Impact of Structured Cabling on AI Network Performance

White Paper

Table of Contents

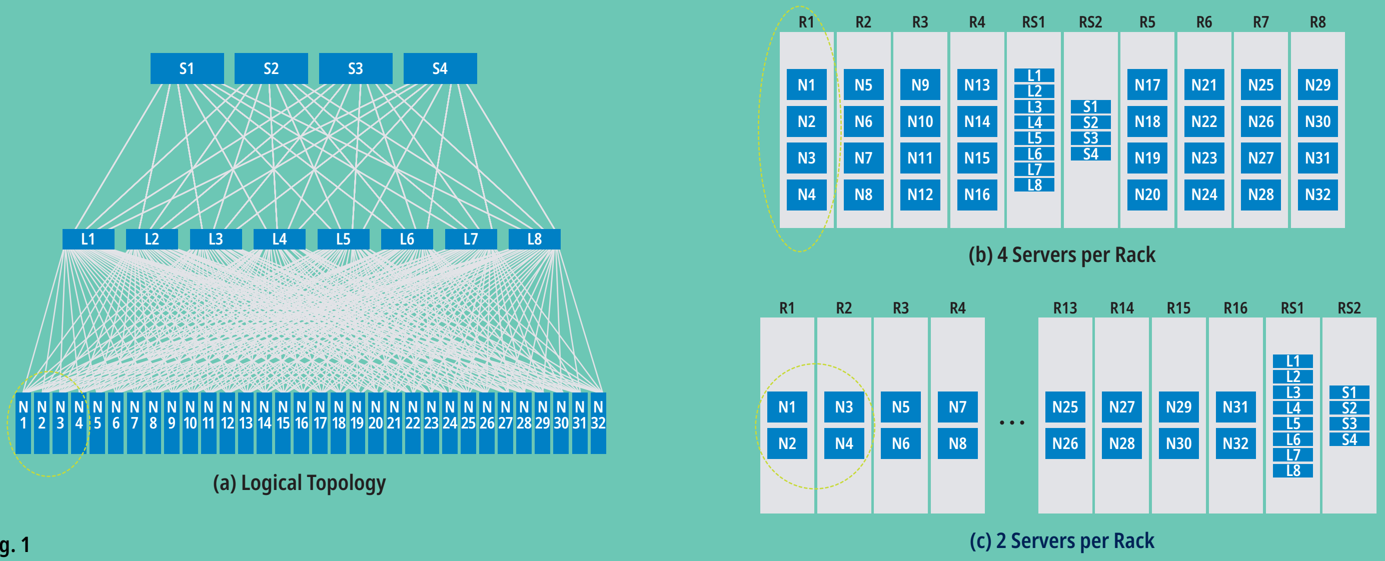
State-of-the-art Artificial Intelligence (AI) and Machine Learning (ML) systems for training or inference require very high bandwidth, low (tail) latency, and fabric topologies capable of interconnecting a large number of accelerators (GPUs, TPUs, or other types). Those systems use a specialized optical network, called the back-end network, typically InfiniBand (IB) links, or Ethernet with enhancement functionalities optimizing network traffic. Some Ethernet upgrades were already developed in IEEE task forces while others are currently being developed in the Ultra Ethernet Consortium.

AI training performance heavily depends on the underlying network physical layer performance, as training involves sequential computation and communication phases that must be fully completed before moving to the next one. Tail latency, which is dictated by the time to deliver the slowest messages in this computational sequence, significantly impacts training efficiency. It's observed that GPUs can be idle up to 50% of the time due to communication delays.

Communication networks consist of both electrical and optical links. Electrical links (e.g. NVIDIA's proprietary NVLink) connect GPUs within server nodes, providing very high bandwidth and low latency communication. However, the number of GPUs that can be interconnected using electrical links is limited due to the high signal losses in copper conductors at the high frequencies required for these data rates, restricting effective distances to just a few meters. As a result, the scalability of GPU interconnection via electrical links is constrained by these physical limitations.

AI optical back-end networks, which can potentially scale to tens of thousands of GPUs, typically use Spine and Leaf switches and rail-optimized topologies between Leaf switches and servers. Those complex topologies as shown in Fig. 1 (a) can be difficult to deploy, especially with direct point-to-point connections between server nodes to Leaf switches, and Leaf-to-Spine switches. The difficulty of implementing AI networks with direct connections will be even worse when future upgrades are needed. For many data centers, future infrastructure upgrades are necessary to keep pace with the growth of AI models and the services they provide. As the network grows from one to multiple clusters, scaling and maintaining the network becomes challenging when direct point-to-point connections are used.

This is where structured cabling can play an important role since its modular characteristics facilitate deployment, documentation, future upgrades and network scaling, while simplifying maintenance and cable management. This paper will also examine concerns regarding latency, reliability, and physical layer performance.



