From Fast to Faster

An Overview of Network Architecture for AI Workloads

Romeo Lazar
Sales Manager Eastern Europe
Corning Optical Communications
Mai 2024

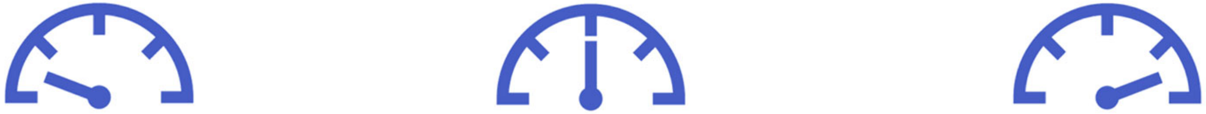
CORNING

The Ethernet and InfiniBand (Technology) Roadmap

CORNING

What Are We Going to Talk About?

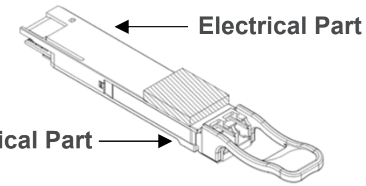
How much data can you transfer in 1 second?



- 10G
- 40G
- 100G
- 200G
- 400G
- 800G
- 1.6T

In simple words, how do we do it?

Switches



Transceiver

Optical Fiber Components

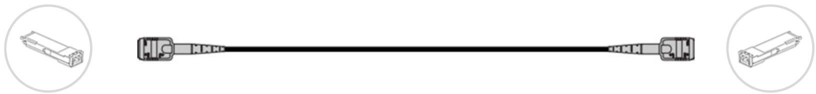
Transmit / Receive

CORNING

How Can The Cabling be Done?

Point-to-Point Cabling (Unstructured Cabling)

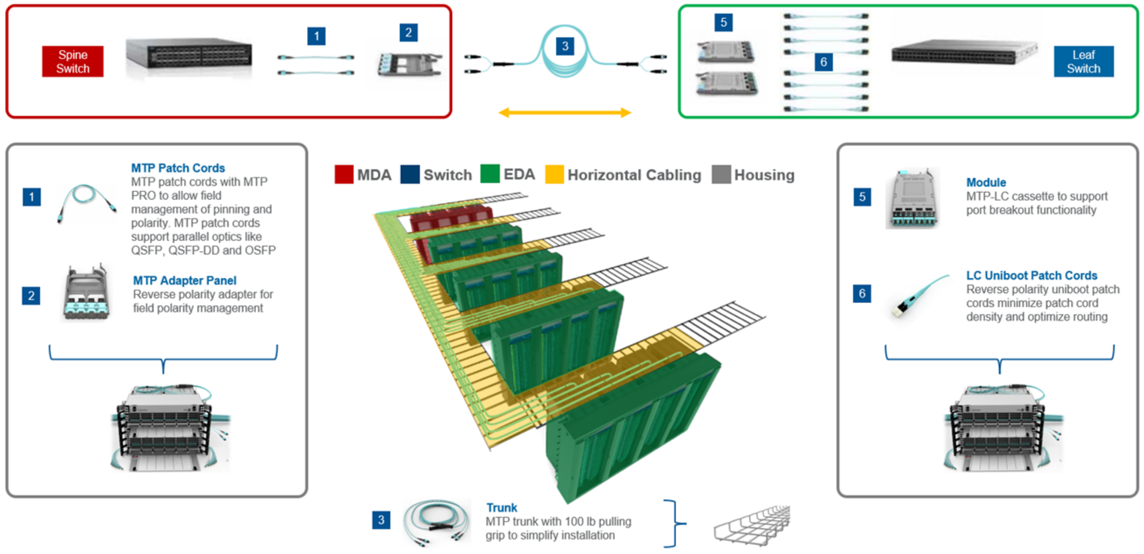
The cabling starts with a few connections . . .



And this is how it ends up ...

- No defined cable paths
- Problem determination difficult
- Changes made at active equipment
- System growth can be impacted

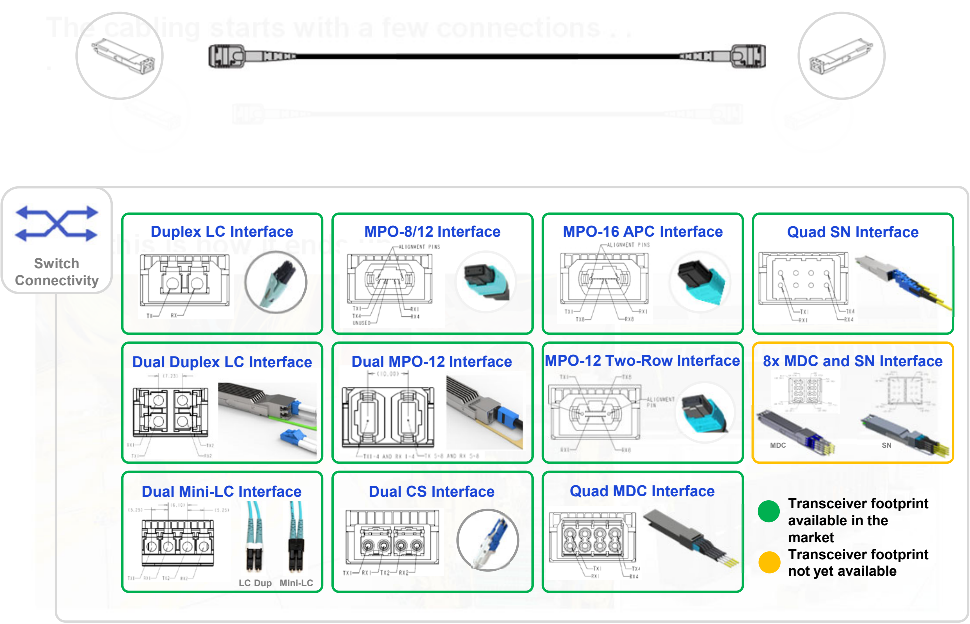
Structured Cabling



- Maximizes space and reduces installation time and cost
- Moves, adds, and changes (MACs) can be made easily
- A structured cabling system will provide the extra space needed for future growth
- A well-planned infrastructure can last 15-20 years and remain operational through multiple generations of system equipment and data-rate increases

How Can The Cabling be Done?

Point-to-Point Cabling (Unstructured Cabling)



