Uptime Institute
INTELLIGENCE

PLANNING & STRATEGY

UI Intelligence report 48

Demand and speculation fuel edge buildout

Authors
Tomas Rahkonen
Research Director of Distributed Data Centers, Uptime Institute
Rhonda Ascierto
Vice President of Research, Uptime Institute

Demand for private and shared edge data centers is starting to grow from low numbers. Certain edge workloads, the strategies of large cloud providers, and the speculative deployments of supporting 5G networks are fueling edge data center buildouts. However, the complexity of edge business cases and other factors threaten to suppress demand.

30-45 minutes to read

UII-48 v1.0M published May 25, 2021 last updated May 25, 2021

This Uptime Institute Intelligence report includes:

ABOUT UPTIME INSTITUTE INTELLIGENCE

Uptime Institute Intelligence is an independent unit of Uptime Institute dedicated to identifying, analyzing and explaining the trends, technologies, operational practices and changing business models of the mission-critical infrastructure industry. For more about Uptime Institute Intelligence, visit uptimeinstitute.com/ui-intelligence or contact research@uptimeinstitute.com.

EXECUTIVE SUMMARY

Uptime Institute's research shows that demand for edge data centers is expected to grow across the world in the near term, particularly for shared data centers and especially in North America. Organizations and suppliers alike anticipate growth (from low numbers) across different industry verticals. Large cloud providers extending their platforms to the edge, in particular, are playing a key role.

KEY FINDINGS

- **The top factors driving demand for edge data centers are a need to reduce latency to improve or add new IT services and a need to reduce network costs and/or bandwidth constraints in data transport over wide distances.**
- **The digitization effects of COVID-19, the expansion by big clouds to the edge, and the mostly speculative deployments of 5G are among other key factors driving demand.**
- **Large-scale edge data center projects are becoming more common. Data center owners and operators, as well as suppliers, expect more edge data center deployments in the next two to three years. While most owners/operators expect to deploy their own private edge data centers, a growing portion say they would prefer leasing space in shared edge facilities.**
- **Edge data center demand growth may be slowed by the complexities of new edge data center business cases and a need to ensure consistent deployments across multiple locations.**

Introduction

Edge computing can be seen as a natural outgrowth of cloud computing. Instead of storing and processing data on servers in large remote data centers, edge computing brings processing closer to where the data is used or generated by a growing number of users and connected devices. Edge computing can help reduce network latency and the amount of data transferred to remote cloud and other core data centers, along with the associated bandwidth costs. We view edge computing as an extension of data center infrastructure to many more locations for the purpose of application optimization and integration, not as a displacement of cloud and core sites.

Together with the many innovations in hardware and software (including artificial intelligence), edge computing in the coming years promises greater data center and IT efficiencies and agility.

An edge data center can be defined by its purpose — to process data closer to a population of local users and local devices. Edge data centers are used to provide local compute and storage, and reliable network connectivity, including on-ramps to clouds and other services. Some use cases for edge data centers are well established at a large scale. They include:

- Video streaming and other services delivered over content distribution networks (CDNs).
- Local and low-latency workloads in regional offices and server rooms, including where compute is embedded into an existing facility (as opposed to a purpose-built data center).
- “Off-cloud” processing, whereby workloads and data are sited close to users/devices for reasons of compliance or security, such as certain types of high-performance computing (HPC) and high-security applications.

In recent years, edge computing has become a focus for retail stores, factory automation, online gaming, and for new and more stringent security applications than previously required — to name just a few use cases.

After many years of market hype and limited deployments of single or a few units, large multisite edge data center projects are now becoming more common. This report, the first in a series on edge data centers, discusses the role of edge data centers and the reasons behind the growth, ranging from an increase in IT workloads to the buildout of networks supporting edge, as well as factors that could inhibit edge growth. It is not, however, a numerical analysis or market sizing.

Please see **Appendix A** for a list of key companies currently active in edge development and **Appendix B** for our research methodology.

Edge data centers: Overview

Where is the edge? What is an edge data center? There is no single response to these types of questions because one organization’s approach to edge computing will likely differ from another’s. The edge is not a fixed place, and edge data centers can have many different characteristics.

