Data Centers in the Age of AI

Article first published 14th June 2023, updated 29th May 2024.

Artificial intelligence (AI) is changing the technology landscape in various industries, and data centers are no exception. AI algorithms are computationally heavy and will increase data centers' power consumption and cooling requirements, deeply affecting data center infrastructure.

The Constraints of the Power Grid

Data centers famously consume a significant amount of energy, and power-hungry AI algorithms will lead to a further increase in data center power consumption. The world's major data center providers are already gearing up for this increase. For example, a recent Reuters report explains how Meta computing clusters needed 24 to 32 times the networking capacity. This increase required redesigning the clusters and data centers to include new liquid cooling systems.

Despite the best efforts of the world's tech giants to rethink their architectures, it's clear that data centers and their new AI workloads are hitting electrical power grid limitations. The capacity of the power grid is now increasingly regarded as the main chokepoint that prevents AI clusters from being more widely implemented in data centers.

Since changes in the power grid distribution would take decades to materialize, data center providers know they cannot continue to centralize their data center architectures. To adapt to the power grid constraints, providers are thinking about how to transfer data between decentralized data center locations instead.

For example, data centers can relocate to areas with available spare power, preferably from nearby renewable energy sources. Efficiency can increase further by sending data to branches with spare capacity. The Dutch government has already proposed this kind of decentralization as part of its spatial strategy for data centers.

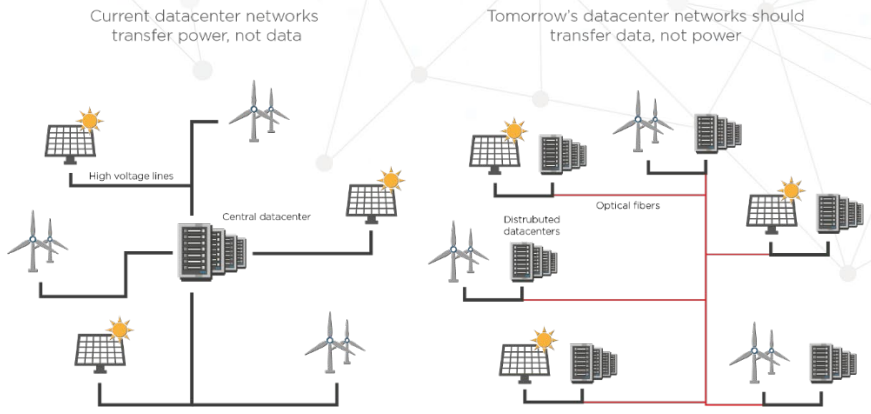


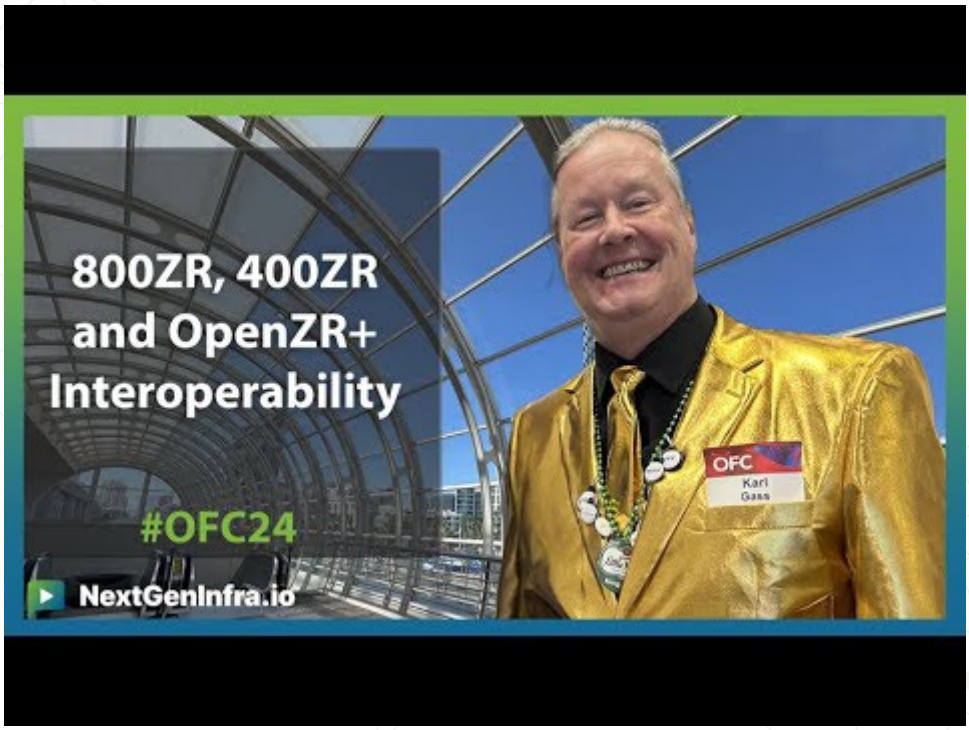
Figure 1: High-speed fiber-optic connections allow data processing and storage to be moved to locations where excess (green) energy is available. Data can be moved elsewhere if power is needed for other purposes, such as charging electric vehicles.

Interconnecting Data Centers over Long Distances

Longer data center interconnects enable a more decentralized system of data centers with branches in different geographical areas connected through high-speed optical fiber links to cope with the strain of data center clusters on power grids.