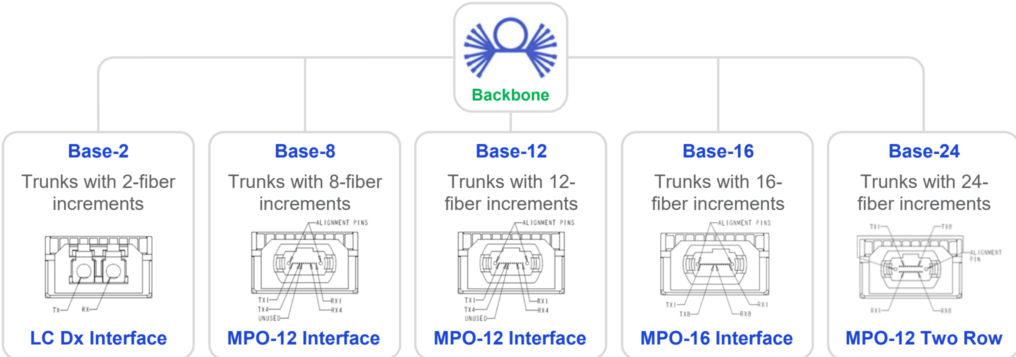
Structured Cabling

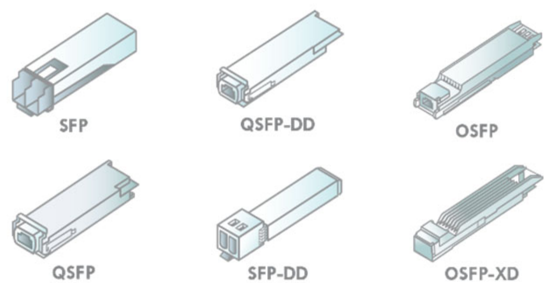
CORNING

Transceiver Roadmap and Backbone of Choice

Transceiver Speed	10G	25G	40G		50G	100G			200G			400G				800G				1.6T
Pluggable Module	SFP	SFP	SFP / QSFP		SFP / QSFP	SFP / SFP-DD / QSFP / QSFP-DD / OSFP			QSFP / QSFP-DD / SFP-DD			QSFP / QSFP-DD / OSFP				QSFP / QSFP-DD / OSFP				QSFP / QSFP-DD / OSFP / OSFP-XD
SMF	LR	LR	LR4 FR4	PLR4 PLRL4	LR FR	LR FR DR LR4 CWDM4	*N/A*	PSM4	LR4 FR4 FR DR	*N/A*	DR4	LR8 FR8 FR4 LR4-6 LR4-10	2FR4	DR4 DR2 DR4-2	*N/A*	LR8 FR8	2LR4 2FR4 FR4	DR4 DR4-2	2DR4 2PLR4 8FR DR8 DR8-2	DR8 DR8-2
MMF	SR	SR	BiDi SWDM4	SR4 eSR4	SR	BiDi SWDM4 VR SR	SR2	SR4 eSR4	*N/A*	VR2 SR2	SR4	*N/A*	*N/A*	SR4.2 VR4 SR4	SR8	*N/A*	*N/A*	VR4.2 SR4.2	SR8 VR8 2VR4 2SR4	VR8.2 SR8.2
Fibers per transceiver	**2**	**2**	**2**	**8**	**2**	**2**	**4 (2x2)**	**8**	**2**	**4 (2x2)**	**8**	**2**	**4 (2x2)**	**8**	**16 (16x1)**	**2**	**4 (2x2)**	**8**	**16 (8x2 or 16x1)**	**16 (8x2 or 16x1)**
Base-2	●	●	●	○	●	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Base-8	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Base-12	●	●	●	○	●	●	●	○	●	●	○	●	●	○	○	●	●	○	○	○
Base-16	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	○ (yellow)	●	○ (yellow)	○ (yellow)	○ (yellow)	●	●
Base-24	●	●	●	○	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○

- ● Allow full scalability, 100% fiber utilization and migration
- ○ (yellow) Allow scalability and migration. Limited backward compatibility with existing Base-8 and Base-12 backbones / installations
- ○ Scalability and migration complexity in some degree (base conversion components, partial fiber utilization)
- – Not recommended due to scalability limitations and high complexity





Picture source: Ethernet Alliance

CORNING

The connector in the backbone is relevant for: Flexibility, Migration to new technologies, Scalability, TCO

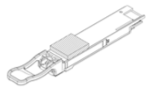
Corning's Way of Working

Voice of Customer + **Voice of Technology** + **R&D** = **Future-ready**

Voice of Customer

- Hyper
- MTDC
- Enterprise

Voice of Technology



- CISCO
- ARISTA
- DELL EMC
- BROADCOM

R&D

- EDGE™
- EDGE8®
- Clean Advantage™
- EDGE™ Rapid Connect
- EDGE™ MDC

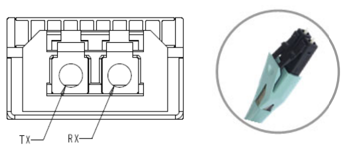
Future-ready

- TCO
- 10G
- 40G
- 100G
- 200G
- 400G
- 800G
- 1.6T

OSFP Optical Interfaces

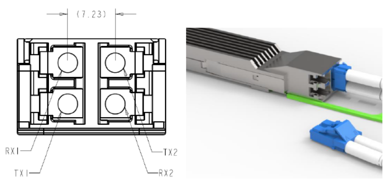
2-Fiber Transceivers

Duplex LC Optical Interface

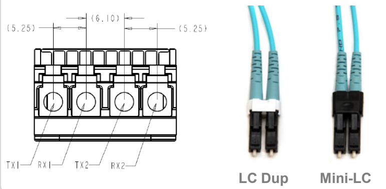


4-Fiber Transceivers

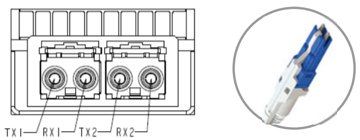
Dual Duplex LC Optical Interface



Dual Mini-LC Optical Interface

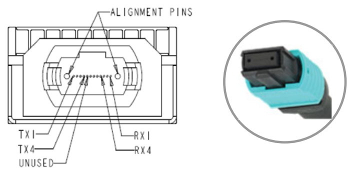


Dual CS Optical Interface

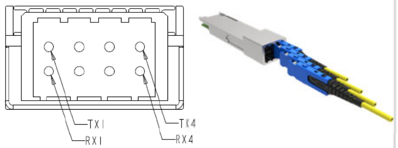


8-Fiber Transceivers

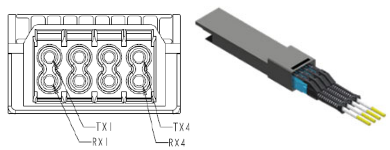
MPO-8/12 Optical Interface



Quad SN Optical Interface

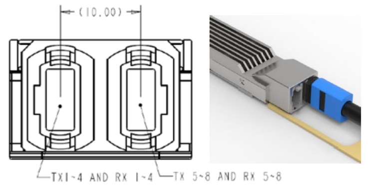


Quad MDC Optical Interface

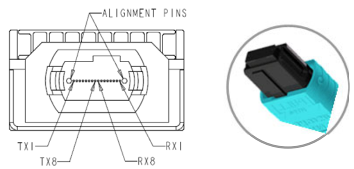


16-Fiber Transceivers

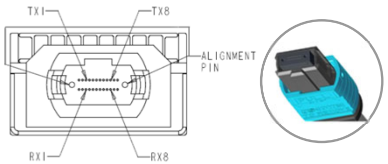
Dual MPO-12 Optical Interface



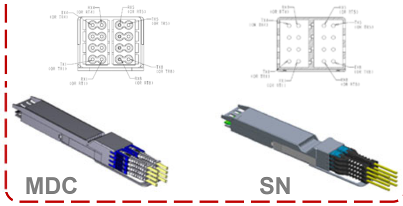
MPO-16 Optical Interface



MPO-12 Two Row Optical Interface



8x MDC and SN Optical Interface



- Footprint available in the market
- Footprint available and high adoption expected
- Transceiver footprint **not** yet available

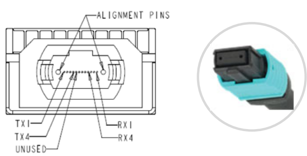
CORNING

Understanding the Numbers of Connectivity

Base - 8 Backbone

Backbone

MPO-8/12 Interface



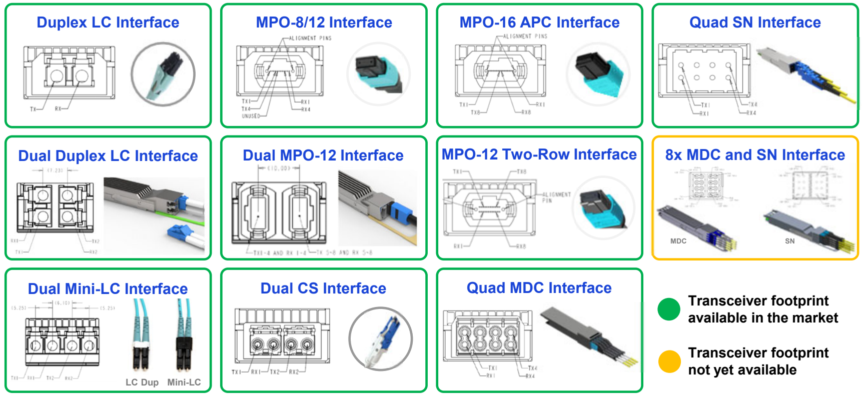
Trunks with 8-fiber increments

Migration

Base-8 supports the following data rates

10G | 40G | 100G | 200G | 400G | 800G | 1.6T*

Switch Connectivity

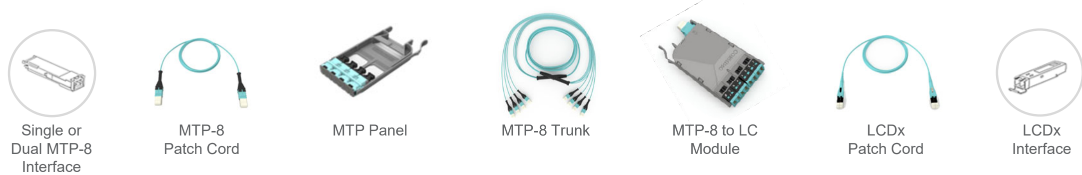


- Transceiver footprint available in the market (green)
- Transceiver footprint not yet available (yellow)