Edge computing and edge infrastructure can be defined as any combination of approaches that brings IT capacity close to where data is generated or consumed (by users and/or by devices).

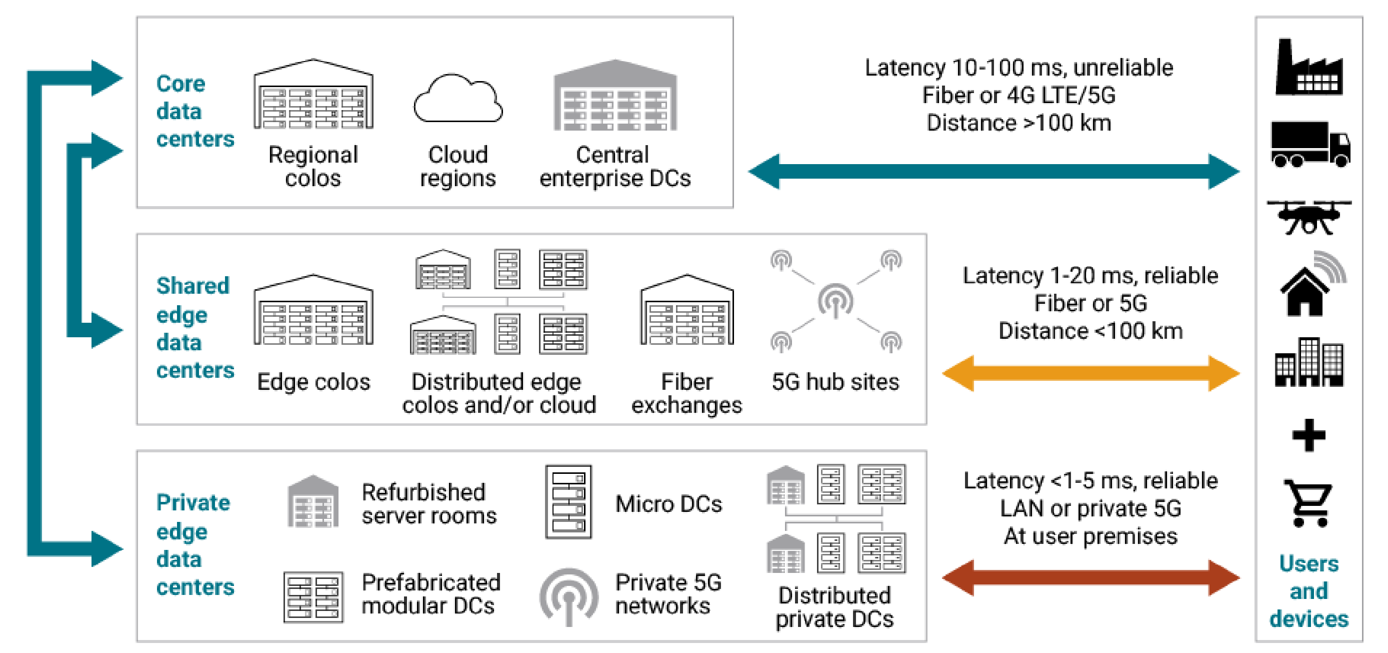
Why is this important?

The two most common drivers for edge data centers today are:

- Reducing latency to improve or add new IT services.
- Reducing network cost and/or bandwidth constraints.

Edge data centers can also help alleviate data security and business continuity concerns. While public cloud computing may be economical for many, an overconcentration of workloads under management of just a handful of major cloud operators is a risk for some. Even enterprises with a cloud-first strategy often retain control over some workloads (either in their own private data centers or in a leased colocation facility), most commonly for disaster recovery and backup. Edge workloads are also becoming increasingly business-critical; the internet protocol (IP) networks that support many workloads must have very high availability and reliability.

There are various types of data centers that may be deployed for edge computing, as shown in Figure 1.



