NVIDIA Cable Management Guidelines and FAQ

Exported on Nov/06/2023 02:39 PM

Table of Contents

Introduction

High-speed interconnects are playing an increasingly more important role in data centers.

This note provides:

- An introduction to the terminology related to high-speed copper and optical cables and transceivers in general
- An introduction to NVIDIA's interconnect product families
- Recommendations for cables and transceivers installation

Data centers often use several different types of high-speed interconnects matching each interconnect type to specific requirements. Modern, high-speed data centers have focused on the following interconnects:

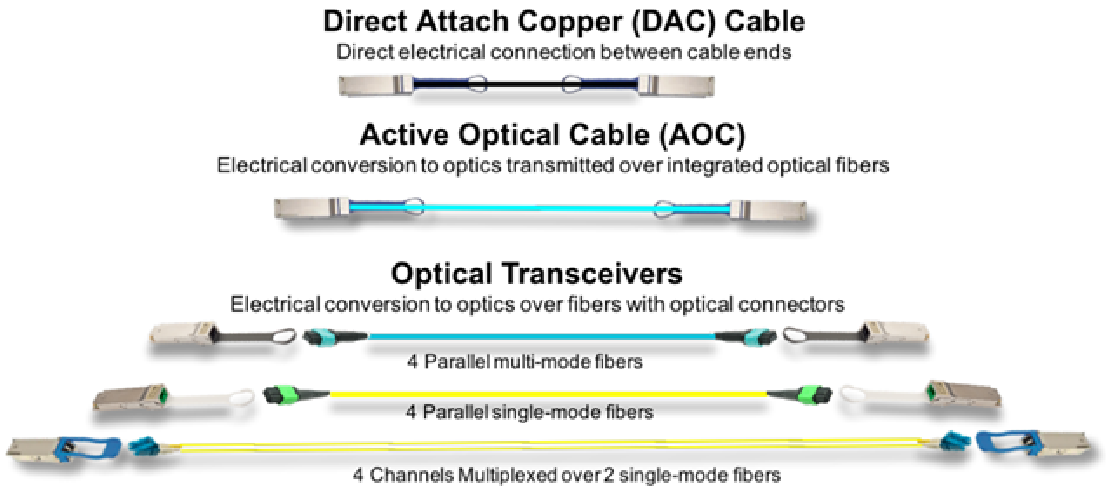
DACs (Direct Attach Copper) is the lowest cost, but after 2-5 meters (rate dependent) the attenuation of the signal is significant and becomes unrecognizable at the receiver.

AOCs (Active Optical Cable) are used from 3 meters to about 100 meters. It is not practical to install AOCs that are longer than 100 meters due to the trouble replacing them in case of issues.

Multi-mode Optics: More expensive SR (Short Range), SR4 (Short Range 4 Channels) multi-mode transceivers can be used up to 100 meters after which the signal degrades due to dispersion = the light bouncing around inside the large core multi-mode fiber and the signal pulse becoming too distorted to recognize at the receiver.

Single-mode Optics: Parallel single-mode transceivers (PSM4) are used up to 500 m. After 500 meters the cost of 8 fibers adds up with each meter, so multiplexing the four channels signals into one single fiber is more economical using CWDM (Coarse Wavelength Division Multiplexing)/FR up to 2 km and LR (Long Reach, 10 km), ER (20 km) or ZR (80 km).

Figure 1 – Different types of interconnects



8-channel devices are being added to the portfolio for 400G and 800G links on DACs, AOCs and transceivers.

Cable and Connector Definitions

Terminology and Basic Definitions

Cable Form Factors and Connector Types

SFP (Small Form Factor Pluggable) – A transceiver or cable with a one or two lanes (channel) in each direction. All cables and transceivers commonly used in datacenters are bidirectional.

SFP+ denotes the 10 – 14 Gb/s type of AOC/transceivers, while **SFP28** is the notation for the 25-28 Gb/s products with an SFP form factor. The noted data rate is the data rate in each direction.
SFP-DD, a double-density version of SFP, with 2 lanes in a form factor with same width as the SFP is defined, but are not part of Nvidia's product portfolio at the time of release of this paper.
SFP transceivers are part of the Ethernet architecture, but not used in InfiniBand systems.

QSFP (Quad Small Form Factor Pluggable) – A bidirectional transceiver or cable with 4 lanes in each direction.

Standards: Electrical pinout, memory registers, and mechanical dimensions for both **SFP and QSFP** devices are defined in the public MSA (Multi-source Agreement) standards available at: www.snia.org/sff/specifications.

QSFP+ denotes cables/transceivers for 4 x (10 – 14) Gb/s applications, while **QSFP28** denotes the 4 x (24…28) = 100 Gb/s product range with QSFP form factor, used for InfiniBand EDR 100Gb/s ports and 100Gb/s Ethernet (100GbE) ports. The **QSFP28** interface is specified in SFF-8679.

QSFP56 denotes 4 x (50…56) Gb/s in a QSFP form factor. This form factor is used for InfiniBand HDR 200Gb/s and 200/400GbE Ethernet cables/transceivers in Nvidia's portfolio.

QSFP-DD refers to a double-density version of the QSFP transceiver supporting 200 GbE and 400 GbE Ethernet. It employs 8 lanes operating at up to 25Gb/s NRZ modulation or 50Gb/s PAM4 modulation. QSFP-DD cables will in general not work in standard QSFP cages, but switches/NICs with QSFP-DD cages may support the older QSFP transceivers/cables.

OSFP (Octal Small Form Factor Pluggable) is wider and longer than QSFP and accommodates 8 lanes side-by-side. This form factor is used for 200/400/800G transceivers in Nvidia's InfiniBand NDR portfolio. More info on https://osfpmsa.org

AOC (Active Optical Cable) – An optical fiber cable with an optical transceiver with the fibers bonded inside and not removable. The optical transceiver converts the host electrical signals into light pulses and back. Bonding the fiber inside means the AOC only needs to be tested electrically and eliminates the costly optical testing.

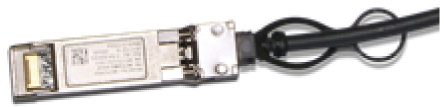
Transceiver (transmitter and receiver) is a converter with an electrical connector in one end and optical connector in the other end. It can have one or more parallel lanes in each direction (transmit and receive).

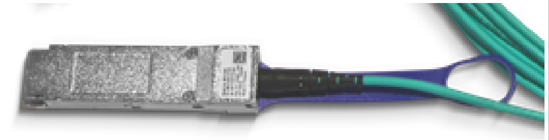
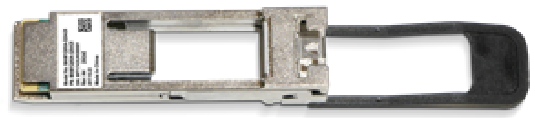
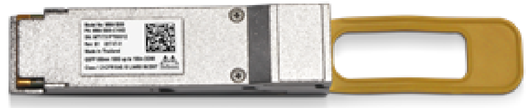
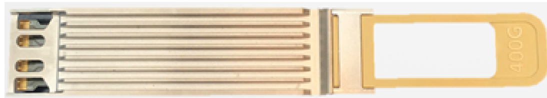
Transceiver or AOC? – You can argue that two transceivers connected with a patch cable replace an AOC. However, if you don't have cleaning tools and experience with optical connectors, it is safer to use an AOC where the optical cable is fixed inside the 'connector'. The AOC's 'connectors' are actually similar to detachable transceivers, but they work as a kit with a well-known transceiver at the other end. AOCs don't have any issue with multi-vendor interoperability. Nevertheless, it is easier to replace a pair of transceivers than an AOC since you don't have to install a new cable as the cable is already in place.

Traditionally, AOCs are more common in InfiniBand installations, while transceivers with patch cables are more common in Ethernet systems with structured cabling.

DAC (Direct Attached Copper) cable or **PCC (Passive Copper Cable)** – A high-speed electrical cable with an SFP or QSFP connector in each end, but no active components in the RF connections. The term 'passive' means that there is no active processing of the electrical signal. The DACs still have an EEPROM, a memory chip in each end, so the host system can read which type of cable is plugged in, and how much attenuation it should expect.

