Converge! NETWORK DIGEST | AvidThink®

2020 Edge and Beyond Report

RESEARCH BRIEF

Table of Contents

Research Briefs are independent content created by analysts working for AvidThink LLC. These reports are made possible through the sponsorship of our commercial supporters. Sponsors do not have any editorial control over the report content, and the views represented herein are solely those of AvidThink LLC. For more information about report sponsorships, please reach out to us at **research@avidthink.com**.

About AvidThink™

AvidThink is a research and analysis firm focused on providing cutting edge insights into the latest in infrastructure technologies. Formerly SDxCentral's research group, AvidThink launched as an independent company in October 2018. Over the last five years, over 110,000 copies of AvidThink's research reports (under the SDxCentral brand) have been downloaded by 40,000 technology buyers and industry thought leaders. AvidThink's expertise covers Edge and IoT, SD-WAN, cloud and containers, SDN, NFV, hyper-convergence and infrastructure applications for AI/ML and security. Visit AvidThink at **www.avidthink.com**.

DELIVER MORE SERVICES FROM THE EDGE TODAY.

A unified edge solution enabling Communications Service Providers to simplify 5G, Public/Private Multi-Access Edge Computing and Telco cloud networks. Accelerating the time to market new services while reducing the costs of edge infrastructure by a factor of 10x.

www.kaloom.com

2020 Edge and Beyond Report Part I: Overview and Analysis

Introduction

The edge is a hot area of research and investment today. It is an overhyped term, with some analysts declaring that the edge will be bigger than the cloud market and that a majority of cloud workloads will migrate to the edge. The estimates of the edge market size vary wildly, going up to tens of billions by 2025, dependent on what is included in that estimate.

AvidThink's view on this, having discussed 5G and edge with many industry experts and end-users, is that edge computing is a natural evolution of cloud computing. However, there are irrational expectations for the edge and misconceptions about its importance and what it will enable and what it won't.

In this first part of our edge report (Part I), we'll provide a quick overview of the edge, followed by our analysis. The latter half of the report (Part II) will take a broader look across the edge market. Readers already familiar with edge computing, especially the components and players that make up the ecosystem, may choose to skim the latter half of this report.

This report will not attempt to rehash work done by other analysts and industry groups. There's plenty of excellent material out there, not the least of which is the **State of the Edge (SOTE) report** from the Linux Foundation. Our analysis will instead provide our perspectives of the reality on the ground, which has been informed by recent and ongoing interactions with cloud providers, carriers, MSOs, application vendors, and enterprises. For instance, we recently interviewed dozens of global industry insiders as part of a project on 5G and storage.

In addition to our perspectives, we will provide recommendations on how cloud providers, enterprises, carriers, and other ecosystem members should position themselves to take advantage of the cloud's next evolution. As always, we welcome your feedback at **research@avidthink.com**!

Data and Application Trends Driven by Digital Transformation

Data has been called the new oil: the world's most valuable resource and the fuel for new industries, innovations, and services. The 2020 pandemic, and the associated shifts by both businesses and consumers from physical to online activities, has accentuated the value of data-driven companies. Alphabet, Amazon, Apple, Facebook, and Microsoft's financial performance with their vast troves of data about customers, devices, and applications has outpaced that of the average enterprise. They collectively **represent one of the most significant concentrations of wealth and capital (22 percent) in the history of the S&P 500**. However, like oil, data is a raw material useless until processed, and enterprises now face such a gusher of data that processing using only centralized data centers and cloud infrastructure is impracticable.

On the demand side, consumption of data and applications in both enterprise and consumer markets is accelerating. Because enterprise and consumer applications need to process more data from an increasing number of sources and render immediate, useful results, they are expected to do so with decreasing latency. Likewise, applications are pushed to provide access to data regardless of where users are — in the office, at home, at a coffee shop, on-the-go in airports, and on the road. The desire for real-time applications that respond instantly to users is stronger than ever, all in service of improved user experience and greater productivity. To serve today's and tomorrow's users, no matter where they are, will require a shift in computing architecture.

There and Back Again

Ironically, the hyperscale cloud platforms that have accelerated digital transformation across all industries by pushing centralized utility computing architectures will soon see a re-architecture wrought by this same digital transformation. In a partial reversal of the trend to migrate all computing and data into centralized clouds, today's data-centric enterprises, cloud operators, and carriers are moving to distributed designs with local processing at or near the data source. AvidThink agrees that, with the rise of IoT and digital transformation, enterprises will generate the vast majority of data at or near the edge. Whether processing happens at the edge or in centralized clouds depends on the cost of processing versus data transport and storage cost, as well as data jurisdiction and security considerations.

Edge computing not only minimizes the cost and latency of backhauling data to a central facility but facilitates rapid scaling of capacity and near real-time analysis of results. This distributed architecture is an effective way of filtering and summarizing raw data before aggregation to a central repository for further analysis and long-term preservation.

The embrace of edge infrastructure is merely the latest in the ongoing cyclical nature of IT architectural norms. The industry has been through several periods of infrastructure consolidation and disaggregation, most recently in the 1980s and 90s as minicomputers, Unix workstations, and PCs displaced mainframes for many enterprise tasks. As is in these prior cycles, the trick in today's evolution of edge infrastructure is finding the right balance between consolidation and distribution.

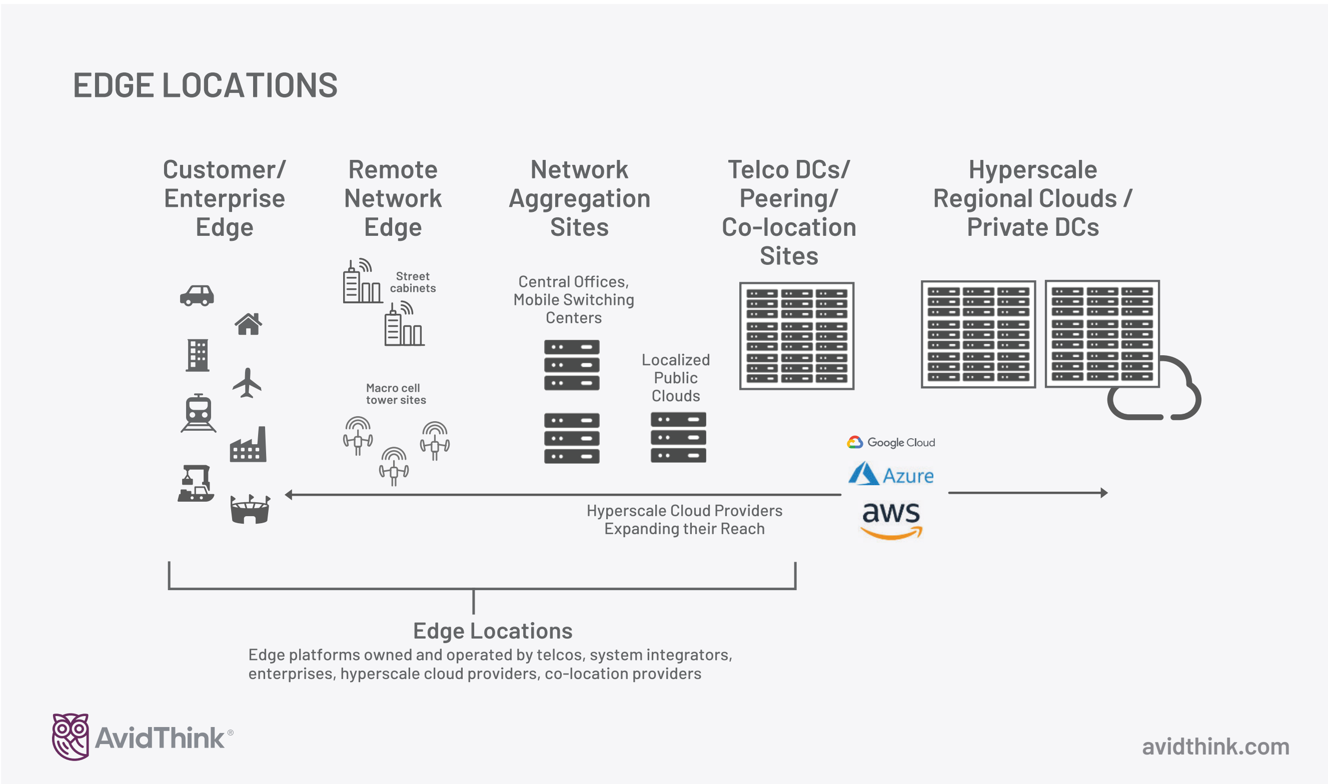
One significant and salubrious difference between then and now is that, unlike the first era of distributed computing sparked by PCs and workstations, today's edge infrastructure doesn't displace centralized systems, but coexists with them. Instead, enterprises and carriers stitch edge infrastructure, private data centers, carrier central offices, and public cloud infrastructure into a coherent, centrally managed, and governed application fabric with data and workloads freely moving between resource nodes as needed.

