COMMSCOPE®

Cabling Considerations for AI Data Centers

For decades, the danger of malicious artificial intelligence (AI) has been a trope in science fiction. Film antagonists like HAL 9000, the Terminator, the Replicants, and the robots from The Matrix are opposing forces to the plucky humans who must overcome the dangers of technology. Recently, the release of DALL-E 2 and ChatGPT has captured the wider public's imagination of what AI can do. This has led to discussions on how AI will change the nature of education and work. AI is also the main driver for current and future data center growth.

There are three main aspects to AI:

- During training, a large data set is fed into the algorithm that consumes the data and "learns" from it.
- The algorithm is then exposed to a new data set and tasked with deriving new knowledge or conclusions based on what it learned during training. For example, is this a picture of a cat? This process is known as "inference AI."
- The third (and perhaps most exciting) aspect is what's known as "generative AI." Generative AI is when the algorithm "creates" original output—text, images, videos, code, etc.—from simple prompts.

AI computation is handled by graphical processing units (GPUs): specialized chips designed for parallel processing and well suited to AI. The models used to train and run AI consume significant processing capacity—typically too much for a single machine. Figure 1 shows the historical growth of AI models in petaFLOPS (quadrillions of floating-point operations per second). Processing

these large models requires multiple interconnected GPUs spread over many servers and racks. An AI data center deploys dozens of these AI clusters and the cabling infrastructure that ties everything together to keep the data flowing. This presents the subsequent challenges and opportunities in cabling AI data centers, with a few best practices and tips for success.

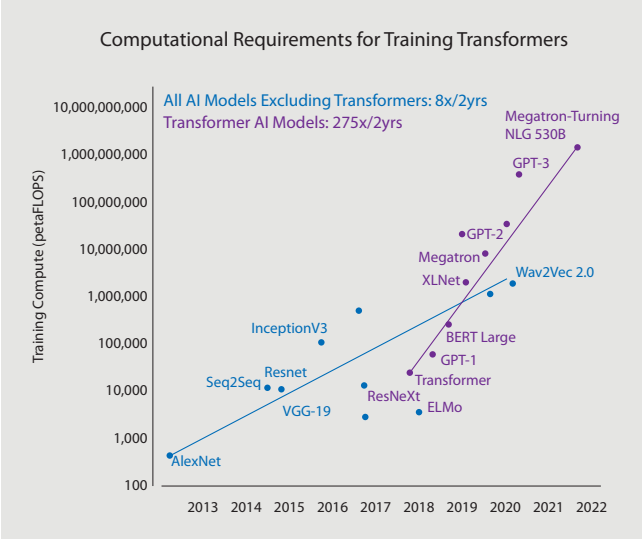

Figure 1: AI model size in petaFLOPS. Source: NVIDIA

How does AI architecture differ from typical architecture?

Nearly all modern data centers, especially hyperscales, use a folded Clos architecture in their data halls. Also called a "leaf-and-spine" architecture, Clos architecture is where all the leaf switches in a data center connect to all the spine switches. Server racks typically connect to a top-of-rack (ToR) switch. The ToR connects to a leaf switch at the end of the row or in another room via fiber cable. The servers in the rack connect to the ToR with short copper cables—1-2 meters long—carrying 25G or 50G signaling.

This typical architecture uses few fiber cables in the data hall. For example, Meta data centers that use the F16 architecture (see Figure 2) will have 16 duplex fiber cables from each server rack in a row. These cables run from the ToR switch to the end of the row, where they connect with modules that combine duplex fibers to 24-fiber cables. The 24-fiber cables then connect to leaf switches. Data centers implementing AI will typically house AI clusters next to compute clusters that use this traditional architecture. Traditional compute is sometimes called the "front-end network," and AI clusters are sometimes called the "back-end network."

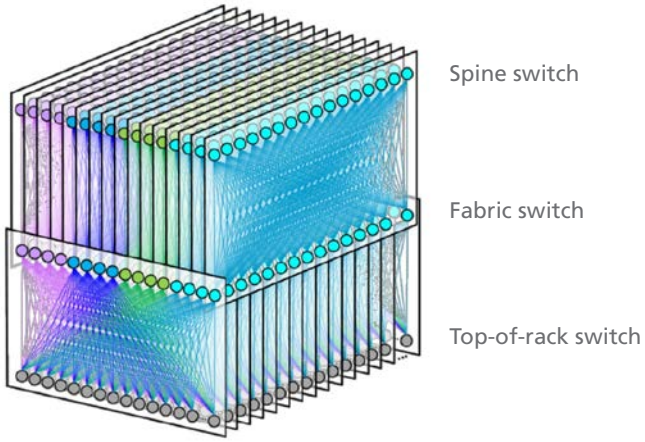

Figure 2: Facebook F16 data center network topology. Source: Engineering at Meta

As noted, AI clusters have unique data processing requirements and, thus, require a new data center architecture. GPU servers require much more connectivity between servers, but, due to power and heat restraints, there are often fewer servers per rack. The result is more inter-rack cabling in an AI data center architecture than in a traditional architecture. Each GPU server connects to a switch within the row or room. These links require 100G to 400G at distances that copper cannot support. In addition to the switch fabric, each server requires connectivity to storage and out-of-band management.

