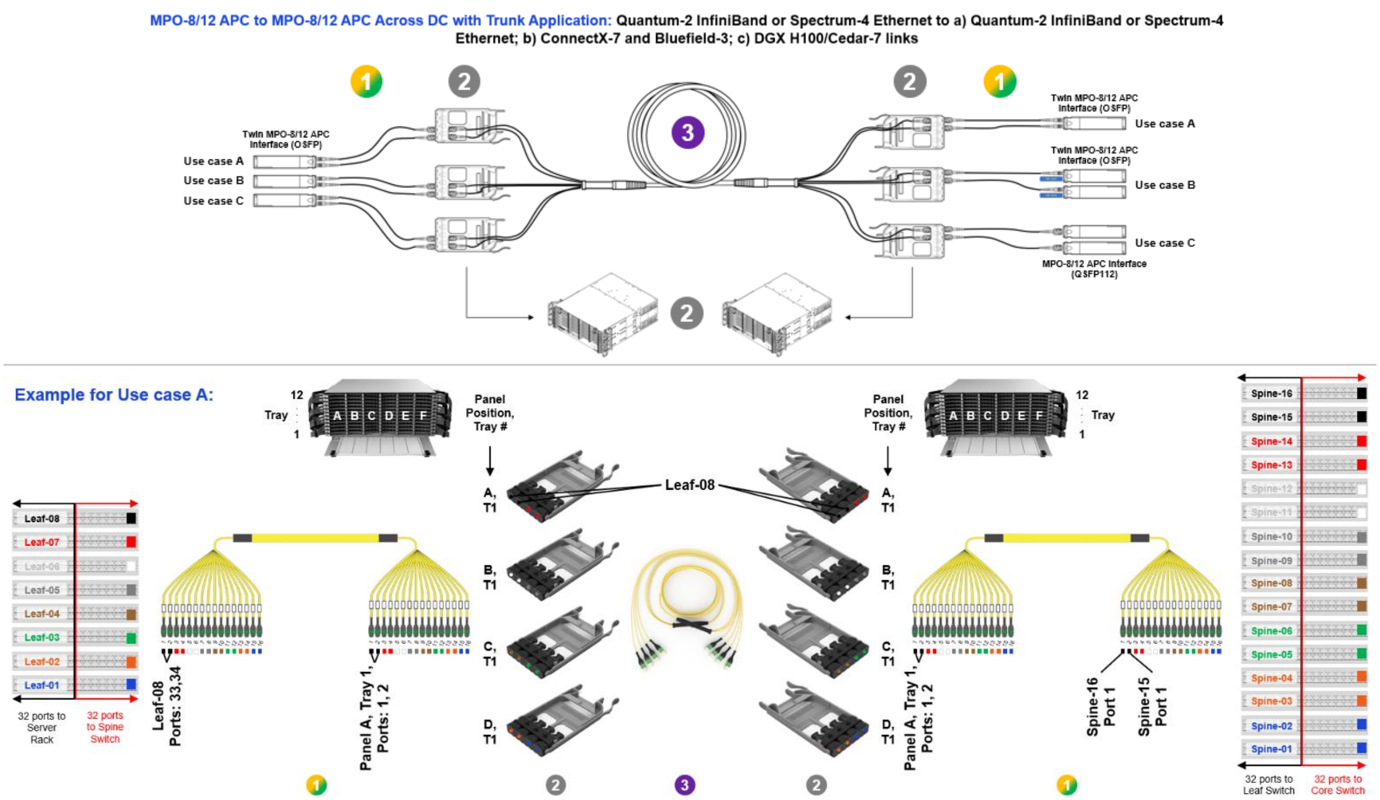


Figure 16. Level B – Leaf to Spine cabling using Structured Cabling with bundled jumpers - Example based on 32 SU Cluster/POD.

Still using the Black Rail (Rail 8) as a reference, if we take the connections from Leaf-08 located in Leaf Rack 1 and use Leaf Switch ports 33 and 34 (as we did before), we can route these using bundled jumpers that follow the same mapping we described in the point-to-point cabling. However, structured cabling components, such as adapter panels and housings (fiber enclosures), will be added as described in Figure 2.

The objective is to run high-fiber-count connections between different areas using a multifiber trunk serving as a backbone. We will use the adapter panels to map our connections from Leaf to Spine.

In the example in Figure 16, we see that the bundled jumper carrying the connections from Leaf-08, Ports 33 and 34, enters the adapter panel located in Position A, Tray 1, where the first two adapter ports are allocated to the Black Rail. Multiple adapter panels then use backbone cabling to route all those connections to a different housing located in another section of our data center, following the same mapping. Finally, we use a bundled jumper to make the last connection to Spine-15, Port 1 and Spine-16, Port 1, resulting in the same outcome as with point-to-point cabling.

If we follow this type of mapping, positions E and F from our patch panel may remain unused. However, it is common practice in the industry to utilize spare adapters and connections, even when they are not necessarily used in our mapping example. These spare positions can be used to manage spare connections that may be needed by customers, allowing for flexibility and accommodating their specific requirements.

The main advantage of implementing structured cabling in this connection is that it maximizes space and reduces installation time. It also allows for easy moves, adds, and changes (MACs), especially when Leaf and Spine Switches are physically located in different areas. Furthermore, it provides flexibility for future growth, which point-to-point cabling cannot support.

2.4.3. Level C – Spine to Core cabling.

Spine-to-Core Switches can be cabled using point-to-point connections. In some specific custom designs, implementing structured cabling at the Spine to Core connection level is also possible; this will depend on the physical location of the Core in relation to the Spine. An additional option is the use of Direct Attach Copper cables (DAC).

As we are using a 32 Scalable Unit Cluster or POD as an example, this design includes a total of 128 Core Switches distributed across 8 Core Racks, with each rack containing 16 Core Switches. Utilizing the Logic Diagram in Figure 14 as a reference, for exemplification purposes, the Core Rack is positioned in the middle of two Spine Racks.

Let’s use a point-to-point connection with a bundled jumper to simplify the port mapping explanation. In Figure 17, on the left, we can see Spine Rack 1, and on the right, we find Spine Rack 2. Each of these racks, as we know, will have 16 Spine Switches, two for each Rail color. In the middle, we have our Core Rack, containing 16 Core Switches. Each Core Switch will aggregate all the different Rails coming from the various Spine Switches.

For the purpose of this example and based on our port mapping depicted in Figure 9, let’s assume that Spine Rack 2 is physically “mirroring” Spine Rack 1. This means that the ports connected to the Leaf Switches in Spine Rack 2 are from 33 to 64 instead of 1 to 32, and the ports connected to the Core Switches are from 1 to 32 instead of 33 to 64. By positioning our Core Switch Rack in the middle, we can have straightforward routing from both Spine Racks using the bundled jumper.

First, let’s focus on the connections coming from Spine Rack 1. Ports 33 and 34 in all the Spine Switches will be routed to Ports 1 to 32 in Core-01; Ports 35 and 36 in all the Spine Switches will be routed to Ports 1 to 32 in Core-02, and so on, until Ports 63 and 64 in all the Spine Switches are routed to Ports 1 to 32 in Core-16.

Next, we need to handle the connections from Spine Rack 2. Ports 1 and 2 in all the Spine Switches will be routed to Ports 33 to 64 in Core-01; Ports 3 and 4 in all the Spine Switches will be routed to Ports 33 to 64 in Core-02, and so on, until Ports 31 and 32 in all the Spine Switches are routed to Ports 33 to 64 in Core-16.