As we describe the factors underlying edge infrastructure and its technology, categories, and advantages, keep in mind this overarching vision of cloud-edge coexistence and cooperation.

As we describe the factors underlying edge infrastructure and its technology, categories, and advantages, keep in mind this overarching vision of cloud-edge coexistence and cooperation.

What is the Edge?

Edge infrastructure is an umbrella term that encompasses a broad range of implementations and locations. Which locations we call the **edge** depends on one's vantage point — after all, one person's edge is another person's core location. Our conversations with different stakeholders in this ecosystem has surfaced a multitude of edge locations, which we depict in the diagram below:



The consensus is that anything to the left of hyperscale regional clouds, including more localized data centers, or even co-location sites, down to on-premises infrastructure (e.g., enterprise, home, venue, port, oil rig, cruise ship, airplane) is viewed as the edge. This is independent of who owns or operates the infrastructure: hyperscale cloud providers, colocation providers, mobile network operators (MNO), wireline network operators, multi-service operators (MSOs), cloud hosting companies, enterprises, system integrators, and other entities. Most analysts including us do not consider end-user devices as part of the edge. Therefore, laptops, desktops, tablets, etc. are not included. However, automobiles can also be thought of as part of the edge, as increasingly vehicles look like micro-datacenters on wheels.

In AvidThink's discussions with participants in the edge economy, there are two general areas of the edge that are seeing immediate interest:

- **The telco/cloud edge:** the last mile or near last-mile edge close to the users (enterprises, consumers). This includes the carrier edge at central offices (COs) or mobile switching centers (MSCs), and upstream aggregation sites that serve multiple COs or MSCs. This includes public cloud presence in these carrier edges, such as Amazon's Wavelength offering co-located at carrier MSCs. The cloud edge includes public cloud local installations, such as Amazon Web Services' (AWS) Local Zones, meant to be located in the same geographic region (e.g., city-level as customers). Google Cloud Platform (GCP) has similar installations, given their history with placing content caches for YouTube within or near carrier environments. Microsoft Azure is embarking on similar initiatives. Both cloud providers and communication service providers are working hard to figure out how to utilize computing and storage assets in these locations to improve enterprise and consumer application performance while exploring viable business models.

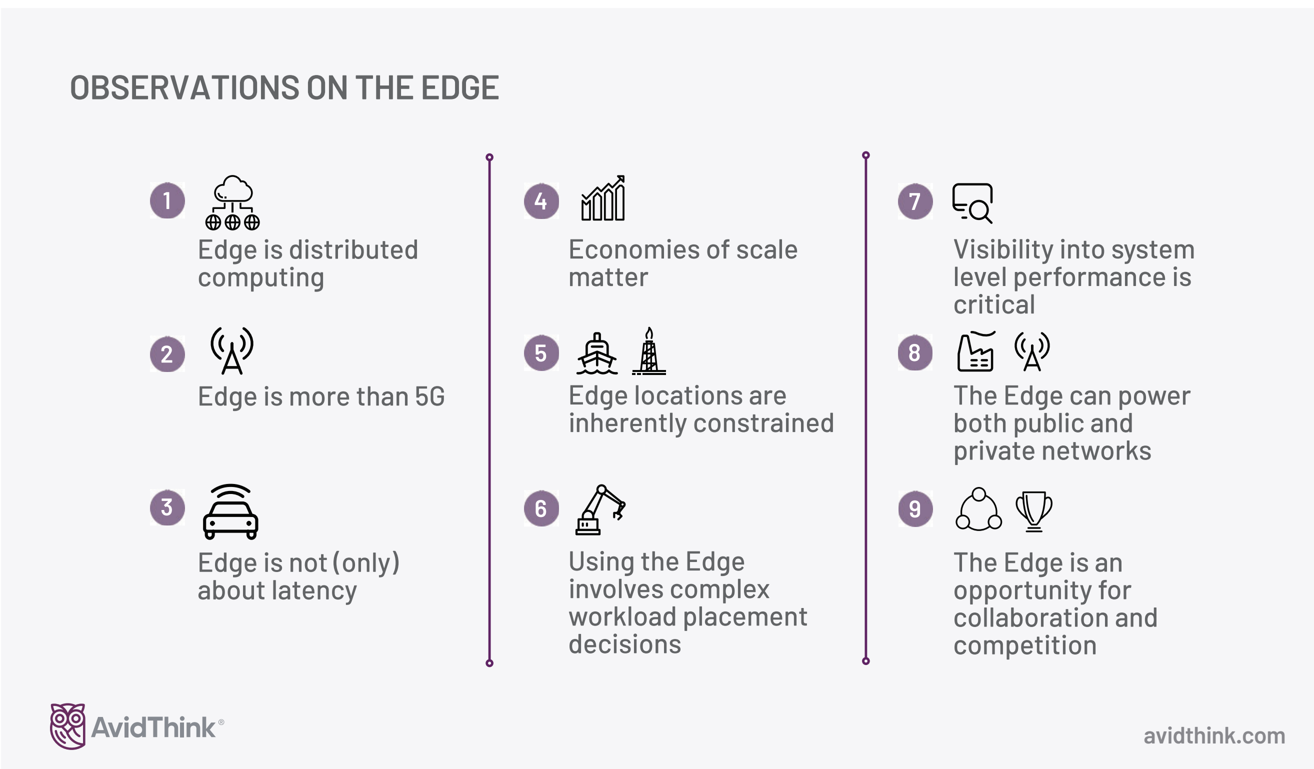
- **The on-premises edge (on-prem edge):** refers to infrastructure equipment located within a customer site — enterprises and consumers — but managed and operated by a third-party (cloud provider, carrier, system integrator, or other managed service provider). This includes racks or half-racks of computing and storage, customer premises equipment (CPE) in the form of universal CPE (uCPE) for SD-WAN deployments, or remote office/branch office (ROBO) or industrial manufacturing or warehouse servers.

The other edge locations, such as cell-site or street-level cabinets, will be relevant in the future. However, carriers are still actively seeking enterprise-centric use cases beyond carrier virtualized and open RAN. These locations, sometimes call the **far edge** or the **radio edge**, will be open to future innovations of the radio access network (RAN). In the meantime, carriers who have run early proofs-of-concept in the far edge have realized there will be a limited return on investment (ROI) on building out and operating infrastructure in these remote locations for now.

Our views align with the Linux Foundation's State of the Edge (LF SOTE) project's perspective on what they term as the "infrastructure edge" and the "user/device edge." LF SOTE indicates that the vast majority of edge-related activities will center around these two edge locations. For simplicity, we'll term our telco/cloud edge as the cloud edge and the on-premises edge as the on-prem edge to remove ambiguity, and call the potential edge infrastructure around the RAN as radio edge.

Nine Observations on the Edge

Now that we've established the basics, we'll dive into AvidThink's observations and analysis of the edge computing market. We've grouped our thoughts under nine headings below, resisting the urge to make up one more to hit that perfect ten. We hope you find them useful.



One: Edge is Distributed Computing

In the early days of edge computing, there were attempts at strictly delineating different aspects of the edge. Fog computing, mobile edge computing, followed by multi-access edge computing (MEC), distributed cloud were all terms from major vendors and standards bodies, including Cisco, ETSI/IEEE, Ericsson, and multiple consortiums (OpenFOG, etc.). Efforts at standardizing the edge and frameworks for building applications at the edge continue today; for example, the GSMA's recent Operator Platform project. Also, Akraino from Linux Foundation, StarlingX from OpenStack, and commercial alternatives from MobiledgeX and Volterra.