Cabling Infrastructure

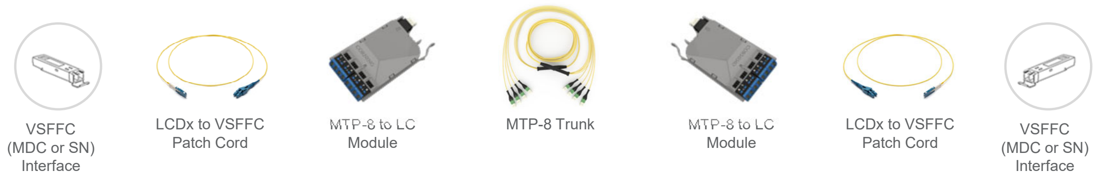
Examples of components used with different optical interfaces for different data rates

40G | 100G | 400G | 800G | 1.6T



400G | 800G | 1.6T

10G | 40G | 100G | 200G | 400G | 800G



*1.6T Transceivers using LC Duplex are also expected to be launched to the market

Where Used

- **Backwards compatible** with existing Base-8 and Base-12 architectures.
- Used in **small to large data centers**, enabling **migration to new transceiver technologies** with minimal to no change in existing structured cabling
- Widely believed to be the **most flexible option** to accommodate future industry trends, supporting deployments of new varieties of connectors at the transceiver, with full fiber utilization

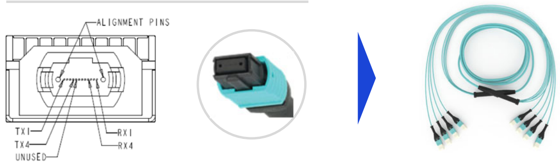
CORNING

Understanding the Numbers of Connectivity

Base - 8 Backbone

Backbone

MPO-12 Interface



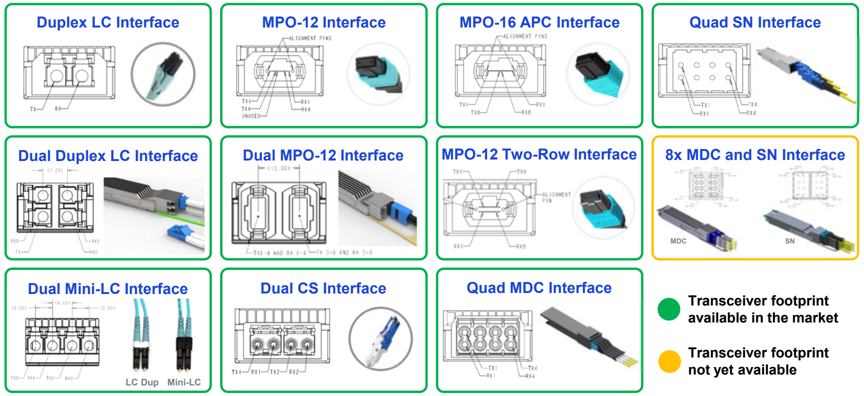
Trunks with 8-fiber increments

Migration

Base-8 supports the following data rates

10G · 40G · 100G · 200G · 400G · 800G · 1.6T

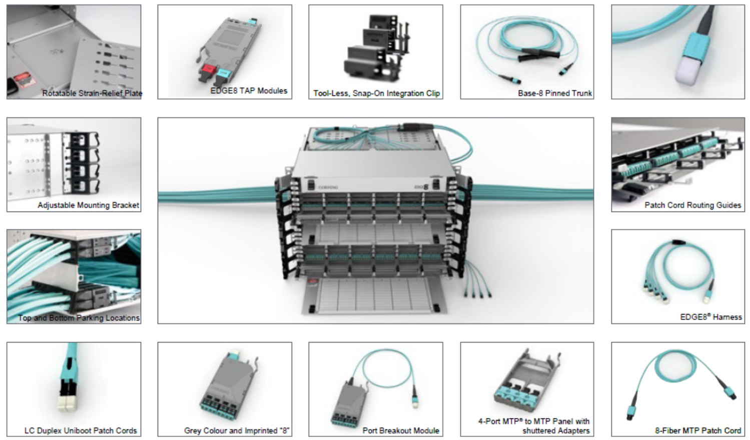
Switch Connectivity



- Transceiver footprint available in the market
- Transceiver footprint not yet available

Cabling Infrastructure

EDGE8 Solutions



- The best option supporting **migration** from 10G to 1.6T
- Supports Base-2, Base-8 and Base-16 connectivity with **duplex and parallel architectures**
- Support port **breakout solutions** to save space, power and cooling
- Supports **network monitoring** without adding separate space consuming hardware
- Supports keyed connectivity for **Secure Solutions**
- Supports **latency sensitive** applications
- **High Density** supporting **144F** per RU using **LC Dx** or **576F** per RU using **MTP-8**
- **Optical frames** available in single and dual versions: **5,760 duplex** or **23,040 parallel fibers**

Where Used

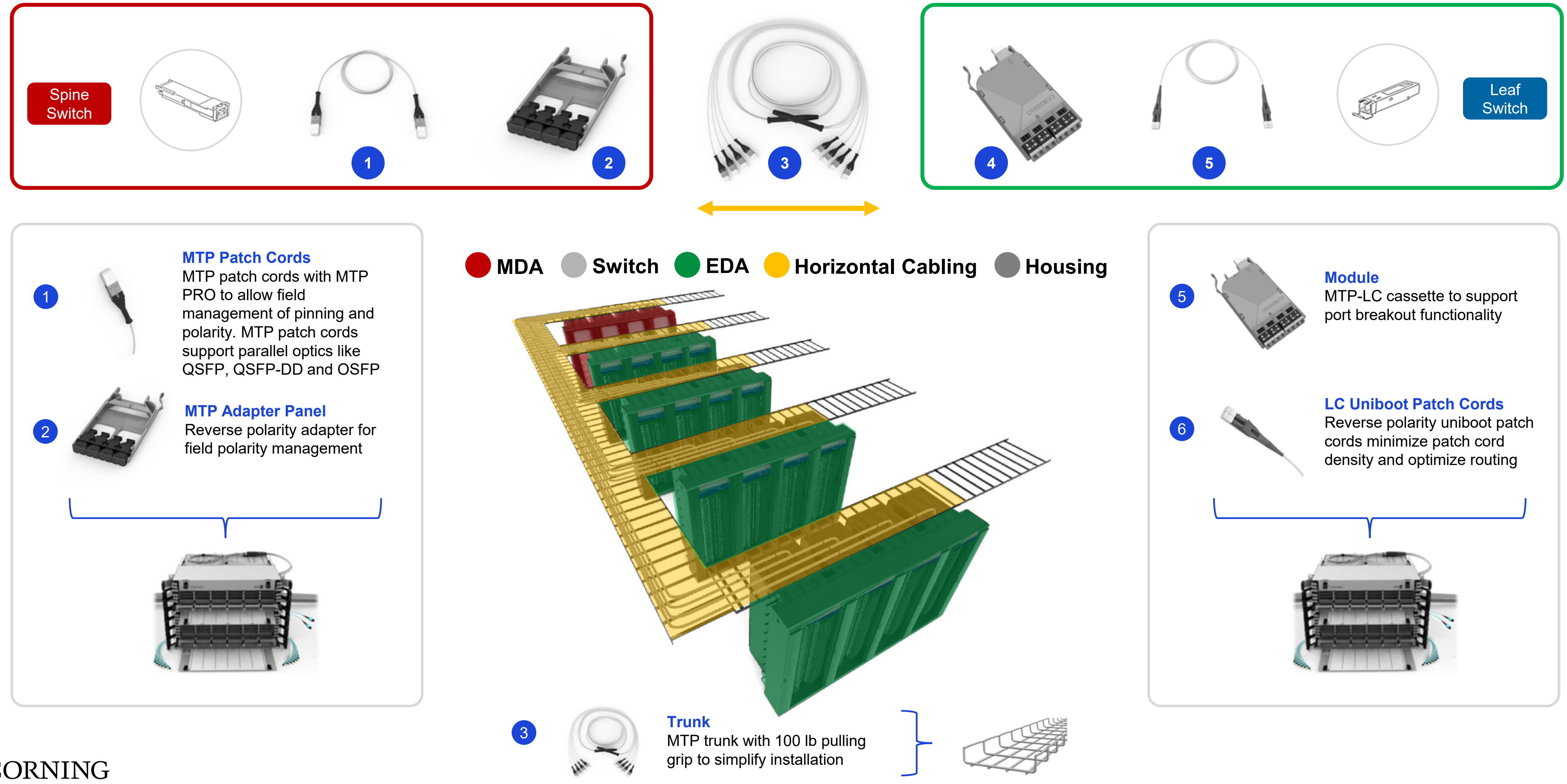
- **Backwards compatible** with existing Base-8 and Base-12 architectures.
- Used in **small to large data centers**, enabling **migration to new transceiver technologies** with minimal to no change in existing structured cabling
- Widely believed to be the **most flexible option** to accommodate future industry trends, supporting deployments of new varieties of connectors at the transceiver, with full fiber utilization