Uptime Institute Intelligence, 2021

Figure 1. **Data centers at the edge**

Core data centers are typically large or very large and provide services over large areas (e.g., for several smaller countries, such as in Europe or Africa, or for multiple metropolitan areas in a large country, such as in the United States). Core data centers that are owned or operated by cloud providers are typically sited in groups of at least three at different locations (often up to 100 miles apart) to create a cloud region. Enterprise and government core data centers are typically used as a centralized IT location; colocation core data centers are often used as regional hubs.

While a core data center can provide sufficiently low latency for some edge workloads, many are sited in remote locations or are simply too far away from users or devices to provide low-enough latency. Network speed and capacity are common issues: the growth of high-speed connections and traffic volume means network access points and switches can become clogged, slowing transmission. Network paths are also growing more complex, increasing latency. Data traveling through multiple switches/routes (hops) will often be slowed down, especially if networks are shared.

Multiple network connections and paths also increases the likelihood of data being lost or corrupted due to component failures.

For these reasons, core data centers are mostly used to process, integrate and store data generated by edge applications. Typically, core data centers do not have low latency requirements.

Edge data centers, on the other hand, are used to carry out processing that is both low latency and reliable. At a high level, there are a two basic types of edge data centers, private and shared:

- **Private edge data centers** are owned by organizations whose primary business is not to provide cloud or IT services to other organizations. Private edge data centers can be used to deploy workloads with very low latency requirements, below 1 millisecond (ms) of round-trip time on a local area network (LAN) or corporate network. Today, small, private prefabricated modular (PFM) edge data centers are commonly installed as upgrades to existing server rooms. They often have between two and 25 IT racks or are single-rack, fully self-contained micro data centers.

  Private 5G networks, operated together with private edge data centers, are emerging as an enabler for edge workloads, especially for industrial internet of things (IoT), such as automated factories and major logistics hubs. Some organizations, particularly in retail, are connecting multiple private edge data centers to create a distributed edge network. By shifting workloads between sites, this architecture can improve business agility and increase data center resiliency.
- **Shared edge data centers** are owned by a service provider (colocation or cloud) and leased or used to provide services to multiple organizations. Today, most shared edge data centers are bricks-and-mortar (traditionally built) colocation facilities with multi-megawatt IT capacity that serve clients within a 100 kilometer (km; 62 mile) area (e.g., in a larger city or a remote region). These facilities, which can meet the latency requirements of most edge workloads, can be retrofitted buildings, telco central offices (including those that have been retrofitted for standard IT), or purpose-built data centers. Some edge colocation providers deploy smaller PFM data centers with IT capacities up to 200 kilowatts (kW), including in remote regions.

There are examples where edge colocation and cloud providers deploy, or are evaluating the deployment of, mesh networks using multiple edge data center sites, located near last-mile fiber networks and/or 5G tower sites. A mesh network supports full failover and continuous operation if one edge data center fails – thanks to replicating data – while leveraging real-time telemetry to detect the need for failovers and making use of diversified network routes. These distributed edge data centers are predominantly small PFM facilities that can process IT loads of between 20 kW and 200 kW each. Typically, up to 10 edge data centers can be used to cover a metropolitan area. (More on this in our upcoming report on edge data center resiliency.)

Some 5G operators provide data center capacity to edge computing partners, such as cloud providers and large internet companies, at their radio network hub sites. These hub sites have direct communication links to multiple 5G radio sites, which enable deployment of low-latency IT services covering a metropolitan area.

Edge data centers in practice

Following are examples of edge computing using core data centers and shared or private edge data centers (and of a cloud provider extending its platform to the edge):

- In Los Angeles, Boston, Houston and Miami (all in the US), Amazon Web Services (AWS) has implemented "Local Zones" to provide local cloud computing services with reliable single-millisecond latency. Customers can integrate data from a Local Zone to cloud workloads running in an AWS core data center. This is an edge use case as illustrated by the yellow arrow in **Figure 1**. Local Zones, and similar services from other cloud providers, are typically hosted at edge colocation data centers.
- A German automaker deploys industrial IoT in some of its factories using a private 5G network for device connections and refurbished server rooms to host AWS Outposts. Outposts is an IT hardware/software combination from AWS that runs local cloud instances and integrates with the automaker's data running in AWS core data centers. This is an edge use case as exemplified by the red arrow in **Figure 1**.

