Practical Considerations for Designing a Future-Ready AI Data Center

Designing, implementing and managing data cabling in an AI data center involves unique challenges due to the introduction of high-performance processor chips to the servers (i.e., Accelerator Application Specific Integrated Circuits like GPUs and TPUs). These high-performance integrated circuits are capable of parallel processing huge amounts of data, and when combined with High Bandwidth Memory componentry in a rack-scale system (RSS) are capable of achieving 3.6 exaFLOPS of compute performance.

Today, some of the highest-performing AI superchips can process at speeds of up to 40 petaFLOPS[1] (floating point operations per second). For comparison, the performance of a traditional central processing unit (CPU), of the type commonly found in data center servers, has been found to provide theoretical peak performance of around 1,200 times slower than the latest GPU.[2]

As mentioned previously, in an AI cluster, an Accelerator Application Specific Integrated Circuit (AASIC) does not operate on its own but rather works in parallel with multiple other AASICs and supporting hardware like High-Bandwidth Memory (HBM), scale-up and scale-out switching hardware, power supplies and cooling infrastructure. Often, an RSS will be configured with up to 72 GPUs working together in a chassis or rack, each GPU requiring huge amounts of data bandwidth flowing in and out via 400G, 800G and even 1.6T channels to maintain peak operating performance.

The cooling challenge

With increased performance comes new cooling challenges that can impact the compute cabinet and, more specifically, the space available to introduce and manage fiber-optic cabling. Nothing comes for free, and the increased performance of these systems comes at the cost of higher power consumption that in turn creates heat as a byproduct. This heat must be evacuated from the servers and racks to keep the data center hardware running at optimum performance levels.

Previously, compute rack power amounted to 8 to 10 kW, and the resulting heat could be easily transferred away from active hardware using chilled air to maintain a constant ambient temperature of around 77°F (approximately 25°C). For the past 20 years, most data centers have been designed to support such air-cooled operations.

As higher-performing compute hardware with higher thermal densities have been introduced, rack power requirements have moved beyond 10 kW and are tracking upwards—as high as 80 or 100 kW—with some cloud providers forecasting rack-level power loads of 0.5 MW, levels that air cooling alone cannot efficiently or cost-effectively support. This trend has pushed designers to consider liquid cooling for their operations. Water and other fluids can be 50 to 100 times more efficient than cooling with air alone, which promises to address the thermal challenges of modern AI clusters. That fluid must be pumped around the data halls, into the rows and racks, and finally into the server chassis hosting the powerful accelerators.

It's at this point where the physical layer cabling design takes on enhanced importance, because we have to consider how to physically connect all those accelerators to one another in a coherent way that doesn't impede the performance of the overall AI architecture. Plus, this task is complicated by the introduction of additional cooling hardware (in the form of fluid piping), alongside a dramatic increase in the number of fibers required to fit into a small footprint, whether that be a server chassis or an entire rack. **Figures 1 and 2 below illustrate the challenge facing designers when regarding cooling AI clusters.**

Additional DTC pipework infrastructure consumes real estate inside the rack that was once available for structured cabling. This further stresses the interior space's capacity as the number of fiber cables deployed in AI cabinets is also on the rise.