*1.6T Transceivers using LC Duplex are also expected to be launched to the market

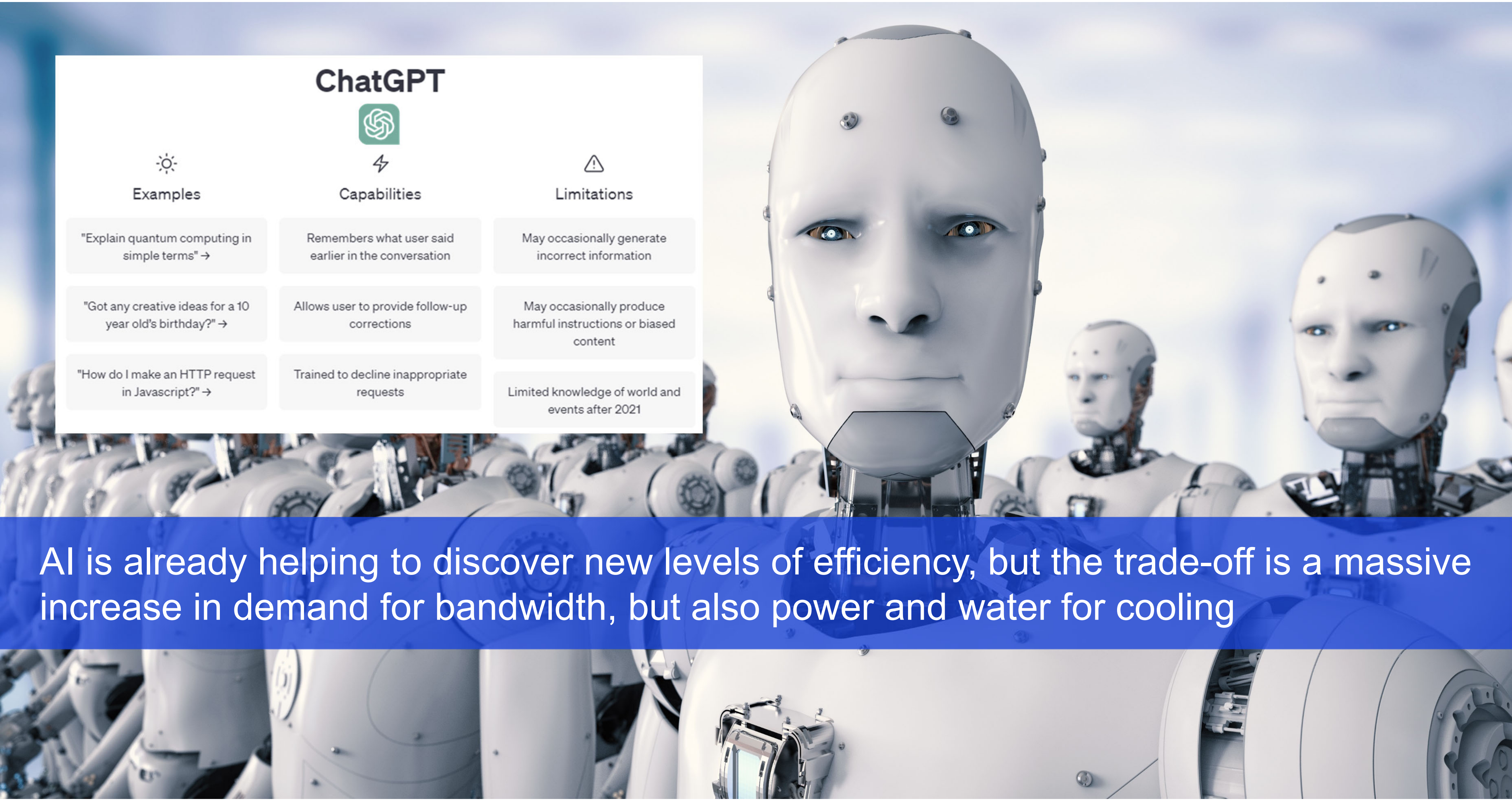
CORNING

An Overview of Network Architecture for AI Workloads

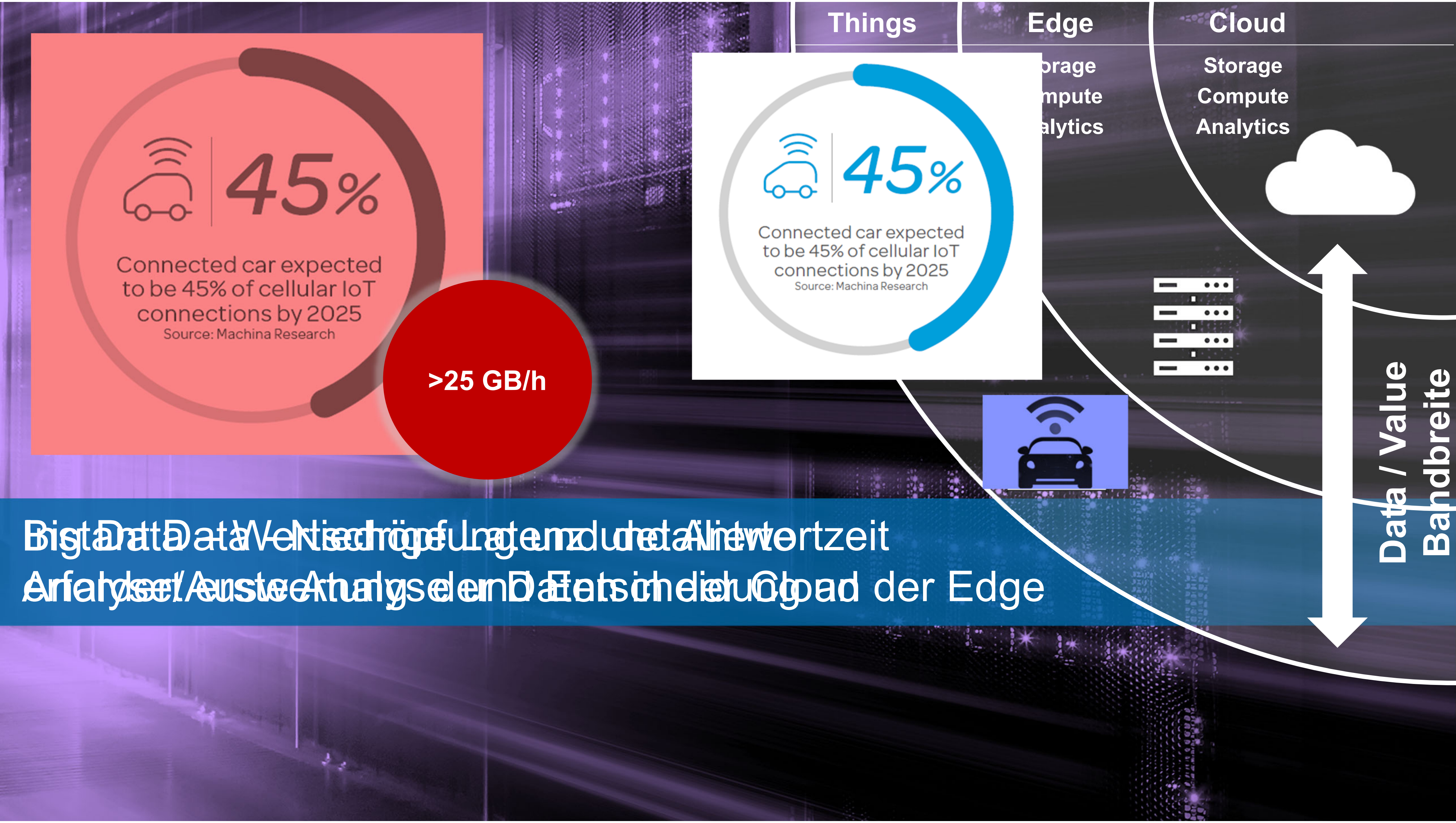
Interconnecting MDA to EDA with EDGE8®



CORNING



AI is already helping to discover new levels of efficiency, but the trade-off is a massive increase in demand for bandwidth, but also power and water for cooling



Things
Edge
Cloud
Storage
Compute
Analytics
45%
Connected car expected to be 45% of cellular IoT connections by 2025
Source: Machina Research
>25 GB/h
Data / Value
Bandbreite

High Performance Computing, Artificial Intelligence & Machine Learning

Image Source: ChatGPT

ChatGPT was trained using **10,000 of Nvidia's GPUs** clustered together in a supercomputer on Azure.