(a) Logical Topology

(b) 4 Servers per Rack

(c) 2 Servers per Rack

Fig. 1
Example of one NVIDIA scalable unit.
(a) Logical Topology with 32 server nodes (N)
(b) Physical layout of the 32 server nodes 8 Leaf switches and 4 Spine assuming 4 servers per rack.
(c) Physical layout of the same logical topology using 2 servers per rack.

Key Test Result Discovered

Latency is not adversely affected by structured cabling

Latency

Structured cabling introduces more connection points compared to point-to-point cabling. Adding optical connectors can increase optical losses but it does not increase latency. On the contrary, a structured cabling system provides more flexibility for optimizing and managing routing paths using less cable slack than point-to-point direct connections while also providing equal or better propagation latency.

Due to the relatively short distances used in AI networks (<50m for SuperPods), which correspond to around 250 ns of light propagation delay, other sources of latency produced at the transceivers and switches can become more relevant. For example, FEC encoding and decoding can take 100s of nanoseconds alone. Other switch processes like frame buffering and queueing contribute more to propagation latency (100s-to-1000s ns). Therefore, minimizing the the number of hops that packets go through the network reduces delays in the data shared among GPUs.

AI workloads, particularly in large-scale distributed systems, rely on the communication performance across multiple interconnected GPUs. Consequently, the network segments with longer communication delays have a significant impact on the AI system operation. In these conditions, tail latency becomes more important than the absolute or mean value of the latency.

Network topologies such as Spine-and-Leaf and rail-optimized, are designed to flatten the network to reduce the number of hops required for GPU-to-GPU communication, reducing tail-latency. Rail-optimized fabrics improve network performance by leveraging the high-speed internal links within nodes (e.g. NVLINK) to reduce the number of hops in the scale-out of the network. In rail-optimized topologies, specific GPUs of all the servers have to be connected to the same Leaf switch. For example, the GPU 0 of server A and GPU 0 of server B are connected to Leaf 0. The same order is followed for the other GPUs, meaning that GPUs 7 are connected to Leaf 7 as shown in Fig 2.

Fig. 2 (a-b) also illustrates how this configuration reduces the latency of the interconnection. In part(a) the traditional method of communication between GPU 0 of server A to GPU 7 of server B, requires multiple hops through Leaf and Spine switches. The paths the signal follows on the network are highlighted with yellow lines for Paths 1a, 2a, 3a, and 4a. Three hops are required for that communication. A hop through Leaf 0 that connects Path 1a to Path 2a, a hop through a Spine that connects Path 2a to Path 3a, and another hop through Leaf 7, to connect Path 3a to Path 4a. Each hop requires electrical-to-optical, optical-to-electrical conversions, FEC encoding/decoding, and switch queuing, all of which add latency.

In contrast, using this rail-optimized configuration shown in Fig. 2(b), to communicate the same GPUs requires only one hop at the optical network.

To enable this, GPU 0 of server A sends the data directly to GPU 7 within the same server using its internal high-bandwidth link (Path 1b). Then the communication between GPU 7 in Node A and GPU 7 in Node B requires one hop at Leaf 7 to connect Path 2b and Path 3b as shown in the figure.

The advantages of flat network topologies to scale out AI workloads are well understood. However, deploying flat topologies such as a rail-optimized network requires precise connection and cable mapping which creates a complex deployment if using direct point-to-point cabling. Structured cabling facilitates a more simple deployment and management of those networks.