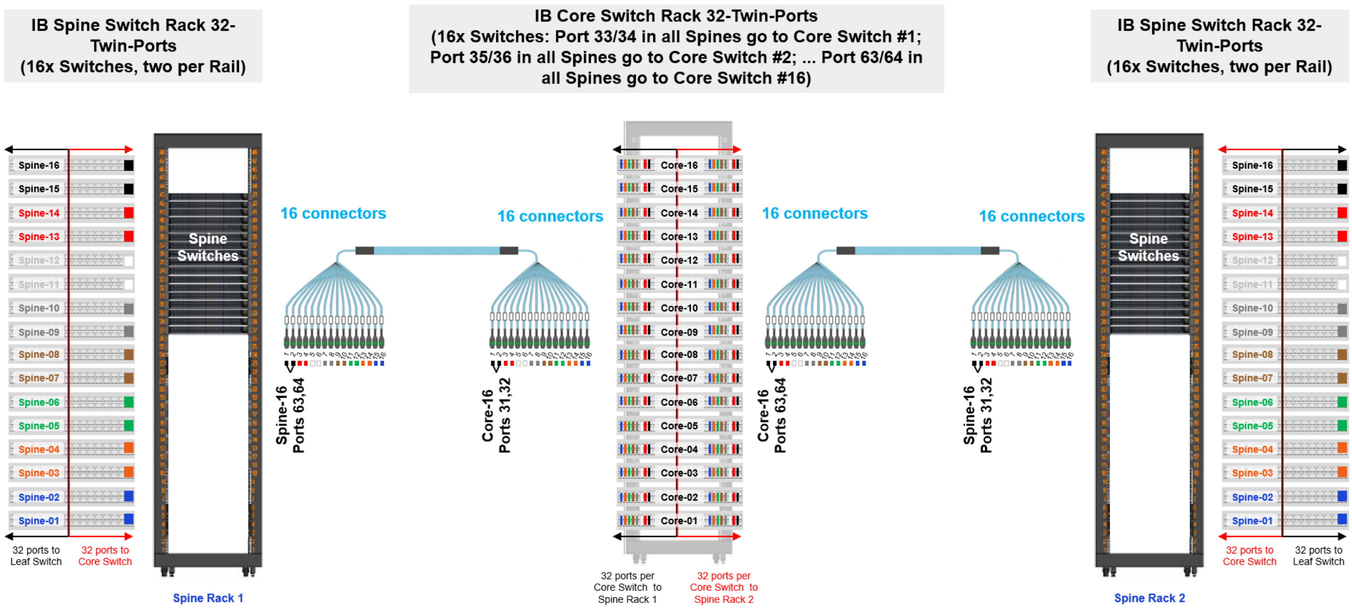


Figure 17. Level C – Spine to Core cabling using Point-to-Point connection with a bundled jumper - Example based on 32 SU Cluster/POD.

In Figure 18, we can take a closer look at how the cabling utilizing the bundled jumper could be done. This shows Spine Rack 1, with Ports 33 and 34 on all the Spine Switches being routed to Ports 1 to 32 on Core-01. A vertical cable manager (depicted as a rack in Figure 18 only for exemplification) can be used to facilitate the bundled jumper routing from Spine to Core.

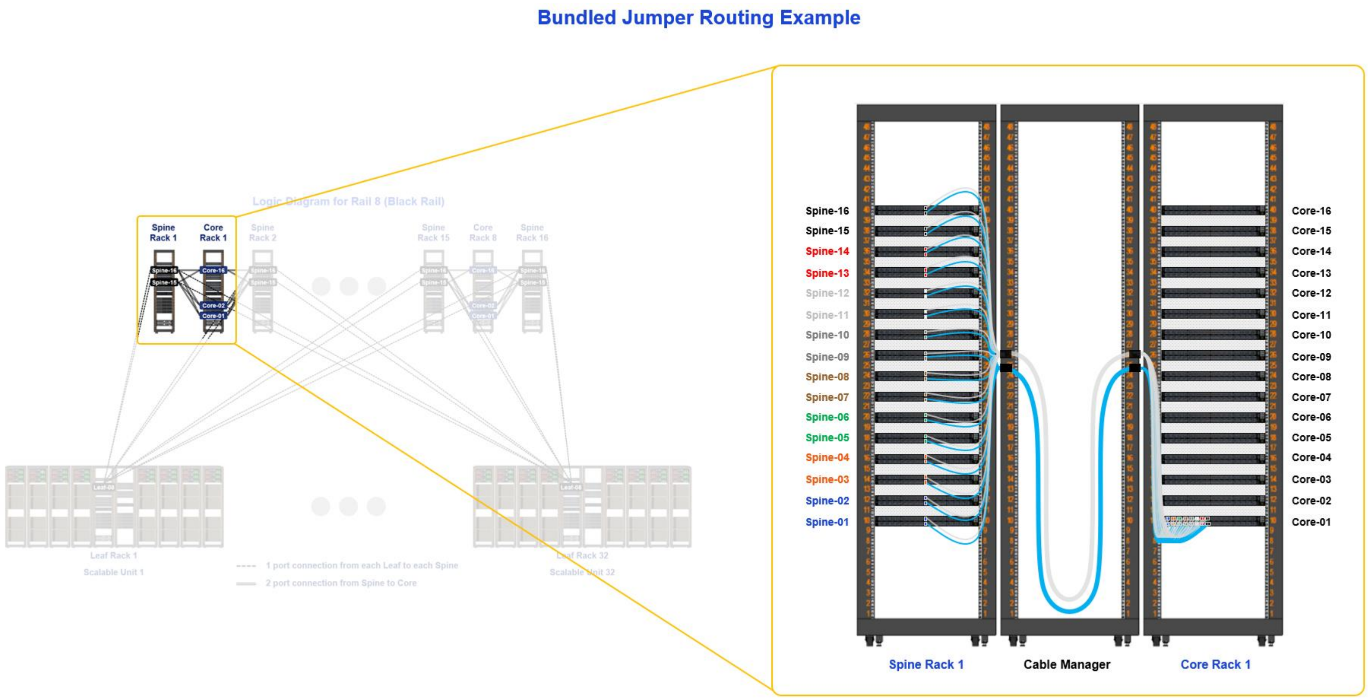


Figure 18. Level C –Ports 33 and 34 on all the Spine Switches routed to Ports 1 to 32 on Core-01.- Example based on 32 SU Cluster/POD.

2.5. Multimode vs Single Mode.

The decision to use either Multimode or Single Mode fiber within the network will depend on specific design considerations. Multimode fiber has a reach of up to 50 meters, which typically limits its application to Scalable Unit or Spine to Core connections. However, this may not always be practical, as the design of the Scalable Unit might not follow a Middle-of-Row layout or be physically close to each other. Similarly, the Spine and Core Switches might not be physically close to each other, which supports the case for using Single Mode fiber across the entire design.

2.6. The Big Picture.

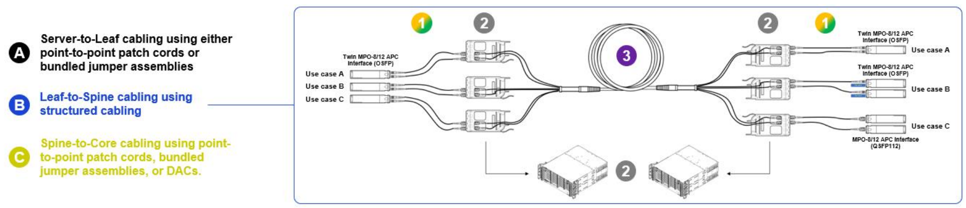
Now that we have learned about the different Cluster or POD sizes, as well as how cabling can be done between the active equipment in the Compute Fabric, let's visually summarize the components that can be used, these components will be dependent on the specific designs but would be mostly based on the different products and part numbers we have reviewed in this document.

2.6.1. Cabling a 4 Scalable Unit (SU) Cluster/POD.

In Figure 19, cabling a 4 SU POD with 128 servers, 32 Leaf Switches, and 16 Spine Switches requires:

- 1,024 connections from Node to Leaf.
- 1,024 connections from Leaf to Spine.
- Optionally, if a Core is available, it would also require 1,024 connections.

SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	--	1024	1024	1024
8	256	2048	64	32	--	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
64	2048	16384	512	512	256	16384	16384	16384



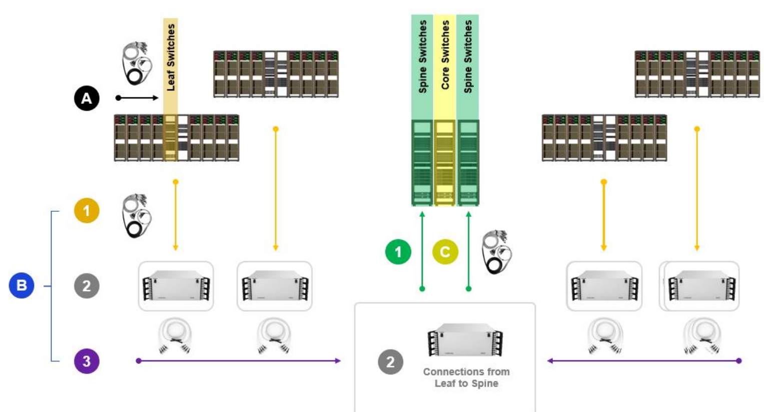


Figure 19. 4 Scalable Unit (SU) Cluster/POD with cabling Levels A, B, C.

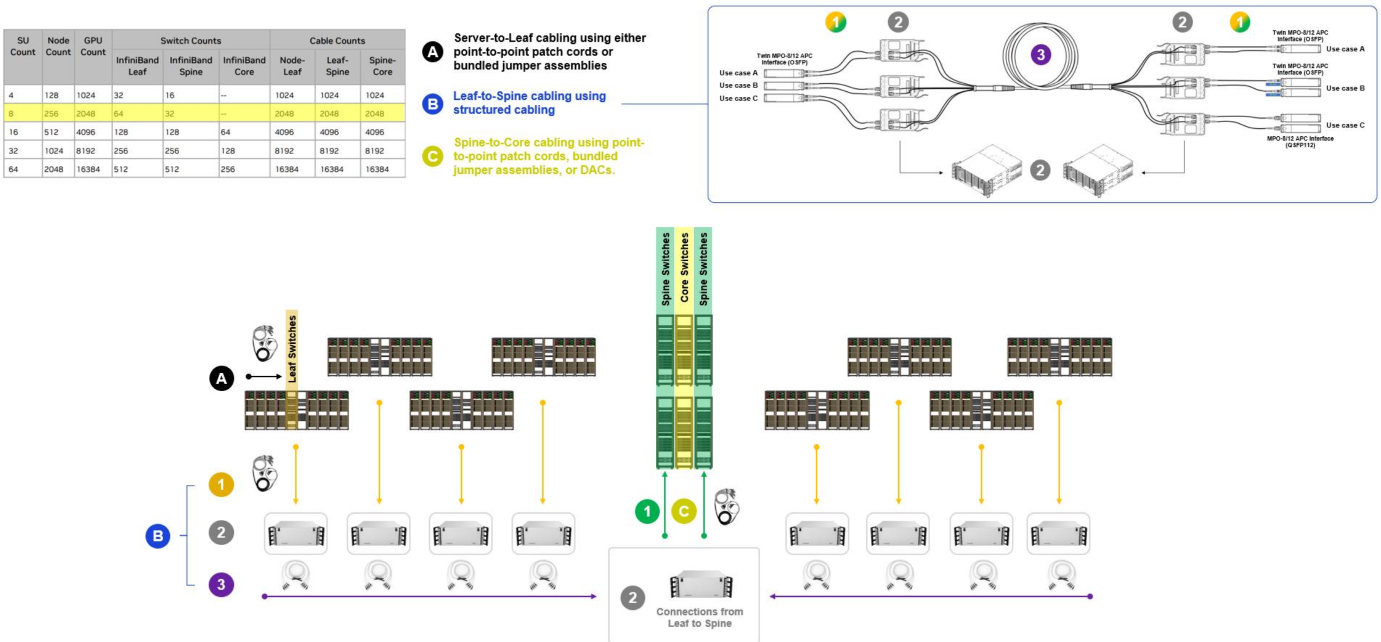
By utilizing the different levels of cabling that we have reviewed, we can understand the following:

- Data Center Operators can use Level A at the Scalable Unit to perform Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies. This would require either 256 individual 8-fiber jumpers of varying lengths or only 8 bundled jumpers of 256 fibers each, with varying lengths, per SU.
- Data Center Operators can use Level B to perform Leaf-to-Spine cabling using structured cabling. This would require the implementation of individual jumpers or bundled jumpers, adapter panels, housings, and multifiber trunks to serve as backbone cabling. With this backbone cabling, we can easily use 64 trunks (16 trunks of 144-fibers per Leaf), which would replace individual patch cords and bundled jumpers in a point-to-point connection. This method allows for spare ports, better management of complexity, and improved pathway space across the Data Center.
- If a Core is implemented, Data Center Operators can use Level C to perform Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

2.6.2. Cabling an 8 Scalable Unit (SU) Cluster/POD.

In Figure 20, cabling an 8 SU POD with 256 servers, 64 Leaf Switches, and 32 Spine Switches requires:

- 2,048 connections from Node to Leaf.
- 2,048 connections from Leaf to Spine.
- Optionally, if a Core is available, it would also require 2,048 connections.



SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	--	1024	1024	1024
8	256	2048	64	32	--	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
64	2048	16384	512	512	256	16384	16384	16384

Figure 20. 8 Scalable Unit (SU) Cluster/POD with cabling Levels A, B, C.

By utilizing the different levels of cabling that we have reviewed, we can understand the following:

- Data Center Operators can use Level A at the Scalable Unit to perform Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies. This would require either 256 individual 8-fiber jumpers of varying lengths or only 8 bundled jumpers of 256 fibers each, with varying lengths, per SU.
- Data Center Operators can use Level B to perform Leaf-to-Spine cabling using structured cabling. This would require the implementation of individual jumpers or bundled jumpers, adapter panels, housings, and multifiber trunks to serve as backbone cabling. With this backbone cabling, we can easily use 128 trunks (16 trunks of 144-fibers per Leaf), which would replace individual patch cords and bundled jumpers in a point-to-point connection. This method allows for spare ports, better management of complexity, and improved pathway space across the Data Center.
- If a Core is implemented, Data Center Operators can use Level C to perform Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

2.6.3. Cabling a 16 Scalable Unit (SU) Cluster/POD.

In Figure 21, cabling a 16 SU POD with 512 servers, 128 Leaf Switches, 128 Spine Switches and 64 Core Switches requires:

- 4,096 connections from Node to Leaf.
- 4,096 connections from Leaf to Spine.
- 4,096 connections from Spine to Core.

SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	--	1024	1024	1024
8	256	2048	64	32	--	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
64	2048	16384	512	512	256	16384	16384	16384

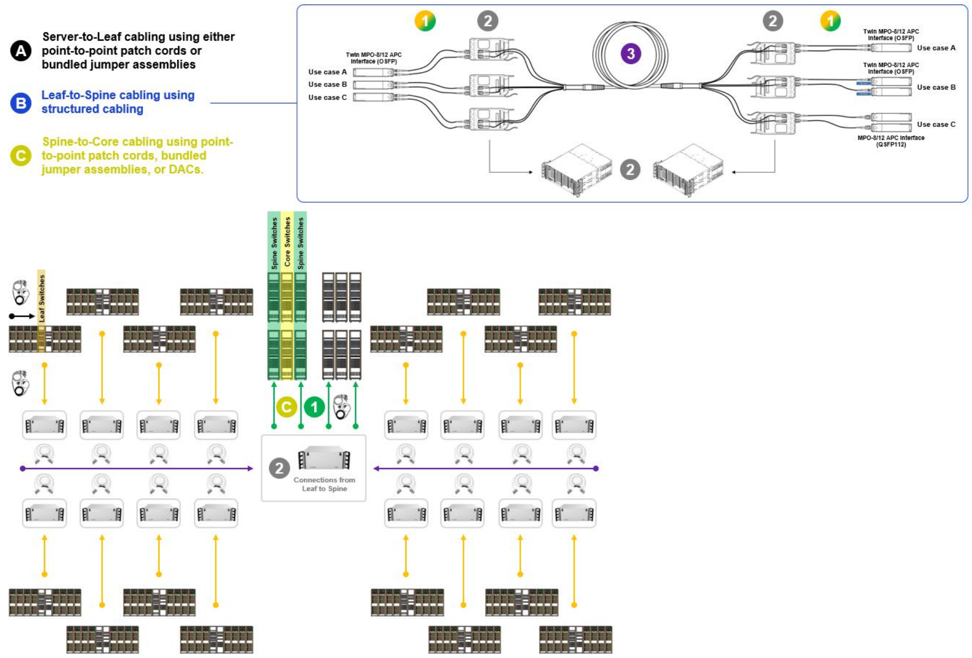


Figure 21. 16 Scalable Unit (SU) Cluster/POD with cabling Levels A, B, C.