Solutions to manage and orchestrate edge platforms, many based on Kubernetes, abound: from Red Hat's OpenShift, VMware's Telco Cloud Platform, Google's Anthos, and Mutable.io's unique AirBNB approach. Mutable.io, a startup, intends to round-up idle servers across carriers, MSOs, and hosting companies into a federated edge cloud.

Ultimately, the edge, in combination with the central clouds, represents a distributed computing platform. We expect a large proportion of edge applications, especially those originating in the enterprise, to be extensions of central-cloud-hosted applications. AvidThink's view is that aside from carrier RAN applications (virtualized RAN, cloud RAN, open RAN), a large proportion of initial edge workloads will be satellite components that collaborate with core cloud applications. Pure edge application frameworks designed primarily for deployment at the carrier edge will take a while before they find traction. Separately, the underlying orchestration and bare metal management platforms remain in play.

Two: Edge is More Than 5G

The edge is more than wireless. Recent pronouncements from carriers, sometimes in combination with cloud providers, would have us believe that the edge is all about 5G and wireless. We're bombarded by messages touting multi-Gbps bandwidth, ultra-low-latency sub-1 ms networks capable of unlocking and unleashing new use cases like remote surgery, autonomous vehicles, and industrial robots powered by 5G and the edge. The fact is that the edge was there before 5G, the edge added value even on 4G networks, and many business-focused edge use cases will start with wireline networks. In other words, the edge is not dependent on 5G.

> **The fact is that the edge was there before 5G, the edge added value even on 4G networks, and many business-focused edge use cases will start with wireline networks.**

Many of the coolest demos around 5G and the edge will likely remain just that for a while. Take for example, autonomous vehicles dependent on the edge for control and remote surgery over a 5G network? Both use cases belong in regulated industries and involve human lives, so until the network and edge are 100% ubiquitous and available, use cases like these will not be generally available.

Industrial automation enabled by the edge will first deploy using on-prem edges, but reducing OpEx through cloud-managed orchestration. We might even see near-premises cloud-edge locations hosting this industrial automation control software, but these sites will likely be connected via fiber links, not 5G fixed wireless access (FWA). Today, enterprises connected by fiber to cloud edge or carrier CO locations will already find reliable round-trip times of sub-5ms. They can potentially benefit from running their industrial automation and control systems in telco/cloud edge platforms hosted in these near-prem locations less than 50-100km away. Eventually, as 5G FWA becomes available, and as costs drop by at least an order of magnitude, enterprise links may see a transition to FWA.

Even though the edge isn't directly dependent on 5G, 5G is dependent on the edge to achieve its full promise of ultra-reliable low-latency networking. Without the ability to run user plane functions (UPF) that do the bulk of packet handling at the edge, sub-10ms and sub-5ms times will not be achievable, much less the holy grail of sub-1ms. Likewise, the innovation expected with a virtualized and disaggregated RAN will require the ability to run the RU (radio unit) and the DU (distributed unit) close to where the radio heads are. And it is likely in these locations (radio edge) where the RU and DU will be hosted. It will almost certainly be where the RAN intelligent controller (RIC) of open RAN initiatives will be placed. Depending on the use case, UPFs might be one of the first functions hosted at the radio edge or at least the telco/cloud edge.

Three: Edge is Not (Only) About Latency

There are use cases where the edge is critical in achieving the ultra-low latency that specific applications need. However, in our conversations with end-users, they value the other attributes of the edge as much as latency. In addition to low latency, the edge provides:

- **Improved isolation and security**, enabling regulatory or corporate compliance: workloads running at on-prem edges don't have to send data into the public cloud and can comply with any necessary data jurisdiction or compliance controls. These edge locations can also be used to obscure, transform, or encrypt the data prior to sending it upstream. Cloud storage can handle the data opaquely for archival and backup while the edge handles unencrypted data processing.
- **Improved jitter** (consistency in latency): the relative proximity of the edge to where data is being generated or consumed reduces the variance in latency (at least in the network transport component). The more network hops between where data is processed and data are generated, the greater the jitter. Having consistent latency can allow the software to effectively mitigate the latency without having to contend with large variations in delay. For some applications, like collaboration and communication tools, low jitter can be helpful.
- **Reduced uplink costs:** the more data processing that occurs closer to where the data is generated, the more opportunities for data reduction, which can result in reduced bandwidth consumption on uplinks.
- **Resilience against intermittent connectivity:** by having computing and storage on-prem, workloads can continue to run even if wide-area network links are down. For environments like cruise ships, airplanes, oil rigs, or vehicle-based services, an on-prem edge provides some level of autonomy and reliability in computing especially when coupled with the ability to synchronize data and state when connections are available.

In many use cases, latency matters much less than these other edge benefits, and enterprises should not walk around with blinders focusing on latency as the only important metric for considering the edge.

Four: Economies of Scale Matter

While using edge infrastructure can provide improved latency, isolation, and performance, the cost to install and operate edge infrastructure will almost always be higher than in centralized clouds.

One of the reasons that hyperscalers can offer competitive prices for on-demand infrastructure in their clouds is due to **economies of scale**. Hyperscale cloud providers have figured out how to run large complexes of computing, storage, and networking at scale. They leverage their market power to build out massive data centers and procure equipment and connectivity at much higher volumes and lower costs than individual enterprises. Hyperscalers invest in developing orchestration systems and software platforms that enable much lower operation expenses, allowing a smaller staff to design, deploy, and operate data centers at scale.

While a few of these innovations and savings will translate to cloud edge locations operated by hyperscalers, many will not. Edge locations are much smaller than centralized data centers, meaning that HVAC, power, networking, and other costs cannot be amortized over a large number of servers. Likewise, the decentralized nature of these edge locations increases the operational expense to populate, update, and service them. When replacing servers at the edge involves a truck roll or even a helicopter ride to an oil rig, the expense will undeniably be higher.

All this translates to a higher cost of computing and storage at the edge when compared to centralized clouds. The cost premium will vary, but the initial pricing analysis of AWS Wavelength (located in Verizon's mobile aggregation sites) shows a

20-30% premium for similar machine types compared to a nearby AWS region. This comparison is consistent with looking at **AWS Outposts (which sit on-prem) versus AWS region instances**. In that situation, we're looking at 20-50% premium based on utilization rates of AWS Outpost racks and cost of hosting the Outpost rack on-prem or near-prem.

When we cost out radio edge sites, the premium will likely be even higher. Those sites are much harder to reach, more distributed, and have constrained power budgets, limited space, and cooling. This implies expensive high-resiliency components.

Beyond the premium, running in a central cloud provides high-performance and cost-effective access to supporting cloud services, like AI engines, data analytics, managed databases, and more. Edge components can access the same services over the wide-area network but will incur network transport costs and increased latency and reduced throughput.

What AvidThink believes this will drive is a re-architecture of applications to have different component tiers. As many components as possible will run in lower-cost central clouds, and only applications or application components that absolutely need to be at the edge will be hosted at the edge. Nevertheless, all these components will be orchestrated through a centralized orchestration mechanism, which will reduce the enterprise complexity in managing distributed applications.

Five: Edge Locations Will Be Inherently Constrained

As we discussed, there's wider recognition now that many edge locations are not just smaller data centers with fewer racks of servers, but that space constraints might be more severe. In CO or MSC locations, there might be limited space, little cooling, and small power budgets. This could mean different types of servers: shorter-depth, smaller form factor. It could imply the need for lower-power CPU systems and reliance on hardware accelerators like SmartNICs with FPGAs and ASICs to perform high-speed network processing with a lower power budget. Or specialized AI chipsets and GPUs to perform fast inferencing on lower power budgets. On enterprise premises, such as in smaller branch locations, or in street-level cabinets, higher-reliability systems that can sustain occasional power outages, unconditioned power, more extreme temperature ranges will be needed. This means that edge management and orchestration systems will have to deal with a greater diversity of hardware and platform types than has historically been present within larger data centers.

With space, power, cooling at a premium, converged architectures (hyperconverged systems or HCI) might prove to be a better fit. Further, composable machine architectures could allow more efficient use of different hardware components. Other examples include a recent Linux Foundation demonstration of a virtual central office by IBM, Kaloom, Red Hat. The demo participants showed off a high-performance edge-based user-plane function (UPF) stack for 5G networks running in 3 rack-units. By leveraging a programmable switch fabric with software controllers running on a switching platform, the joint solution conserved precious rack space in edge locations while maintaining high network-throughput with low latency.