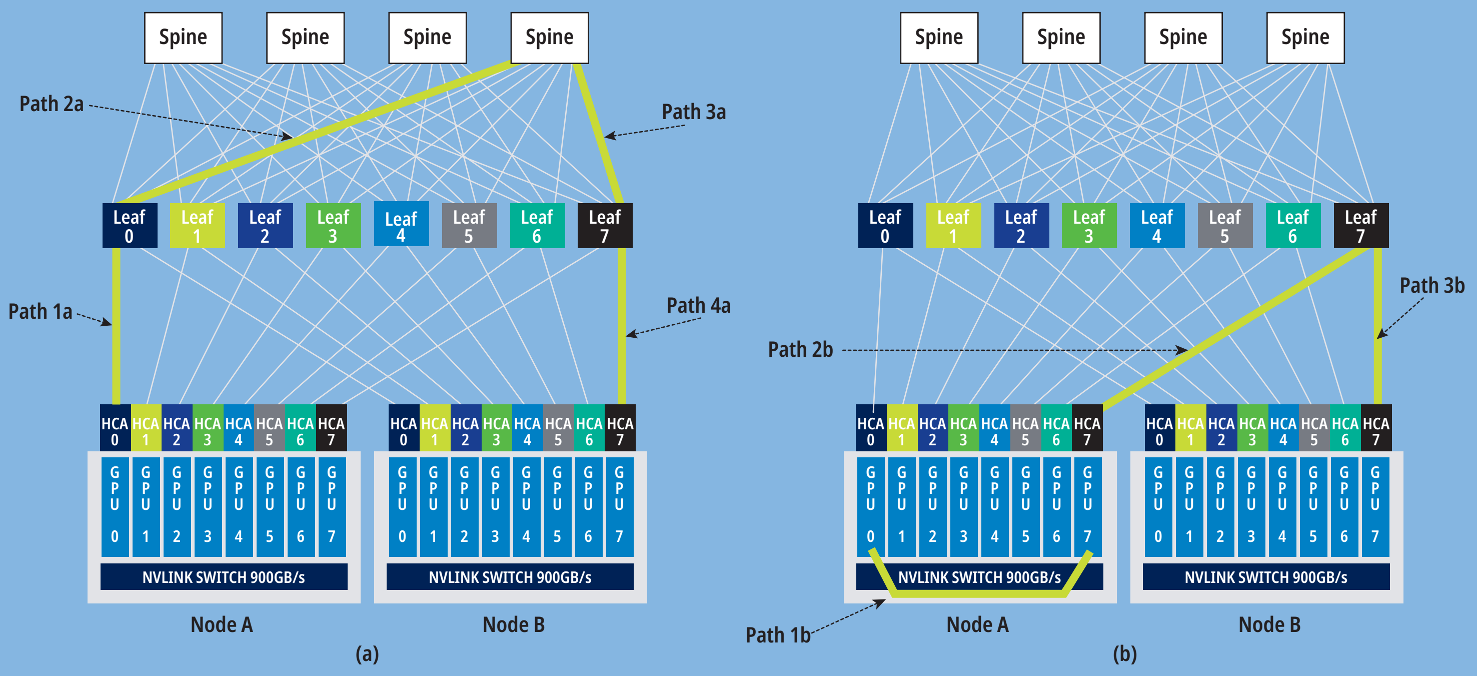


Fig. 2
llustrates a rail-optimized topology configuration. Yellow traces show the signal patch to compare latency when using (a). The Leaf and Spine switches and (b) The NVLINK and Leaf switches.

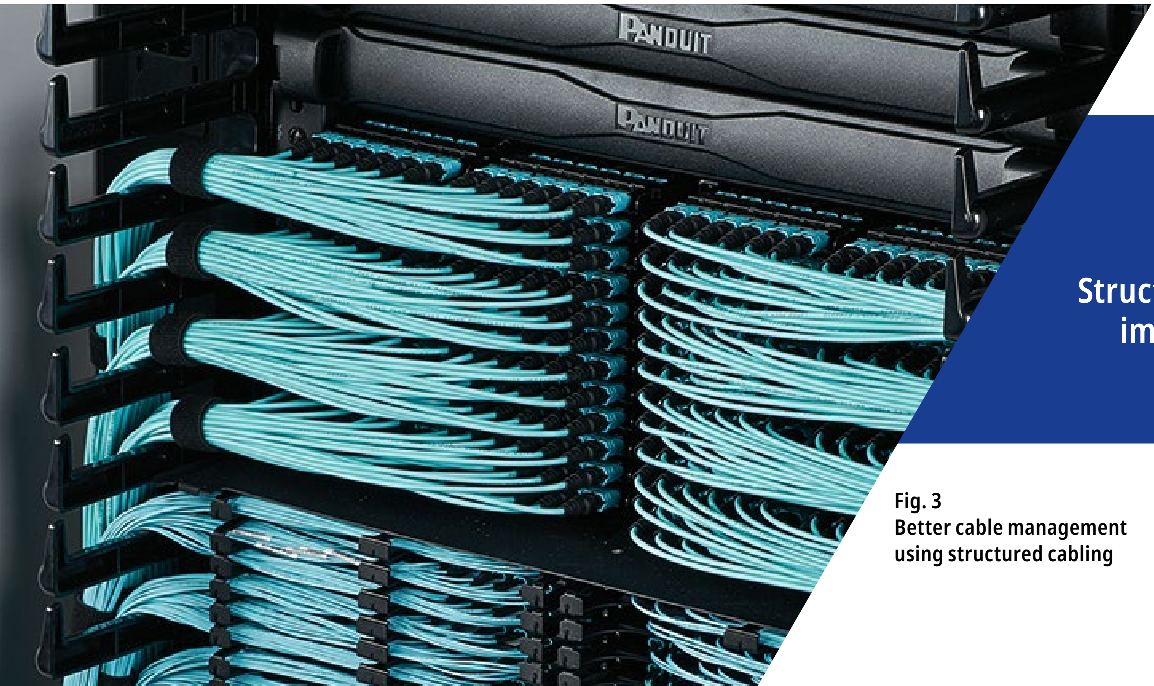
Reliability

Direct connections on large AI networks can often result in messy, and disorganized cabling. Cable slack for interconnects and trunks is typically longer in direct connections increasing the likelihood of finding tangled cables deployed with a smaller than recommended bend radius. This can cause stresses, higher losses, or permanent damage to the fiber leading to higher failure rates. In some cases, the stress on the fiber can develop into cracks in the glass over time. When issues like these arise, they can be challenging to identify and resolve swiftly. In contrast, structured cabling, shown in Fig. 3 with trunks separated from patch cords via patch panels, simplifies cable management and reduces slack. Trunks are routed through cable trays, while patch cords connect to servers or switches, improving organization, ease of maintenance, and upgrades critical for future network scaling.

