Market Insight Report Reprint

Living on the edge: A primer on hybrid cloud and edge infrastructure

October 11 2021

by Owen Rogers, Ian Hughes

Without the internet, the cloud is nothing. But few of us really understand what is inside the internet. What is the so-called 'edge' of the internet, and why does it matter? And how does cloud play into the edge story? This beginner's guide to hybrid cloud and edge infrastructure seeks to explain these issues to a non-tech audience.

451 Research

S&P Global
Market Intelligence

Introduction

Without the internet, the cloud is nothing. But few of us really understand what is inside the internet. What is the so-called 'edge' of the internet, and why does it matter? And how does cloud play into the edge story? This primer seeks to explain these issues to a non-tech audience.

THE TAKE

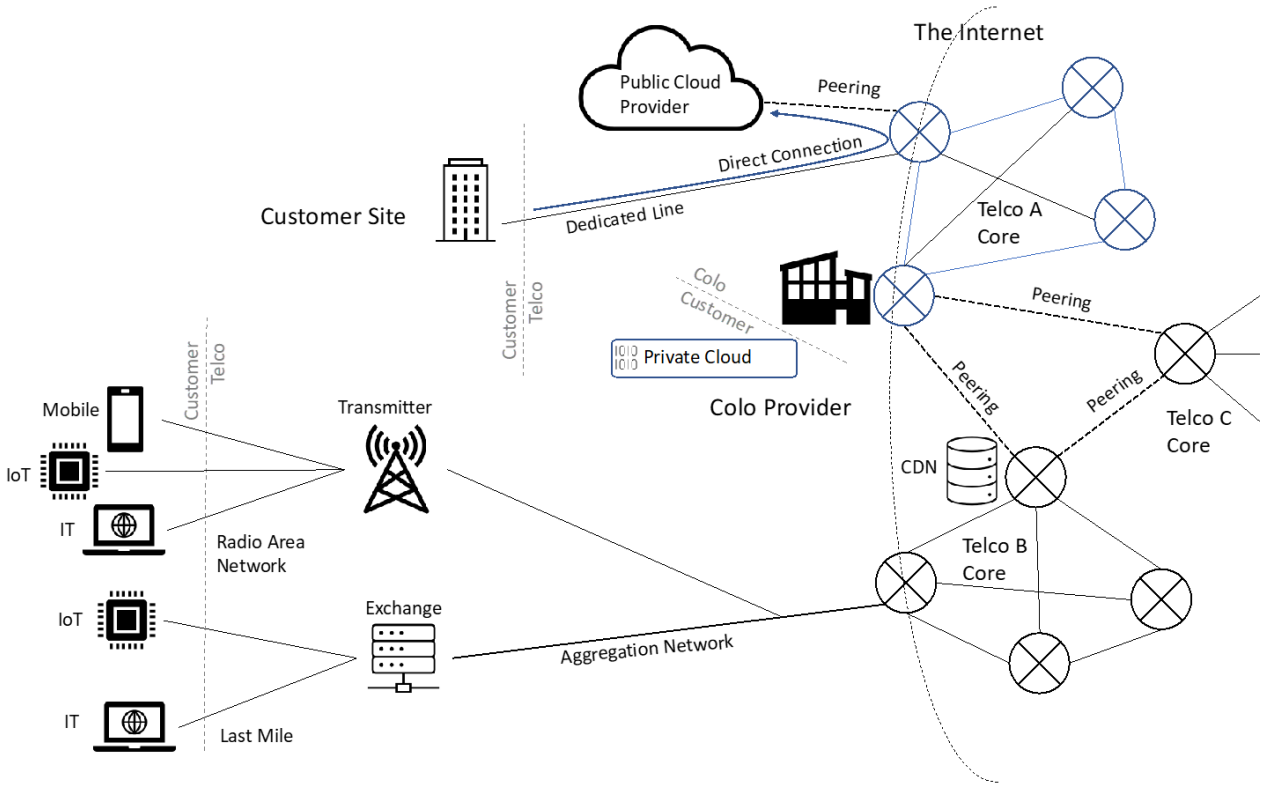
Hybrid IT is more than just public and private clouds. It is the collection of tools, products and services that allow enterprises to choose the best location for each workload's particular requirements. Edge offers low latency for applications where either a fast response is critical or where transporting data to centralized locations is too costly or not possible due to poor connectivity. But edge isn't one single thing. Just as cloud providers offer a vast array of services, the edge offers different capabilities for different needs. The network is of critical importance to any application, but it's often not at the forefront of decision-making. As consumers demand a faster and more impressive experience, and enterprises look to improve efficiency and take advantage of innovations in AI and IoT, latency will become increasingly important. 5G is one piece of the puzzle, but applications will need to be architected across public, on-premises and edge locations to optimize the experience using the latest innovations without adding unnecessary expense.