Six: Using the Edge involves Complex Workload Placement Decisions

To maximize the value of edge computing, existing applications that have been migrated into the cloud will need to be rearchitected. This is not a small task, and placement of workloads across multiple computing locations (the spectrum of sites from hyperscale clouds to on-prem edges) needs to take into account, at minimum, the following factors:

- Where the data is generated
- Where data should be processed
- Where processed data needs to be consumed (by end-users or other services)
- Where information is eventually stored
- Performance needs of the application in terms of response times, throughput
- Security and compliance
- Capabilities of the underlying infrastructure platform, especially if specialized hardware like GPUs or FPGAs are needed

- Cost of transportion of data
- Cost of transient and permanent storage of data
- Cost of computing and memory at each location
- Availability constraints for the application and reliability of underlying uptime SLAs for the infrastructure at each location
- Overall budgetary goals

Extending cloud applications to take into account edge locations will take a significant effort and will likely be a multi-phased project, focusing on addressing the most pressing needs first (e.g., moving components that provide low-latency response times out into the edge).

Seven: Visibility into System-Level Performance is Critical for Success

As part of the evaluation, users will want to make sure that they are utilizing the edge in a way that makes architectural and business sense. For instance, if a user-facing application takes 1-2 seconds to process specific inputs before responding to a user, reducing the network latency from 50ms or 100ms to 10ms makes little difference to the overall user experience.

However, if a sensor on a high-speed assembly line detects a fault in an industrial control situation, the enterprise customer would want the system to pause immediately before the fault impacts other items in the pipeline. In this case, placing the control logic, either on-prem or near-prem, would likely have significant benefits.

It is, therefore, essential to measure and understand where latencies accumulate in a system and understand which delays make sense to mitigate. Likewise, understanding the data processing pipeline, from where data is generated to where it is processed and stored, and the size of the data at each stage is essential.

In the use of edge systems, ongoing monitoring of performance metrics across all the workload locations will be necessary to ensure optimal use of the different resources spread throughout the network. Having an appropriate telemetry system that integrates with the overall orchestration system to ensure ongoing closed-loop monitoring will be even more critical with these distributed applications. As recognition of the importance of the edge, performance monitoring vendors like Accedian, Keysight, Spirent and Viavi have expanded their product lines to accommodate edge (in addition to 5G) infrastructures.

It is, therefore, essential to measure and understand where latencies accumulate in a system and understand which delays make sense to mitigate.

Eight: The Edge Can Power Both Public and Private Networks

When we discuss the edge, we often envision the public 5G network. However, one of the key value propositions for on-prem (and potentially near-prem) edge is as a host for running a private network core (both for 4G LTE and 5G). There's a significant market brewing that will see an uptake in enterprise use of mobile wireless technologies, as an adjunct for WiFi networks, and a replacement for existing wired ethernet networks (especially industrial ethernet). 4G LTE and subsequently, 5G can provide many of the capabilities that WiFi today cannot, and approach performance levels generally associated with wired networks.

For large event venues, factories, warehouses, airports, shipping ports, and even universities, having an edge-powered private network can bring significant benefits and provide better ROI than further investment in wired networks or in expanding today's WiFi installations. We won't detail the opportunity for private enterprise networks based on mobile technologies in this report, but that will be a topic for the future.

Nine: The Edge is an Opportunity for Collaboration and Competition

Across the entire spectrum that is the edge, there are numerous parties that have shown interest in carving a out piece of the edge market for themselves. These include the carriers (fixed and mobile) who own key edge locations near enterprise customers and consumers. Of course, the hyperscale cloud players are looking to expand out of their centralized data centers. But in addition, multi-service operators (MSOs) in the form of cable companies, system integrators, enterprises themselves, and even tower companies have an interest in this space. There are also pure-play solutions targeting the edge, like Packet (recently acquired by Equinix), Vapor.io, and StackPath (partnered with Juniper Networks).

The edge market is not going to be a winner-take-all market, given the diverse nature of the ecosystem. We've summarized the some of the key strengths and weaknesses, as well as the opportunity for each of the different classes of players in the table below.

	Carriers	Cloud Providers	Colo/Inter-exchange Companies/Tower Companies	System integrators (SI)/ Network Equipment Providers (NEP)/ Property owners
Strength	• Control the last mile of network and localized real-estate that can host edge compute. • Have direct relationships with enterprises and already sell them connectivity.	• Control the public clouds that host IaaS/ PaaS/SaaS workloads. • Own the application stacks and developer ecosystem. • Cost efficiency via massive scale and high automation. • Virtually unlimited access to capital and strong cash flow. • Agile and fast to react to customer demands.	• Topologically advantaged on the Internet — close to or at the last mile and interconnect with the major carriers and cloud services. • Viewed as a potentially neutral ground that facilitates tie-ups with both cloud and carriers, or as a meeting-room between multiple carriers and clouds.	• SIs have business relationships, know-how, and capacity to build edge stacks on-premises • NEPs design and develop the infrastructure and network stacks that can host edge applications (relevant for ultra-reliable low-latency or tightly regulated/data sovereignty situations). • Property owners control and rent real-estate to multiple enterprises and can credibly provide a multi-tenant edge.
Weakness	• Limited experience running cloud environments at scale and hosting third-party applications. • No demonstrated success with cultivating a developer ecosystem. • Many culturally stuck in the legacy telco world.	• Do not currently control last-mile assets. • Have limited experience with the underlying mobile network (or wireline network) that delivers connectivity to the enterprise.	• Limited experience in running cloud software infrastructure (know how to manage physical DC at scale — HVAC, power, monitoring, etc.) • No direct relationship with the developer ecosystem.	• Edge offering will only be on campus and likely managed individually. Leads to location lock-in with a limited scale. • Limited access to the developer ecosystem.

	Carriers	Cloud Providers	Colo/Inter-exchange Companies/Tower Companies	System integrators (SI)/ Network Equipment Providers (NEP)/ Property owners
Opportunities	• Have to build out and host edge for 5G roll-outs for low-latency UPFs in any case, can leverage infrastructure build-out. • Virtualized RAN is an opportunistic use case that can be extended to third-parties.	• Partner with carriers to combine cloud platform expertise with last-mile assets, use the experience as a guide to potential acquisitions (or to directly compete in the future). • Foster development ecosystem to figure out novel applications for labs. Participate with carriers in incubation projects.	• Tower providers own strategic locations and can explore partnership with carriers to build cell-site micro-DCs. • IXC/Colo can partner with carriers to offer joint cloud-agnostic fiber links from COs/ MSOs into local clouds to unlock low-latency for cloud apps. • IXC/Colo can partner with hyperscale clouds directly and reduce dependence on carriers.	• Partner with cloud providers to leverage their cloud technology for hybrid offerings, particularly automation, orchestration, and app stacks. • Partner with carriers to provide connectivity and offload options for local apps and data.
Threats	• Deep-pocketed hyperscalers don't need carriers to build hyper-local zones. • Provider-neutral edge co-los working with tower companies can provide alternative wireless edge infrastructure without assistance from carriers. • Hyperscalers build out native capabilities and mobile functionality to reduce carrier's leverage with them (e.g. acquiring telco talent and telco assets like Azure's acquisition of Affirmed and Metaswitch).	• Resistance from enterprises to private edge infrastructure that locks into a particular provider. • Carriers or other parties creating multi-cloud options by building abstraction layers on top of hyperscale clouds to reduce lock-in.	• Carriers, hyperscalers partnerships result in alternative edge locations close enough to wireless-wired connection points that don't require being at the tower. • Carrier and hyperscaler partnerships succeed, reducing the need for a more neutral third-party that this category represents.	• NEPs who move too far into the on-prem edge, competing with telcos, incur the wrath of carriers and end up impacting their carrier sales. • Hyperscale edge platforms are able to directly meet the vast majority of enterprise needs without needing SI assistance.

The above are some of the more prominent talking points that have come up in our conversations with ecosystem players. There are other considerations and observations which we would be happy to share in more details conversations. Note that the edge is evolving rapidly and the insights above will likely evolve as the market shifts.