Cable/Transceiver Form Factors and Connector Definitions

Definition	Photo
DAC (Direct Attach Copper) cable with **QSFP** connector	
DAC with SFP connector	

Definition	Photo
AOC (Active Optical Cable) with **QSFP** connector	
QSA (QSFP to SFP Adapter)	
QSFP transceiver QSFP28 Transceiver for 100G transmission QSFP56 Transceiver for 200G transmission QSFP112 Transceiver for 400G transmission	 
QSFP-DD transceiver 8 lane 200/400G transceiver	
OSFP transceiver Single/Dual 8 lane 1/2 x 400G transceiver	
SFP transceivers 25G SFP28 Transceiver (~1 W)	

- QSFP56/SFP56 has 4/1-channels like the QSFP28/SFP28 generation but twice the data rate.
- Same Duplex LC and MPO-12 optical connector as QSFP28/SFP28 generation
- QSFP56 offers more space and thermal dissipation capacity
- 50G PAM4 doubles the data rate
- SFP56 ports accept SFP28 devices; QSFP56 ports accept QSFP28 devices
- QSFP28/SFP28 ports will NOT accept newer QSFP56/SFP56 devices
- SFP-DD ports will accept SFP+, SFP28, and SFP56 devices

SFP-DD is a 2-channel device, and hence requires a new optical connector scheme. Two types are currently (2019) supported by the SFP-DD MSA: Corning/US Conec MDC, and Senko SN.

Optical Transmission and Fiber Types

MMF (Multi-Mode Fiber) – The type of fiber used for VCSEL (Vertical Cavity Surface Emitting Laser) based transmission, normally operating at 850 nm wavelength. Its maximum reach is 100 m for 25 Gb/s line rates. Multi-mode fiber has a large light carrying core (50 µm) and matches the diameter of VCSEL lasers and PIN detectors making assembly very low cost.

OM2, OM3, OM4 (Optical Multi-mode) are classifications of MMF for different reach and speeds. Higher number indicates lower degradation of the optical signal, and longer reach. MMF cables commonly have the colors shown below, but standards are not fully consistent.

- **OM2** – **orange** – used for data rates at 1-14 Gb/s, 62.5 µm fiber core diameter
- **OM3** – **aqua** – 70 m reach for 25/100 Gb/s transceivers, 50 µm core diameter
- **OM4** – **aqua** – 100 m reach for 25/100 Gb/s transceivers, 50 µm core diameter
- **OM5** – **aqua green** – not commonly used (2023)

Multi-mode fiber patch cords

SMF (Single-Mode Fiber) – The type of fiber used for Indium Phosphide or Silicon Photonics based transceivers, operating at 1310 or 1550 nm wavelength. Single-mode fiber usually has a yellow jacket and can reach 100s of km. The tiny 7-9 µm light carrying core makes building single-mode optics much more expensive than multi-mode optics.

CWDM, WDM, DWDM, **(Coarse Wavelength Division Multiplexing, Normal, Dense)** – a technology for transmitting multiple optical signals through the same fiber. All signals have different wavelengths (colors). WDM transceivers make it possible to reduce the number of fibers in the link to two, one for transmit, and one for receive.

Dense WDM employs a very narrow 0.78 nm laser wavelength spacing used in single-mode links. The laser needs to be temperature controlled so these devices usually employ an electrical cooler – which adds cost.

Coarse WDM employs a wide 20 nm laser wavelength spacing used in single-mode links and because of the wide wavelength spacing does not require a cooler, so less expensive.

Short WDM (SWDM) employs 4 different wavelengths multi-mode VCSEL lasers.

PSM4 (Parallel Single-Mode 4 fiber) is the opposite of WDM in the sense that each signal is transferred in its own fiber. This requires 4 fibers in each direction but enables simpler transceiver design since all signals can have same wavelength and no optical MUX/DeMUX (AWG) is required and no TEC (Thermo Electric Cooler) to stabilize the laser wavelengths. PSM4 is a MSA (Multi Source Agreement), i.e. a standard supported by a number of transceiver vendors.

Reach of Transceivers

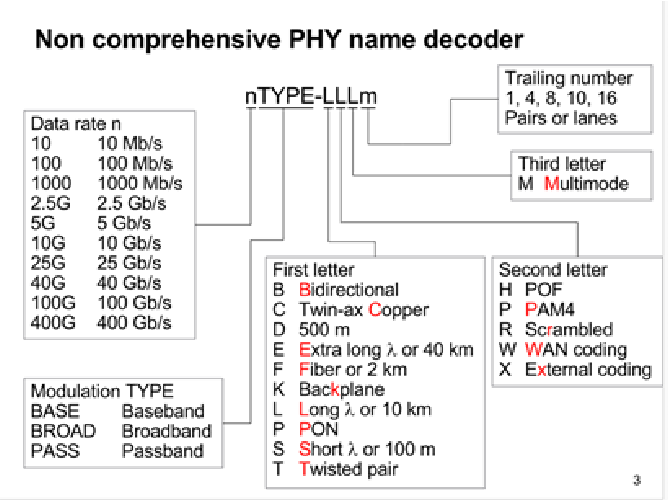
Transceivers are classified with data- rate and reach, governed by the IEEE Ethernet standards. For 100 - 400 Gb/s transceivers the most common definitions are:

- 100GBASE-**CR4** – 100 Gb/s, standard for DAC cables (twisted pair) for short reaches, up to about 7 m.
- 100GBASE-**SR4** –100 Gb/s, SR4=Short Reach (100 meters on OM4 multimode fiber), 4 fibers
- 100GBASE-**LR4** – 100 Gb/s, LR=Long Reach (10 km using WDM on SMF), 2 fibers
- 100GBASE-**ER4** – 100 Gb/s, ER=Extended Reach (30-40 km using WDM on SMF), 2 fibers
- 100GBASE-**ZR** – 100 Gb/s, ZR is not an IEEE standard, 80+ km reach.
- 200GBASE-CR4 – 200 Gb/s on DAC (passive copper) twisted pair cable, up to 3 m
- 200GBASE-SR4 – 200 Gb/s, SR4=Short Reach (100 meters on OM4 multimode fiber), 4 fibers
- 200GBASE-DR4 – 200 Gb/s, DR4 = 500 meters on single mode fibers, 4 fibers per direction
- 200GBASE-FR4 – 200 Gb/s, FR4 = 2 km, single mode fibers using WDM, 1 fiber per direction
- 200GBASE-LR4 – 200 Gb/s, LR4 = long reach, 10 km, single mode fibers using WDM, 1 fiber per direction
- 400GBASE-DR4 – 400 Gb/s, 500 meters on single mode fiber, 4 fibers each direction
- 400GBASE- FR4 – 400 Gb/s, WDM, 2 km on 1 single mode fiber/direction, 4 electrical lanes
- 400GBASE-FR8 – 400 Gb/s, WDM, 2 km on 1 single mode fiber/direction, 8 electrical lanes

All 200/400 Gb links use PAM4 signaling which implies that Forward Error Correction (FEC) is required.

The interface types listed above are examples for 100, 200, and 400 GbE links. The IEEE 802 standards define a wide range of standards for different Physical Media Devices (PMDs), see https://en.wikipedia.org/wiki/Terabit_Ethernet#200G_port_types. and PMD Naming Conventions figure below. Some of the transceiver types are not IEEE standards but separate industry MSAs (Multi-Source Agreements) usually formed by a leading transceiver company.– PSM4, SWDM4, CWDM4 and 400G FR4, are examples.

PMD Naming Conventions



Ref. https://ieee802.org/3/cn/public/18_11/anslow_3cn_01_1118.pdf

In the Data rate block, 200G (200 Gb/s) was added after 2018 when the above figure was published.

Optical Connector Types

High-speed cables make use of edge ‘gold-finger’ connectors on the electrical side which attaches to the host system (switch, network card on server/ storage). On the optical side, the following connector types are the most common:

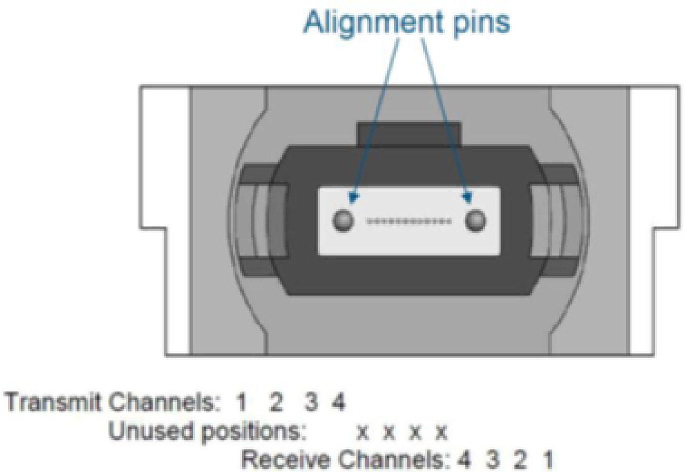
MPO (Multi-fiber Push On), is a connector standard supporting multiple rows with up to 12 fibers in each. A QSFP transceiver with MPO receptacle uses the outermost 4 positions on each side. The center 4 positions are not used.