The internet

What is the internet? And who owns it? Essentially the internet is a collection of networks which share traffic. A network is a collection of routers connected via a massive series of communications links, often via fiber-optic cables. The routers' job is to determine over which link data should be sent to reach its destination. The internet is designed to be resilient, ensuring this routing can dynamically cope with any outages in network nodes. A core network is the very fast, very resilient backbone of the telco, which distributes data over long distances.

Each core network is typically owned by a telco (or a network provider). The telcos peer with each other, meaning they share traffic. This peering architecture creates a very large, very interconnected network that isn't owned by a single entity – the internet.

Figure 1: Diagram of Standard Internet Network Architecture



Source: 451 Research, Cloud Price Index

A cloud provider will typically have its own core network, which links its datacenters across regions and the globe. The cloud provider also peers with other networks so it too can share traffic and thus connect to the internet. Rather than draw every component of a network, we often represent a network as a cloud – hence the term 'cloud computing.'

Mobile and broadband users pay a fee to their telco to have access to the internet. The telco then works among peering partners to share costs for traffic exchanged, but this is totally invisible to the consumer. The role of the core network is to move data as quickly as possible close to where it is required. But if the core network were to directly connect every single user, the cost would be astronomical, and efficiency might be compromised by data having to traverse many routers to find its destination. Instead, an aggregation network branches off the core network, taking data closer to exchanges or mobile transmitters. From here, users access the network through 3G, 4G and now 5G connections to masts and broadband connections to exchanges.

The core handles the bulk of overall data transmission. The aggregation network transmits data closer to individuals. The access network handles the 'last mile' to the user's device.

Great! So what?

The problem is time, or more specifically, latency. The more parts of the network that data flows through (and the physical distance between those parts), usually the longer data takes to get there. Sometimes this delay isn't problematic. Even if a streaming video takes one second to get from the website to your cell phone, you wouldn't even notice after the initial delay. A subscription coffee club delivers your daily coffee every day. After the initial wait for the coffee to be sent from the warehouse to your home, you'd expect it every day without delay, assuming the club delivers coffee each day. The problem arises when there is interaction. If, after the first delivery, you must write back to the club stating what you need, not only do you suffer the delay in your message getting to the warehouse, but also the delay in the warehouse responding to you.

Such sensitivity is particularly apparent in applications like online multiplayer, mixed reality and cloud-based gaming. When in a client/server environment a player moves their rifle, jumps to a platform or invades a country, the game needs to respond in kind. Many small packets of data for these actions flow from each online game client to servers, and then back out again to fellow player clients. Cloud gaming has a greater data challenge because everything the player sees and hears is rendered by a cloud resource; the game will constantly be creating a new high-resolution image and audio and sending it back to the player over 60 times a second, responding to inputs from player actions. Our report on the evolution and complexity of the games industry applied at the edge and in the cloud describes this further.

In most cases, a delay not only ruins gameplay, but it can also render the game impractical due to poor user experience. On a more industrial scale, robots or machinery are affected by latency issues. While many devices have a degree of autonomy, they are part of an integrated process often including much older machinery. For a fleet of machines or an IoT-connected production line, they all need timely data interchange. A production line that analyzes the quality of a part can be significantly affected if communication between a sensor and an algorithm is delayed. Control signals to shut down processes when a dangerous situation is detected need to be as fast as possible, especially in process industries such as chemical and petroleum, descriptively labeled as the 'boomable' industries.

Latency performance is one of the big negatives of cloud computing. By its very nature, it is accessed via a network over a distance where physics, with its maximum speed-of-light limitation, can't be ignored. To access a cloud resource, a user faces several potential delays. A mobile or fixed broadband connection is shared by many users and may be oversubscribed – we say the network is 'contended' when multiple users must share the connection. The aggregated network from the mast or exchange to the core network too is shared. Once at the core network, the data may still have to connect across multiple telcos before finally arriving in the hyperscaler's cloud. The speed of the transmission isn't related just to the telco's network – it's related to how many users are using it at that time, and how other telcos' networks are performed. Understanding this, cloud providers are increasingly creating more regions in more datacenters – a big driver is to reduce latency. And some providers are even creating smaller 'metro' clouds in cities, such as AWS with its Local Zones service.