In any case, given the nature of the market structure, we expect to see more collaboration from the carriers and the cloud providers. AvidThink's research indicates that in markets with few major players — e.g., North America, each of the major carriers will be forced to partner with the three major hyperscale cloud players (AWS, Azure, GCP), and likely the next tier down (IBM, Oracle) as well. The pressure on the cloud providers is similar — to ensure that an application can tout low-latency for all mobile subscribers or even wireline subscribers, the cloud providers will need to give application developers platforms situated near or within the carriers that service the majority of subscribers.

In markets that are more fragmented, there may be room for aggregators (which could be one of the incumbents) to play a role in providing federated access to various subscriber pools, and vice versa, access to multiple hyperscale clouds.

Getting Ready for the Edge

Despite our best effort at providing structure and order to edge infrastructure trends and ecosystems, as **we recently wrote in an article on Fierce Telecom**, the business and monetization models for edge infrastructure and services remain unclear. The number of companies pursuing edge strategies spread across several markets that don't traditionally compete makes predicting the market's evolution a multidimensional problem that defies facile conclusions.

The future of edge infrastructure, in part, depends on the choices and preferences of both enterprise and consumer software developers since it is their products that will drive demand for edge services. Cloud providers have a decided advantage on this front since they provide the development platforms and services critical to next-generation applications. Furthermore, no companies have proven to be more innovative and agile than these hyperscalers. The carriers, NEPs, and other edge competitors will have great difficulty keeping up with the stream of new products and services being developed by their cloud competitors.

However, real estate is critical to edge infrastructure, and here, it's all about location, location, location. When it comes to positioning edge infrastructure, the advantage shifts to the carriers, IXCs, and tower companies. REITs, co-lo, and tower companies have assets critical to cloud providers, and if they came together, could counter the inherent topological advantages of the carriers. AvidThink expects lots of partnerships, and trial balloons as each group in the edge ecosystem seek to exploit its strengths and bolster its weaknesses through joint ventures without ceding too much to their 'frenemies.'

In the next few sections, with this highly dynamic playing field as a backdrop, we'll offer our recommendations to each of the major groups of stakeholders based on our current market observations.

Recommendation for Enterprises

AvidThink recommends that enterprises interested in the edge take the following steps to determine if the edge can add value to their infrastructure strategy.

- Identify enterprise applications that might benefit from one or more of the attributes of running at the edge (as we've covered earlier in the report): latency, security, and privacy, bandwidth savings, etc.
- Determine if there are any regulatory and compliance requirements that impact the application and data being considered for the edge.
- Catalog and understand the nature of the application and its components — it's vital to figure out if the application can be architected to run at the edge, and understand the magnitude and value of potential performance improvements.
- As part of this process, figure out where data is generated, where it can be processed, as well as where and when it is consumed. Also, determine where the data may stored - both short-term and long-term.
- Understand the underlying topology of locations involved in data generation, transport, and consumption and map out appropriate edge locations.

- Investigate edge platforms available to the enterprise, regardless of who provides them (carriers, cloud providers, equipment providers, system integrators) and determine the best partners in the enterprise's edge strategy.
- Understand the underlying edge software platforms and software services the edge platforms provide.
- Look at the re-architecture efforts for the application, figure out if there are enterprise developer resources available, and understand the costs (and how long it will take).
- Figure out the costs for orchestrating the application across multiple computing locations and various networks, and figure out troubleshooting and telemetry strategies.
- Do a comprehensive tally of the costs involved in running parts of the application at the edge, calculate the ROI, and determine if the overall performance improvements are worthwhile.

For the other members of the edge ecosystem, their success is determined by their ability to best meet the enterprises' needs or by leveraging their edge presence to offer innovative solutions to consumers directly. Regardless, it is imperative for these other ecosystem members to understand how they can reduce the enterprise's friction in moving to the edge to help enterprises maximize their ROI.

We'll now share recommendations pertaining to each category in the ecosystem.

Recommendations for Telecommunications Providers

For the telecommunications providers (including MSOs), they need to determine their role in the edge and cloud value chain. Should they simple provide network connectivity and secure access or expand beyond the typical role of a carrier? Regardless, we stand by our recommendation above to understand their enterprise customers' journey to the edge and reduce friction in that journey.

In any case, because of the carriers' lack of relationship with enterprise application owners and developers, they will need to partner with cloud providers and parts of the ecosystem. Here are some recommended steps that carriers should take:

- Understand the role that cloud providers in the market will play and which ones can be strategic partners. In markets with few competitors, AvidThink anticipates partnerships between all pairs of major carriers and cloud providers (as we've explained above).
- Figure out the value that carriers brings into the partnership equation. Whether unique and trusted relationships with their customers, valuable real estate locations close to consumers and businesses (owned or leased from tower companies, etc.), content or platform capability or just superior network QoS, telcos needs to leverage their assets to ensure a fair relationship with cloud providers and especially the hyperscalers.
- Consider the relationship with other cloud providers (hosting providers), colocation companies, and next-generation edge providers as alternatives or adjunct to their hyperscale cloud relationships.
- Evaluate if any of a carrier's assets can be leveraged via the wholesale platforms, especially in situations where they can offer unique services with cost advantages but lack the expertise or wherewithal to directly serve end-customer segments.
- Develop in-house capabilities to understand cloud platforms better and foster a culture around modern application development. Whether a carrier decides to provide their cloud platform (directly or with vendor assistance) or partner with hyperscale clouds to offer edge-related services, they will need to enhance their DNA to become conversant with cloud and modern application infrastructure.
- Determine if system integrator relationships can help accelerate their go-to-market with enterprises and if so, recruit and foster those relationships.

Recommendations for Cloud Providers

Cloud providers, particularly hyperscale cloud providers, have a significant amount of assets when it comes to the application infrastructure for the edge. They will likely have orchestration and automation capabilities beyond those of the other players in the ecosystem. Often, the missing element is the last-mile connectivity to the enterprises and end-consumers. In some use cases, they might also lack familiarity with on-premises connectivity technologies, especially when it comes to mobile-technologies (private LTE, private 5G). Many of the hyperscalers have already acquired assets to fill this hole (Microsoft acquiring Affirmed Networks and Metaswitch, for instance), and have been recruiting carrier executives and experts. For cloud providers, considering some of the following steps might prove advantageous in their go-to-market efforts for the edge:

- Ensure a complete portfolio of offerings, expanding beyond regional clouds to have local cloud zones, on-prem managed cloud options, and continue to partner with carriers (both wireless and wireline providers) to offer carrier-based cloud zones (or low-latency cross-connects into hyperscale local cloud locations). Enterprises will likely want choices as they rearchitect their applications to be edge-enabled.
- Determine what relationship the cloud provider wants with the carriers in the region and how to structure win-win partnerships for edge offerings. Make credible commitments to reassure carriers that the partnership with them will be stable in the mid- to long-term. Carriers are concerned about hyperscalers extracting value (as happened with the over-the-top situation with both 4G LTE and broadband expansion) at the carriers' expense, and being relegated to commodity connectivity provider.
- Work with the other partners in the ecosystem to foster innovation and ideas around edge deployment — the market is still immature. Many enterprises are unclear as to how and what to use the edge for. Likewise, rearchitecting applications or writing new applications to take advantage of edge locations will not be a trivial task, and experimentation and the development of best practices will be crucial to the joint success.

Recommendations for the Rest of the Ecosystem

For the other players in the ecosystem, they will want to focus on the enterprises (and consumers) and determine how they bring value to the process — how each player can reduce the time and effort spent by enterprises in **edgification**, just as many of them provide enterprises with assistance with **cloudification**.

In the end, whether and how to expand to the edge should be a business decision, and it's critical to cut through the current surrounding hype with the edge (and 5G) and focus on making a sound, not fashionable, decision.

