The Evolving Role of the Data Center in Today's 5G-Enabled World

For decades, the data center has stood at or near the center of the network. For enterprises, telco carriers, cable operators, and, more recently, service providers like Google and Facebook, the data center has remained the heart and muscle of IT. The emergence of the cloud has undoubtedly emphasized the central importance of the modern data center. But if you listen closely, you'll hear the rumblings of change.

As networks plan for migration to 5G and the internet of things (IoT), IT managers are now focusing on the edge and the increasing need to locate more capacity and processing power closer to the end users. As they do, they are re-evaluating the role of their data centers. According to Gartner, by 2025, 75% of enterprise-generated data will be created and processed at the edge—up from just 10% in 2018.[1] At the same time, the volume of data is getting ready to hit another gear. A single autonomous car will churn out an average of 4 terabytes of data per hour of driving.

Networks are now scrambling to figure out how best to support huge increases in edge-based traffic volume and the demand for single-digital latency performance without torpedoing the investment in their existing data centers. Heavy investment in east-west network links and peer-to-peer redundant nodes is part of the answer, as is building more processing power where the data is created. But what role will data centers play?

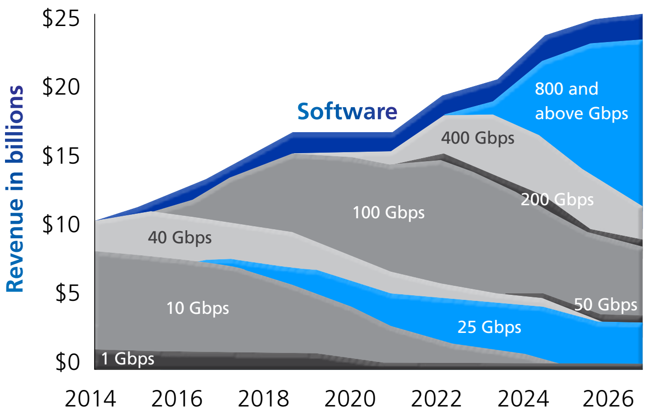


Figure 1: Data center Ethernet switch revenue by speed. Source: 650 Group, Market Intelligence Report December 2020

The AI/ML feedback loop

The future business case for hyperscale and cloud-scale data centers lies in their massive processing and storage capacity. As activity heats up on the edge, more power will be needed to produce the algorithms that enable the data to be processed. In an IoT-empowered world, the importance of artificial intelligence (AI) and machine learning (ML) cannot be understated—and neither can the role of the data center.

Producing the algorithms needed to drive AI and ML requires massive amounts of data processing. Core data centers have begun deploying larger CPUs teamed with tensor processing units (TPUs), graphical processing units (GPUs), or other specialty hardware. In addition, AI and ML require very high-speed, high-capacity networks featuring an advanced switch layer feeding banks of servers—all working on the same problem. AI and ML models are the product of this intensive effort.

On the other end of the process, AI and ML models need to be located where they can have the greatest business impact. For enterprise AI applications like facial recognition, for example, the ultra-low latency requirements dictate they be deployed locally, not at the core. But the models must also be adjusted periodically, with data collected at the edge and fed back to the data center to update and refine the algorithms.

Playing in the sandbox or owning it?

The AI/ML feedback loop is one example of how data centers must work to support (not dominate) a more expansive and diverse network ecosystem. Adapting to a more distributed, collaborative environment for players in the hyperscale data center space will not come easily. Hyperscalers want to ensure that AI and ML processing and accessing the edge takes place on their platform, but not necessarily in their facilities.

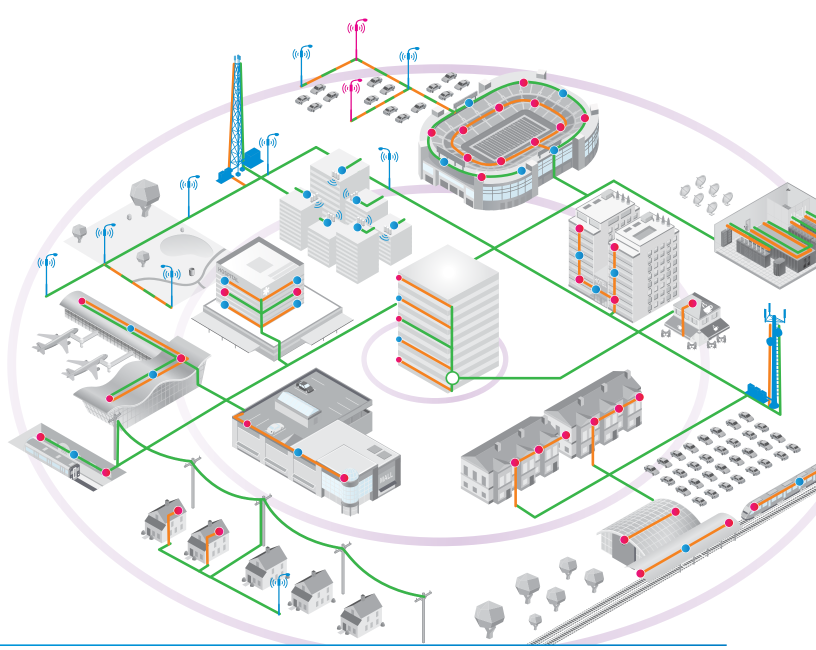
Providers like AWS, Microsoft and Google are now pushing racks of capacity into customer locations—including private data centers, central offices and on-premises within the enterprise space. This enables customers to build and run cloud-based applications from their facilities, using the provider's platform. Because these platforms are also embedded in many of the carriers' systems, the customer can run their applications anywhere the carrier has a presence. While still in its infancy, this model provides more flexibility for the customer while enabling providers to control and stake a claim at the edge.

Meanwhile, other models hint at a more open and inclusive approach. Edge data center manufacturers are designing hosted data centers with standardized compute, storage and networking resources. A smaller customer, such as a gaming company, can rent a virtual machine to host their customers, with the data center operator charging via a revenue-sharing model. This is an attractive model for a small business competing for access to the edge—and potentially the only way for them to compete.

Foundational challenges remain

As the vision for next-generation networks comes into focus, the industry must confront the challenges of implementation. Within the data center, we know what that looks like: Server connections will go from 50G per lane to 100G; switching bandwidth will increase to 25.6T; and migration to 100G per lane technology will take us to 800G pluggable modules.

Less clear is how we design the infrastructure from the core to the edge—specifically, how we execute data center interconnect (DCI) architectures and metro and long-haul links and support the high-redundancy, peer-to-peer edge nodes. The other challenge is developing the orchestration and automation capabilities needed to



manage and route massive amounts of traffic. These issues are front and center as the industry moves toward a 5G and IoT-enabled network.

Getting there together

What we do know for sure is that, in today's 5G-enabled world, the job of building and implementing next-generation networks will involve a coordinated effort. The data center—whose ability to deliver volume compute and storage cannot be duplicated at the edge—will certainly play a role. But, as responsibilities within the network become more distributed, that role will become subordinate to that of the larger ecosystem.

Tying it all together will be a faster, more reliable physical layer, beginning at the core and extending to the furthest edges of the network. This cabling and connectivity platform, powered by traditional Ethernet optics and coherent processing technologies, will fuel capacity. New switches featuring co-packaged optics and silicon photonics will drive more network efficiencies. And with that comes more fiber everywhere—packaged in ultra-high-count, compact cabling—that will underpin the evolution of network performance.

References

1 What Edge Computing Means for Infrastructure and Operations Leaders; Smarter with Gartner; October 3, 2018

CO-118689-EN (03-24)