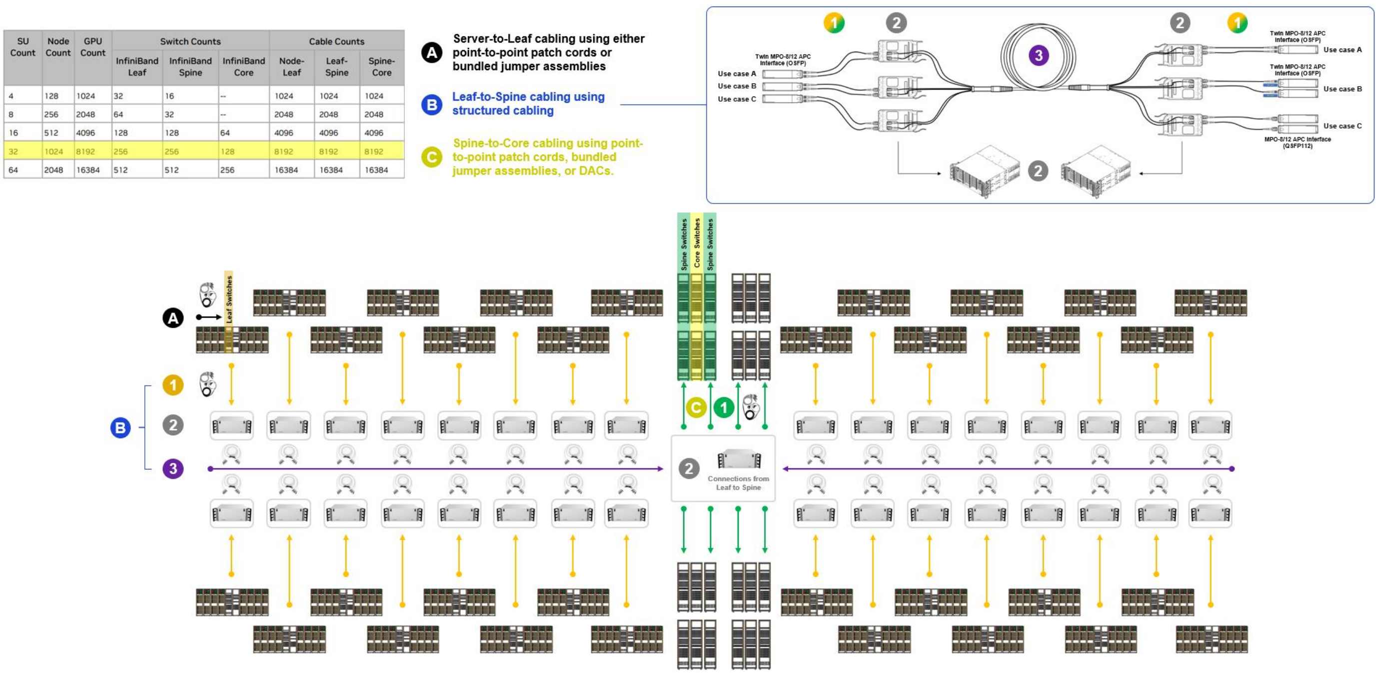
By utilizing the different levels of cabling that we have reviewed, we can understand the following:

- Data Center Operators can use Level A at the Scalable Unit to perform Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies. This would require either 256 individual 8-fiber jumpers of varying lengths or only 8 bundled jumpers of 256 fibers each, with varying lengths, per SU.
- Data Center Operators can use Level B to perform Leaf-to-Spine cabling using structured cabling. This would require the implementation of individual jumpers or bundled jumpers, adapter panels, housings, and multifiber trunks to serve as backbone cabling. With this backbone cabling, we can easily use 256 trunks (16 trunks of 144-fibers per Leaf), which would replace individual patch cords and bundled jumpers in a point-to-point connection. This method allows for spare ports, better management of complexity, and improved pathway space across the Data Center.
- Data Center Operators can use Level C to perform Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

2.6.4. Cabling a 32 Scalable Unit (SU) Cluster/POD.

In Figure 22, cabling a 32 SU POD with 1,024 servers, 256 Leaf Switches, 256 Spine Switches and 128 Core Switches requires:

- 8,192 connections from Node to Leaf.
- 8,192 connections from Leaf to Spine.
- 8,192connections from Spine to Core.



SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	--	1024	1024	1024
8	256	2048	64	32	--	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
64	2048	16384	512	512	256	16384	16384	16384

Figure 22. 32 Scalable Unit (SU) Cluster/POD with cabling Levels A, B, C.

By utilizing the different levels of cabling that we have reviewed, we can understand the following:

- Data Center Operators can use Level A at the Scalable Unit to perform Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies. This would require either 256 individual 8-fiber jumpers of varying lengths or only 8 bundled jumpers of 256 fibers each, with varying lengths, per SU.
- Data Center Operators can use Level B to perform Leaf-to-Spine cabling using structured cabling. This would require the implementation of individual jumpers or bundled jumpers, adapter panels, housings, and multifiber trunks to serve as backbone cabling. With this backbone cabling, we can easily use 512 trunks (16 trunks of 144-fibers per Leaf), which would replace individual patch cords and bundled jumpers in a point-to-point connection. This method allows for spare ports, better management of complexity, and improved pathway space across the Data Center.
- Data Center Operators can use Level C to perform Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

2.6.5. Cabling a 64 Scalable Unit (SU) Cluster/POD.

In Figure 23, cabling a 64 SU POD with 2,048 servers, 512 Leaf Switches, 512 Spine Switches and 256 Core Switches requires:

- 16,384 connections from Node to Leaf.
- 16,384 connections from Leaf to Spine.
- 16,384connections from Spine to Core.

SU Count	Node Count	GPU Count	Switch Counts			Cable Counts		
			InfiniBand Leaf	InfiniBand Spine	InfiniBand Core	Node-Leaf	Leaf-Spine	Spine-Core
4	128	1024	32	16	--	1024	1024	1024
8	256	2048	64	32	--	2048	2048	2048
16	512	4096	128	128	64	4096	4096	4096
32	1024	8192	256	256	128	8192	8192	8192
64	2048	16384	512	512	256	16384	16384	16384