Moreover, there plans for significantly increased GPU usage, with speculation that their **upcoming AI model** may require as many as **10 million GPUs**.

NVIDIA dominates the market for chips used in AI systems, with about 90% of the GPU market for ML.

Current AI Training workloads require large GPU clusters (32k) driving need for power and cooling efficiency and high bandwidth in the MTDC and Cloud

Two Different Approaches to AI/ML

E

Ethernet (Ultra Ethernet Consortium starting)

BROADCOM, CISCO, ARISTA, intel, EVIDEN an atos business, Hewlett Packard Enterprise, AMD

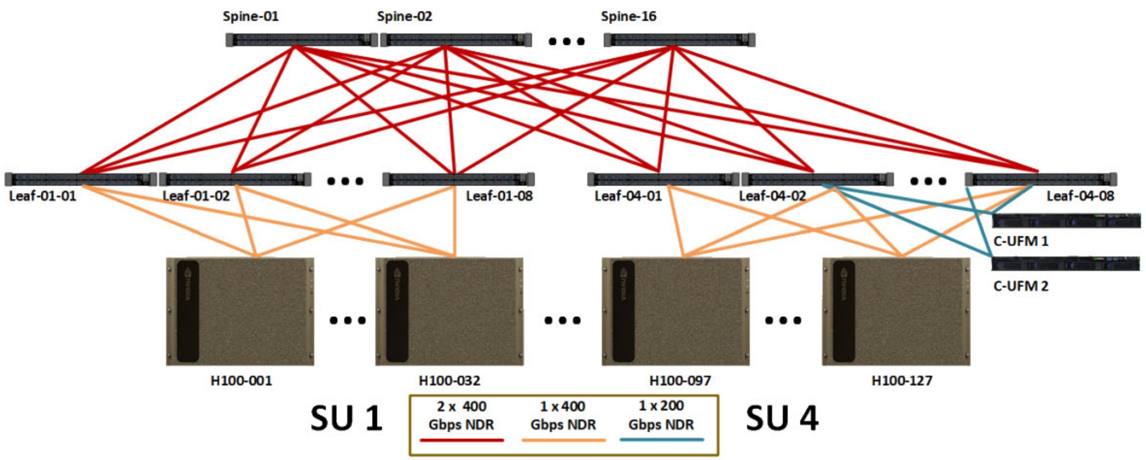
IB

InfiniBand

NVIDIA, IBM, Hewlett Packard Enterprise, intel

Dedicated Network for AI/ML

NVIDIA design as example:



Source: NVIDIA DGX SuperPOD. Reference Architecture Featuring NVIDIA DGX H100 Systems

An Overview of Network Architecture for AI Workloads

NVIDIA DGX-H100 SuperPOD

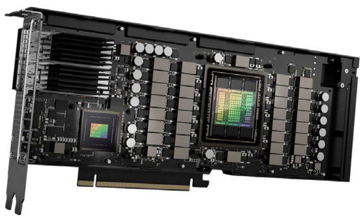
- **Compute Network Fabric**
- **Storage Network Fabric**
- **In-Band Network Fabric**
- **Out-of-Band Network Fabric**

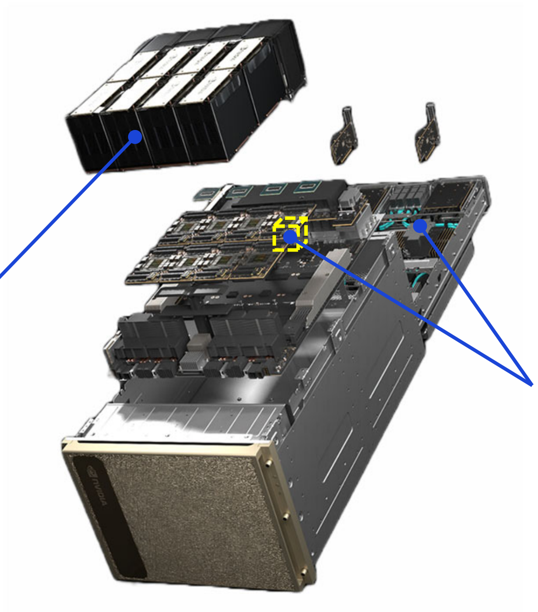
Image Source: NVIDIA DGX SuperPOD https://www.youtube.com/watch?v=NoCdoBl9vPw

NVIDIA DGX-H100 Compute Node (DGX-Data Center GPU Accelerated)

- For most machine learning workloads, both GPU and CPU work together to maximize performance.
- The CPU performs data cleaning on raw datasets before training models.
- Once this data is pre-processed, the CPU sends it to the GPU for parallel training/inference.
- After which, the GPU accelerates parallelizable math operations during training.
- Both are necessary for high-performance machine learning.

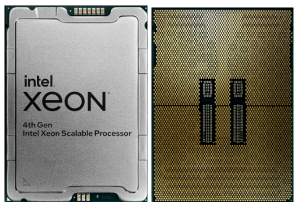
8x GPU





DGX-H100 Node/Server

2x CPU

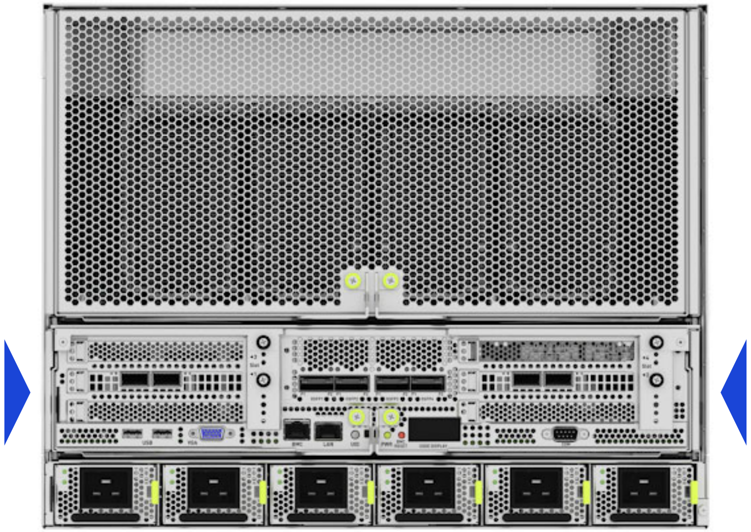


Feature	Description
Form Factor	8U Rack mount
System Weight (max)	287.6 lbs 130.45 kg
Input (200–240-volt AC) (max)	**10.2 kW**

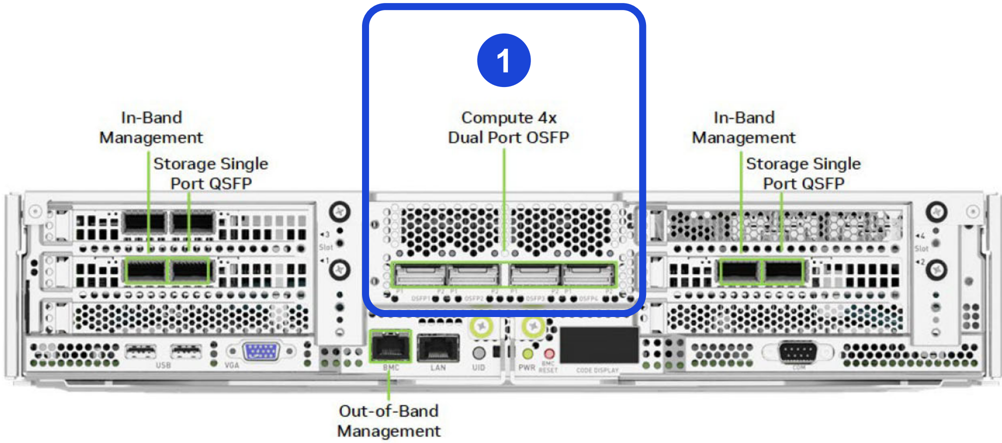
~$482,000 USD at release

Images Source: 4TH Gen Xeon Scalable https://www.intel.com/; H100 User Guide https://docs.nvidia.com/