Hosting a website on a server on-premises, i.e., in an office or warehouse, doubles the problem; not only do users face delays from their location to the core, but from the core to the office. Situating the server in a colocation provider can help resolve this because there is usually a direct connection into a core network. Alternatively, the user may purchase dedicated capacity from the remote office to the internet – in other words, the network isn't shared by anyone else, or the user is given definite capacity on a shared network.

In a hybrid cloud, network is critical. If clouds in different locations are going to work together, then they need to be able to communicate rapidly with each other. As a result, many enterprises choose to host their private cloud(s) in a third-party datacenter, close to core networks and therefore closer to its public cloud ('close' as in time to transmit, rather than physical distance). But to securely communicate between public and private clouds, a direct connection is needed – this provides encryption between a private cloud (on-premises or in a third-party datacenter) and the hyperscaler. All hyperscalers offer such capability, including AWS Direct Connect, Microsoft ExpressRoute and Google Dedicated Interconnection, IBM Cloud Direct Link, Oracle FastConnect and Alibaba ExpressConnect. The hyperscalers now also offer private clouds that include hardware specified by them that can be installed in a customer's choice of datacenter (at the customer's expense) and that provides a secure connection back to the private cloud, providing an integrated hybrid cloud (so-called 'cloud-to-ground'). These include AWS Outposts and Microsoft AzureStack. More can be found in Cloud Price Index: Cloud-to-Ground and Cloud-Around. Companies such as Equinix, Megaport and Interxion/Digital Realty have built brands around acting as interconnection hubs to reduce cloud latency.

Sometimes these edge offerings are packaged with private network products that have gained 'mindshare' due to the promised performance of 5G and a global push to make spectrum available for enterprise use.

Solved, right?

Interactions between private cloud and public cloud can be minimized through dedicated connectivity, proximity to core networks and network prioritization. But this is feasible only for companies that have a large budget and, more important, fairly static IT requirements (such as between two datacenters).

Consumers using cell phones or home broadband aren't going to spend on dedicated lines to core networks. Nor is it even feasible for a user roaming from place to place to make such commitments. The consumer's experience is fully controlled by their network provider. This isn't just problematic for end users – it can also be problematic for companies selling services over broadband. If an online game with fast, high-quality graphics can't be played by a large percentage of users, then those users become lost revenue opportunities. Netflix has been successful because of economies of scale – if it couldn't reach most of its target audience, would it have been so successful? Not likely.

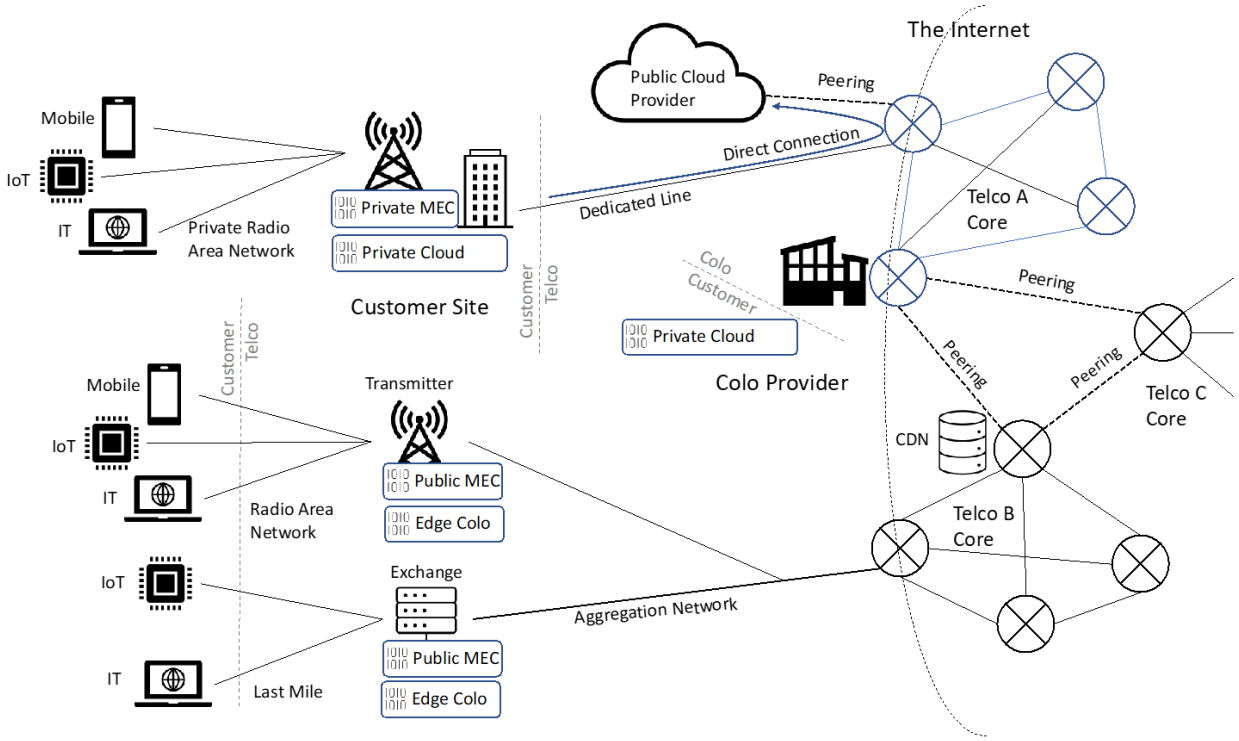
Connections between users and clouds are only going to get quicker over time. But there is a physical limit, in that data still takes some time to go from A to B and back again. A content delivery network helps to some degree. Essentially, a CDN is a collection of servers situated where networks peer and thus exchange traffic (an internet exchange). The server caches content so that, rather than having to route data across many networks from the source for every request, the CDN server can serve the data closer to where it is needed. But CDNs work best with static data – data that is constantly changing due to a user interaction (such as an online game) isn't helped much by CDNs because data cached is unique for each request. And the problem of the delay between the user and the core network still exists. All the hyperscalers offer CDNs as do third parties such as Cloudflare, Stackpath and Rackspace. Our Economics of CDN report provides more detail.