Single-row MPO Connectors used in QSFP Transceivers

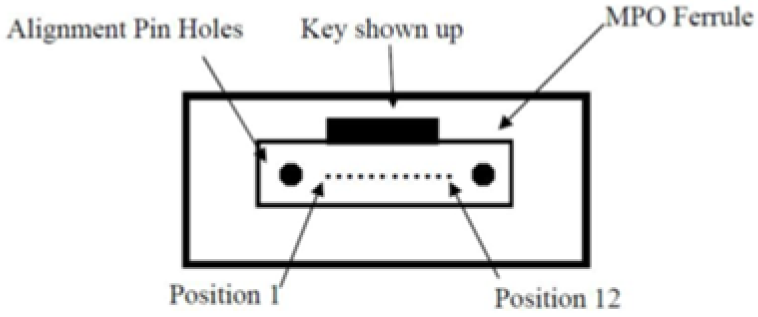
MTP connectors are a vendor specific proprietary high-precision version of MPO connectors.

The optical port in the parallel 2 x 4-lane QSFP optical transceiver is a male MPO connector with alignment pins, mating with fiber-optic cables with female MPO connector. The connector contains a 12-channel MT ferrule (allows to bundle multiple channels into a single connector).

QSFP28 Optical Receptacle and Channel Orientation for Male MPO Connector



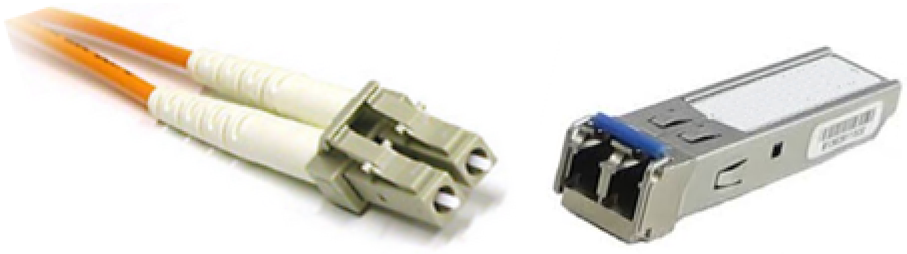
Female MPO Cable Connector Optical Lane Assignment



Reference: IEC specification IEC 61754-7.

LC connectors are used for both single-mode and multi-mode fibers and are used in both SFP and QSFP MSA transceivers.

Duplex LC Connector and SFP Transceiver with LC Receptacles

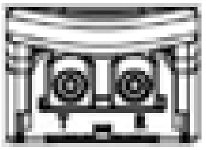
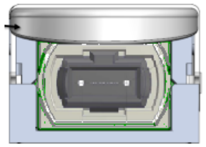
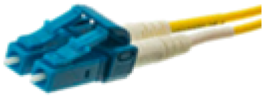
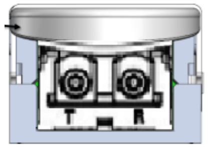


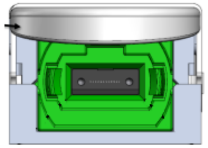
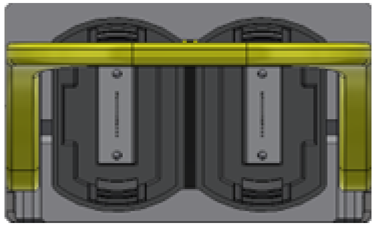
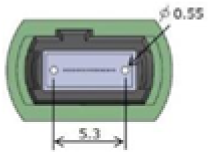
There are many other optical connector standards. MPO and LC are commonly used for data center patch cables and transceivers.

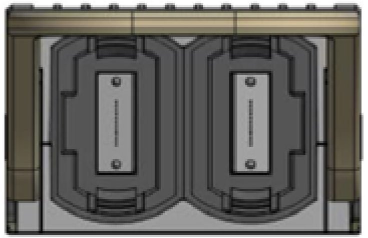
Optical Patch Cables

The choice of Optical patch cable depends on the type of transceivers you need to connect.

Transceivers and Cable Connectors

Transceiver	Reach and Type	Connector on Transceiver	Connector on Patch Cable
MMA2P00- MFM1T02A	25G SR SFP 10G SR SFP 2 fiber multimode	Multimode Duplex LC/UPC 	Duplex LC/UPC 
MC2210411-SR4 MMA1B00-xxxx MMA1T00-VS	40G SR4 QSFP 100G SR4 QSFP 200G SR4 QSFP 2x4 fiber multimode	Multimode Male MPO/UPC (with pins) 	Female MPO/UPC (with holes) 
MMA1L20-AR	25G LR SFP 2 fiber Single mode	Single mode Duplex LC/UPC	Duplex LC w single-mode fiber 
MC2210511-LR4 MMA1L30-CM MMA1L10-CR	40G CWDM, QSFP, 100G CWDM, QSFP, 2km 100G LR4 QSFP 2 fiber Single mode	Single mode Duplex LC/UPC 	

Transceiver	Reach and Type	Connector on Transceiver	Connector on Patch Cable
MMS1C10-CM MMS4X00	PSM4, QSFP, 500m 2x4 fiber single mode	Single mode MPO/APC (8 fiber, Angle polished connector) 	Female MPO/APC with single-mode fiber. The key is centered 
	Two 8-fiber Single mode in one unit		
T-DQ8FNS-N00 MMA4U00-WS-F	QSFP-DD SR8 2x8 fiber Multi-mode	Male MPO16/APC (16 fiber Angle Polished Connector) 	Female MPO16/APC with multi-mode fiber. The key is offset. 

Transceiver	Reach and Type	Connector on Transceiver	Connector on Patch Cable
MMA4Z00	OSFP SR8 Two 8-fiber Multi-mode in one unit	Male MPO12/APC (12 fiber Angle Polished Connector) 	Female MPO12/APC with multi-mode fiber 

Recently, NVIDIA devices with OSFP form factor have been expanded to work in both Ethernet and InfiniBand systems.

NV Link

A third type of application is NV Link (used in NVIDIA DGX systems). The DGX systems are equipped with either ConnectX-6 or ConnectX-7 HCAs (network adapters). Systems with ConnectX-6 adapters can use the MMA4U00-WS-F transceiver. Systems with ConnectX-7 adapters have OSFP connector and can use MMA4Z00 and MMS4X00 transceivers listed above.

UPC vs APC connectors

In the past, longer-reach single-mode applications like 100GBASE-LR4 allowed for greater insertion loss. With less-expensive transceivers entering the market comes a reduced insertion loss allowance. Compared to the 6.3 dB allowed for 100GBASE-LR4 which supports 100G up to 10 kilometers, the short-reach 100GBASE-DR applications up to 500 meters comes at just 3 dB. Just like 100G multimode applications, designers need to be aware of their loss budgets that could limit the number of connections in the channel.

With single-mode fiber and higher data rates, return loss is more of a concern. Too much light reflected back into the transmitter can cause bit errors and poor performance. The reflections can be significantly reduced using angled physical contact (APC) connectors, where an 8-degree angled end face causes reflected light to hit and be absorbed by the cladding.

Generally, there are some basic considerations related to the use of single-mode fiber. A single mode is more difficult to keep clean than multimode. A speck of dust on a 62.5 or 50 µm multimode fiber core blocks a lot less light than on a 9µm single-mode fiber core.

When inspecting APC single-mode connectors, you want to make sure to use an APC inspection probe tip designed to match the angle of the APC connector. This is required as part of the inspection equipment.

