Edge Data Centers in the Age of AI

Cloud computing offers scalable and flexible IT resources over the Internet, allowing businesses to avoid the upfront cost and complexity of owning and maintaining their own IT infrastructure. Cloud services allow organizations to access a wide array of computing resources on demand, such as servers, storage, databases, and software applications.

Edge computing complements cloud computing by processing data near the source rather than relying on a central data center. This is important for applications requiring real-time processing and low latency, such as autonomous vehicles, industrial automation, and smart city technologies. By minimizing the distance data must travel, edge computing reduces latency, increases data processing speed, and could also increase data privacy and reliability.

These advantages are causing the global market for edge computing to explode, with STL Partners predicting that the total edge computing addressable market will grow from $9 billion in 2020 to $445 billion in 2030. They also forecast that the number of edge data center sites will triple, from 491 in 2023 to 1584 in 2025.

Various trends are driving the rise of the edge cloud:

- **5G technology and the Internet of Things (IoT):** These mobile networks and sensor networks need low-cost computing resources closer to the user to reduce latency and better manage the higher density of connections and data.
- **Content delivery networks (CDNs):** The popularity of CDN services continues to grow, and most web traffic today is served through CDNs, especially for major sites like Facebook, Netflix, and Amazon. By using content delivery servers that are more geographically distributed and closer to the edge and the end user, websites can reduce latency, load times, and bandwidth costs as well as increasing content availability and redundancy.
- **Software-defined networks (SDN) and Network function virtualization (NFV).** The increased use of SDNs and NFV requires more cloud software processing.
- **Augment and virtual reality applications (AR/VR):** Edge data centers can reduce the streaming latency and improve the performance of AR/VR applications.

However, in this article, we will discuss arguably the most important trend powering the explosive growth of the edge data center sector: artificial intelligence.

The Impact of AI on Edge Data Centers

AI will enhance many of the applications we mentioned before, including real-time analytics, autonomous vehicles, augmented reality, and IoT devices, which require low-latency data processing and immediate decision-making capabilities.

As AI workloads grow, traditional centralized data centers face challenges in handling the sheer volume of data generated at the edge. Due to these challenges, there is increased motivation to move these AI workloads towards the network edge. Edge data centers provide a decentralized solution, distributing computing resources closer to end-users and devices, which reduces the strain on centralized cloud infrastructures.

This proximity to the end user allows for more real-time data processing and decision-making, which is critical for applications that require immediate responses. Industries such as manufacturing, healthcare, retail, and smart cities could benefit from edge AI. For instance, in manufacturing, edge AI can monitor machinery in real-time to predict and prevent failures, enhancing operational efficiency and reducing downtime. In healthcare, edge AI enables real-time patient monitoring, providing immediate alerts to medical staff about critical changes in patient conditions.

The integration of AI at the edge also addresses the growing need for data privacy and security. By processing data locally, sensitive information does not need to be transmitted to centralized cloud servers, reducing the size and risk of data breaches and streamlining compliance with data protection regulations. Moreover, edge AI reduces the bandwidth required for data transfer, as only the necessary information is sent to the cloud, optimizing network resources and reducing costs.

The Value of Edge Data Centers

Several of these applications require lower latencies than before, and centralized cloud computing cannot deliver those data packets quickly enough. As shown in Table 1, a data center on a town or suburb aggregation point could halve the latency compared to a centralized hyperscale data center. Enterprises with their own data center on-premises can reduce latencies by 12 to 30 times compared to hyperscale data centers.

Type of Edge		Data center	Location	Number of DCs per 10M people	Average Latency	Size
On-premises edge		Enterprise site	Businesses	NA	2-5 ms	1 rack max
Network (Mobile)	**Tower edge**	Tower	Nationwide	3000	10 ms	2 racks max
	Outer edge	Aggregation points	Town	150	30 ms	2-6 racks
	Inner edge	Core	Major city	10	40 ms	10+ racks
Regional edge		Regional	Major city	100	50 ms	100+ racks
Not edge		Hyperscale	State/national	1	60+ ms	5000+ racks

Table 1: Types of edge data centers and their characteristics. Source: *STL Partners*

Cisco estimates that 85 zettabytes of useful raw data were created in 2021, but only 21 zettabytes were stored and processed in data centers. Edge data centers can help close this gap. For example, industries or cities can use edge data centers to aggregate all the data from their sensors. Instead of sending all this raw sensor data to the core cloud, the edge cloud can process it locally and turn it into a handful of performance indicators. The edge cloud can then relay these indicators to the core, which requires a much lower bandwidth than sending the raw data.

Distributing data centers is also vital for future data center architectures. While centralizing processing in hyper-scale data centers made them more energy-efficient, the power grid often limits the potential location of new hyperscale data centers. Thus, the industry may have to take a few steps back and decentralize data processing capacity to cope with the strain of data center clusters on power grids. For example, data centers can be relocated to areas where spare power capacity is available, preferably from nearby renewable energy sources. EFFECT Photonics envisions a system of datacentres with branches in different geographical areas, where data storage and processing are assigned based on local and temporal availability of renewable (wind-, solar-) energy and total energy demand in the area.