5G, in particular, is poised to reduce this latency, slicing the network to provide software-defined quality of service paths for different types of data. In other words, dedicating capacity to types of data to meet performance requirements. But even with 5G, there is still a delay from the base station to the cloud and back.

Return of the MEC

Multi-access edge computing, or MEC, is one such approach. It essentially lets companies locate their applications in a server in a physical network location closer to end users, thus substantially reducing latency. When combined with 5G, wireless latency can be 1-5ms instead of 100ms or more.

Figure 2: Diagram of Standard Internet Network Architecture With MEC



Source: 451 Research, Cloud Price Index

In a public MEC, the customer accesses compute and storage capacity on the telco, close to where it is required and consumed as a service. These multi-tenant approaches can be as close to the user as the exchange or cell base station or might be at a point of metro aggregation. Examples of such offerings include AWS Wavelength and Azure Edge Zones with Carrier. Using these AWS or Azure interfaces, compute can be provisioned at a partner's edge locations (e.g., Verizon, KDDI, SK Telecom and Vodafone with AWS, or AT&T with Azure), closer to end users. In edge colocation, the user can provision their own edge device into a shared 'mini' colocation facility closer to end users. AWS has its Snowball Edge device and Azure has AzureStack Edge. These services also provide direct integration with public cloud. Partners are also offering compatible edge and private cloud devices, such as Hewlett Packard Enterprise, Lenovo and Dell EMC.

Such edge devices can also be deployed in a customer's premises in a single-tenant private MEC architecture. With a private 5G network, the customer then has dedicated access to and ownership of its own ultra-low latency cellular network with a rapidly responding platform. Why would anyone want this? In an industrial setting, there may be thousands of devices measuring and actuating in a production process. An algorithm may be used to provide feedback and instructions to machinery in response to measurements from sensors and videos. These devices may need to be wireless, for example, if they operate over a large area or are attached to vehicles or components on a production line. This algorithm may need to send some data back to the cloud for monitoring or to take advantage of specialized capabilities.

For example, imagine a high-resolution x-ray is used in the production line to check for defects. There are multiple lines in the warehouse, with each line having multiple x-ray quality checks. Sending photos from each part of the process to the cloud for analysis takes time and expense, especially if each camera has to wait to transmit data due to many users sharing the bandwidth. With a MEC approach, these photos can be processed on the edge, speeding up the whole production process. A dedicated 5G network ensures the warehouse doesn't experience delays due to other users' demands. Most hyperscalers now offer edge software to enable efficient processing and transmission of data on devices and sensors (e.g., AWS Greengrass and Azure IoT Edge) and IoT platforms that track devices and manage communication. There is more detail in our Economics of IoT report.