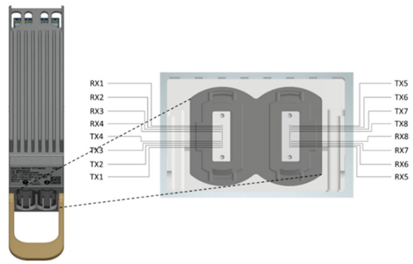
Compute Network Fabric



DGX-H100 Node (Sever) Networking



1. **Compute Fabric:** There are 4xDual Port OSFP on each Node (8x400G connections)
2. There are 8 Leaf switches for each Scalable Unit (POD)
3. The fabric is rail-optimized, meaning that all the same Host Channel Adaptors (HCA) from each node are connected to the same leaf switch.
4. The fabric is built using Quantum 9700 Infiniband switches using 800Gbps/ Twin port OSFP transceivers

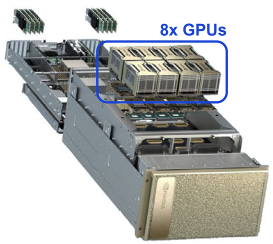


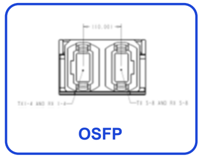
1 800Gb/s Twin-port OSFP, 2x400Gb/s

NVIDIA's Reference Architecture as Example

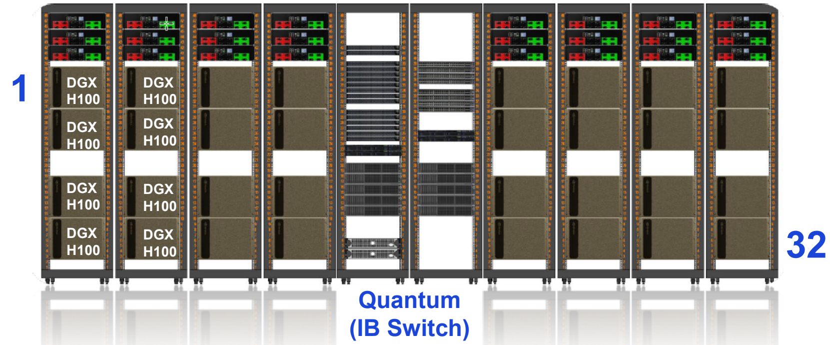
The system is built upon building blocks of **scalable units (SU)**, each containing **32 DGX H100** systems, which provides for rapid deployment of systems of multiple sizes.

Each SU has **256 GPUs**, this 32 DGX H100 in 8 racks





NVIDIA DGX H100 Server = Compute Node



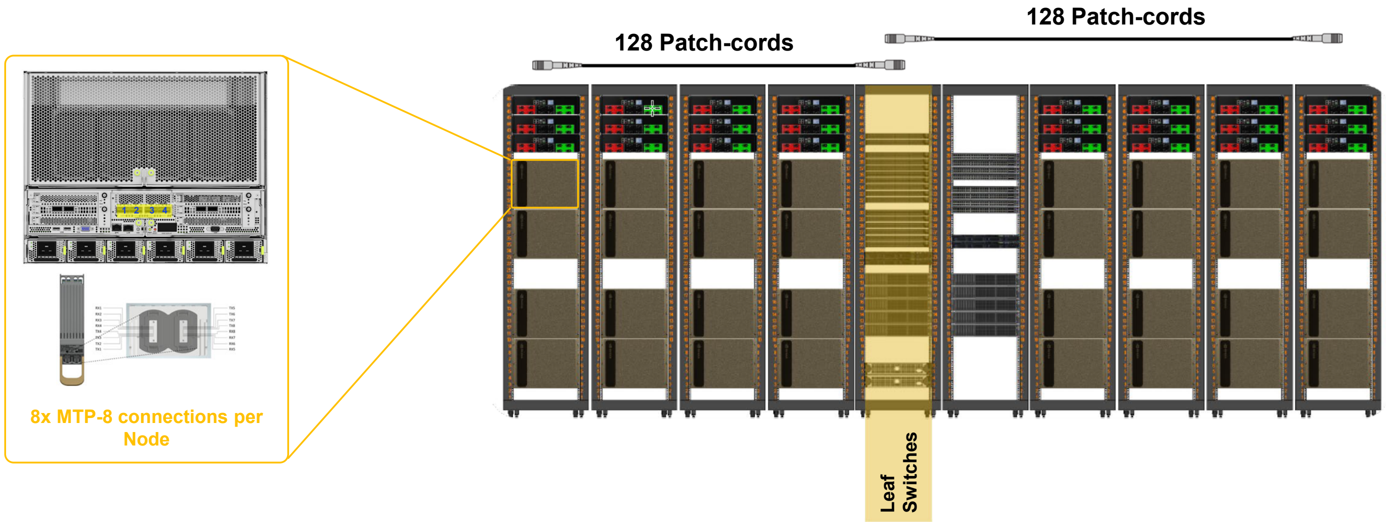
Example of Dedicated Network for AI/ML utilizing NVIDIA InfiniBand

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

 Source: NVIDIA DGX SuperPOD. Reference Architecture Featuring NVIDIA DGX H100 Systems

Cabling a Scalable Unit (POD)

Each POD requires 256 MTP-8 connections (8 per H100 Node) to the Leaf Switches



Innovative Solutions in Progress: Partnering with Customers for Future Success

Jumpers (Patch-cords)
SMF, MMF

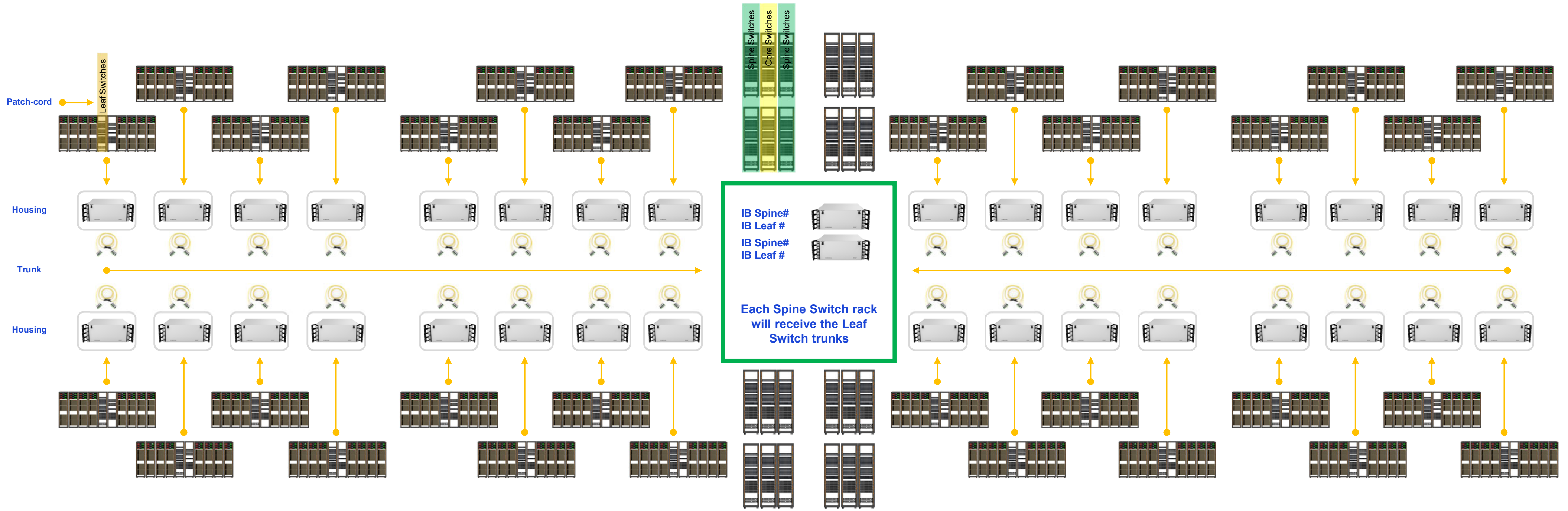
Bundle Jumpers
SMF, MMF

Trunks (w/pulling grip)
SMF, MMF

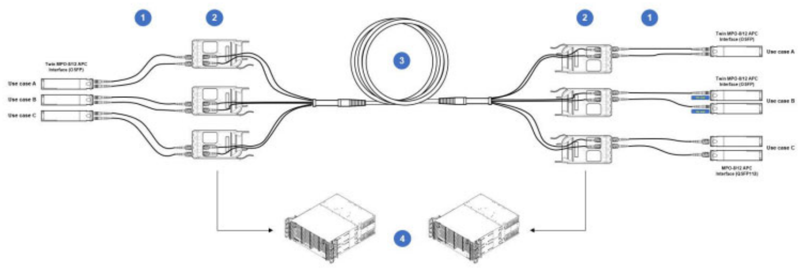
EDGE Distribution System
SMF

Cabling a “Medium Cluster”