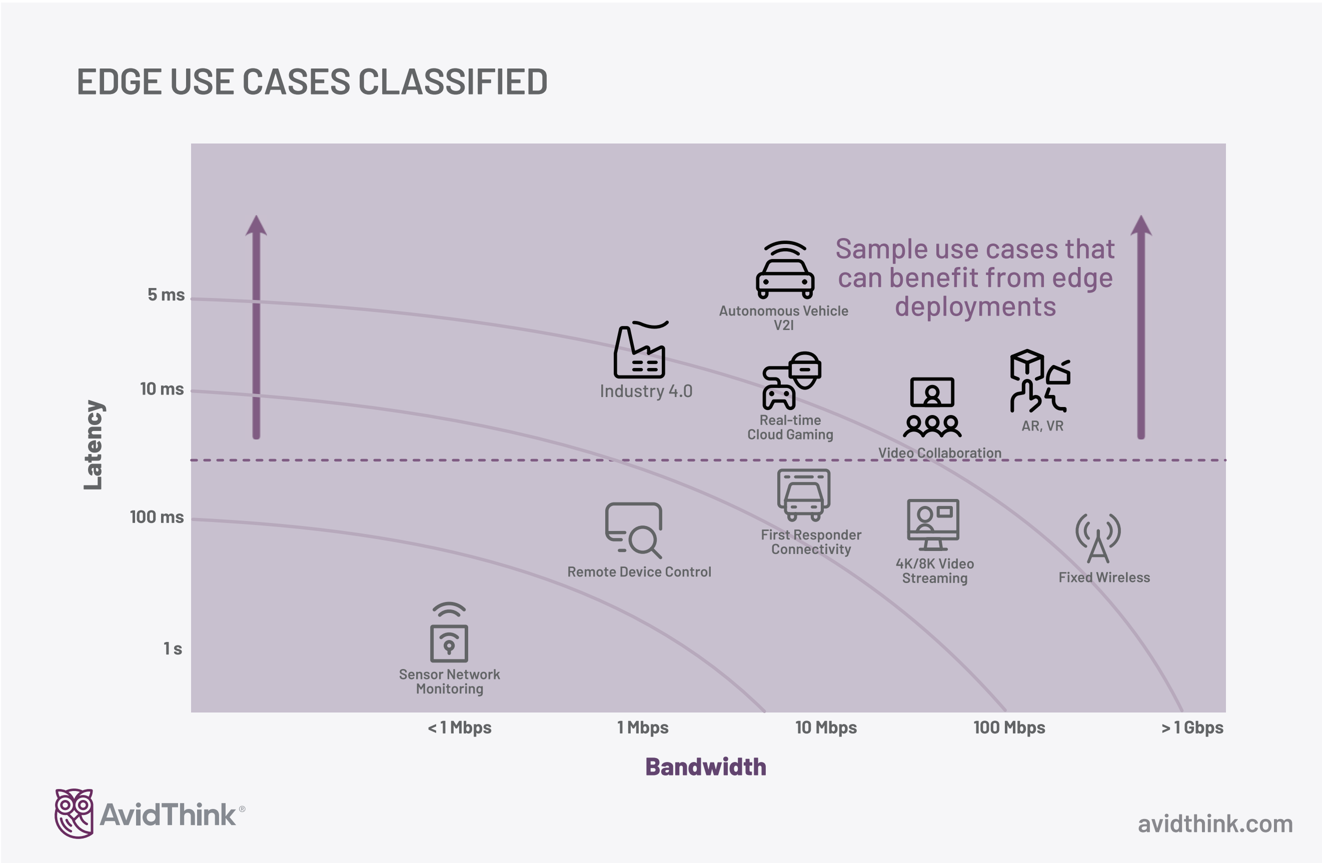
2020 Edge and Beyond Report Part II: The Expanded View

In this section, we will take a quick stroll through the expanse that is the edge. The area has grown so much that there's no easy way to cover all the elements that make up the edge. However, for our readers who might be coming up to speed on the topic, we provide additional discussion and material so that you can appreciate the diversity of topics and components that now come under the edge umbrella.

Typical Edge Use Cases and Drivers

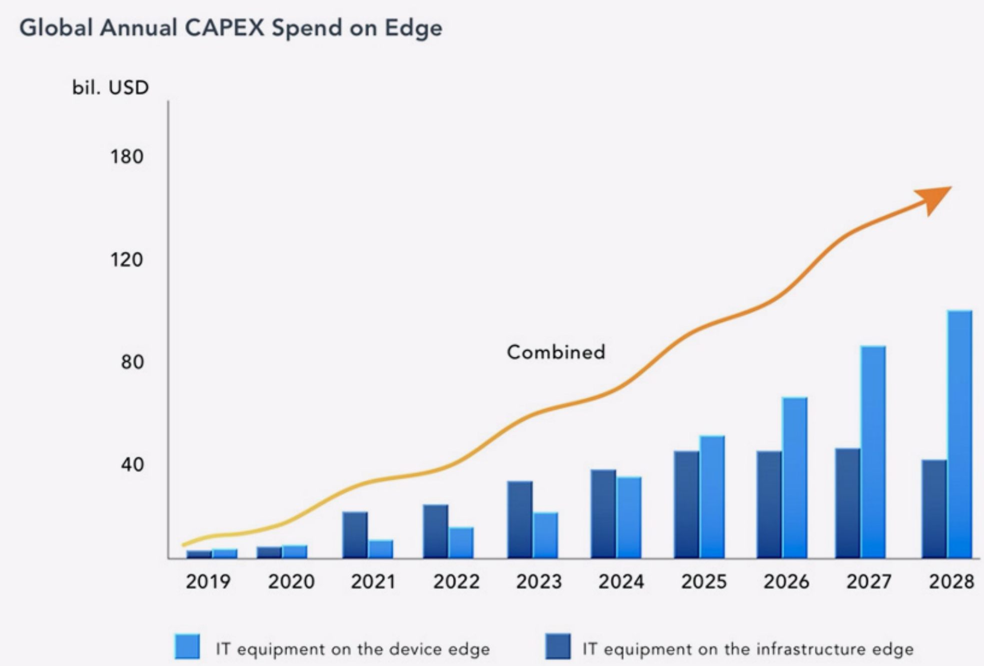
There are many business needs, spanning industries and situations, fueling the growth in edge infrastructure. The primary factors underlying a surge in edge infrastructure are:

- **Hybrid cloud architecture** and the need to balance control, data privacy, and performance in enterprise cloud infrastructure designs and where edge infrastructure provides a hybrid deployment option tuned for low-latency applications and Industry 4.0 manufacturing and logistical processes.
- **Low latency application and network services** including CDNs, real-time data analysis, deep learning applications, online gaming, and video streaming. The intense competition among online services, mostly operated by mega-cap technology companies, puts pressure to differentiate based on performance and reliability, both of which can be addressed by scale-out edge infrastructure.
- **IoT data analysis** where edge infrastructure enables enterprises to use data from intelligent sensors, retail POS terminals, location tracking systems, and other IoT devices to gain competitive advantage by analyzing and acting upon data at its source and in real-time. Analyzing data in situ not only provides near real-time information but reduces costs for the storage and transport of raw data to cloud infrastructure. Furthermore, significant improvements in the performance and capabilities of low-power embedded processors, many of which include GPUs or other neural network acceleration hardware, enables edge systems to run the type of advanced analysis and AI software that once required data center servers. Edge computing and IoT are also the basis for **smart city infrastructure**, public and private **video surveillance systems**, and emerging **V2X communication and AV systems**.
- **Network carrier service expansion** via **5G infrastructure and software-defined network services** using an NFV, VNF/CNF service architecture. Most carriers spent the last decade migrating to soft-defined infrastructure and virtual network services (NFV and VNF/CNF) in preparation for pervasive digitization of enterprise processes, an explosion in enterprise and consumer data, and network usage, and the hefty technical and capacity requirements of 5G infrastructure. Aside from being software-defined, i.e., separating network control from data flow, these next-generation networks are much more distributed, employing edge infrastructure to deliver content to 5G customers, network services to businesses, and OTT video streaming and conferencing to households that have become home offices, classrooms and movie theaters in the wake of coronavirus closures.