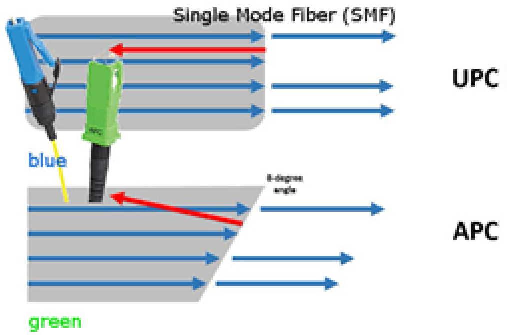
For APC connectors, note that not the entire end face of the connector is in contact with the cleaning device. It cleans the middle portion of the connector where the fibers are located and does not catch contamination at the outer parts.

While no damage will occur if you connect an APC connector to the input, you will get a warning about the received power being too low. To test products with APC connectors, you will need two hybrid UPC-to-APC cords and two APC-to-APC cords to make the connection. For Tier 2 OTDR testing, since reflections are absorbed by the cladding and return loss is very small when using APC connectors, the OTDRs will show APC connections as a non-reflective loss like a good fiber splice.

For 200GBASE-DR4 and 400GBASE-DR4 short-reach single mode applications, MPO connectors are in use as they require 8 fibers, with 4 sending and 4 receiving at 50 or 100 Gb/s. That's where a tester like Fluke Networks' MultiFiber Pro or Viavi's Sidewinder with dedicated on-board MPO connector which scan all fibers simultaneously is highly recommended to avoid time-consuming use of MPO to LC fan-out cords to separate the multiple fibers into single fiber channels.

For testing single mode fiber systems, you also want to make sure you're testing at both the 1310 and 1550nm wavelengths. Not only if these two wavelengths pass so will everything in between, but slight bends might not show up at the 1310 nm wavelength.

UPC vs APC connector



Connecting a server with QSFP network card/transceiver to a QSFP port in a switch

The fiber that connects with the transmitter's lane 1 must end at receiver lane 1 at the far end of the cable. Position 1 of the MPO connector at the near end of the cable connects to position 12 of the opposite MPO connector.

Use a patch cable with MPO connectors at both ends, and with crossed connections as shown below.

MPO to MPO Patch Cable Fiber Position

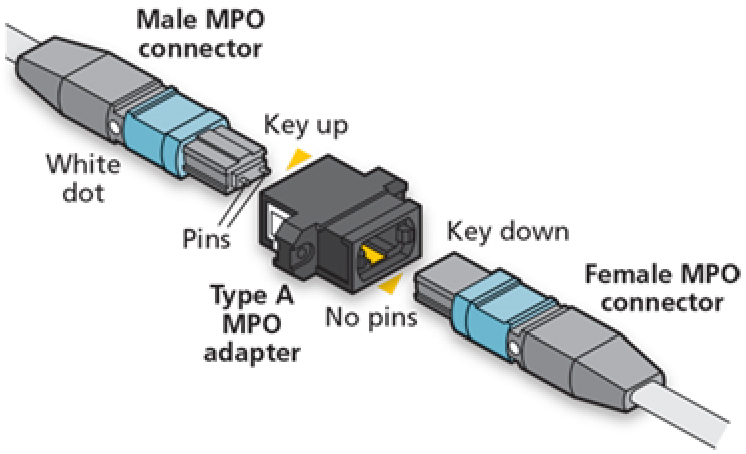
Left Cord	Connection	Right Cord
1	→	12
2	→	11
3	→	10

Left Cord	Connection	Right Cord
4	→	9
5	Not Connected	8
6	Not Connected	7
7	Not Connected	6
8	Not Connected	5
9	←	4
10	←	3
11	←	2
12	←	1

This is sometimes referred to as a ‘Type B cable’,
ref. https://www.flukenetworks.com/blog/cabling-chronicles/101-series-12-fiber-mpo-polarity

Multiple MPO patch cables can be connected in series, but each added connector pair increases modal dispersion in the link which again impairs performance. An odd number of ‘crosses’ must be used between transceivers at the two ends to get transmitters connected with receivers.

Connecting MPO Cables with an MPO adapter

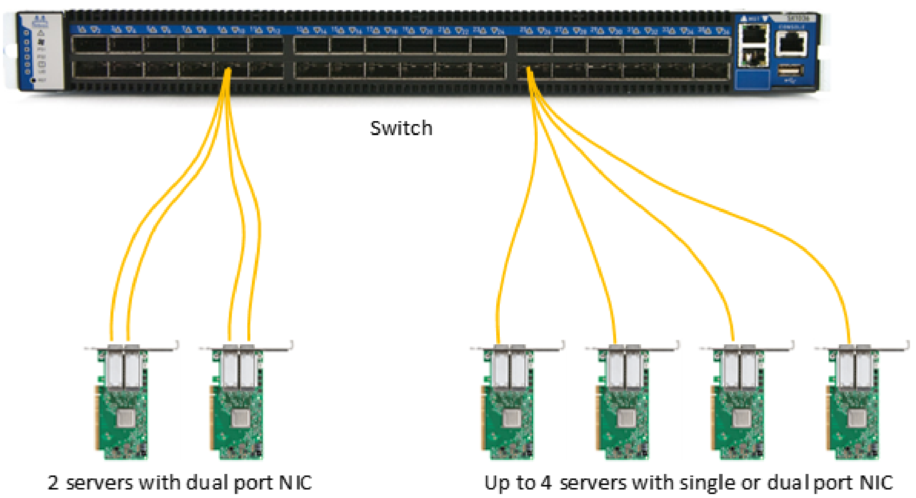


If two transceivers are to be directly connects, a “cross-over” fiber cable must be used to align the transmitters on one end to the receivers on the other end.

Connecting servers with SFP network card/transceivers to a QSFP port in a switch

A QSFP port and transceiver contains four independent transmit/receive pairs. I.e. you can connect 4 servers with SFP cards/transceivers to a single QSFP port in a switch. This enables connection of four 10GbE NICs to one 40GbE port, or four 25GbE NICs to one 100GbE port.
In either case you need an MPO to four Duplex LC splitter (breakout) cable. Either multi-mode or single-mode optics can be used depending on the reach needed.

Servers sharing QSFP Switch ports



The QSFP ports of the switch must be configured to work in split mode, with the 4 lanes working in ‘split’ mode; that is, the lanes operate as independent channels instead of operating as a single logic port. This can be achieved with passive copper splitter cables (DACs) or with optical splitter cables. Switch ports (not NIC ports) can be configured to operate in split mode.

Optical transceivers for the optical solution are not shown in the figure above.

Splitter cable examples: 25/100 GbE

- MCP7F00 – 100 Gb 1:4 splitter DAC, max 3 m
- MCP7H00 – 100 Gb 1:2 splitter DAC, max 3 m
- MFA7A20 – 100 Gb 1:2 optical splitter, up to 20 m long tails
- MFA7A50 – 100 Gb 1:4 optical splitter, up to 30 m long tails

Splitter cable examples: 50/200 GbE

- MCP7H50 – 200 Gb 1:2 optical splitter DAC, max 3 m
- MFS1S50 – 200 Gb 1:2 optical splitter, up to 30 m long tails
- MFS1S90 – 200 Gb 2:2 optical H-cable, up to 30 m long tails

Note 1: network adapter card ports cannot be split – only switch ports.

Note 2: The total number of ports that can be split with cables is based on the specific number of MACs inside the switch chip. See the switch documentation for specific configuration limits.

Optical splitter cables are available in the market for use between SR4 and SR transceivers.

InfiniBand

Port splitting/sharing a switch port across multiple servers was originally implemented for Ethernet, but is also available with the latest NDR generation of InfiniBand networking products. A wide variety of copper and optical cables have been developed for splitting 400/800G port capacity across 2 or 4 servers/hosts. Contact the NVICIA Networking team for more information on this topic.

Multi-mode splitter (breakout) cable

For longer reaches, a single-mode QSFP PSM4 transceiver can be connected to up to four NICs with LR transceivers using a single-mode splitter cable. Today, a common split is a 100G PSM4 split to 2x50G PSM4 transceivers used in large servers or storage systems.

Single-mode splitter (breakout) cable (not an NVIDIA product)

⚠ You cannot split the channels of a WDM transceiver using simple splitter cables. WDM transmitters use a single pair of fibers with the four channels carried on light of different wavelengths.

Networking Standards

LinkX® is the product line brand for NVIDIA's DAC, AOC and transceivers products that supports InfiniBand and Ethernet.

