Independent market research and
competitive analysis of next-generation
business and technology solutions for
service providers and vendors

HEAVY READING
WHITE PAPER

5G's Impact on Edge Infrastructure

A Heavy Reading white paper produced for CoreSite

CORESITE

AUTHOR: STERLING PERRIN, SENIOR PRINCIPAL ANALYST, HEAVY READING

INTRODUCTION

For several years, data center trends have pointed to a shift from a few highly centralized mega facilities to more distributed data centers populating large metros. Edge computing is a new evolution of the distributed processing and storage trend that brings high performance compute, storage, and networking resources closer to users and devices than ever before. 5G is one of the key drivers accelerating edge computing.

This Heavy Reading white paper provides an in-depth analysis of the relationship between 5G and edge computing, describing how 5G will drive the edge and how the edge will enable 5G. The paper details some of the promising early 5G edge use cases and their requirements, and then defines the players in the emerging 5G edge ecosystem. Lastly, the paper focuses on the pivotal role that colocation providers will play in the emerging edge.

DEFINING THE EDGE AND DRIVERS FOR EDGE DEPLOYMENT

The "edge" can be complicated to describe and definitions abound, but The Linux Foundation provides a good starting point in defining edge computing:

The delivery of computing capabilities to the logical extremes of a network in order to improve the performance, operating cost and reliability of applications and services. By shortening the distance between devices and the cloud resources that serve them, and also reducing network hops, edge computing mitigates the latency and bandwidth constraints of today's Internet, ushering in new classes of applications.

– Open Glossary of Edge Computing Version 2.0, Linux Foundation, 2019

The edge computing requirement to distribute computing close to users runs directly counter to the historical cloud computing architecture that houses resources in a small number of massive (or hyperscale) data centers. For economic reasons, hyperscale data centers are often placed in remote locations. Thus, with edge computing, the pendulum swings from a centralized cloud architecture to a highly distributed one.

Significantly, the "edge" is a location, not a thing, and the location of the edge will vary depending on the use cases/applications as well as the edge owner. As a result, there will not be a single edge, but rather, multiple edges operating throughout the network. Omdia has defined three major classes of edge based on owners:

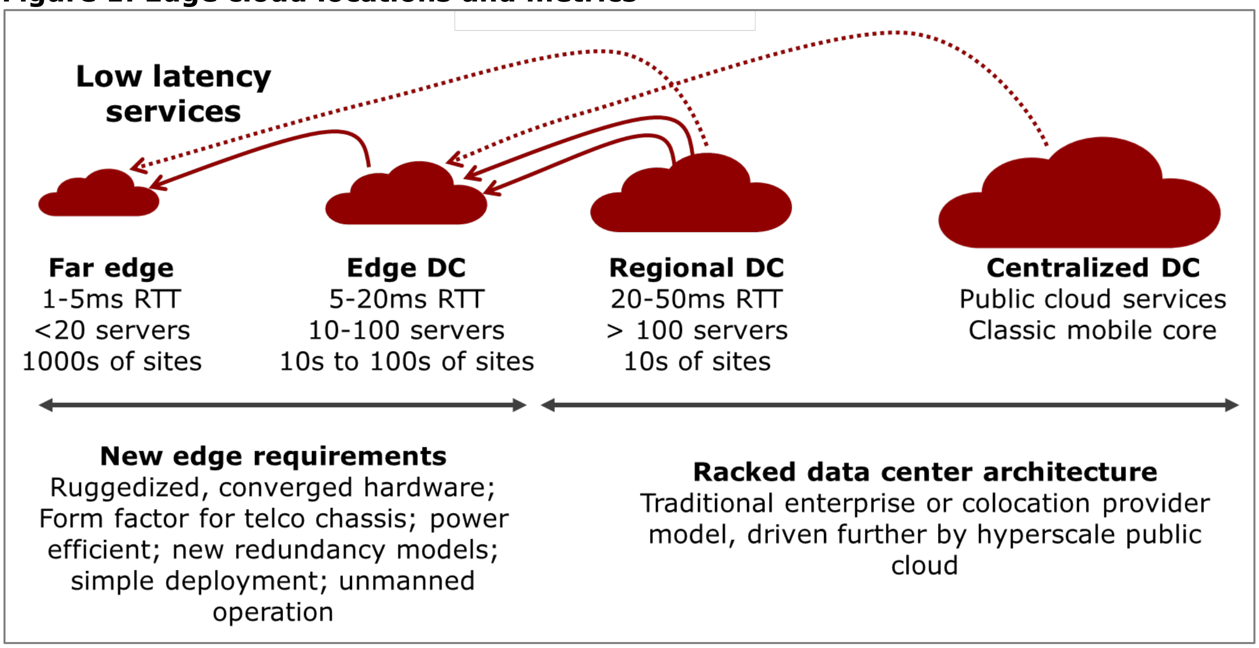
- **Service provider telco edge:** Telco IT–operated sites, including central offices, regional data centers, and leased space in colocation provider data centers
- **Service provider cloud edge:** Cloud service provider (CSP) IT–operated sites, including leased space in colocation provider data centers
- **Enterprise edge:** Enterprise-operated sites, including branch offices, industrial locations, regional data centers, and leased space in colocation provider data centers