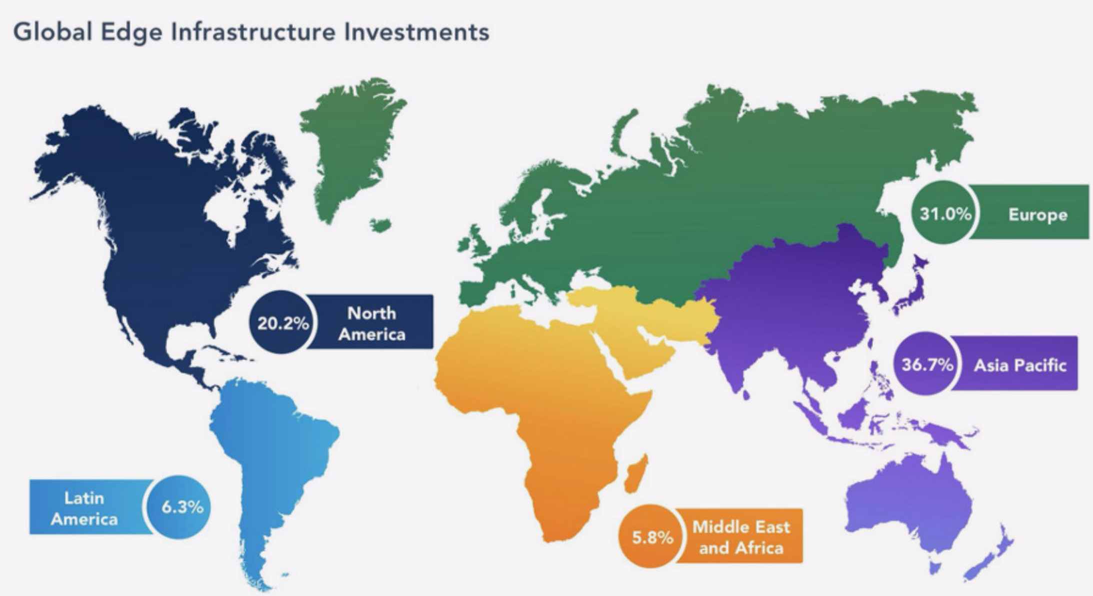


Edge Market Size

As we've indicated in the first part of the report, the edge market size varies wildly across analyst estimates. There are several ways of analyzing the size and scope of the edge infrastructure market. The two picked by the LF SOTE project are capital spending and the total power consumption of deployed edge infrastructure. Its **2020 Report** estimates at least $700B in aggregate capital spending on edge IT and data center facilities over the next decade, growing at a 35 percent CAGR nearly 10-fold from about $15 billion in 2020 to 2028. The Asia-Pacific and European regions will account for about two-thirds of the spending, with North America adding another 20 percent. Over the next few years, most of the investment will be on infrastructure edge equipment, spending on the device edge side ramping up later to account for about two-thirds of the total CAPEX by the end of the decade.



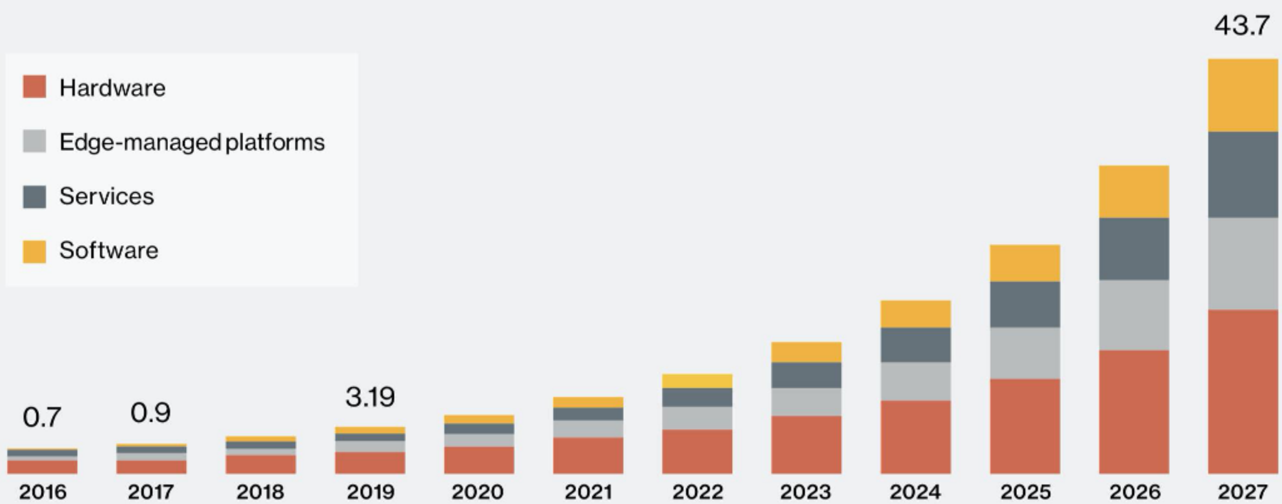
Source: LF State of the Edge 2020 Report



Source: LF State of the Edge 2020 Report

Other estimates of the edge computing market arrive at similar growth rates of 30-35 percent annually. However, work by Market Research Future, Research and Markets and Grandview Research (as cited in the Panduit-MIT Technology Review report and replicated in the chart below) wildly differ in their projection of the total market size. These varying estimates illustrate the difficulty of forecasting an industry where there aren't universally accepted market categories since each arrives at significantly lower estimates of overall capital spending. Thus, while the totals differ depending upon how a particular firm defines the edge market, everyone agrees that *spending on edge infrastructure will roughly double every two-three years until later in the decade, reaching tens of billions in annual CAPEX*.

US Edge Computing Market Size by Component, 2016-2017 (USD billion)



Source: MIT Technology Review Insights, based on data from Grandview Research, 2020

The LF SOTE report sees 4G and 5G mobile and residential demand and supporting infrastructure by network operators accounting for the largest share of edge spending over its forecast period. However, enterprise spending won't trail far behind as IDC estimates that the percentage of enterprise infrastructure budgets allocated to the edge will grow from less than 10 percent today to more than 50 percent by 2023.

A Closer look at Edge Scenarios

Consumer, municipal, and industrial (Industry 4.0) devices and applications provide the demand for edge infrastructure met by network carriers, colocation operators, and enterprises building out the infrastructure edge. As we examine these industrial and service provider scenarios, our view is that last-mile infrastructure at cellular base stations and remote cabinets (far edge) will eventually follow initial edge deployments at network aggregation points, COs, and remote enterprise facilities like manufacturing sites, logistics hubs, and warehouses.

Over the next few years, we expect several industries to be responsible for most of the significant deployments of edge infrastructure. The following is a summary of the business segments, use cases, and associated technology and market factors behind telco/cloud edge investments.

Telecommunications and Other Network Service Providers

Software-defined networking never lived up to the early hype promising to revolutionize enterprise networking. However, it profoundly changed carriers as they built next-generation networks around virtual resources and network functions. Among the many benefits of decoupling network data and control planes is the telecommunication carriers' ability to disaggregate network services from particular pieces of hardware and deploy services from various locations using hardware sized for the expected workload.

As we indicated earlier in the report, deployments will eventually become more granular and decentralized, with smaller server deployments in cellular base stations, office and apartment buildings, and neighborhoods coming after carriers gain experience and comfort with remotely managing edge servers and VNFs/CNFs.

Carriers, broadband providers, and ISPs plan must accommodate explosive growth in network traffic fueled by enterprise process digitization (aka digital transformation), consumer preference for streaming video and online gaming, and increased wireless demand from 5G. A principal strategy for greatly expanding capacity is via a scale-out infrastructure design using NFV with VNFs delivered from edge infrastructure.

The disaggregation will typically start small, with edge infrastructure in mini-data centers at COs or regional exchange points. In some cases, these COs, mobile switching centers, or exchange points can be augmented with high-speed fiber connectivity to data centers located 20-50km away, which provides a balance between low-latency and size of the facility (and therefore, cost of computing, which would be lower in more extensive facilities).

Regardless, as we indicated earlier in the report, deployments will eventually become more granular and decentralized, with smaller server deployments in cellular base stations, office and apartment buildings, and neighborhoods coming after carriers gain experience and comfort with remotely managing edge servers and VNFs/CNFs.

Cloud Service Providers, Both IaaS and SaaS

Like carriers, hyperscalers face the challenges of accommodating rapid growth coupled with new services such as those using machine learning, neural networks, and data analytics that require more computing, storage, and network resources. Cloud growth has been compounded by businesses that responded to COVID-19 lockdowns and a remote workforce by expanding their use of cloud-based productivity and communication services. Thus, ongoing digital transformation merged with an immediate transition to a distributed workforce to accelerate the cloud provider's need for expanded capacity.