Network documentation plays a crucial role in improving the reliability of large-scale networks, especially when dealing with thousands of optical links. When network circuit paths and cable routes are clearly documented, network engineers and service technicians can quickly trace connections when problems arise. Therefore, critical AI network downtime is reduced by avoiding time spent on identifying points of failure, while saving maintenance and service costs and optimizing large capital spent on the AI system. Telecommunications Industry Association (TIA) standard TIA-606-C (Administration Standard for Telecommunications Infrastructure) provides guidelines for labeling and recording data for telecommunications and network systems in commercial buildings, in alignment with structured cabling standards such as TIA-568.3-D. Solutions like the RapidID™ Network Mapping System, where network components, including both ends of the cable assemblies, are pre-labeled at the factory are of great help in documenting the network.

While structured cabling offers better organization and easier maintenance, it comes with its own challenges. Structured cabling generally requires more connection interfaces. Most installation environments are contaminated with dust and debris, which can find ways to travel to the end-faces of connectors. This could increase connector insertion loss (IL) and return loss (RL) which can cause network performance issues.

These risks, however, can be minimized by using high-quality products that fully implement standard guidelines, and through proper installation practices to avoid contamination of the connector end face. Ultimately, the benefits of structured cabling, such as improved scalability, and manageability outweigh these concerns when proper care is taken.



Key Test Result Discovered

Structured cabling is even more important in high-density AI environments

Fig. 3
Better cable management using structured cabling

Network Performance

AI networks are designed to leverage the highest available physical layer bandwidth and minimize tail latency, to achieve optimal performance for model training and inference.

Tail latency increases with increased packet re-transmission. Therefore, a low channel bit error ratio (BER) is critical for AI network performance to eliminate packet losses and subsequent re-transmission of packets so that tail latency can be controlled. Channel BER depends on signal impairments due to optical fibers, connectors, and transmit and receive performance of transceivers. Worst-case channel performance requirements are specified by Ethernet, Fiber Channel, and InfiniBand standard organizations.

For example, the latest Ethernet specification published in March 2024, (IEEE 802.3df) includes aggregated 800G data rates over 8 duplex lanes (16 fibers) for multimode (800GBASE-SR8) and single mode (800GBASE-DR8) channels. IEEE 802.3df describes the quality of the signals after propagating over the longest allowed fiber reach when worst-case transceivers are used, which results in worst-case BER before forward error correction (FEC). IEEE 802.3df also specifies the photodetector receiver electronics bandwidth and sensitivity, among other parameters. Standards based transmitter and receiver performance tests are widely used in manufacturing as a fast and relatively accurate estimator of channel BER. These channel performance specifications make it possible for vendor interoperability.

One of the concerns against structure cabling in AI networks is the added connector loss that may cause channel performance risk. This argument is easily refuted for transceivers fully compliant with Ethernet channel specs which allocate connectivity losses of 1.5 dB for MMF channels (800GBASE-SR8) and about 2.5 dB for SMF channels (800GBASE-DR8). However, some of today's transceivers used in AI networks are proprietary solutions and it cannot be directly assumed that they follow IEEE to address this concern. To understand this, we need to provide actual channel performance test data in comparison to the IEEE 802.3df worst-case channel specifications. To that end, Panduit evaluated single-mode and multimode 800Gbps OSFP (octal small form factor) transceivers that support both InfiniBand and Ethernet protocols and are used in deploying NVIDIA DGX or HGX server-based AI clusters (Fig. 4 Top). Testing was done with both DR8 compliant and proprietary NVIDIA transceivers in a direct connect architecture as shown in Fig. 4 (TOP).

Key Test Result Discovered

NVIDIA and other IEEE-compatible transceivers have enough headroom to use structured cabling

We noted that NVIDIA offered single-mode 800G 2xDR4 and multimode 800G SR8 transceivers at reduced reaches of 100m (SMF) and 50m (OM4 MMF) respectively, instead of IEEE 802.3df specified reaches of 500m (SMF) and 100m (OM4 MMF). Based on evaluated NVIDIA specs, we found that NVIDIA 800G-SR8 transceivers are compliant with IEEE 800GBASE-SR8 transmitter and receiver specifications, and they should be capable of operating with 1.5 dB connector losses for up to 100m. To validate this, we measured BER performance of an off-the-shelf NVIDIA 800G-SR transceiver using several different MMFs, one of which represented the worst-case modal bandwidth. We also used a modally independent Keysight attenuator to simulate different connector loss conditions. This direct BER measurement, although more time-consuming than the oscilloscope-based tests used by manufacturers, is more representative of the channel performance.