These trends push the data center industry to look for interoperable solutions for longer interconnects over 80 to 120 km distances. The advances in electronic and photonic integration allowed coherent technology for metro DCIs to be miniaturized into QSFP-DD and OSFP form factors. This progress allowed the Optical Internetworking Forum (OIF) to create the 400ZR and ZR+ standards for 400G DWDM pluggable modules. With modules that are small enough to pack a router faceplate densely, the datacom sector could profit from a 400ZR solution for high-capacity data center interconnects of up to 80km.

After the success of 400ZR standardization, the data center industry and the OIF are starting to promote an 800ZR standard to enable the next generation of interconnects. In OFC 2024, we started seeing some demos from several vendors and the OIF on this new standards initiative.



Optical Interconnects Inside Data Centers

The increasing complexity of AI processing will impact not just the interconnections between data centers but also the architectures inside the data center. AI nodes inside data center racks are normally connected via electrical or RF signals, while the racks are connected via optical fiber interconnects. However, as AI systems do more parallelized processing, data center racks run into electrical memory and power consumption constraints. These electrical interconnects between AI nodes are increasingly becoming a bottleneck in the ability of data center architectures to scale and handle the demands of AI models sustainably.

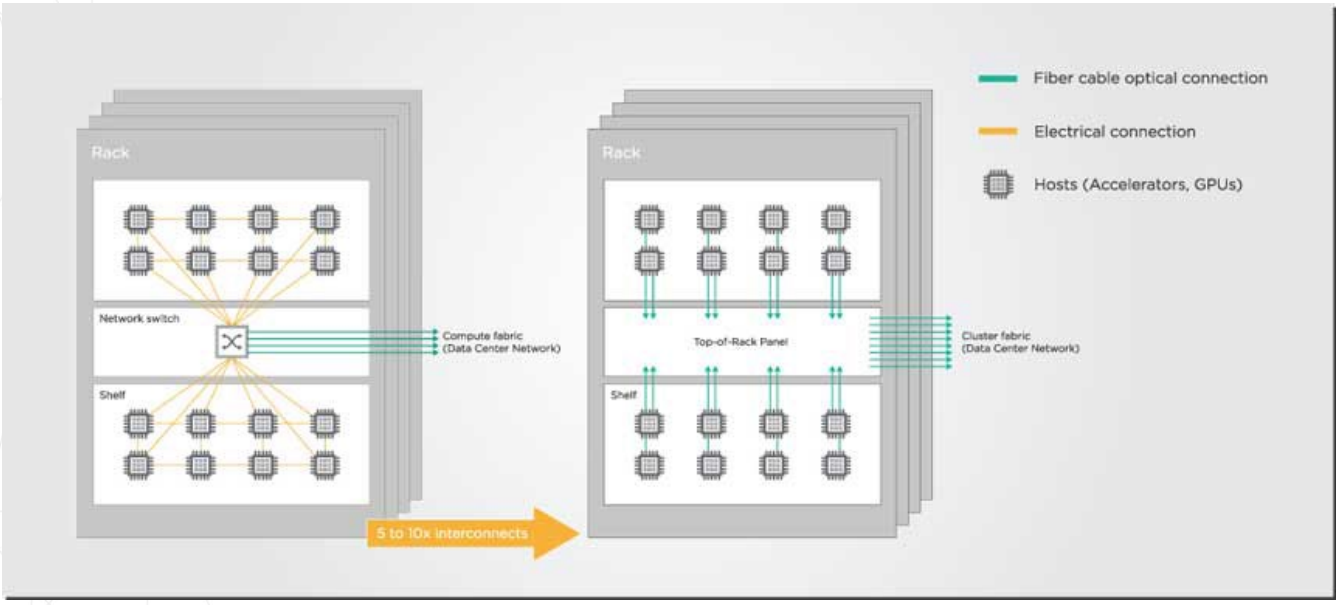


Figure 2: How architectures might change inside the AI data center. AI nodes on a single rack are connected via electrical or RF connections to an electro-optic switch that connects to other racks with optical fiber. In the future (on the right side of the figure), these AI nodes might be interconnected via optics. Figure inspired by Rob Stone's (Meta) presentation.

Optics will play an increasingly larger role in this area. As explained by Andrew Alduino and Rob Stone of Meta in a talk at the 2022 OCP Global Summit, interconnecting AI nodes via optics will be vital to decreasing the power per bit transmitted inside data center racks. This means changing the traditional architecture inside these racks. Instead of using an electro-optical switch that converts the electrical interconnects between AI nodes to optical interconnects with another, the switching inside the data center rack would be entirely optical.

Avoiding the power losses of electrical connections and electrical-optical conversions will improve the cost and power per bit of connections inside the data center. As more data center capacity is needed, these new optical interconnections might also need co-packaged coherent optics to scale effectively. This is the argument made recently by our very own Joost Verberk, EFFECT Photonics' VP of Product Management, at the 2024 OCP Regional Summit in Lisbon.

Takeaways

As AI continues reshaping the technological infrastructure, data centers undergo significant transformations to meet the new demands. The shift towards AI-intensive operations has exacerbated the existing strain on power grids, pushing data center providers towards more decentralized solutions. This includes relocating to areas with spare power and transferring data with optical interconnects across geographically dispersed locations.

The adoption of advanced optical interconnections is happening also inside data centers, as data center racks might transition to all-optical switching to connect their AI nodes. These evolving strategies not only address the immediate challenges of AI workloads but also set the stage for more sustainable and scalable data center operations in the future.

More Articles from EFFECT Photonics