InfiniBand (IB) is a computer-communications standard used in high-performance computing that features very high throughput and very **low latency**. InfiniBand is commonly used in HPC (High-Performance Computing) and hyperscale datacenters. InfiniBand is promoted by the InfiniBand Trade Association (IBTA), http://www.infinibandta.org/. See InfiniBand: Introduction to InfiniBand for End Users for an introduction.

Ethernet (ETH) is a family of general computer networking technologies commonly used inside and outside datacenters. It comprises a wide number of standards, commonly referred to as IEEE 802.3, which is promoted by IEEE (www.ieee.org).

Form Factors, power classes, connector definitions and management interface specifications are found in https://www.snia.org/sff/specifications2.

InfiniBand (IB) and Ethernet (ETH) Cables Differences

The main differences between the two protocols are as follows:

- InfiniBand links up to Nx25 Gb/s; generally, don't use Forward Error Correction to minimize link latency. For higher data rates, FEC is a necessity.
- CDR (Clock and Data Retiming) default state:
  - IB EDR: Clock/data recovery (CDR, or retiming) is bypassed/disabled except for AOCs 30 m or longer running Nx25 Gb/s or lower rates.
  - IB HDR: Clock/data recovery (retiming) is as well as FEC are necessary for error free transmission due to the physical nature of PAM4 signaling.
  - Ethernet 100G: The CDR is default on.
    The CDR must be disabled to pass data at lower rates, for example 40 Gb/s.
  - 200/400 GbE (Ethernet) – CDR and FEC are both required for error free transmission. 25/100 GbE is supported but lower data rates are not generally supported.
- Copper cables:
  - IB EDR: The cable length and related attenuation determines if the operation can be achieved without FEC.
  - Ethernet 25/100GbE: Reed Solomon Forward Error Correction or RS-FEC is enabled by default for cables denoted CA-25G-L which are longer than 3 m. FEC is not required for cables denoted CA-25G-N which are up to 3 m long.

The EEPROM memory map of QSFP28 (100 Gb/s cables/transceivers) is defined in specification SFF-8636 for 4-lane transceivers, and for SFP28 (25 Gb/s cables/transceivers) in SFF-8472 for 1-lane transceivers.

Management of transceivers with more than 4 lanes is defined in the Common Management Interface Standard (CMIS), http://www.qsfp-dd.com/wp-content/uploads/2021/11/CMIS5p1.pdf

Transceivers with QSFP formfactor and 4 lanes can also be CMIS compatible. You need to read the memory map to tell if a given transceiver is the SFF or the CMIS type.

Memory map differences summary (informative):

- A summary is given in IB Vol 2 Annex A3.2: InfiniBand vs. Ethernet Memory Map Differences – QSFP/QSFP+ https://cw.infinibandta.org/document/dl/8125 (membership required).
- IB EDR loss budget (asymmetric): IB Vol 2 Annex A2.5 EDR Overall Link Budget for Linear Channels (informative)
- Ethernet: IEEE 802.3 clause 92 – copper cables, clause 83 – Physical Medium Attachment (PMA) including CDRs

LinkX Product Qualification

All LinkX® cables and transceivers for data rates up to InfiniBand EDR and 25/100 GbE (Ethernet) are tested in Nvidia end-to-end systems for pre-FEC BER of 1E-15 as part of our product qualification; more specifically, as part of the System Level Performance (SLP) test.

IB HDR, 200 GbE and higher data rates, cables and transceivers are different from previous generations. Due to the nature of physics of the PAM4 modulation used in these cables and transceivers, error-free transmission is only achievable with the use of FEC. This type of cables/transceivers are qualified at 1E-15 effective BER in Nvidia InfiniBand/Ethernet end-to-end systems.

Cable Installation and Management Guidelines

Unpacking and Handling

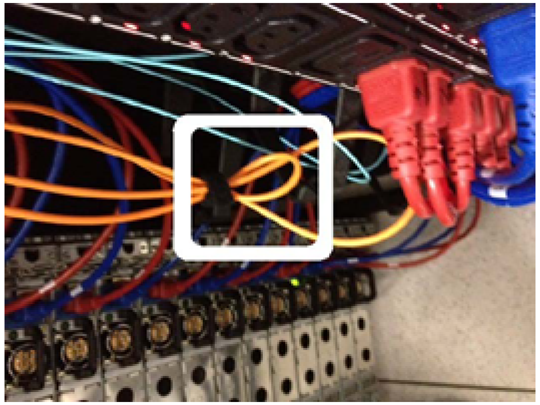
Unpacking of cables and transceivers should be done in an environment which complies with the storage humidly specified in the product datasheet. Cables and transceivers contain active electronic components. i.e. the products are designed with sufficient ESD tolerance to make them fit for use in datacenters. Please refer to the product datasheet of each cable/transceiver for specific ESD information and observe standard ESD precautions when unpacking and plugging the cables/transceivers into the host systems.

Warnings

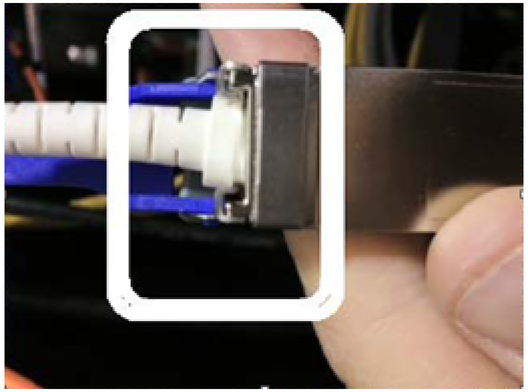
This section is based on experience from damages found on cables returned to Nvidia, which in most cases cannot be repaired.

- Do not twist the cables.
- Do not pull the cables.
- Do not staple the cables.
- Do not uncoil the cable, as a kink might occur.
- Do not step on the cable or connectors. Plan cable paths away from foot traffic or rolling loads.
- Do not pull the cable out of the shipping box through any opening or around any corners. Unroll the cable as you lay it down and move it through turns.
- Do not open a kink by twisting the cable. If it is not severe, open the kink by unlooping the cable.
- Do not pack the cable to fit a tight space. Use an alternative cable route.
- Do not hang the cable for a length of more than 2 meters (7 feet). Minimize the hanging weight with intermediate retention points.
- Do not drop the cable or connectors from any height. Gently set the cable down, resting the cable connectors on a stable surface.
- Do not cinch or fix the cable with hard fasteners or cable ties. Use soft hook-and-loop fasteners or Velcro ties for bundling and securing cables.
- Do not drag the cable or its connectors over any surface. Carry the entire cable to and from the points of connection.
- Do not force the cable connector into the receptacle by pushing the cable. Apply connection or disconnection forces at the connector only.
- Avoid over-bundling the cables or placing multiple bundles on top of each other. This can degrade the performance of the cables underneath.
- Do not bend the cable beyond its recommended radius. Ensure that cable turns are as wide as possible.

Extreme Bend Radius



Strain on the Connector



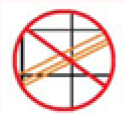
CAUTION: Do not kink cables.

CAUTION: Do not bend cables beyond the recommended minimum radius.

CAUTION: Do not pull cables without using the proper pulling equipment to eliminate strain on the cable or connector.

CAUTION: Do not twist the cables.

CAUTION: Do not step, stand or roll equipment over the cables.

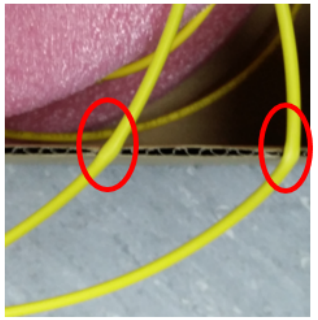


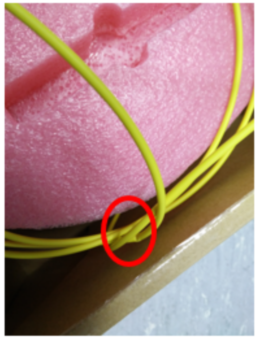
CAUTION: Do not lay cables on the floor.

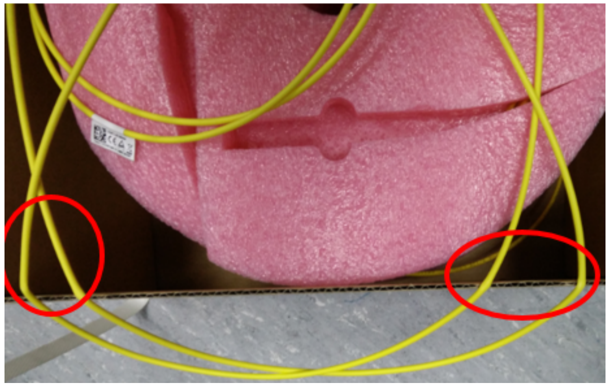
CAUTION: Do not pull cables without using the proper pulling equipment to eliminate strain on the cable or connector.