Fig. 5 shows the test setup which comprises the transmitter/receivers under test, the fiber under test (FUT), and a variable optical attenuator to simulate the connector losses. For the multimode setup, the FUT consisted of 50m and 100m of worst-case standard compliant OM4 MMF and our best-performance OM4 fiber known as OM4+ Signature Core™. The latter is a fiber that corrects for the dispersion of the channel and has been used for many years for extended-distance channels.

Typically, acceptable BER for data center applications is well below 1-bit error per trillion transmitted bits (<1e-12). Up to 25Gbps (per lane), transceivers used simple two-level signaling to transmit zeros and ones, achieving BER <1e-12 without error correction schemes. Today's PAM-4 transceivers require FEC schemes to achieve a BER greater than 1e-12. Those FEC codes are capable of correcting error rates up to 240 errors per million of transmitted bits (2.4e-4), achieving better than 1 error per trillion of transmitted bits (1e-12).

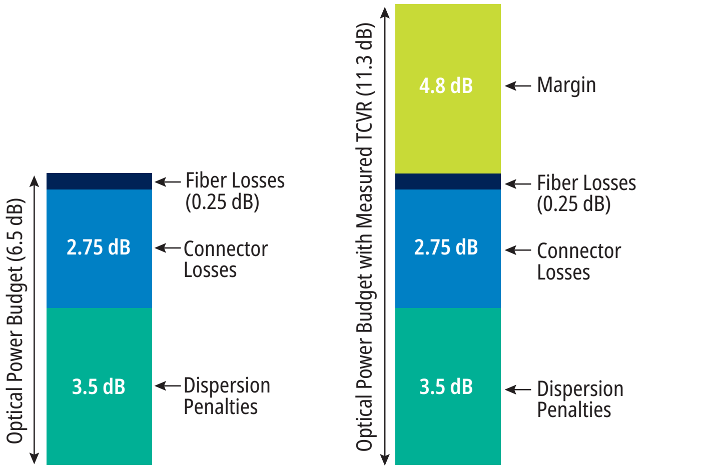
Our experiments with 50m of worst-case compliant OM4 fiber showed that more than 5 dB of connector loss is required to degrade the BER to the IEEE-specified worst-case value of 2.4e-4. We also found that the transceiver operated over 100m of worst-case OM4 with up to a 3.5 dB margin. When using 100m of Panduit Signature Core fiber, which offers dispersion compensation, the margin increased to over 5 dB.

Discover how Panduit delivered the ultimate in reach and performance for a US Military Data Center with 400G connectivity

These results demonstrate that 800G SR8 transceivers can tolerate 1.5 dB of connectivity loss over the specified 50m reach, with a large margin (>3.5 dB) to account for aging and temperature variations.

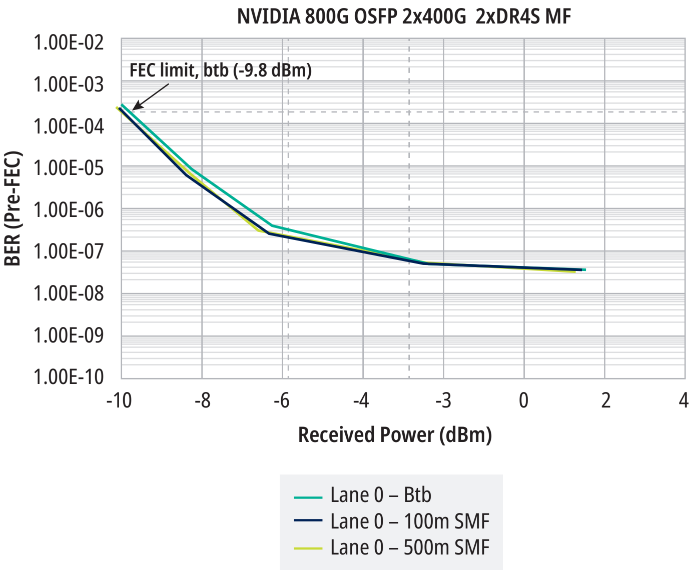
Similarly, evaluations of NVIDIA 800G DR8 transceivers over 100m and 500m of standard SMF revealed that more than 2.5 dB of connector loss can be tolerated, providing significant headroom for laser aging. Additionally, the transceivers specified for 100m operation performed well, even at 500m, with negligible penalties.

800G DR4 – 100m and 500m SMF



IEEE 802.3df Minimum Requirements for SMF (500m) 800GBASE-DR4

Measure Margin using NVIDIA Transceivers with Minimum Compliant **(worst case)** 100m and 500m SMF

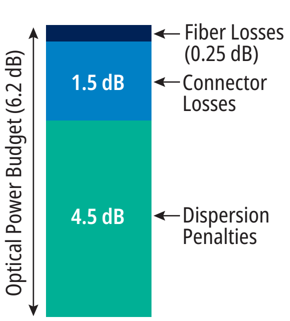


Note: NVIDIA transceivers offer 4.8 dB more power than the IEEE 802.3df specification.