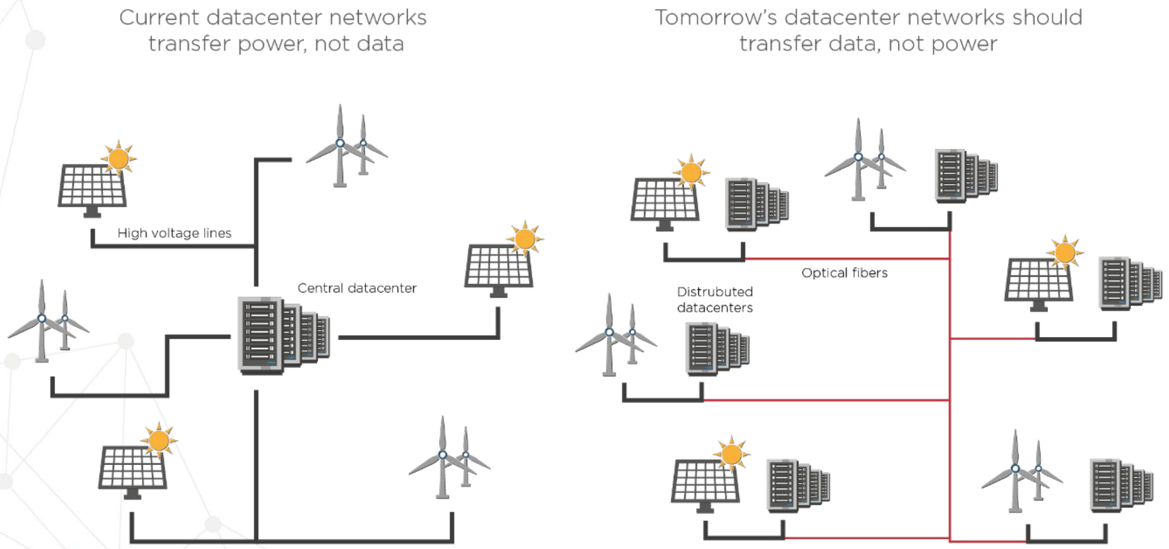

Figure 1: High-speed fiber-optic connections allow data processing and storage to move to locations where excess (green) energy is available. If power is needed for other purposes, such as charging electric vehicles, data can be moved elsewhere

Coherent Technology Simplifies the Scaling of Edge Interconnects

As edge data center interconnects became more common, the issue of how to interconnect them became more prominent. Direct detect technology had been the standard in the short-reach data center interconnects. However, reaching the distances greater than 50km and bandwidths over 100Gbps required for modern edge data center interconnects required external amplifiers and dispersion compensators that increased the complexity of network operations.

At the same time, advances in electronic and photonic integration allowed longer reach coherent technology 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 small enough modules 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. If needed, extended reach 400ZR+ pluggables can cover several hundreds of kilometers. Cignal AI forecasts that 400ZR shipments will dominate in the edge applications, as shown in Figure 3.

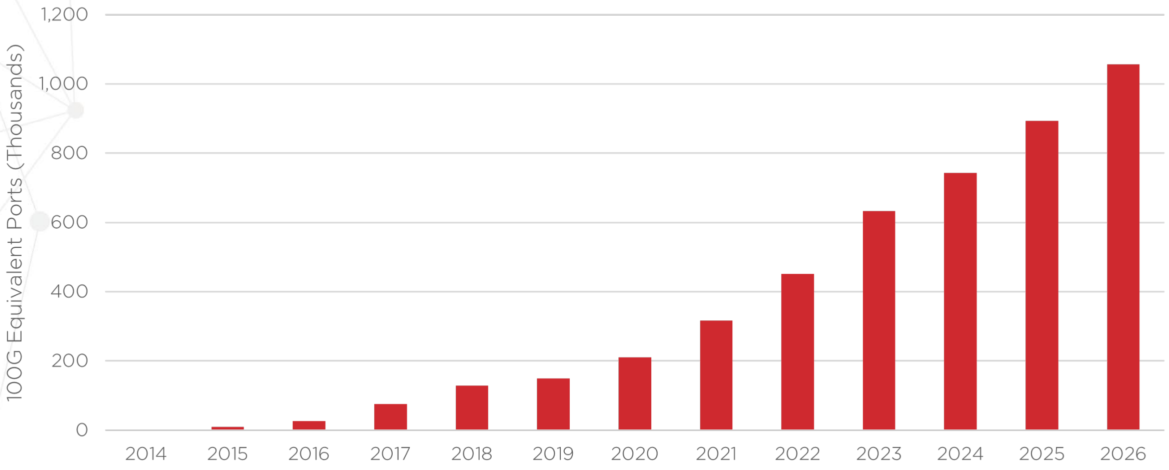

Figure 3: Forecast of 100G port equivalents shipped for edge applications. These shipments are overwhelmingly 400ZR standard technology. Source: Cignal AI Transport Applications Report Q42021.

Further improvements in integration can further boost the reach and efficiency of coherent transceivers. For example, by integrating all photonic functions on a single chip, including lasers and optical amplifiers, EFFECT Photonics' optical System-On-Chip (SoC) technology can achieve higher transmit power levels and longer distances while keeping the smaller QSFP-DD form factor, power consumption, and cost.

Takeaways

Edge computing is becoming increasingly critical for supporting AI-driven applications, such as autonomous vehicles, real-time analytics, and smart city technologies. The proximity of edge data centers to end-users allows for faster data processing, lower latency, and improved data privacy. This shift is driven by the growing need for real-time decision-making capabilities and compliance with data protection regulations. As AI workloads continue to expand, edge data centers provide a decentralized solution that alleviates the pressure on centralized cloud infrastructures and optimizes network resources.

More Articles from EFFECT Photonics