Deformed Optical Cable

Do not Kink Cables while Packing and Unpacking







Suitable Working Environment

Small impurities like dirt and dust can destroy fiber optics. Airborne particles are about the size of the core of a single-mode fiber. They dampen the signal and may scratch the connectors if not removed.

> CAUTION: Dirt is the most common cause of scratches on polished cable connectors and high loss measurements.

- Work in a clean area. Avoid working around heating outlets, as they may blow dust on the optical connectors.
- Always keep dust caps on connectors, bulkhead splices, patch panels or any other connection inlets.
- Avoid exposing cables to direct sunlight and areas of condensation.
- Remove unused cables which can restrict air flow. This is to prevent overheating.
- Avoid placing copper cables near equipment that may generate high levels of electromagnetic interference, e.g. power converters or air conditioners.
- Avoid running electrical cables near power cords, fluorescent lights, building electrical cables, and fire prevention components.

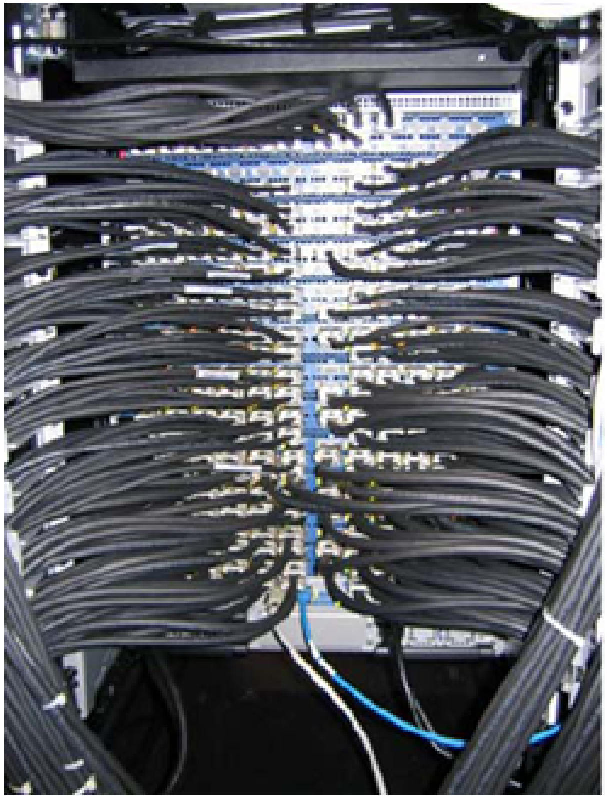
Installation Best Practices

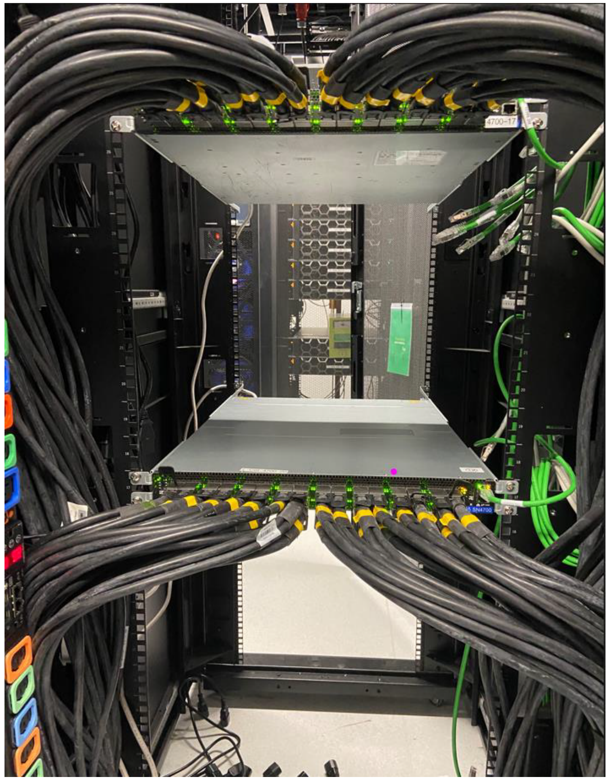
For the cable installation best practices within the rack, make sure of the following:

- Ample space and easy access to replace any transceiver and system units (e.g. fan units) in the switch.
- The rack is wide enough to place the cables between the switch and the rack's side walls.
- The cables do not block air flow.
- The cables do not block transceivers or system unit extraction.
- The cables are tied to the rack structure to remove strain and tension on the connectors.
- The weight of the cables is supported by the cable management system and the rack floor.
- Provide strain relief on the cable. Support the cables every 2 meters or place cables in a tray.
- Do not place cables and bundles where they may block other equipment.
- Avoid routing cables through pipes and holes, as this may limit additional future cable runs.
- Use Velcro based ties every 12” (30cm) to 24” (60cm).
- Mount the cables to the rack, one by one, before plugging them into the switch to ensure optimal use of space. Use velcros in the process to eliminate weight burden.
- Plug the cables in from the sides towards the middle. Bundle each half with velcros, then place the left half to the left side of the rack, and right half to the right side of the rack.
- Pre-bend the cable tips before plugging them into the switch. This will significantly reduce the strain on the port, especially thicker cables, such as 26 AWG copper cables.
- NVIDIA switch railkit allows recessing the switch deeper into the rack to minimize cable protrusion. The rack depth and air escape must be considered in this type of setup.

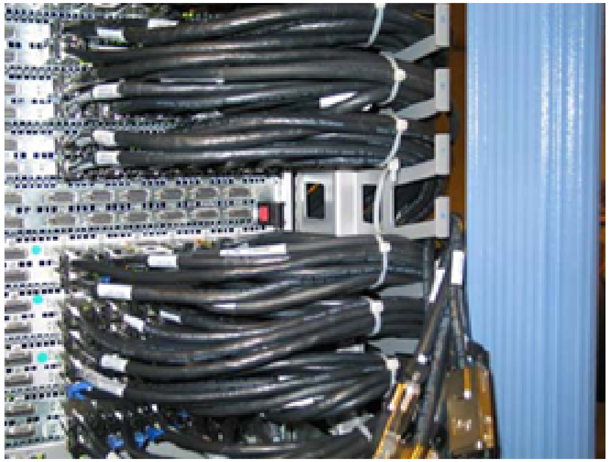
⚠ Test every cable as it is installed. Connect both ends and make sure that it has a physical and logical link before connecting the next cable.

Correct Installation Examples





Plan for Leaf Extraction



Cable Management Best Practices

- Multi-mode fibers come in two dimensions, 50 and 62.5 µm core diameter. Do not mix 50 µm cables with 62.5 µm cables in the same link.
- To ease management and troubleshooting, bundle cables together in groups of relevance (for example, ISL (Inter-Switch Link) cables and uplinks to core devices).
- Use cables of correct length. Leave only a little slack at each end. Keep cable runs under 90% of the max distance supported for each media type, as specified in the applicable standard.
- Keep copper and fiber runs separated.
- Install spare cables in advance for future replacement of damaged cables.
- Use color coding of the cable ties. The colors should indicate the endpoints. Place labels at both ends, as well as along the run.
- Locate the main cabling distribution area in the middle of the data center.