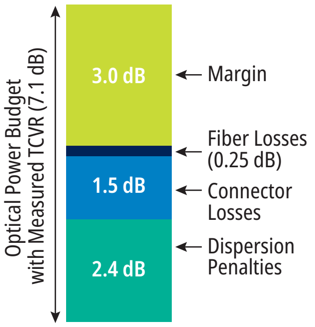
Transceiver power will degrade over time so a margin of 1.5 dB is commonly accepted as the requirement for uninterrupted performance.

Customers should expect actual results will be significantly better than the measured data using worst case cabling.

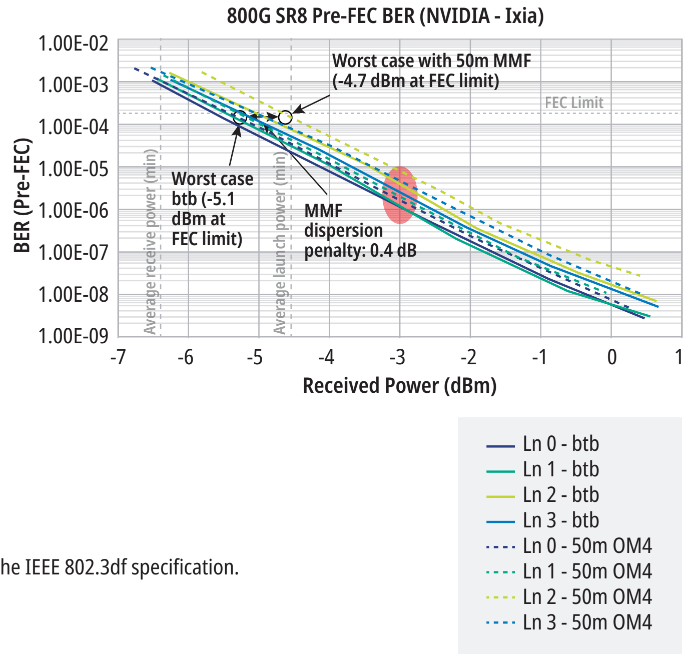
50m MMF



IEEE 802.3df Minimum Requirements for OM4 (50m) 800GBASE-VR8



Measure Margin using NVIDIA transceivers and minimum compliant **(worst case)** Panduit OM4 fiber (50m)

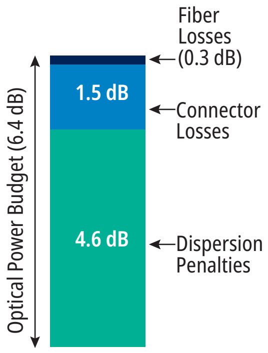


Note: NVIDIA transceivers offer 0.9 dB more power and 2.1 dB less dispersion penalty than the IEEE 802.3df specification.

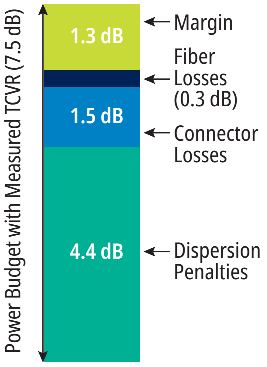
Transceiver power will degrade over time so a margin of 1.5 dB is commonly accepted as the requirement for uninterrupted performance.

Customers should expect actual results will be significantly better than the measured data using worst case cabling.

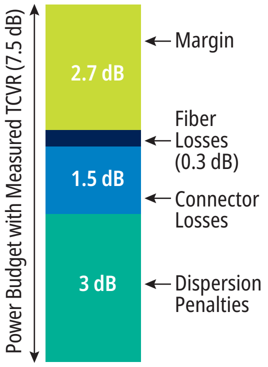
100m MMF



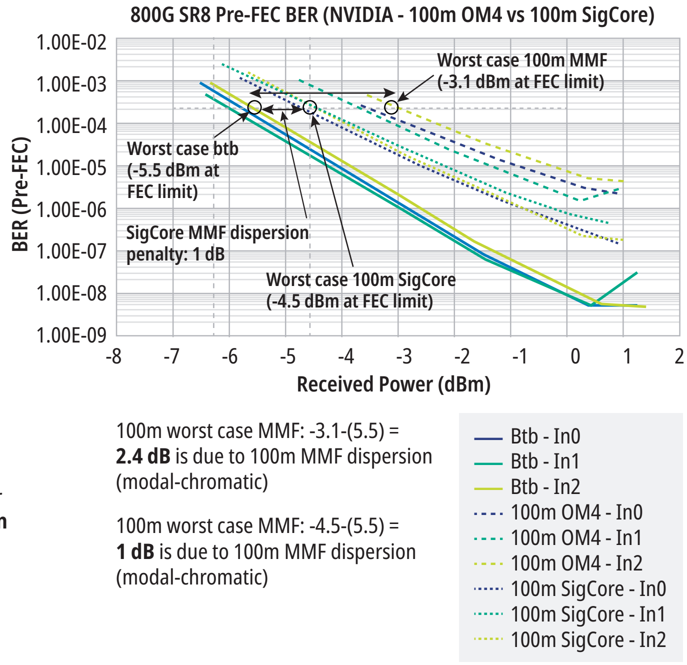
IEEE 802.3df Minimum Requirements for OM4 (100m) 800GBASE-SR8