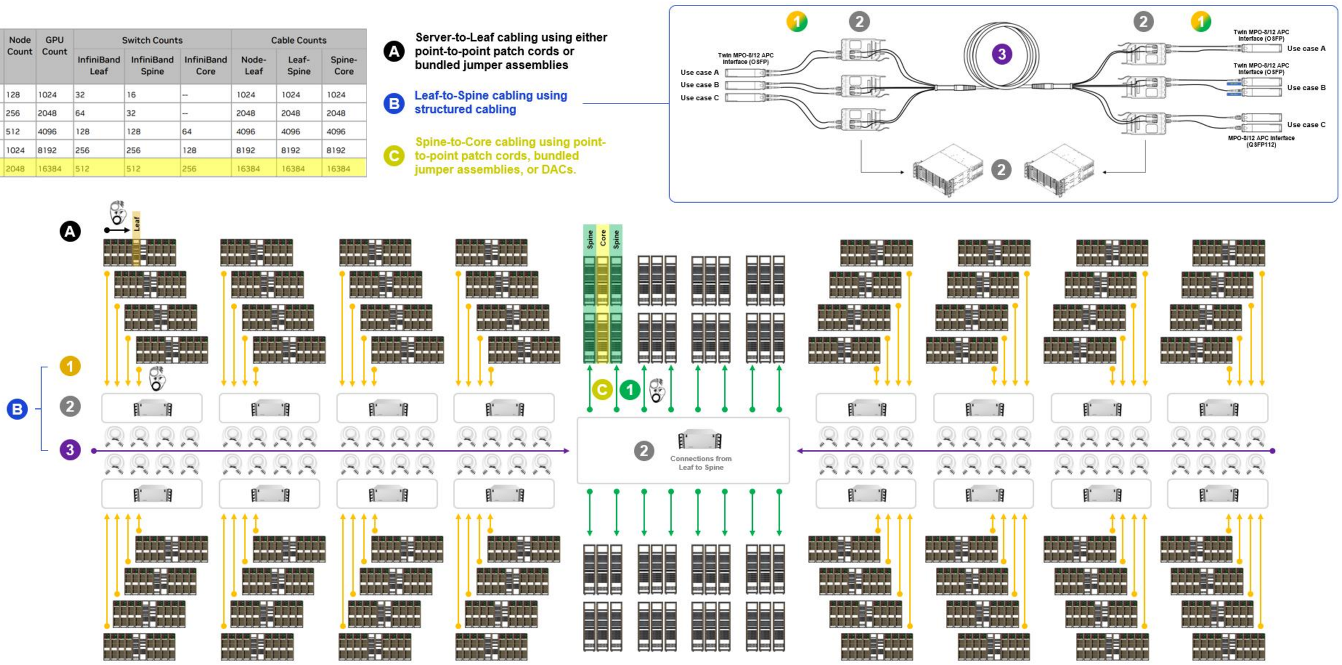


Figure 23. 64 Scalable Unit (SU) Cluster/POD with cabling Levels A, B, C.

By utilizing the different levels of cabling that we have reviewed, we can understand the following:

- Data Center Operators can use Level A at the Scalable Unit to perform Server-to-Leaf cabling using either point-to-point patch cords or bundled jumper assemblies. This would require either 256 individual 8-fiber jumpers of varying lengths or only 8 bundled jumpers of 256 fibers each, with varying lengths, per SU.
- Data Center Operators can use Level B to perform Leaf-to-Spine cabling using structured cabling. This would require the implementation of individual jumpers or bundled jumpers, adapter panels, housings, and multifiber trunks to serve as backbone cabling. With this backbone cabling, we can easily use 1,204 trunks (16 trunks of 144-fibers per Leaf), which would replace individual patch cords and bundled jumpers in a point-to-point connection. This method allows for spare ports, better management of complexity, and improved pathway space across the Data Center.
- Data Center Operators can use Level C to perform Spine-to-Core cabling using point-to-point patch cords, bundled jumper assemblies, or DACs.

2.7. Conclusion.

In summary, understanding the detailed cabling requirements for each level (A, B, and C) is crucial for optimizing the deployment of a DGX SuperPOD. In addition, implementing structured cabling where possible can enhance manageability and efficiency, particularly in large-scale setups.

Engaging with Corning engineering teams during the design phase ensures that cabling strategies align with specific data center layouts and customer requirements.

Annex 1 – High-Density Housings.

EDGE8® HD housings mount in 19-inch racks or cabinets and provide industry-leading ultra-high-density connectivity when combined with EDGE8 modules, panels, harnesses, trunks, and jumpers.
As each customer and project has specific needs, please add the housing that best suits your needs to the BOM:

	Part Number (Global)	Maximum Number of Modules or Panels	Maximum Fiber Density		Height
	EDGE8-01U-SP	18	LC	144 fibers	1U
			MTP®	576 fibers	
	EDGE8-02U	36	LC	288 fibers	2U
			MTP®	1,152 fibers	
	EDGE8-04U	72	LC	576 fibers	4U
			MTP®	2,304 fibers	

Table 11 - High-Density Housings.

Annex 2 – Polarity Drawings.

Polarity drawings, often referred to as fiber optic polarity diagrams, are essential when designing and implementing data center links using fiber optic cabling. They play a crucial role in ensuring proper connectivity, signal integrity, and compatibility between different network components.

This section will cover the specific polarity drawings applicable to each one of the scenarios previously described.

Scenario 1 - 800G and 400G - Server to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Point-to-Point Cabling.

Use Case A.

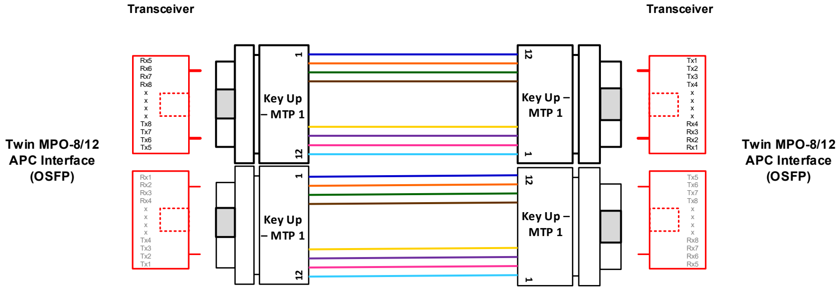


Figure 24 - Scenario 1 - 800G and 400G - Switch to Server Local - Use Case A.

Use Case B.

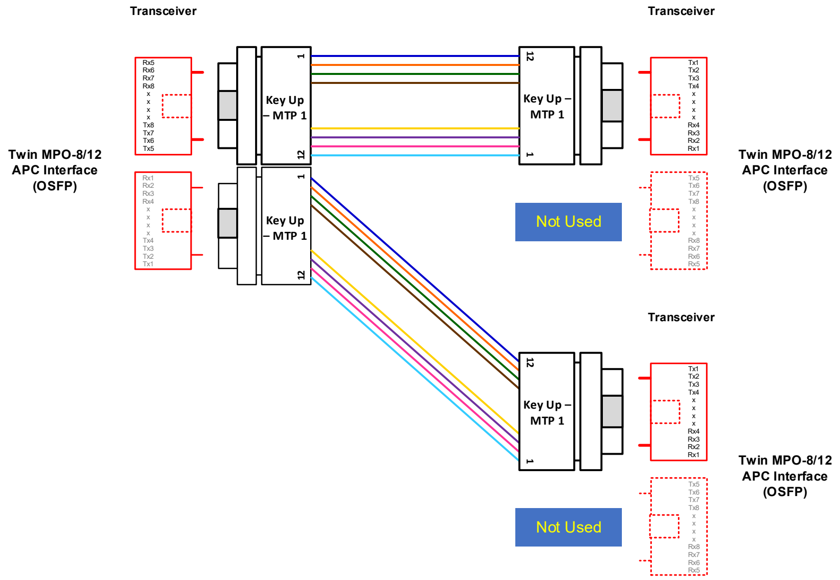


Figure 25 - Scenario 1 - 800G and 400G - Switch to Server Local - Use Case B.

Use Case C.

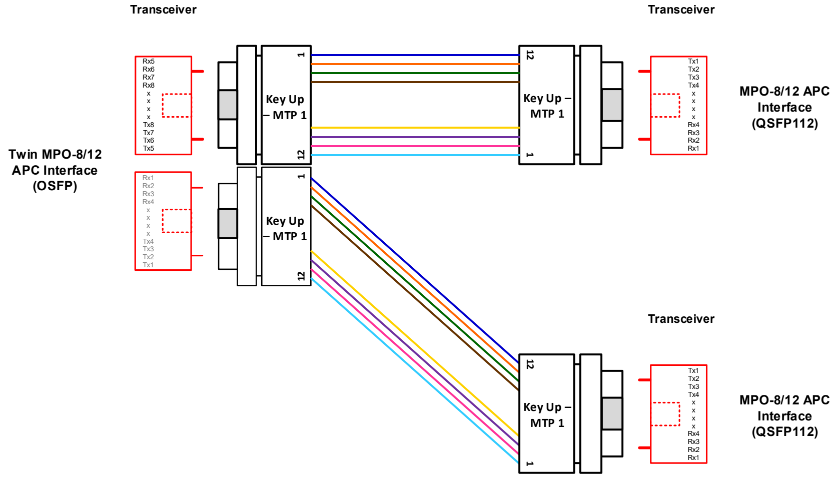


Figure 26 - Scenario 1 - 800G and 400G - Switch to Server Local - Use Case C.