Different applications/use cases will dictate a range of latencies, but there is a broad industry consensus that the edge requires less than 20ms round-trip time (RTT) between the location and the user or machine. Heavy Reading further subdivides locations between "far edge" requiring <5ms RTT and "near edge" requiring 5–20ms RTT from the edge location to the user/machine.

Note that distance and latency are tightly linked. Application RTT requirements dictate the maximum distance that the edge can reside from the end user or machine, but that distance will vary based on the transmission medium—such as transmission over fiber, coax, or air—as well as the number of hops (and processing) along the way. Because of these distance variables, RTT/latency metrics are typically used to describe the edge, but distance is equally crucial.

Figure 1 shows edge cloud locations matched with various metrics.

Figure 1: Edge cloud locations and metrics



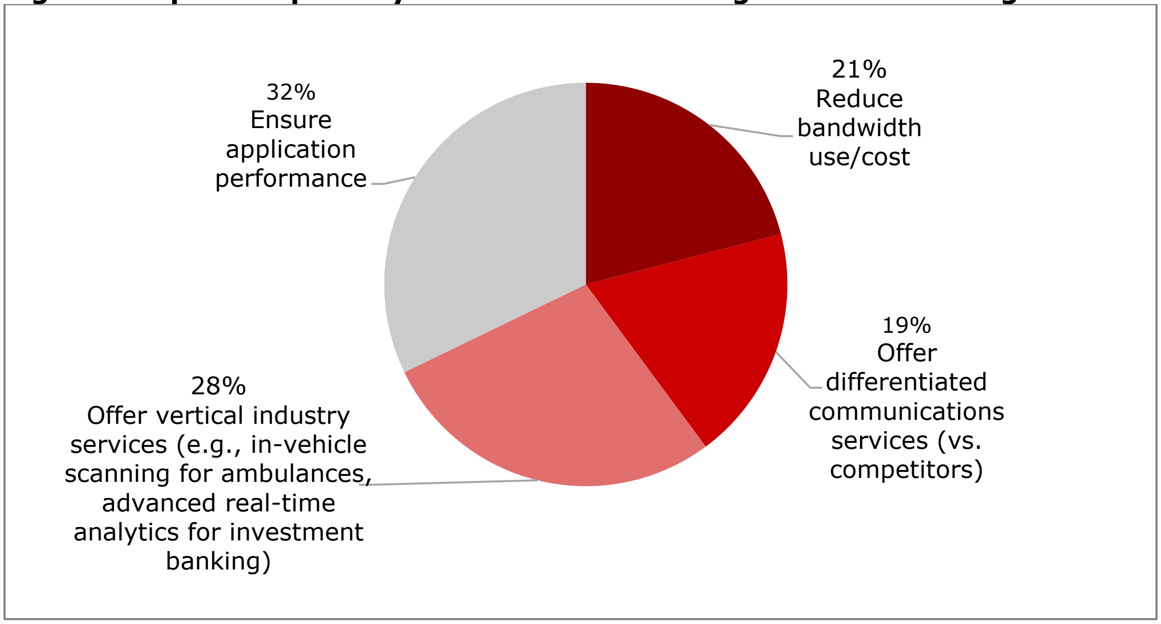
Source: Heavy Reading

Relation between 5G and edge

Edge is not defined by 5G, and vice versa, but 5G and edge are intricately intertwined. In many cases, 5G and edge computing will be combined to deliver advanced applications for enterprises and consumers, particularly those that require the ultra-high reliability and low latency (URLLC) functionalities that are enabled by 3GPP's Release 16 standardization (released in July 2020). 5G alone cannot deliver low latency without the proper location of the compute, storage, and networking resources on which advanced 5G applications and services depend.

In the 2020 5G Network & Service Strategies Survey, Heavy Reading asked more than 140 operator respondents globally to identify their primary motivation for moving workloads to the edge. The primary drivers were (1) ensuring application performance and (2) offering specific vertical industry services (see **Figure 2**).