Measure Margin using NVIDIA Transceivers with Minimum Compliant **(worst case)** OM4 fiber (100m)



Measure Margin using NVIDIA Transceivers with Signature Core OM4+ Fiber (100m) **(lowers dispersion loss by 1.4 dB)**



100m worst case MMF: -3.1-(5.5) = **2.4 dB** is due to 100m MMF dispersion (modal-chromatic)

100m worst case MMF: -4.5-(5.5) = **1 dB** is due to 100m MMF dispersion (modal-chromatic)

Note: NVIDIA transceivers offer 1.1 dB more power and 0.2 dB less dispersion penalty than the IEEE 802.3df specification.

Transceiver power will degrade over time so a margin of 1.5 dB is commonly accepted as the requirement for uninterrupted performance.

Customers should expect actual results will be significantly better than the measured data using worst case cabling.

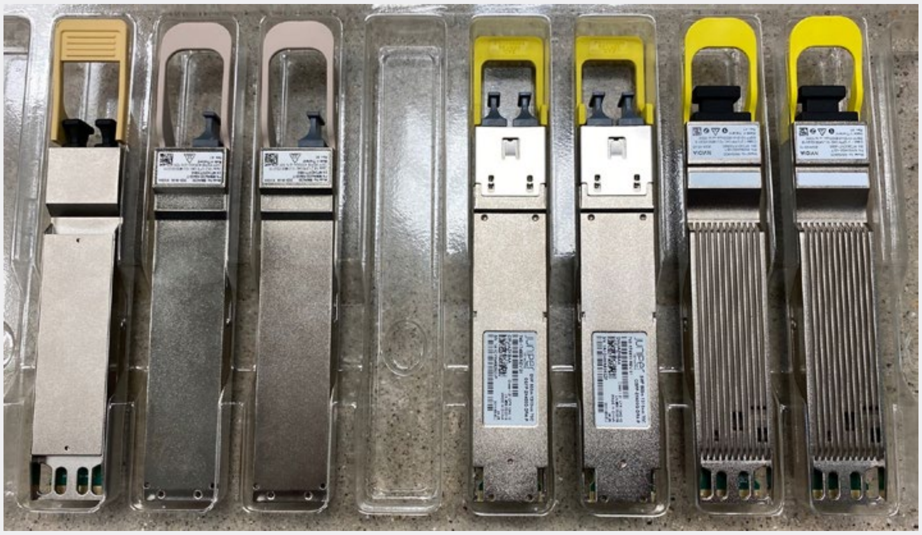


Fig. 4a

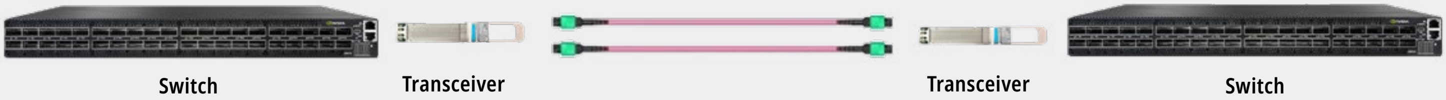


Fig. 4b
(a) NVIDIA transceivers (b) Simple example of the channel connecting switches.
In many cases, the MPO cables go to different switches according to the implemented topology.