Scenario 2 - 800G and 400G - Switch to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Structured Cabling Across DC with Trunk.

Use Case A.

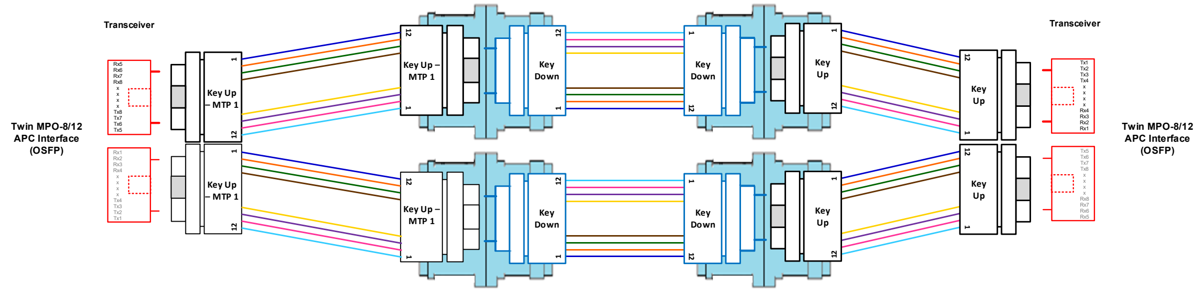


Figure 27 - Scenario 2 - 800G and 400G - Switch to Switch Across DC with Trunk - Use Case A.

Use Case B.

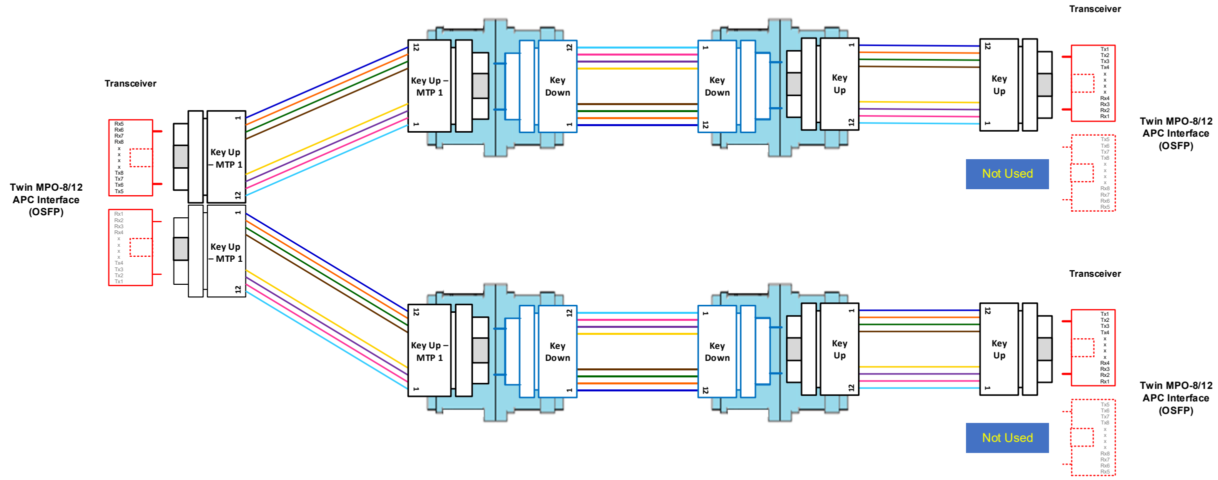


Figure 28 - Scenario 2 - 800G and 400G - Switch to Switch Across DC with Trunk - Use Case B.

Use Case C.

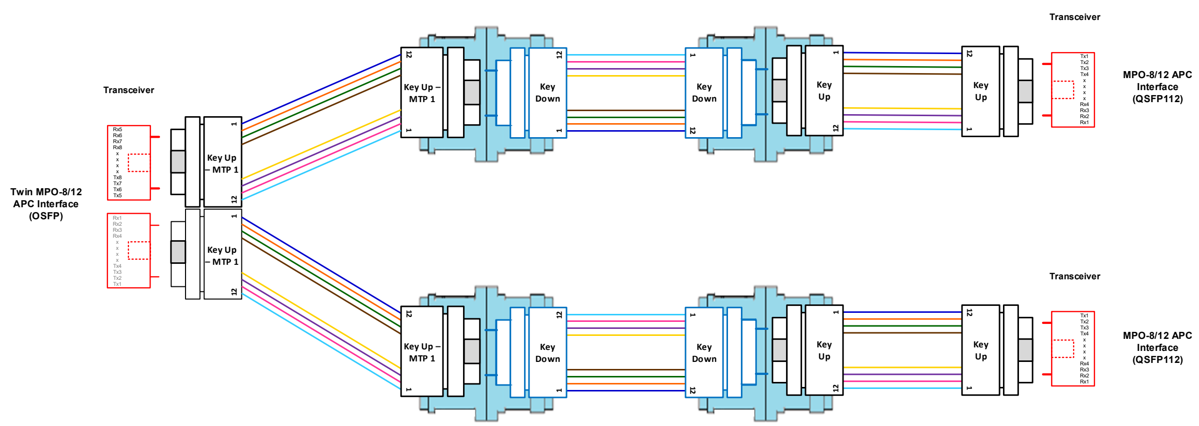


Figure 29 - Scenario 2 - 800G and 400G - Switch to Switch Across DC with Trunk - Use Case C.

Scenario 3 - 800G and 200G - Server to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Point-to-Point Cabling.

Use Case A.