Cable Insertion

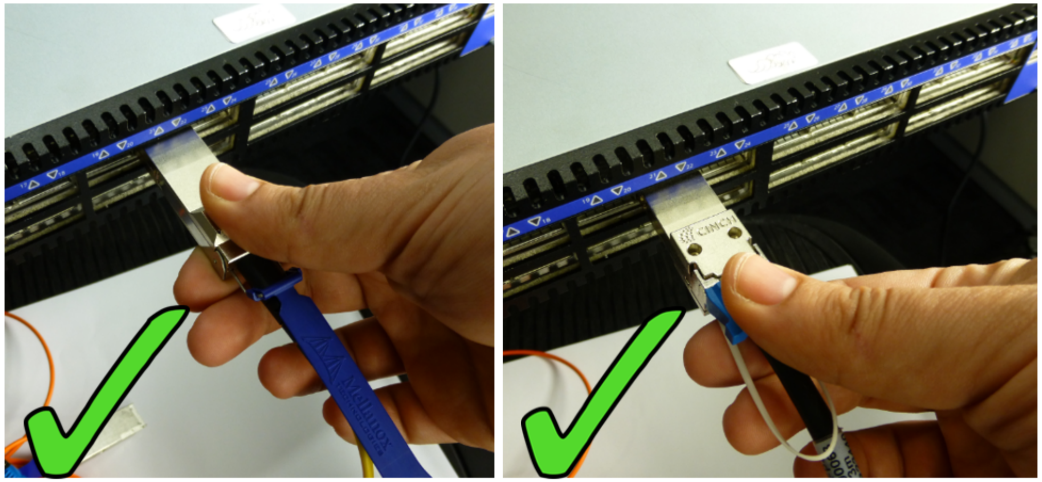
Please follow the instructions below to ensure smooth cable insertion and to avoid damage to the connector or target port.

- Grasp the cable by the lower part of the connector between your thumb and index finger. See figure below

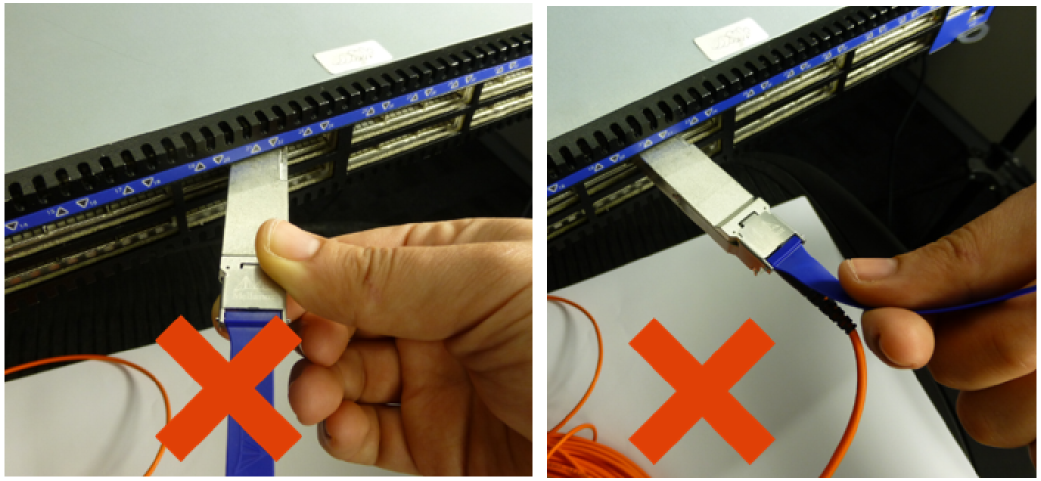
- Hold the cable connector perpendicularly to the ports panel, and gently push the connector into the port cage

> ⚠ Do not hold or use the pull-tab to insert the cable. The pull-tab is intended for cable extraction only.

Insert Connector Perpendicularly to Panel



Incorrect Insertion



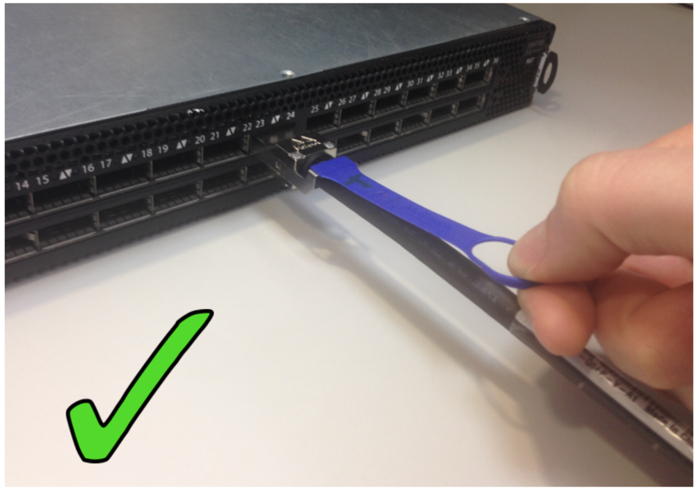
Cable Extraction

For a transceiver with integrated cable and pulltab, perform the following steps:

- Grasp the pulltab and pull it firmly, but gently outwards perpendicularly to the panel’s face, until the transceiver is released from the panel’s port cage
- Slide the transceiver and cable from the port

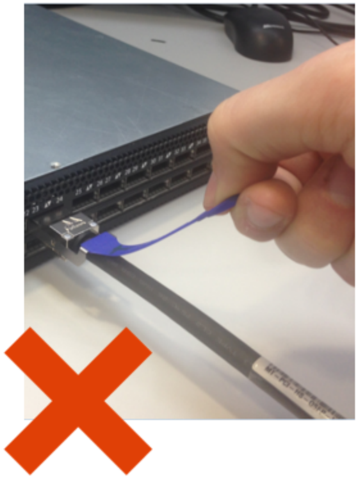
> ⚠ Do not grasp the cable’s jacket for extraction. Use the pulltab as shown in the figure.

Extract Pulltab Perpendicularly to Panel



Incorrect Extraction





Frequently Asked Questions (FAQ)

Optical Cables and Transceivers

Q: Is there any difference between electrical cables and optical cables and transceivers that I should be aware of when using the latter?

A: Optical cables offer a much longer reach for transmission of high-speed data. For AOCs the optical interface is not accessible, and the installation of the AOC is the same as for an electrical cable.

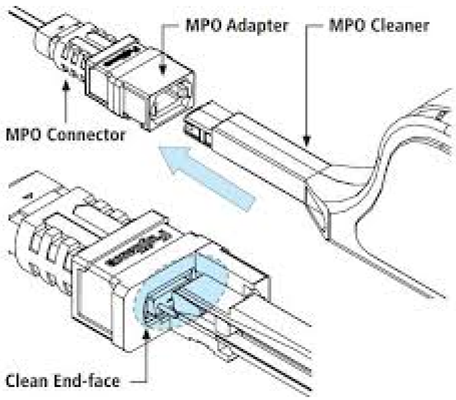
For optimal performance of EDR (25 Gb/s SFP28/100 Gb/s QSFP28) cables and transceivers, it is imperative to optimize the optical transmitter's input equalizer and the optical receiver output pre-emphasis to the host system (server, switch) characteristics. This is normally done by the host system in accordance with the MSA recommendation in SFF-8636. Alternatively, the AOC/transceiver can be specifically programmed to match the actual host system.

Optical links can also be composed of a pair of optical transceivers with an optical fiber cable between them. The transceivers have a detachable, accessible optical interface. Due to the small dimensions of the optical light beams, both the transceivers and the optical connectors on the cable must be cleaned prior to insertion.

Q: How do I clean the transceiver's optical connector?

A: A QSFP transceiver and its mating cable are terminated in MPO or MTP connectors. Fiber cleaning equipment is available in the market for easy cleaning of both cable connectors and transceiver connectors.

Cleaning an MPO or MTP Connector



Similar optical fiber cleaners are available for the LC type connectors that are used with SFP+, and SFP28 transceivers.

Q: With light coming out of the connector, is there a risk of eye damage?

A: The optical transceivers of the LinkX family emit infrared, invisible light. The power is a few mW only, and the connectors have no focusing lens at the emission point, so the light beam is spread out when leaving the connector. However, you should never look directly into a connected optical cable or a transceiver plugged into a host system.

Forward Error Correction

Q: why are CA-N and CA-L added to your latest generation of Nx25 GbE passive copper cables (DACs)? Which cable to choose for the 2.5 m version for which you offer both?

A: For short inter-rack communication links, direct attach copper (DAC) cables are commonly used. IEEE 802.3by specifies 3 different FEC settings, depending on the cable attenuation.

If link latency is critical for the system, then use the “CA-N” type of cable. Host systems (switches, network cards) by default disable forward error correction (FEC) for CA-N type cables.

For longer links, with loss specification > 16.48 dB cables are specified as ‘CA-S’ which require BASE-R FEC, and cables with > 22.48 dB attenuation are specified as ‘CA-L’ cables which require use of RS-FEC.

Length and thickness of copper cables both affect the attenuation. Attenuation increases with cable length, but decreases with cable thickness. Thinner cables have higher attenuation than thicker counterparts, but are less expensive, have lower weight, and are more flexible.