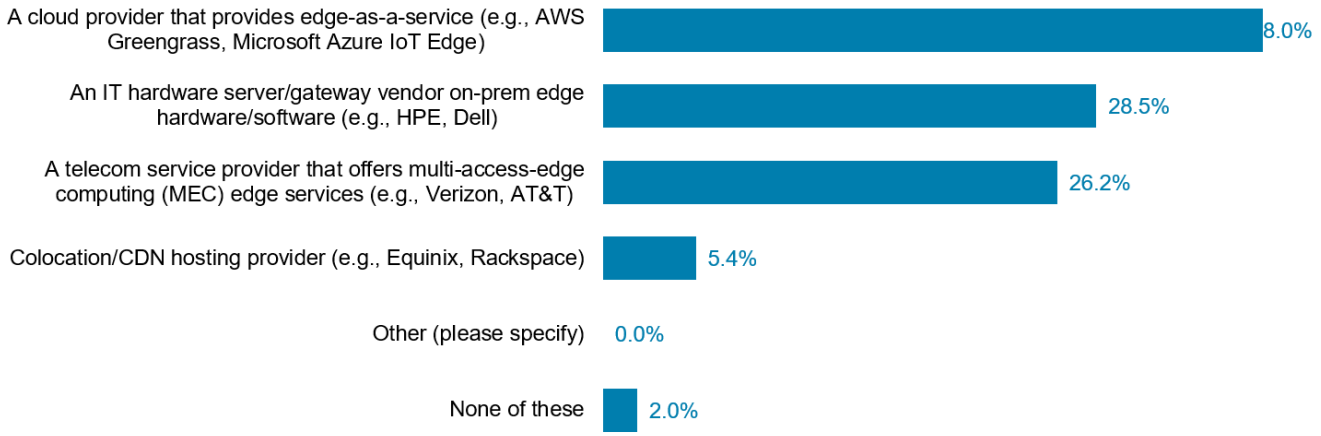
AI and machine learning (ML) are technologies that are poised to have a substantial impact in edge. Instead of sending data back to the cloud for analysis and response, a local processing unit close to where the data is generated determines what action needs to be taken, if any. These may be specific edge servers in an on-site datacenter, but they may also be onboard the machine or device. A typical model is emerging that is based on ML. A factory machine or process data is first delivered to a cloud AI system to develop or train a model. This may be data from traditional data logging, historic data or new data gathered from IoT sensors. Latency is not an issue in this scenario because this training is separated from daily operations.

Once a model is created, it is deployed back down to the edge computing device. In normal operation, it will make machine adjustments or report back only when the device is starting to operate out of normal parameters. Over time, the model will be reevaluated in the cloud or compared with other similar models and then redeployed to the low-latency edge process. Devices such as quality inspection cameras operate in this way with onboard processing, dealing with large amounts of visual information but only needing to specify signals of metadata to indicate a failure in quality of a product.

Outlook

MEC is one example of emerging edge computing architecture that is expected to account for a quarter of IoT project data processing and storage, according to respondents to 451 Research's Voice of the Enterprise: Internet Of Things, Workloads & Key Projects survey conducted in 2021.

Figure 3: Companies Considered as Primary Vendor for Storing, Processing and Analyzing IoT Data at the Edge



Q: Which of the following types of companies does your organization plan to use as their primary vendor for storing, processing and analyzing IoT data at the edge?

Source: 451 Research, Voice of the Enterprise: Internet of Things, Workloads & Key Projects 2021

This device edge example sits on the continuum of multiple edges and out to cloud infrastructure. For more detail on the entire ecosystem, see our Edge Computing Infrastructure and Services Market Map 2021.

For simplicity, we've focused on MEC and the edge as locations to hold virtual machines. But most hyperscalers are offering serverless technologies, built-in ML tools, containers, storage, databases and other capability as-a-service at the edge. The billing mechanism remains on-demand and pay-as-you-go, and the cloud provider continues to manage the underlying technology. The only difference is that, instead of being in a public cloud location, it is situated closer to users. Containers, in particular, are becoming a standard packaging mechanism for compute and allow easy deployment regardless of location. As the fundamental building block of cloud-native architectures, this gives developers and solution architects huge power to build applications that meet all their requirements of performance, scalability, resiliency and cost-effectiveness and to meet users where they are.

CONTACTS

The Americas
+1 877 863 1306
market.intelligence@spglobal.com

Europe, Middle East & Africa
+44 20 7176 1234
market.intelligence@spglobal.com

Asia-Pacific
+852 2533 3565
market.intelligence@spglobal.com

www.spglobal.com/marketintelligence