Figure 2: Operator primary motivation for moving workloads to edge



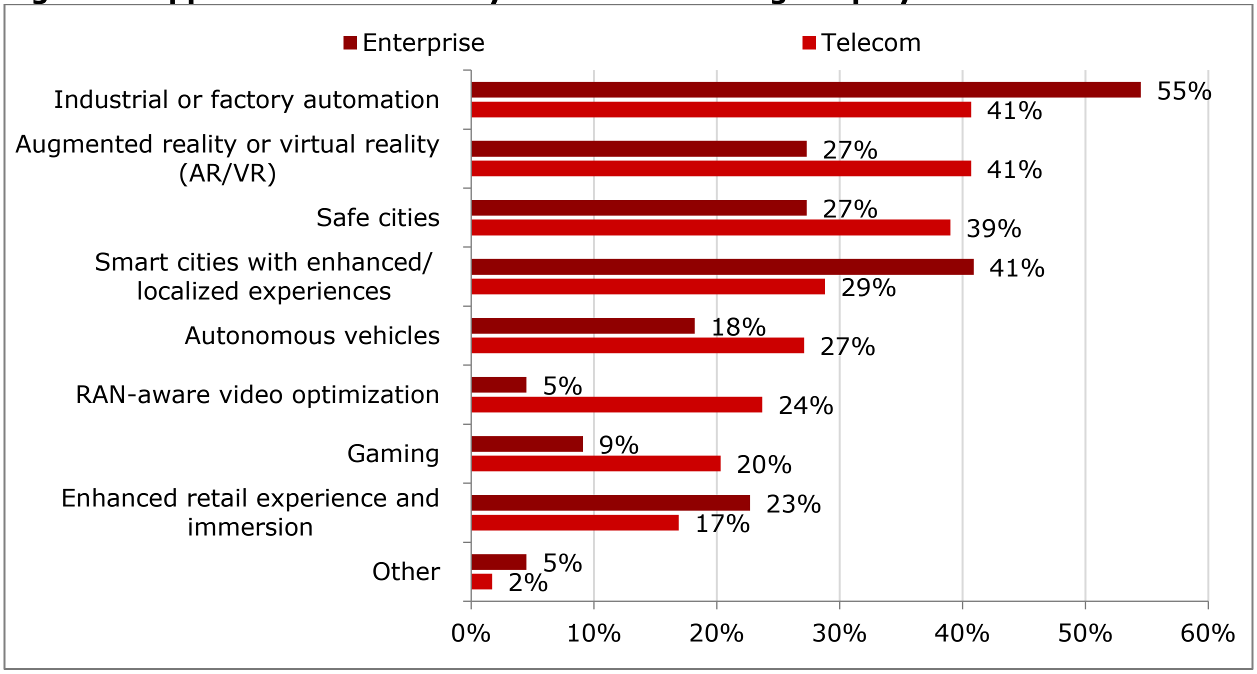
n=143

Source: Heavy Reading 5G Network & Service Strategies: 2020 Operator Survey

5G USE CASES REQUIRING EDGE COMPUTING

Not surprisingly, in a promising but nascent market segment, potential use cases abound. Only time will which use cases will take off, and it is a near certainty that some of the ultimate winners have yet to be envisioned. Heavy Reading surveyed both network operators and enterprises on applications most likely to drive their initial deployments. While not perfect, the data (presented in **Figure 3**) provides an early view of operator and enterprise thinking.

Figure 3: Applications most likely to drive initial edge deployments



n=59 telecom; 22 enterprise

Source: Strategies for Connecting the Edge: 2019 Heavy Reading Survey

Below, Heavy Reading provides additional detail on some of the most promising early use cases, with added insights into why these use cases require edge compute.

Factory automation

The global manufacturing industry is a key edge focus for telecom operators and enterprises alike (shown as the top edge driver for both in **Figure 3**). The manufacturing industry requires high quality, high availability, precision, security, and low latency operations. Today, most factories rely on wired communications in order to address requirements for bandwidth, latency, and reliability, but 5G opens the possibility of meeting these performance requirements while adding flexibility and convenience unavailable in wired networks. For this reason, factory automation was among the first use cases identified in 3GPP for URLLC capabilities (specified in 3GPP Release 16).

Across the spectrum of edge locations, many factory automation applications will fall into the far edge subcategory for two main reasons. The first reason is latency. Robotic manufacturing mandates not just high availability, but also ultra-quick response times. For many factory robotics applications in factory automation, an RTT of <5ms will be required. At the extreme end of automated manufacturing is tactile interaction in which humans wirelessly control the robotic processes. Tactile operations will require real-time reactions in <1ms ranges. Response requirements of less than 5ms (and certainly those less than 1ms) require edge computing and storage on-premises.

Beyond latency, security may also drive far edge (or local) compute for automated manufacturing. Enterprise CIOs may not be comfortable housing such critical data and process control systems offsite. They may mandate an on-premises deployment regardless of latency or reliability specifications.