Backbone cabling (512 trunks – 16 trunks per Leaf) will substitute 8,192 individual patch-cords, managing complexity across the data center



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



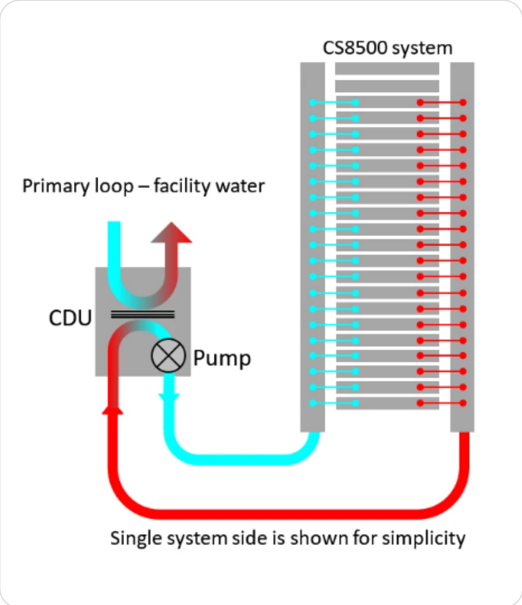
CORNING

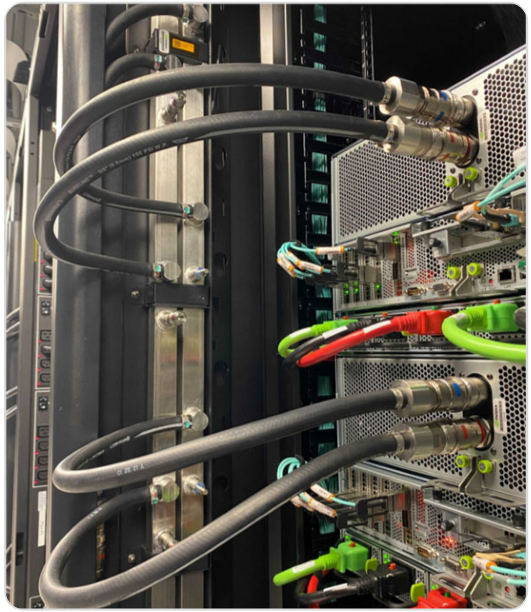
Power Requirement

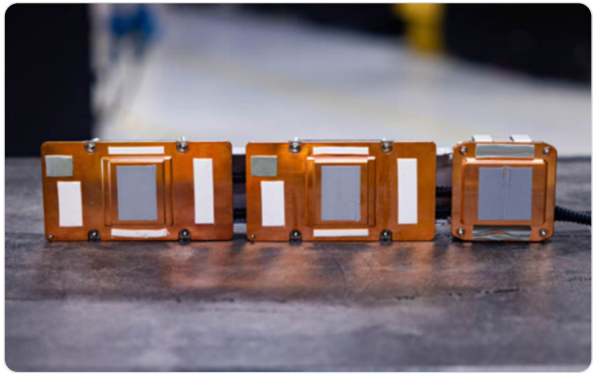
- ✓ **Due to power consumption of server rack (41 kW each) a full row of servers cannot be located in an existing DC design**
- ✓ **Base design is 256 GPUs/SU (Scalable Unit)**
- ✓ **There are 8 Production racks in a POD, meaning 8 racks x 32 GPUs**

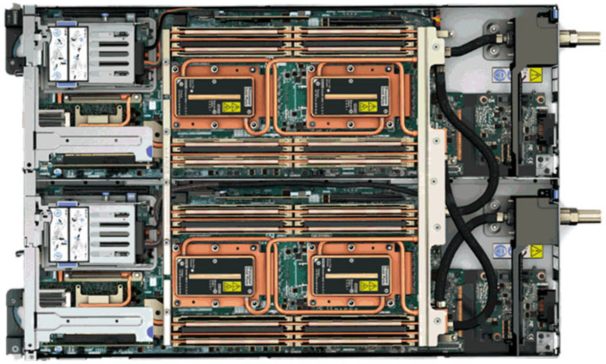
 Source: NVIDIA DGX SuperPOD. Reference Architecture Featuring NVIDIA DGX H100 Systems

Cabling and Direct Liquid Cooling









Sources: NVIDIA DGX SuperPOD. Reference Architecture Featuring NVIDIA DGX H100 Systems; https://docs.nvidia.com/networking/display/cs8500sysum/system+cooling+design+overview; https://lenovopress.com/LP0636#animation

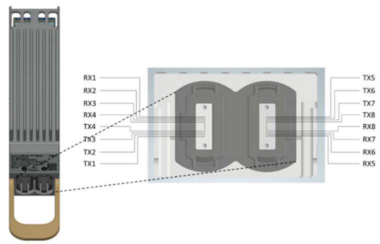
Latest NVIDIA's DGX B200 (Blackwell GPU) Architecture

To achieve the most scalability, DGX SuperPOD is powered by several key NVIDIA technologies, including:

- NVIDIA DGX B200 system—to provide the most powerful computational building block for AI and HPC.
- NVIDIA NDR (400 Gbps) InfiniBand—bringing the highest performance, lowest latency, and most scalable network interconnect.
- NVIDIA NVLink® technology—networking technologies that connect GPUs at the NVLink layer to provide unprecedented performance for most demanding communication patterns.

Component	Technology	Description
Compute nodes	NVIDIA DGX B200 system with eight B200 GPUs	The world's premier purpose-built AI systems featuring NVIDIA B200 Tensor Core GPUs, fifth-generation NVIDIA NVLink, and fourth-generation NVIDIA NVSwitch™ technologies.
Compute fabric	NVIDIA Quantum QM9700 NDR 400 Gbps InfiniBand	Rail-optimized, non-blocking, full fat-tree network with eight NDR400 connections per system

Feature	Description
Form Factor	10U Rack mount
Input (200–240-volt AC) (max)	**14.3 kW**