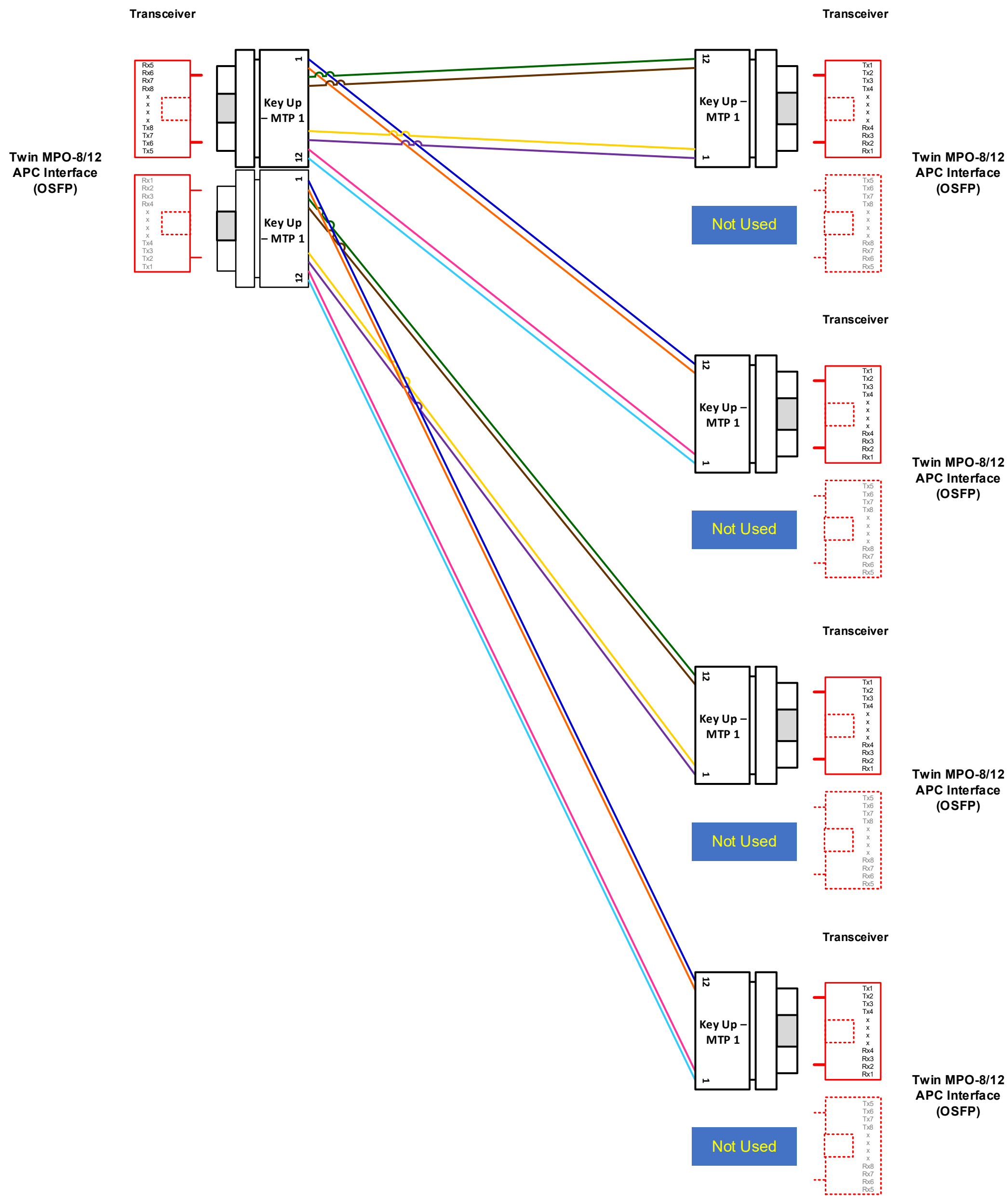


Figure 30 - Scenario 3 - 800G and 200G - Switch to Server Local - Use Case A.

Use Case B.

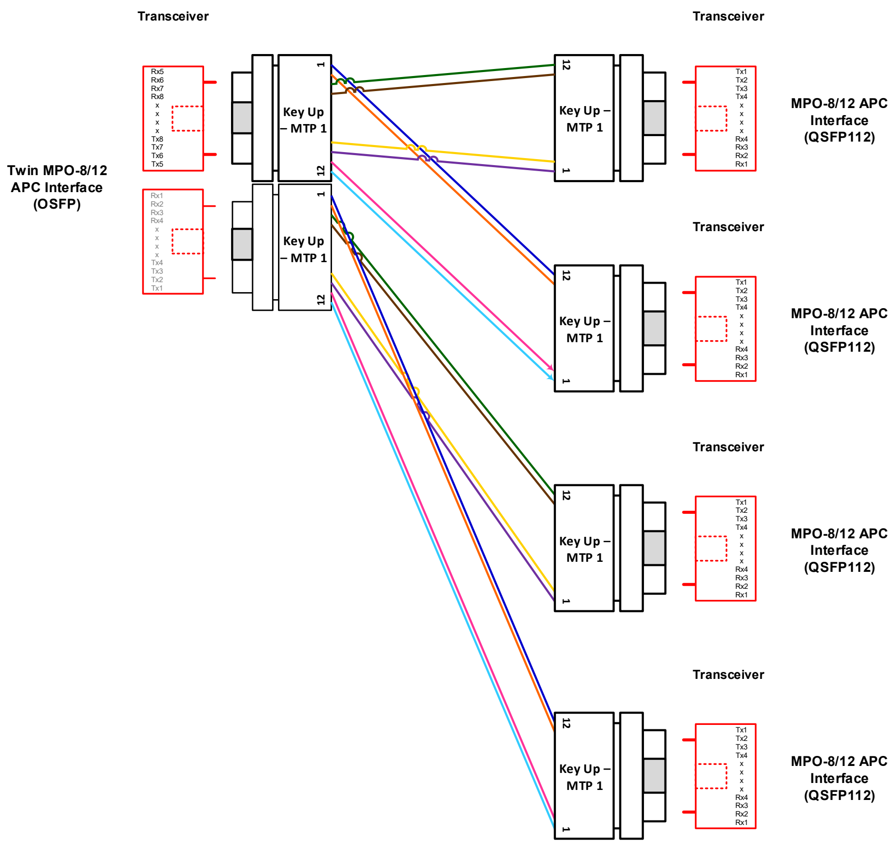


Figure 31 - Scenario 3 - 800G and 200G - Switch to Server Local - Use Case B.

Scenario 4 - 800G and 200G - Switch to Switch Applications.

MPO-8/12 APC to MPO-8/12 APC using Structured Cabling Across DC with Trunk.

Use Case A.

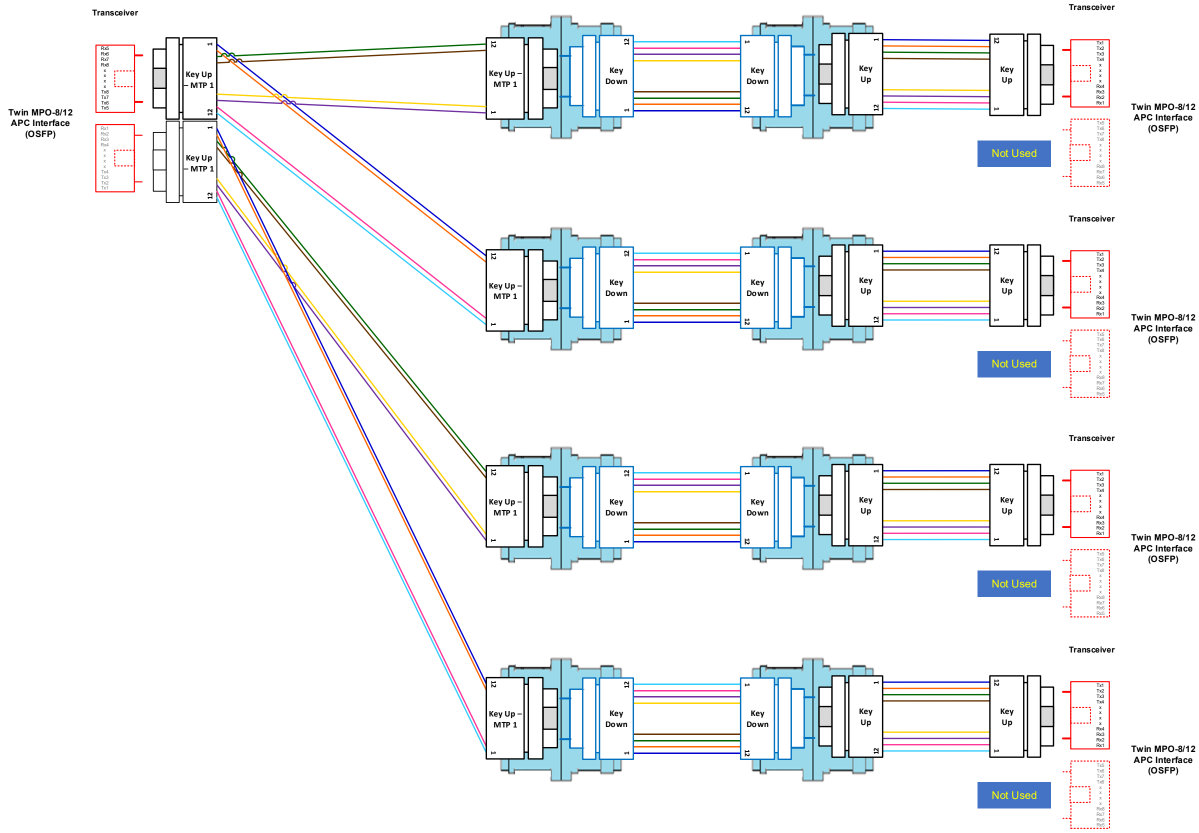


Figure 32 - Scenario 4 - 800G and 200G - Switch to Switch Across DC with Trunk - Use Case A.

Use Case B.