Augmented reality

Augmented reality (AR) is the combination of the real-world environment supplemented by computer-generated input, including graphics, video, and sound. While AR was popularized by the 2016 consumer hit *Pokémon GO*, it actually has wide industry potential across many enterprise verticals and applications. Both 5G and edge connectivity are key enablers for advanced AR applications.

AR applications analyze output from a device's camera to supplement the user's experience. The application is aware of the user's position and direction and is then able to provide information to the user in real time. Crucially, as soon as the user moves, the information must be refreshed. Additionally, for many AR use cases, it may also be necessary to frequently update sensor-driven information in the user's field of view (even when stationary).

Potential use cases run the gamut. Wearing smart goggles, technicians can be provided with real-time maintenance and diagnostic information to quickly and accurately perform their tasks. Sensors in the maintenance environment may need to continuously provide that technician with updated data. In many cases, maintenance and diagnostic tasks may be quite critical in terms of business or employee safety—drawing on the full power of 5G's URLLC capability and the infrastructure that supports 5G.

Alternatively, AR can be used to enhance experiences for users attending museums, city monuments, art galleries, sporting events, or other venues. In these cases, patrons/ attendees point their cameras at areas of interest and receive real-time information related to what is being viewed. Not surprisingly, most of the opportunity here is in stadium sports. Fans can augment the live game with fully immersive media experiences, including real-time stats and unique camera angles, such as from the players' points of view.

Edge computing is required to deliver the AR data—often high bandwidth, delay-sensitive video content—in real time to the users. Beyond the creation of a smooth experience, edge compute reduces the nausea and dizziness sensations that result from high latency and low refresh rates in AR and virtual reality (VR) applications.

In some cases, AR edge placement will be obvious—such as in stadiums for sports or onsite for certain industrial maintenance applications. But other applications may have more flexibility across the far edge to near edge location spectrum.