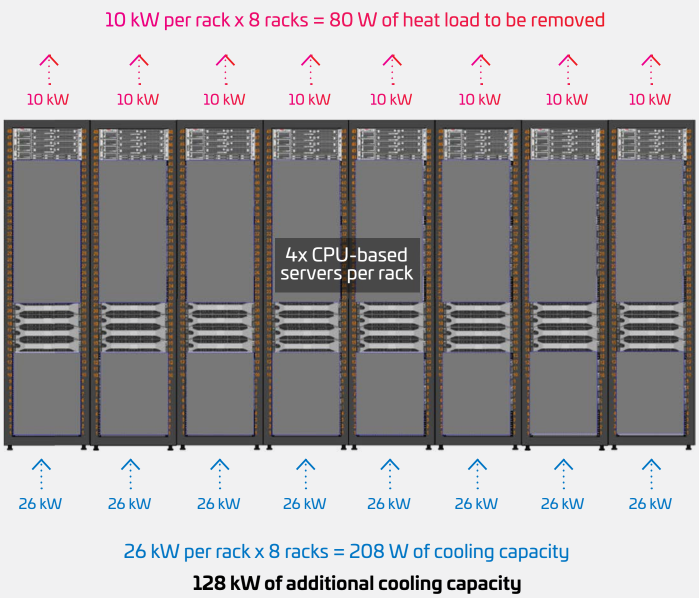


Figure 1: 4x CPU-based servers per rack: Air-cooled capacity in a data center using CPU-based servers. This scenario demonstrates surplus air handling capacity.

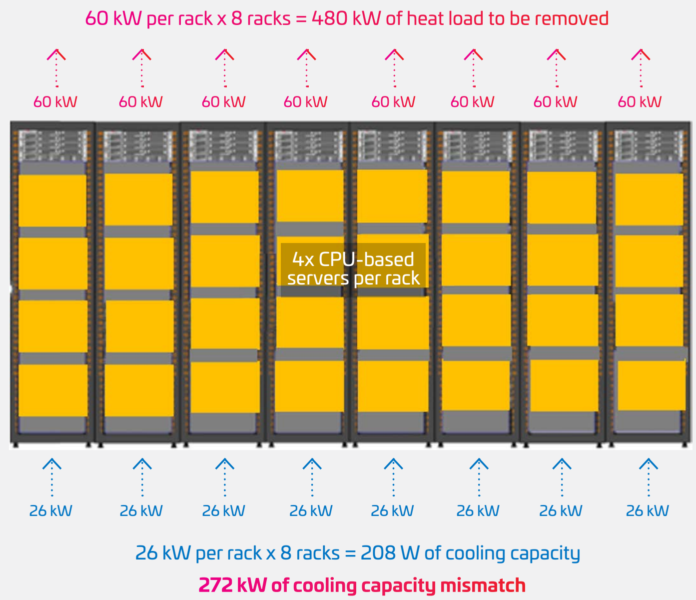


Figure 2: 4x GPU-enabled servers per rack: Air-cooling system is no longer sufficient and requires more cooling capacity, such as from liquid cooling in the form or direct-to-chip (DTC) and its additional pipework for each cabinet, rack and row.

Key points in designing a physical layer to support AI clusters

1. Additional networks must be considered.

In a traditional data center design, we typically have the east-west network fabric cabling (e.g., server to T0 switch, and T1 to T2 switching layers), plus the out-of-band management and dedicated SAN links.

However, an AI cluster deploying the new NVIDIA NVL72 RSS sees each rack typically having the following scale-out fiber network links:

- 72x compute back-end/GPU fabric
- 36x storage + in-band front-end/CPU fabric
- 4x out-of-band management (OOB), an out-of-band management copper link, using Category 6 or Cat 6a structured cabling
- 8x Cat 6 or Cat 6a copper links for the admin network
- 20x additional fiber cabling links to support redundancy