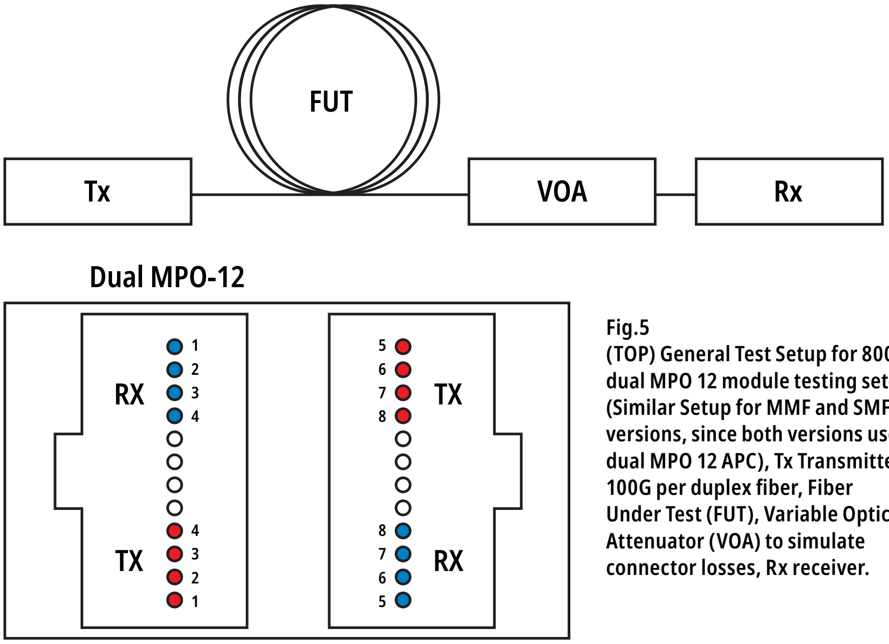
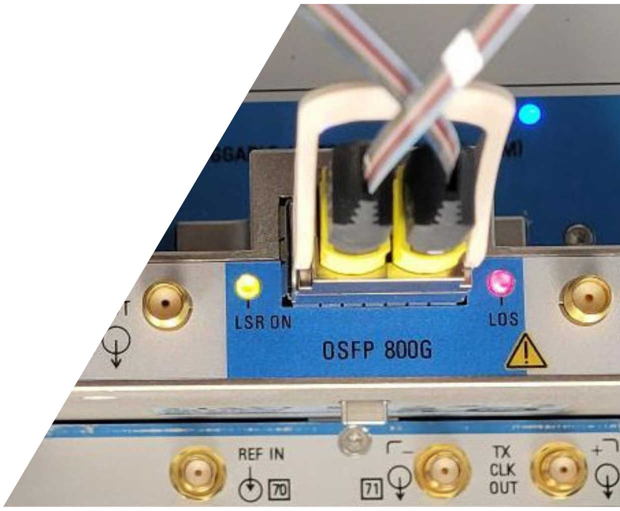


Fig.5
(TOP) General Test Setup for 800G dual MPO 12 module testing set-up (Similar Setup for MMF and SMF versions, since both versions use dual MPO 12 APC), Tx Transmitter 100G per duplex fiber, Fiber Under Test (FUT), Variable Optical Attenuator (VOA) to simulate connector losses, Rx receiver.



MPO connectors on the transceiver picture and detailed interconnection map.

Summary

Large AI networks use several thousand fiber cables which is 4 to 8 times denser than traditional data centers. To manage such a complex network, it is beneficial to utilize standards based structured cabling as it improves organization, protects the fiber, and supplies slack storage. This paper detailed how structured cabling brings these many benefits to AI networks while having no adverse effect on latency or BER.

To evaluate the impact of optical connectivity on the transceiver performance, the headroom of NVIDIA optical transceivers for MMF and SMF channels were calculated based on their specification along with experiments that measured BER testing.

The results of this analysis for MMF channels indicate that the evaluated NVIDIA and third-party transceivers, not included in the report, tolerate connectivity losses of 1.5 dB for MMF without impact on the specified performance while still maintaining an ample headroom for laser aging or temperature dependent penalties. Similarly, for SMF channels, the NVIDIA transceivers, connectivity losses of 2.5 dB can be tolerated with ample headroom for aging.

We conclude, based on theoretical evaluation from specs, our test results, and supplementary data from a third-party vendor, that structured cabling, which is essential for the deployment, maintenance, and scalability of AI networks, can be effectively implemented using 800G NVIDIA transceivers. This implementation is dependent on keeping connector losses within the mentioned limits and following cleanliness guidelines for connectivity.

Key Test Results Discovered

- **Structured cabling has no negative effect on latency**
- **Structured cabling helps alleviate issues in routing cables such as too much slack, fixing/replacing cables with issues, and managing high densities**
- **NVIDIA transceivers have ample headroom for using structured cabling as their transceivers are significantly (1+ dB) above that worst-case insertion loss threshold**

THE INFORMATION CONTAINED IN THIS WHITE PAPER IS INTENDED AS A GUIDE FOR USE BY PERSONS HAVING TECHNICAL SKILL AT THEIR OWN DISCRETION AND RISK. BEFORE USING ANY PANDUIT PRODUCT, THE BUYER MUST DETERMINE THE SUITABILITY OF THE PRODUCT FOR HIS/HER INTENDED USE AND BUYER ASSUMES ALL RISK AND LIABILITY WHATSOEVER IN CONNECTION THEREWITH. PANDUIT DISCLAIMS ANY LIABILITY ARISING FROM ANY INFORMATION CONTAINED HEREIN OR FOR ABSENCE OF THE SAME.

All Panduit products are subject to the terms, conditions, and limitations of its then current Limited Product Warranty, which can be found at www.panduit.com/warranty.

*All trademarks, service marks, trade names, product names, and logos appearing in this document are the property of their respective owners.

REFERENCES
[1] https://docs.NVIDIA.com/networking/display/800gmma4z00ns
[2] IEEE 802.3df: https://standards.ieee.org/ieee/802.3df/11107/

We have the knowledge and experience to help you make the most of your infrastructure investment.

panduit.com

Let’s connect
panduit.com/contact-us

PANDUIT™