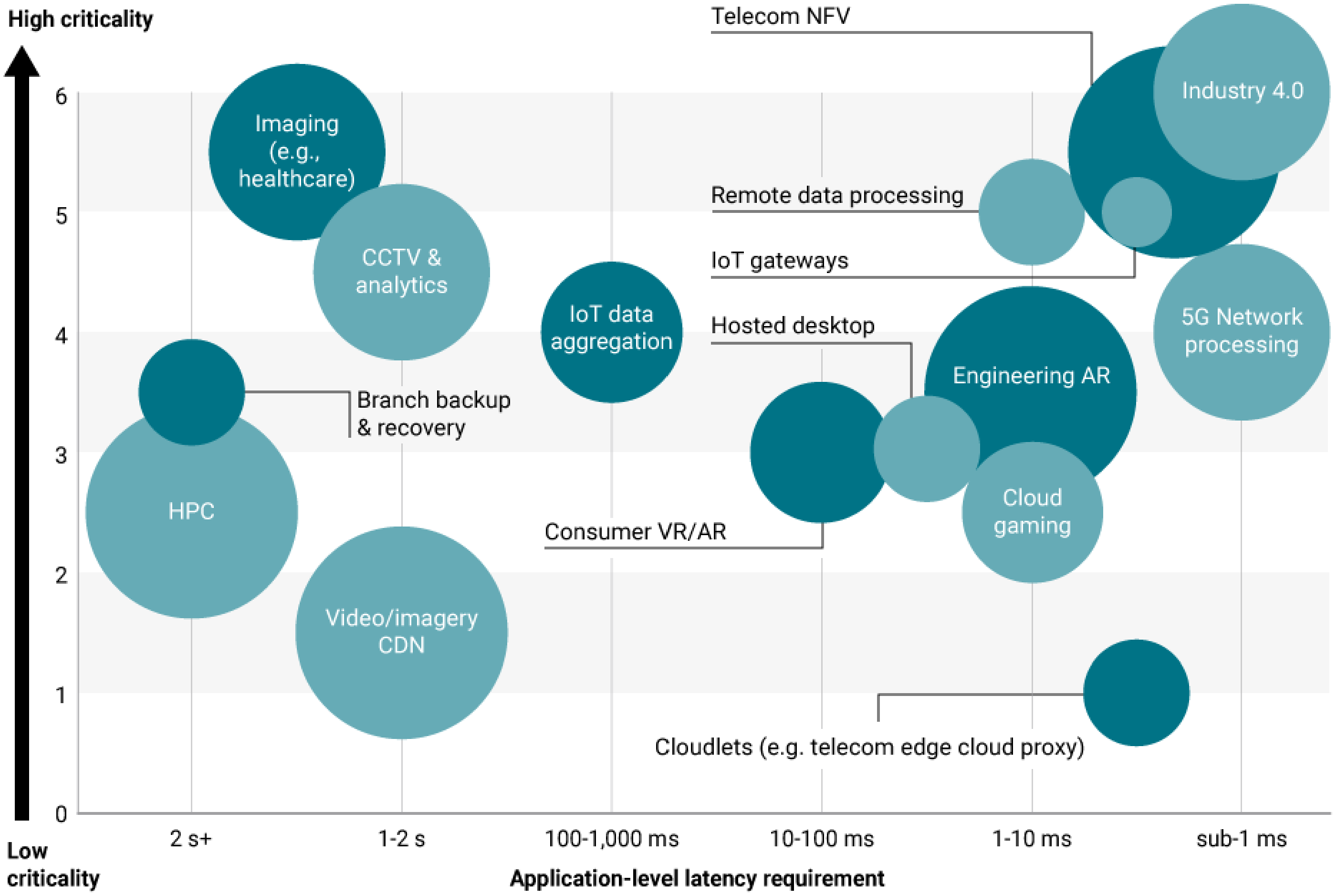
Workloads in edge data centers

The need for edge data centers will ultimately be defined and driven by the requirements of the workloads. At the highest level, there are three primary technical factors determining whether data processing and/or storage should be at the edge or at a more central location:

- Latency, or the speed of workload responsiveness/service delivery.
- Volume of data created or needed locally, which affects bandwidth and data haulage costs.
- Criticality, or the business sensitivity to service failure or delays, and/or the security or compliance requirements of a workload.