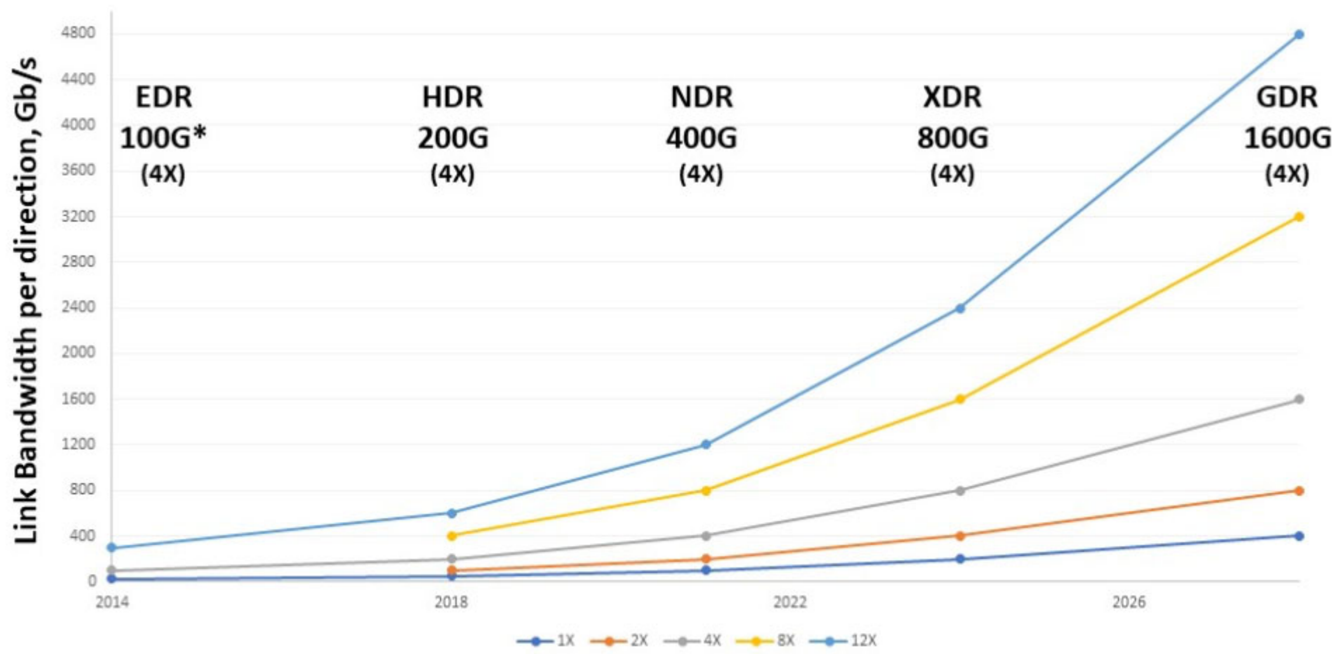


800Gb/s Twin-port OSFP, 2x400Gb/s

CORNING

Sources: NVIDIA DGX B200 https://resources.nvidia.com/en-us-dgx-systems/dgx-b200-datasheet; B200 Architecture https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/

InfiniBand Roadmap



*Link speeds specified in Gb/s at 4X (4 lanes)

Source: InfiniBand Trade Association

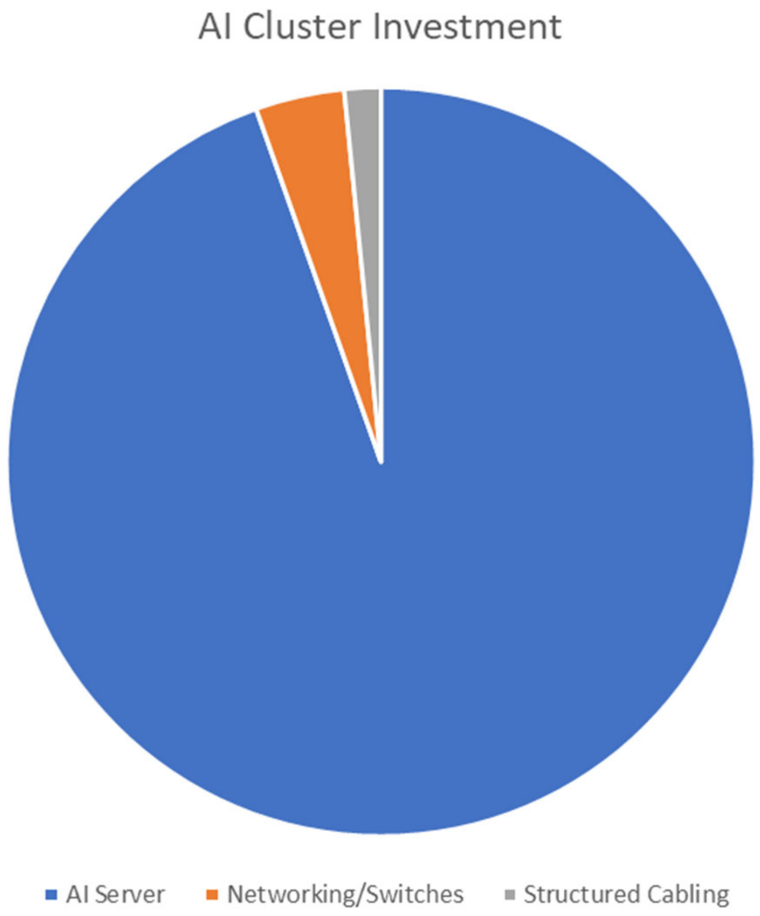
Full name	1X (lane)	4X (lanes)
Enhanced Data Rate (EDR)	25G	100G*
High Data Rate (HDR)	50G	200G
Next Data Rate (NDR)	100G	400G
Extreme Data Rate (XDR)	200G	800G
Gigantic Data Rate (GDR)	400G	1600G

Table 1. Summarized InfiniBand Roadmap

100G per Lambda
4x lanes = 8 Fibers

Current InfiniBand switches utilize **800G OSFP ports**, employing **dual 400G Next Data Rate (NDR) ports**. This configuration uses **8 fibers per port**, resulting in **64x400G ports per switch**. It's highly likely that the forthcoming generation of switches, whatever name they carry, will adopt **Extreme Data Rate (XDR) speeds**. This translates to **64x800G** ports per switch, also utilizing **8 fibers per port** – mostly **single mode fiber**. This 4-lane (8-fiber) pattern seems to be a recurring motif in the InfiniBand roadmap, summarized in Table-1, utilizing even faster speeds in the future.

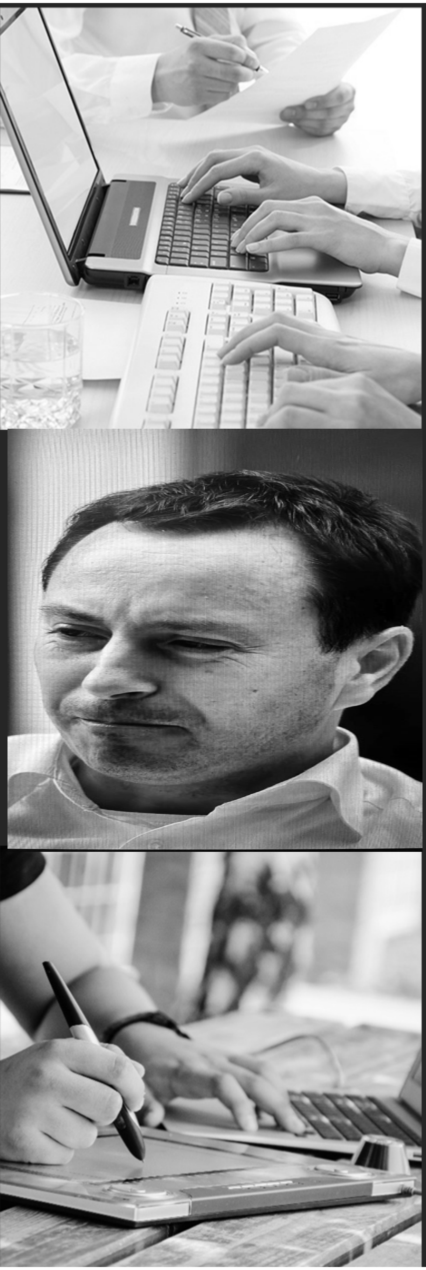
AI/ML and Structured Cabling



- More than **95% of the network cost** for the AI Server cluster is related to the **active gear**
- Power and Cooling investments will also be material
- **Fiber connectivity** is **facilitating** the networking **speed** and processing scalability.
- Fiber requirements for High Performance compute and AI networks is 5 times more than the traditional data center production network
- **Point-to-Point** bundled jumper assemblies can accommodate **smaller clusters, larger** scale-outs need **structured cabling** solutions
- **Roadmaps for Ethernet and InfiniBand transceivers** will scale with Base-8 fiber backbones
- Elements of the structured cabling system (Passive TAPs, Port-Breakout) will enable data center operators to **gain more value from the fiber infrastructure**

Planning for Migration

- **The path to higher speeds** will always depend on your unique needs.
- You may be happy with 40G now but planning to **upgrade to 100G** four years from now. Or maybe you are working with 400G and have your **eyes set on 800G** in five years: Migration will always vary based on your timeline and the available technologies in the market.
- But in most cases, **Base-8 will provide the ideal level of flexiblity to meet your needs throughout your transition**.
- Corning's **EDGE8 structured cabling solutions** will support your transition needs, doesn't matter if we talk about **Ethernet or InfiniBand**



Connect with us:

lazarr@corning.com

linkedin.com/in/RomeoLazar

Corning Optical Communications

Corning Optical Communications

@CorningOpComm

CORNING

CORNING