For example, we can look at the architecture proposed by NVIDIA, a leader in the AI space. NVIDIA's latest GPU server is the DGX H100 and has 4x800G fiber ports for connecting to switches (operated as 8x400G), 4x400G fiber ports for connecting to storage, and 1G and 10G copper ports for out-of-band management. A DGX SuperPOD (Figure 3) can contain 32 GPU servers connected to 18 switches in a single row. Each row would have 384x400G fiber links for the switch fabric and storage and 64 copper links for management. This is a significant increase in the number of fiber links in the data hall. The aforementioned F16 architecture would have 128 (8x16) duplex fiber cables with the same number of server racks.

Figure 3: Rendering of the NVIDIA SuperPOD. Source: NVIDIA

What link lengths are in an AI cluster?

In the ideal scenario illustrated by NVIDIA, all the GPU servers in an AI cluster will be close together. Like high performance computing (HPC), AI and machine learning (ML) algorithms are extremely sensitive to latency. One estimate claims that 30% of the time to run a large training model is spent on network latency, and 70% is spent on compute time. Since training a large model can cost up to US$10 million, this networking time represents a significant cost. Even a latency saving of 50 nanoseconds, or 10 m of fiber, is significant. Nearly all the links in AI clusters are limited to 100-m reaches.

Unfortunately, not all data centers can locate the GPU server racks in the same row. These racks require around 40 kilowatts to power the GPU servers, which is more power than typical server racks. Data centers built initially with lower power requirements will need to space out their GPU racks.

Which transceivers should you use?

Operators should carefully consider which optical transceivers and fiber cables they will use in their AI clusters to minimize cost and power consumption. As explained above, the longest links within an AI cluster are limited to 100 m. Due to the short reach, transceivers will dominate the optics cost.

Transceivers that use parallel optics will have an advantage. They do not require the optical multiplexers and demultiplexers used for wavelength division multiplexing (WDM). This results in both lower cost and lower power for transceivers. The transceiver cost savings of parallel optics more than offset the slight increase in cost for multifiber cable over duplex fiber cable. For example, 400G-DR4 transceivers with eight-fiber cables are more cost-effective than 400G-FR4 transceivers with duplex fiber cable.

Links up to 100 m are supported by both singlemode fiber and multimode fiber applications. Advances like silicon photonics have reduced the cost of singlemode transceivers, bringing them closer to the cost of equivalent multimode transceivers. Our market research indicates that, for high-speed transceivers (400G+), the cost of a singlemode transceiver is still double the cost of an equivalent multimode transceiver. While multimode fiber cable has a slightly higher cost than singlemode fiber cable, the difference is smaller since multifiber cable costs are dominated by MPO connectors.

In addition, high-speed multimode transceivers use 1-2 watts less power than their singlemode counterparts. With 768 transceivers in a single AI cluster (128 memory links + 256 switch links x2), multimode fiber saves up to 1.5 kW. This may seem minor compared to the 10 kW each DGX H100 consumes, but any opportunity to save power is welcome for AI clusters.

In 2022, the IEEE Short Reach Fiber Task Force completed work on IEEE 802.3db-2022, which standardized a new multimode very short reach (VR) transceiver. The new standard targets in-row cabling like AI clusters with a maximum reach of 50 m. These transceivers have the potential to offer the lowest cost and power consumption for AI connectivity.

AOCs vs. transceivers with fiber cable

Many AI, ML, and HPC clusters use active optical cables (AOCs) to interconnect GPUs and switches. An AOC is a fiber cable with integrated optical transmitters and receivers on either end. Most are used for short reaches and are typically paired with multimode fiber and VCSELs. High-speed (>40G) AOCs will use the same OM3 or OM4 fiber as cables that connect optical transceivers. The transmitters and receivers in an AOC may be the same as in analogous transceivers but are the castoffs. Neither the transmitter nor receiver must meet rigorous interoperability specs; they only need to operate with the specific unit attached to the other end of the cable. Since no optical connectors are accessible to the installer, the skills required to clean and inspect fiber connectors are unnecessary.

The downside of AOCs is that they do not have the flexibility offered by transceivers. Installing AOCs is time-consuming, as the cable must be routed with the transceiver attached. Installing AOCs with breakouts is especially challenging. The failure rate for AOCs is also double that of equivalent transceivers. When an AOC fails, a new AOC must be routed through the network. This takes away from the compute time. Finally, when it is time to upgrade the network links, the AOCs must be removed and replaced with new AOCs. With transceivers, the fiber cabling is part of the infrastructure and remains in place for several generations of data rates.

In conclusion, carefully considering the AI cluster cabling will help save cost, power and installation time, enabling organizations to fully benefit from AI

CommScope pushes the boundaries of communications technology with game-changing ideas and ground-breaking discoveries that spark profound human achievement. We collaborate with our customers and partners to design, create and build the world's most advanced networks. It is our passion and commitment to identify the next opportunity and realize a better tomorrow. Discover more at commscope.com.

commscope.com | Visit our website or contact your local CommScope representative for more information.

CO-118691-EN (03-24)