Smart cities

Smart city projects use IT and communications technologies to improve efficiency, manage complexity, and enhance the quality of life of citizens. A smart city is not defined by a single initiative; rather, it encompasses a range of initiatives across governance, safety and security, transportation, energy, physical infrastructure, and healthcare. Goals include increasing citizen satisfaction, improving response times, increasing departmental collaboration, increasing energy efficiency, reducing annual costs, and more.

As noted, there are many use case possibilities under the smart cities umbrella. But the Internet of Things (IoT) is central to most. IoT devices provide connectivity and data through security cameras and sensors densely distributed throughout cities. 5G's IoT capabilities are particularly important, as 5G provides a 50x increase in device density per square kilometer compared to 4G. High bandwidth video is also central to smart city use cases such as safety and security, traffic congestion monitoring, and others. These use cases require the high data rates that 5G networks provide.

Edge computing comes into play through the requirements for high bandwidth (due to video content) combined with the need to make decisions in real time in response to situational data. The traditional centralized cloud model is not suitable. However, placing cloud resources throughout cities in locations with high processing, dense network, and high storage makes many use cases viable. For many of these use cases, near edge response times would suffice.

Cloud gaming

Online games are consumer applications, but they are also big business. Multiplayer online games housed on cloud servers involve many players—often into the hundreds of thousands—competing and cooperating within the same virtual world. Contributing to their success, these games can be accessed on most network-capable platforms, including PCs, video game consoles (e.g., Microsoft Xbox or Sony PlayStation), and smartphones (including those that are powered by 5G). Blizzard Entertainment's *World of Warcraft* and Epic Games' *Fortnite* are just two of many popular titles in this large and rapidly growing genre.

In multiplayer online games, latency is a crucial metric, as most games include extremely fast-paced action. The video game must appear to respond instantaneously to keystrokes and controller movements or the user experience degrades considerably. While games do function with 100ms+ communications, a range of 20–40ms is optimal.

By moving the necessary storage and processing power closer to players, edge computing greatly improves the experience with ultra-low latency delivery and reduced lag time. Edge distribution also enhances the overall fairness of play. Players with low latency connections hold clear advantages over those with high latency—making the difference in the game between win or lose or life or death (virtually speaking). Edge computing enables a more equal footing for all players, regardless of geography. The required response times place online gaming edge compute in the near edge zone (e.g., metro-level deployments), with an emphasis on colocation with communications connectivity.

As a final benefit, edge processing will enable high quality gaming with less processing burden on the user's hardware. Combined with the greater bandwidth of 5G, this will open up new opportunities in the segment for both hardware and software.

THE EDGE CLOUD ECOSYSTEM

Successful edge cloud deployments require an ecosystem. While some industry participants may claim multiple roles, few (if any) are in a position to do it all—not even the hyperscale giants that pioneered cloud computing. As such, strategic partnerships will be key. This is a trend the industry is already witnessing, as evidenced by recent partnerships between hyperscalers and Tier 1 operators (detailed below).

Cloud infrastructure

Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Alibaba Cloud, and IBM Cloud are the leaders in the global cloud infrastructure market, which includes infrastructure as a service (IaaS), platform as a service (PaaS), and hosted private cloud services. This market reached nearly $100bn in 1H20, according to Omdia. In the centralized cloud era, these infrastructure providers relied heavily on colocation providers (see below) for physical space and interconnectivity, but communications service providers have not been particularly strategic for them historically.

The emergence of 5G and edge is changing this dynamic. The industry has witnessed significant partnership activities between hyperscalers and Tier 1 communications service providers over the last year:

- **AWS and Verizon** collaborated to provide 5G multi-access edge computing (MEC) in locations across the US.
- **Google Cloud partnered with Telefónica** to offer MEC in Spain.
- **Google Cloud and Orange** cooperated on cloud, edge computing, and cybersecurity services in Europe.
- **IBM Cloud partnered with AT&T** to enable the management of applications in hybrid cloud environments.