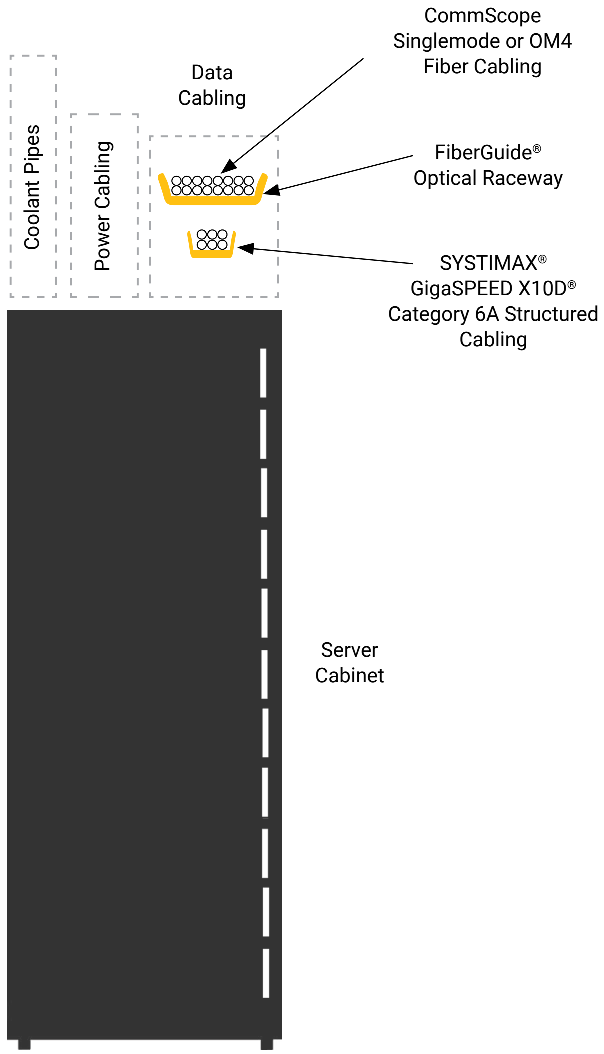
Adding these up, it's quickly apparent that the traditional cabling approach for compute cabinets used in non-AI environments will no longer be a sustainable strategy in an environment where liquid cooling infrastructure is deployed alongside fiber cabling. There is only so much "give" in the network.

2. Infrastructure zoning and overhead pathway positioning.

Having individual zones above cabinet rows can prevent a conflict between data cabling, power feeds and the liquid piping used to transport liquid cooling fluid. Protecting the fiber cabling across the data center hall and having maximum run lengths for AI-enabled servers will help maximize performance. Introducing a fiber raceway like the FiberGuide® system is an effective platform for transporting cabling.

3. Cabinet selection and cable management.

Server racks come in different sizes: 600 mm and 800 mm widths. It might be tempting to select a smaller footprint to conserve space, but in the world of AI, the vertical riser space must be sufficiently large to enclose all the additional infrastructure that is now required. Selecting a wider—and deeper—cabinet is often the more prudent choice. However, this too brings challenges that make traditional cable patching from the rear of a cabinet very difficult, because of the lack of physical space to access the rear of panels. In this instance, designing in-cabinet patching solutions with all-front access platforms, like the Rapid Fiber Connect™ platform, can significantly improve installation time and easier Day 2 moves, adds and changes (MACs).



4. Maximizing GPU performance.

Essential to AI performance is the amount of low-latency, high-bandwidth data provided to the GPUs. High-quality data cabling and optical end-face performance are considered “table stakes” today.

As fabric bandwidths continue to increase, underpinned by higher data rates, singlemode, and OM4 optical fiber, structured cabling is essential. Many optical channels require so much bandwidth that parallel optics must be deployed. These channels require parallel connectivity like MPO16 and MPO12/8 supplied with APC polishing, preventing unwanted optical reflections that impact the performance of the link.

5. Cable organization.

Use structured cabling to have easier MACs and to streamline future scalability for network growth, especially as the pace of evolution in AI hardware is so rapid.

Prioritizing and acting on these areas of focus early on in an AI data center build can help maintain efficient, reliable and scalable data cabling solutions to support their demanding computational workloads. As GPUs demand more resources to power their growing levels of performance, data centers must keep a future-focused view on their buildouts.

By Alastair Waite, Senior Manager,
Global Data Center Market Development, CommScope

This article was originally published as a chapter in CommScope’s 2026 Data Center eBook, available for free here.

[1] https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era/

[2] https://www.heise.de/en/news/AMD-Epyc-9005-server-CPU-First-test-confirms-high-performance-and-efficiency-9976883.html#:~:text=The%20Epyc%209755%20with%20the,AMD's%20Zen%205%20compact%20cores

EB-117988.2-EN (03/26)