Q: What is FEC?

A: Forward error correction (FEC) is a process of adding an error-correcting code (ECC) to a message (‘payload’) so that this can be recovered by a receiver even when multiple errors (up to the capability of the code being used) were introduced during the process of transmission. The ECC is forwarded together with the data message. Since the receiver does not have to ask the sender for retransmission of the data, a backchannel is not required in forward error correction.

To correct errors and protect the data integrity, Forward Error Correction (or FEC) is used in 25G and higher data rates to improve data transfer throughput without the need for re-transmitting the data. FEC is a digital processing technique which effectively reduces the number of errors and helps extend the reach capability of both copper and optical cables. There are a number of FEC modes offering different degrees of error detection and recovery.

The IEEE 802.3by standard defines Clause 91 with Reed-Solomon FEC (RS-FEC) for 25 GbE (ethernet) to support most of the copper and optical interconnects. Clause 74 of the 100 GbE standard specifies BASE-R FEC, also known as fire-code FEC. FC-FEC offers a weaker error correction but with lower latency compared to RS-FEC. To achieve error free communication, the same type of FEC must be enabled on both ends of the link. Different configuration of FEC in the two host systems if one of the common interoperability problems when setting up high-speed communication links.

In Nvidia products the FEC processing takes place in the switches and network cards. Cables and transceivers are not part of FEC processing.

Q: What are N and L in the part numbers?

A: CA-N = FEC is NOT required for 0.5 m to 3 m. In Ethernet links, FEC is normally enabled, but for CA-N DACs it can be turned off by technicians to save power and latency. It is permitted but not required to use FEC.

CA-L = Long cables; require FEC for 3m-5m to ensure error free transmission.

For Nvidia systems, plugging in Nvidia DAC Type CA-N will cause the switch to automatically reduce the FEC algorithm to BASE-R FEC (that is, FireCode), saving ~ 30% in latency.

Customers can manually turn off FEC and save ~120ns per direction. The technician needs to set it up.

Since most of the traffic within a data center is up and down the rack to servers and storage, this is a big deal.

Q: What is the difference between CA-N and CA-L?

A: CA-N uses a thicker wire (26AWG) with enhanced shielding (higher cost), so it attenuates the high-speed signals less and turning off FEC is allowed.

CA-L uses a thinner wire (30AWG) with less shielding (lower cost) because the system knows FEC will be used to correct any data errors.

CA-N & CA-L only apply to 25G/50G/100G Ethernet; they do NOT apply to 100G EDR InfiniBand (which does not use FEC).

CA-N/L applies only to 25G and 100G DACs, but not to 200G or 400G DACs.

Weight of Cables and Transceivers

Q: How much does an AOC typically weigh?

A: In QSFP AOCs and between QSFP transceivers, a fiber cable with 8 fibers is used. This type of fiber cable weighs approximately 10 g/m. The weight per meter is practically the same for single mode and multi-mode fiber cables with the shielding used for indoor cables in data centers.

QSFP AOCs of the same length weigh the same, irrespective of the data rate.

AOC Ordering Part Number, Weight and Description

OPN	Weight [g]	Description
MFS1S00-x005E	166	Nvidia® active fiber cable, 200G, QSFP, LSZH, 5 m
MFS1S00-x010E	210	Nvidia® active fiber cable, 200G, QSFP, LSZH, 10 m

OPN	Weight [g]	Description
MFS1S00-x020E	283	Nvidia® active fiber cable, 200G, QSFP, LSZH, 20 m
MFS1S00-x030E	350	Nvidia® active fiber cable, 200G, QSFP, LSZH, 30 m
MFA1A00-C050	572	Nvidia® active fiber cable, 200G, QSFP, LSZH, 50 m
MFA1A00-C100	1074	Nvidia® active fiber cable, 200G, QSFP, LSZH, 100 m

Q: How much does a PCC (DAC) weigh?

A: Different types of cables are used for different PCCs. In the MCP1600 EDR PCCs, an 8-pair 24 AWG cable is used. This cable weighs 114 g/m. Add 40 g per QSFP connector/shell at each end.

PCC (DAC) Ordering Part Number, Weight and Description (Examples)

OPN	Weight [g]	Description
MCP1650-x00AE30	116	Nvidia® Passive Copper Cable, 200G, QSFP, LSZH, 0.5 m
MCP1650-x001E30	203	Nvidia® Passive Copper Cable, 200G, QSFP, LSZH, 1 m
MCP1650-x01AE30	290	Nvidia® Passive Copper Cable, 200G, QSFP, LSZH, 1.5 m
MCP1650-x002E26	300	Nvidia® Passive Copper Cable, 200G, QSFP, LSZH, 2 m
MCP1650-x003E26	387	Nvidia® Passive Copper Cable, 200G, QSFP, LSZH, 3 m

In the MCP2M00 PCC, a 2-pair 30 AWG cable is used. It weighs 24 g/m. Add 20 g per SFP28 connector/shell at each end. This adds up to approximately 90 g for the 2 m MCP2M00-002 SFP28 cable.

We can provide the weight of other cables on request.

Q: How much does a transceiver weigh?

A: The weight depends on the size, i.e. the form factor.

- An SFP transceiver weighs approximately 20 g.
- A QSFP transceiver weighs approximately 43 g, excluding the protective covers.
- An OSFP transceiver weighs approximately 260 g excluding protective covers.

The QSA (MAM1Q00A-QSA(28), QSFP to SFP) adapter weighs approximately 25 g.

Other Mechanical Characteristics

Q: Why do some of your datasheets have two specifications of a cable's minimum bend radius?

A: Cables typically tolerate a sharper bending (more stress) in the 'middle' of the cable, than by the strain relief and connector. The assembly bend radius indicates the larger minimum bend radius to use when bending the cable where it comes out of the connector shell.

Q: Why do different cables have different definition of their length?

A: When you measure the distance from port to port, you need a cable that is longer since part of the connector is inserted into the switch/NIC. For some cables the length is defined to exclude the inserted part, for others the length is tip-to-tip.

Q: Why don't you have a common fire certification covering all national standards for your cables?

A: Different parts of the world apply different regulatory standards, of which some apply to 'installation material'. Our longer fiber cables can be considered installation material while shorter patch cables and copper cables are not. Certification of cables is costly and time consuming, hence, it is done based on demand from different regions or customers.

Revision History

Rev.	Date	Description	Author
2.3	May. 2023	Updated Cable Installation and Management Guidelines	A. Elbash
2.2	Feb. 2023	NV Link section added, qualification section updated, UPC vs APC section updated, minor updates.	F. Kraemer
2.1	Dec. 2021	Intro, transceiver overview, and product weight updated.	F. Kraemer
2.0	Sep. 2021	Added new part numbers, minor updates and bug fixes.	F. Kraemer
1.9	Aug. 2020	Fixed a mistake in table 2.	F. Kraemer
1.8	Jun. 2020	Updated Section 1.4 w bug fixes + new transceivers added. Added section 1.4.1 about UPC vs APC connectors.	F. Kraemer
1.7	Mar. 2020	Document made available on docs.mellanox.com	A. Elbash
1.6	Aug. 2019	50/200/400G/HDR cables and transceiver section updated; Transceiver pictures updated; QSFP-to-SFP (QSA) adapter added; Minor text updates to all sections, more Q/A's added; CMIS reference added.	B. Smith/ F. Kraemer
1.5	Apr. 09, 2019	Section 1.7 on BER testing updated. 50/200/400/HDR cables and transceiver generation added. Examples of new 200/400G Ethernet standards added. Mellanox part numbers for passive optical cables removed. Explanation on use of splitter cables added. Power class definitions added for optical cables/transceivers.	F. Kraemer

Rev.	Date	Description	Author
1.4	Mar. 11, 2019	A new section with patch cable info added. Mechanical specifications in the FAQ updated.	F. Kraemer
1.3	Jan. 28, 2018	Updated: Table 1 – Cable/Transceiver Form Factors and Connector Definitions, page 8 New section: 2.1 Unpacking and Handling	F. Kraemer
1.2	Nov. 12, 2017	New format. Introduction added. All sections updated.	F. Kraemer B. Smith
1.1	Sep. 20, 2016	Added Figure 15 – Do not Kink Cables while Packing and Unpacking	
1.0	Aug. 16, 2016	Preliminary version	