Hyperscalers built around the economic efficiency provided by automated data centers have continued expanding these power plants of the digital era. However, they also understand that they can't scale their way out of every problem and respond to customers who need lower-latency or local control over sensitive data and applications by building network POPs, introducing new services using micro-sites, and expanding options for on-premises cloud resources. **As demonstrated by Google Cloud's telecommunications services**, Verizon's partnership with **AWS on Wavelength**, or **Microsoft Azure's partnership with AT&T and other carriers**, scale-out expansion is often in collaboration with carriers building out 5G infrastructure.

Enterprise or Campus

Unlike prior cycles of infrastructure decentralization, enterprises in the era of digitization are simultaneously *consolidating* resources in a few central data centers and cloud facilities and *distributing* infrastructure to locations that require local, real-time processing and analysis of data and low-latency application response. Enterprise edge infrastructure is most commonly deployed at:

- ROBO (remote/branch) locations to run a virtual network and hybrid cloud services like SD-WAN, security, and VM or container platforms.
- Manufacturing, logistics, and resource extraction sites target the same network and hybrid cloud services plus data storage and industrial IoT telemetry analysis.
- Logistics, distribution, and transportation hubs with a workload profile similar to manufacturing sites.
- Retail and hospitality locations for network services, point-of-sale, video surveillance, IoT, and VDI workloads.

Digital Service Economy

Digital natives, namely firms that primarily provide digital goods and services, face the same scaling problems as cloud providers. Indeed, video streaming and online gaming sites are most susceptible to performance degradation from consolidating cloud infrastructure workloads. Popular services like Netflix and Google's YouTube have operated edge distribution infrastructure for years; however, as consumer preferences have shifted to OTT video and online gaming, it expands the need for edge infrastructure. Many of these will be operated in partnership with the new generation of edge cloud services mentioned above.

Role of Devices at the Edge

Edge infrastructure is often installed to service intelligent devices pervading both established industries and enabling new services. Edge capabilities are increasingly found in the following situations:

- Transportation infrastructure, ride-sharing, and emerging autonomous vehicle (AV) applications, including:
  - Smart, connected municipal infrastructure like traffic signals, street lighting, and electric grids.
  - Vehicle telematics in conjunction with logistics and transportation edge infrastructure.
- Industrial IoT such as smart sensors and robotics systems paired with edge gateways for data collection and analysis.
- Remote service delivery for healthcare (telemedicine) and education (remote classrooms and distance learning)
- AVs, UAVs (unmanned aerial vehicles, aka drones), often include an embedded server for local processing of sensor and image data.

Edge Infrastructure Components

Edge infrastructure extends across several product categories that are pieced together with virtualization, container, network, and management software into an application platform. These product categories mirror those of data center infrastructure; however, implementations for the edge have notable differences that we discuss here.

Servers

When discussing edge infrastructure, the term "server" is used loosely since it often applies to embedded devices with low power budgets and small form factors. Infrastructure edge deployments typically use traditional 1U-4U rack-mount servers, some with half-width pluggable compute sleds that provide two single- or dual-socket servers per chassis. Infrastructure edge systems are also generally self-contained, with embedded drive bays and multiple network interfaces in configurations resembling conventional HCI systems.

Some infrastructure edge and most device edge servers are SFF (small form-factor) boxes with highly-integrated system boards and a low-power CPU or SoC. Often used in harsh environments, such embedded systems' performance is primarily dependent on the power budget and workload requirements. At one extreme are Lilliputian devices like a Raspberry Pi or

NVIDIA Jetson that can be powered by a USB port. At the high end are devices like this from Supermicro using a Xeon D-series processor with 4-14 cores, 4x10 GbE ports, and up to 256 GB RAM.

Network Switch-Routers

Many infrastructure edge installations aim to consolidate network and application services on a single device running a virtualization or container stack. However, larger edge installations by hyperscalers and carriers are more like small, modular data centers that include multiple servers and switches in a rack or cabinet. In these cases, it's more efficient to run networking on traditional discrete ToR switches.

Servers using SR-IOV devices or DPDK within VMs are very efficient at packet handling; however, they still can't make the raw throughput or price-performance of a dedicated switch. Although ROBO edge deployments often use a server, VM, or container stack and NFV software, such all-in-one appliances are no match for dedicated switches using merchant silicon and a lean network OS. Thus, in multi-server edge deployments, a ToR switch, many of which include basic L3 routing features, might be a possible choice for connecting systems to each other, discrete storage nodes and the WAN.

Device Gateways

Many industrial IoT devices use a diverse set of protocols like DNP3 (SCADA), MQTT (message queue telemetry transport), DDS (data distribution service), AMQP (advanced message queueing protocol), BLE (Bluetooth low-energy), WiFi, Zigbee, licensed cellular (LTE, 5G) and LoRaWAN. Not all of these are IP-based, or Ethernet-based. Device gateways address such a cacophony of protocols by providing connectivity for data acquisition, security enforcement, and local storage and compute capacity.

Gateways are a critical part of a multi-tier IoT design. They serve as a collection and management point for a wide variety of low-power sensors while providing data aggregation, format conversion, filtration, and analysis. A gateway then sends processed results to a data center or cloud service for further analysis and archival storage.

uCPE and SD-WAN/SASE Platforms

We won't cover this in great detail as AvidThink plans to publish an SD-WAN/SASE-focused report in the near future. Suffice to say that universal CPE and SD-WAN platforms can be strong candidates for running other edge workloads, especially since they are already present across multiple enterprise branch and HQ locations. These platforms are increasingly providing the ability to run different multi-tenant workloads from third-parties in response to enterprise demands.

Beyond security and next-gen firewall workloads, other candidates include IoT gateways (which we just discussed) or components of a unified communications service to enable better collaboration.

Software Platforms

Like data center systems, infrastructure edge servers are platforms for hypervisor and/or container software. In edge deployments, the virtualization stack is the foundation for several subsystems, including:

- Container (and perhaps VM) infrastructure
- NFV (VNF and CNF) network services
- IoT device integration and management (see the discussion under device gateways)
- Cloud integration
- Data pipelines and analysis
- Orchestration and automation

Products targeting carrier infrastructure typically include Kubernetes and associated container software that has been tuned for NFV, ETSI-compliant management and orchestration (MANO) systems, a VNF controller, and integration modules for carrier OSS and BSS infrastructure. Other approaches in this space include providing a more complete stack in the form of a container-based or container-hosted environment for developing edge applications. Companies doing this include MobiledgeX and Volterra, and major virtualization and container vendors like VMware and Red Hat are expanding their platforms to accommodate the edge as well.

Edge Colocation Facilities, Management Services

The nature of edge workloads is such that infrastructure can be beneficial anywhere an enterprise operates, whether it's an oil drilling platform or Interstate truck stop. However, cloud providers and carriers' growing need for hyper-local infrastructure to serve low-latency applications has rejuvenated the market for modular data centers.

Also known as data centers-in-a-box, these prefabricated shipping containers or ruggedized cabinets can be filled with racks of edge servers, switches, and storage and strategically placed in network proximity to customers. Such far edge locations are commonly in telecommunications hubs such as interexchange POPs, COs, or 5G base stations; however, smaller installations might be in computer closets at office or apartment complexes.

Carriers and large hyperscalers often own or rent far edge space and centrally manage systems as part of their fleet. However, for organizations without the physical footprint in strategic locations, there's an emerging market in edge colocation providers like Vapor IO, EdgeConneX (acquired by Sweden's EQT), and EdgeMicro. This is in addition to Mutable.io's AirBnB approach to monetize idle servers within carriers, hosting providers and MSOs. These offer rack space, server, network hardware, modular facilities, systems, and network management for organizations needing edge infrastructure.

Edge Ecosystem

As we wrap up the report, instead of putting together and listing a subset of edge products and services, we've decided to take a different approach for the edge. The edge ecosystem is proliferating, and the list of players and vendors grows weekly. To keep the size of this report manageable, we'll refer you to two resources that will be helpful:

- First, a list of Linux Foundations' Edge Projects at **https://www.lfedge.org/projects/**.
- Second, check out the **Edge Landscape at Topio Networks**, which includes an extensive list of vendors and numerous adjacent landscapes you might find helpful.

If you find yourself needing additional details or resources, drop us a line at **research@avidthink.com**, and we'd be more than happy to help!

AvidThink, LLC
1900 Camden Ave
San Jose, California 95124 USA
avidthink.com