Services and applications

Software services and applications run on top of the underlying edge cloud infrastructure. Some cloud infrastructure providers are also major applications providers—including Google and Microsoft—but there are thousands of other applications providers large and small spanning the globe, across enterprise, small business, and consumer markets. These numbers will only increase with 5G and edge cloud.

While the supporting cloud ecosystem rises on the successes of cloud applications, the applications are equally dependent on the strength, innovation, and success of the rest of the ecosystem. For example, AR applications cannot work without ultra-low latency processing and connectivity. And the industry can debate the ultimate market potential for remote surgery, but without the infrastructure to support an ultra-reliable, ultra-low latency tactile internet, the debate is pointless.

Fixed and mobile network connectivity

A cloud is of no value without the connectivity between users and applications and between cloud facilities (i.e., data center interconnection). While cloud infrastructure is primarily global, telecommunications remains primarily regional, meaning that different countries and even regions within countries have dominant providers. As a result, interconnection among carriers has always been critical. The rise of cloud computing has added importance to interconnection within large colocation facilities with densely populated, broad ecosystems.

As noted, 5G and edge cloud add new dimensions to the ecosystem. This is evidenced by the budding interest from hyperscalers in working directly with telecom operators around MEC deployments.

Finally, from a facilities standpoint, edge computing drives an orders of magnitude increase in the number cloud locations (see **Figure 1**). Thus, fixed and mobile operators will need to expand their relationships with colocation providers, tower providers, and emerging entrants (such as micro data center companies) to extend their own edges to reach enterprise and consumer customers.

Facilities/data centers

The term "cloud" suggests a nebulous ether, but in reality, edge applications and services (even virtual ones) must reside in physical locations. Hyperscalers have built their own hyperscale data center facilities for their own centralized cloud, often placing these campuses in remote locations for operational savings. Additionally, a large content-neutral segment has emerged globally to provide secure space, power, cooling, and interconnectivity for multiple industry participants, including cloud providers (such as hyperscalers), communications service providers, and enterprises. Leading content-neutral data center companies include CoreSite, Equinix, Cologix, and CyrusOne. In addition to providing physical real estate, these companies play a critical role in providing a connectivity meeting place for the "who's who" of players. This crucial component is an equally important piece of their tenant value proposition.

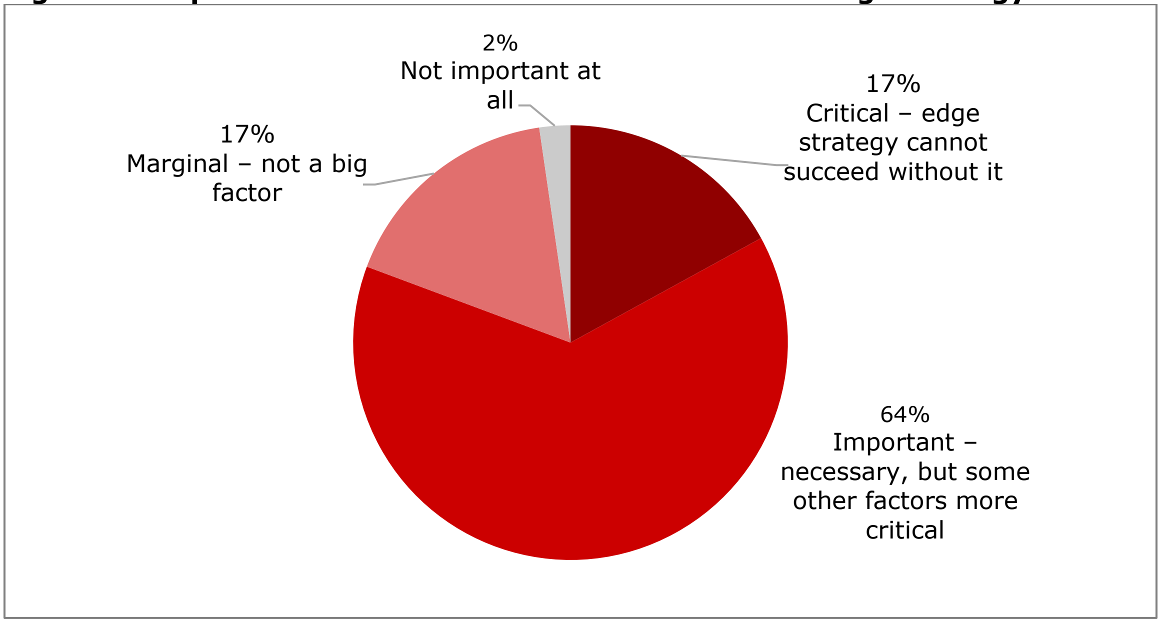
Tower companies that lease cell tower space for mobile network operators have also become important players as the industry moves to 5G and edge connectivity. Major tower companies such as Crown Castle International, SBA Communications, and American Tower see their real estate as value for edge processing and storage—particularly for telco edge cloud.