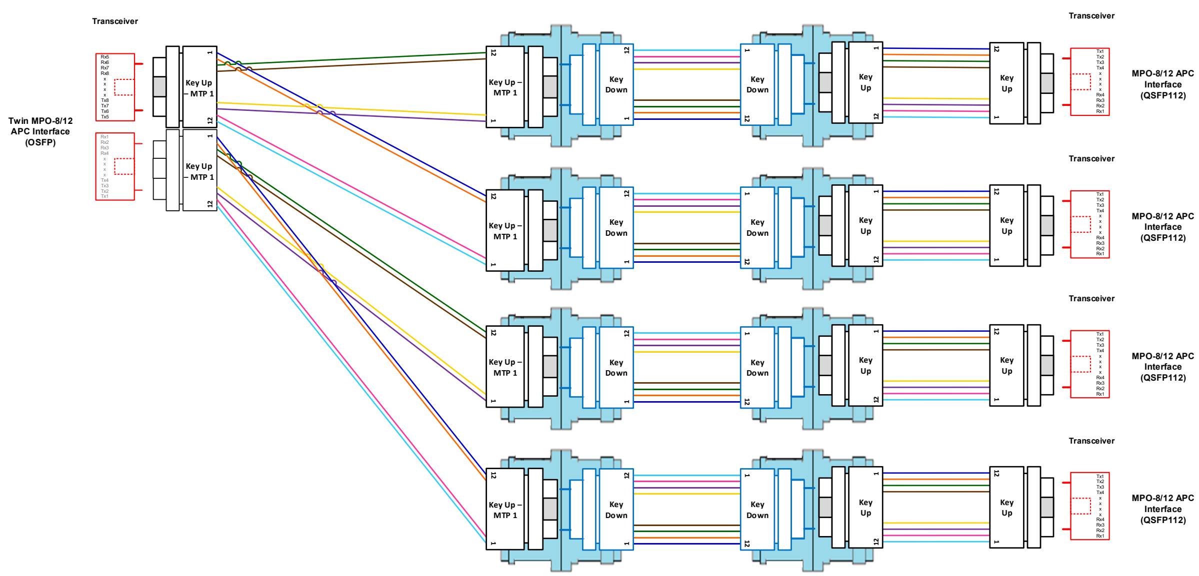


Figure 33 - Scenario 4 - 800G and 200G - Switch to Switch Across DC with Trunk - Use Case B.

Scenario 5 - 800G and 400G - Switch to Switch Applications.

LC-Duplex to LC-Duplex using Point-to-Point Cabling.

Use Case A.

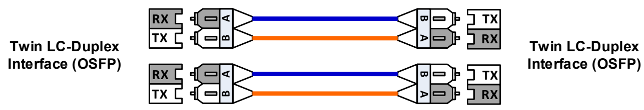


Figure 34 - Scenario 5 - 800G and 400G - Switch to Switch Local - Use Case A.

Use Case B.

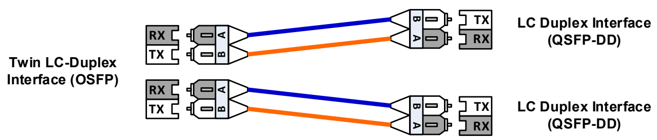


Figure 35 - Scenario 5 - 800G and 400G - Switch to Switch Local - Use Case B.

Scenario 6 - 800G and 400G - Switch to Switch Applications.

LC-Duplex UPC to LC-Duplex UPC using Structured Cabling Across DC with Trunk.

Use Case A.

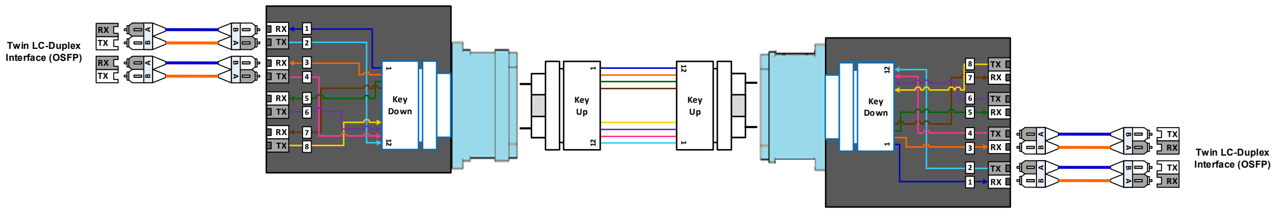


Figure 36 - Scenario 6 - 800G and 400G - Switch to Switch Across DC with Trunk - Use Case A.

Use Case B.

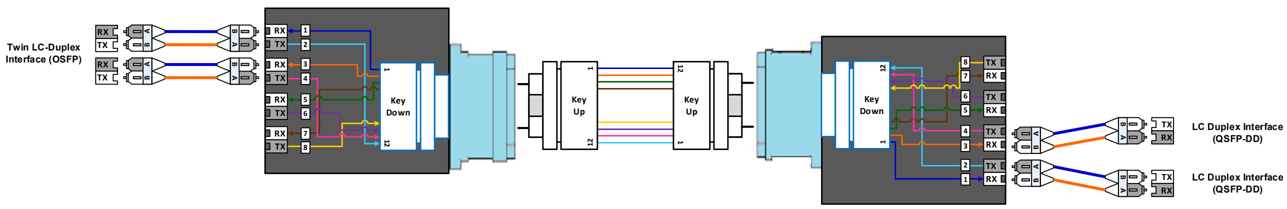


Figure 37 - Scenario 6 - 800G and 400G - Switch to Switch Across DC with Trunk - Use Case B.

Annex 3 – References and Contact Information.

This section contains a partial list of references to NVIDIA Overview White papers. For more detailed information on NVIDIA's products, please visit https://docs.nvidia.com/

Transceivers:

- **MMS4X00-NM** 800Gbps Twin-port OSFP 2x400Gb/s **Single Mode 2xDR4**, 500m https://docs.nvidia.com/networking/display/mms4x00nm800g500m/application+overview
- **MMS4X00-NS** 800Gbps Twin-port OSFP 2x400Gb/s **Single Mode 2xDR4**, 100m https://docs.nvidia.com/networking/display/800gmms4x00ns/overview
- **MMA4Z00-NS** 800Gb/s Twin-port OSFP, 2x400Gb/s **Multimode 2xSR4**, 50m https://docs.nvidia.com/networking/display/800gmma4z00ns/overview
- **MMS4X50-NM** 800Gb/s Twin-port OSFP, 2x400Gb/s **Single Mode 2xFR4**, 2km https://docs.nvidia.com/networking/display/mms4x50nm800g2kmpub

NVIDIA Cables:

- **MFP7E30-Nxxx**, Single Mode, **Straight** Crossover Fibers Cable https://docs.nvidia.com/networking/display/mfp7e30nxxxpub/specifications
- **MFP7E40-Nxxx**, Single Mode, **Splitter** Crossover Fibers Cable https://docs.nvidia.com/networking/display/mfp7e40nxxxpub/specifications
- **MFP7E10-Nxxx**, Multimode, **Straight** Crossover Fibers Cable https://docs.nvidia.com/networking/display/mfp7e10nxxx/specifications
- **MFP7E20-Nxxx**, Multimode, **Splitter** Crossover Fibers Cable https://docs.nvidia.com/networking/display/mfp7e20nxxx/specifications

DGX SuperPOD Architecture:

- **DGX H100** https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/dgx-superpod-architecture.html
- **DGX B200** https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/dgx-superpod-architecture.html
- **DGX GB200** https://www.nvidia.com/en-us/data-center/dgx-superpod-gb200/

Contact Us:

If you find yourself in need of additional support, please don't hesitate to reach out to our dedicated team. Please click here to draft an email. Kindly provide your full name, the name of your organization, and a comprehensive description of the inquiry or requirement for which you seek assistance. A member of our professional service team will respond promptly.

CORNING

Corning Optical Communications LLC • 4200 Corning Place • Charlotte, NC 28216 USA
800-743-2675 • FAX: 828-325-5060 • International: +1-828-901-5000 • www.corning.com/opcomm