CRITICAL ROLE OF COLOCATION IN MEETING EDGE REQUIREMENTS

Colocation is a key component in today's cloud ecosystem for CSPs, cloud infrastructure providers, communications service providers, and enterprises alike. But how does the move from centralized cloud to edge computing change this dynamic?

In the 2019 Heavy Reading *Strategies for Connecting the Edge* survey, we asked operator and enterprise respondents to identify the importance of network-dense colocation for their edge cloud strategies. Eighty-one percent of respondents reported that colocation was at least important for their edge strategy, with 17% reporting that it is critical and that their edge strategy cannot succeed without it (see **Figure 4**; note that operators and enterprises answered similarly).

Figure 4: Importance of network-dense colocation for edge strategy



n=88

Source: Strategies for Connecting the Edge: 2019 Heavy Reading Survey

Expected primary benefits from colocation do differ somewhat for telecom operators and for enterprises. For operators surveyed, the primary colocation benefit is the ability to access the edge where it is needed, followed by the ability to connect to major network service providers and the ability to connect to major content/cloud providers (**Figure 5**).

For enterprises, however, the primary benefit is the time-to-market advantage compared to building their own edge locations. The ability to connect to major network service providers is also important to enterprises (ranked second), followed by the ability to access the edge where needed (ranked third). Connecting to cloud providers is a bit less important for enterprises surveyed, compared to telecom operators.

In general, choosing an edge location is not simply a question of "where." Equally significant is the question of "who's there."

Figure 5: Primary benefits of colocation for edge strategy

Item	Telecom operator rank	Enterprise rank
Ability to access the edge where it is needed	1	3
Ability to connect to major network service providers	2	2
Ability to connect to major content/cloud providers	3	4
Time to market vs. build our own edge	4	1
Cost savings vs. build our own edge	5	5

n=55 telecom, 22 enterprise

Source: Strategies for Connecting the Edge: 2019 Heavy Reading Survey

CONCLUSION

Edge cloud computing and 5G are two of the biggest trends in networks and will continue to be over the next decade. They are intricately intertwined. In many cases, 5G and edge cloud computing will be combined to deliver advanced applications for enterprises and consumers, particularly those that require the URLLC functionalities that are part of the advanced 5G services.

Successful edge cloud deployments require a comprehensive ecosystem. While some participants may claim multiple roles, broad and deep partnerships are essential. At a high level, key players will include the following:

- **Cloud infrastructure** companies, including hyperscalers AWS, Microsoft Azure, Google Cloud, Alibaba Cloud, and others.
- **Services and application** companies that will build innovative content for businesses and consumers enabled by 5G and the edge cloud infrastructure.
- **Fixed and mobile network connectivity** providers to deliver 5G services as well as transport capacity for 5G, as a cloud is of no value without connectivity.
- **Facilities/data centers** to physically house the applications and infrastructure that make up the edge cloud.

Heavy Reading research indicates a crucial role for colocation providers that not only provide physical proximity to users, but also serve as the secure gathering place for all the players in the edge cloud ecosystem.