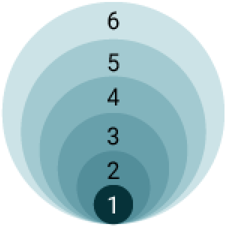
Figure 2 maps select edge workloads according to their typical requirements for criticality (y-axis) and latency (x-axis). The relative volume of data these workloads generate at an individual edge data center site is reflected in the relative size of each bubble.

Demand and speculation fuel edge buildout



Score value	Volume / day / site	Criticality (availability of data or site)
6	100 TB+ (10 Gbps+)	0 downtime/extreme business risks
5	10-100 TB (1-10 Gbps)	few seconds/ high risks
4	1TB-10 TB (0.1-1Gbps)	few minutes / moderate risks
3	100-1,000 GB (10-100 Mbps)	few hours / low risks
2	10-100 GB (1-10 Mbps)	few days / little risk
1	10 GB (1 Mbps)	irrelevant / no risk

Key for volume/day/site



Uptime Institute Intelligence, 2021

Figure 2 . **Edge workloads by latency, criticality and daily data volume**

Note that these are application-level latencies, which are different from the unloaded network latencies in **Figure 1**. Application-level latency (in Figure 2) is based on "loaded networks" and considers expected delays, such as those caused by multiple communication round trips needed to complete a workload operation, as well as network congestion resulting from large data sets (high data volumes).

IoT stands out as the workload in our sample that is most suited for edge data centers. IoT generates relatively little data but typically has high criticality requirements. IoT also often requires millisecond-range latency at the IoT gateway level.

IoT gateways are small industrial IT devices that standardize communication with connected machines and sensors. They can be devices installed in data center cabinets or stand-alone devices. IoT gateways can include storage and compute and, unlike low-powered IoT sensors, can connect to wide area networks (WANs). IoT gateways commonly communicate with edge aggregation sites, which are edge data centers used for data pre-processing, data integration and/or for real-time analytics, such as error detection or security monitoring.

In comparison with IoT gateways, IoT data aggregation sites manage higher daily data volumes and have slightly lower criticality with more relaxed latency requirements — between 100 ms to 1 second. Unlike core data centers, which are sited further away, IoT edge data centers tend to be sited close to networks and, therefore, can provide increased application availability and reliability. (Use cases can differ, based on location and networks. Some IoT gateways can communicate directly to core data centers, including clouds, without the use of an IoT aggregation site.)

There are a few edge workloads that can be considered highly demanding, because they require sub-1 ms latency, have highly criticality, and typically have relatively large data volumes. They include industry 4.0 (the practice of tightly integrating information and communication technology in industrial production), telecom network function virtualization (NFV), and 5G network processing. Two other notable workloads that can have low-latency requirements are cloud gaming and cloudlets (instances of cloud computing at the edge). Augmented reality (AR) and virtual reality (VR) applications, like cloud gaming, depend on low latency to achieve an attractive user experience, making them strong candidates for edge data centers.

Some edge use cases, such as CDN video streaming, make use of edge data centers mainly to overcome network cost and/or bandwidth constraints (device data buffering overcomes any latency issues when these workloads are sited in far-away facilities). High-definition closed circuit television (CCTV) security cameras generate large data volumes and commonly use analytics to make local decisions (in a nearby edge data center) and to reduce data volumes from the edge to a core data center. More advanced capabilities — for example, biometrics, object identification, and cross-comparison of multiple feeds and coordination of cameras — all require additional IT capacity that is local. Imaging, including from medical equipment used by doctors for analysis on-site at medical facilities, is another edge use case because of its high criticality and the requirement to handle large data volumes with a reliable latency of 1-2 seconds.

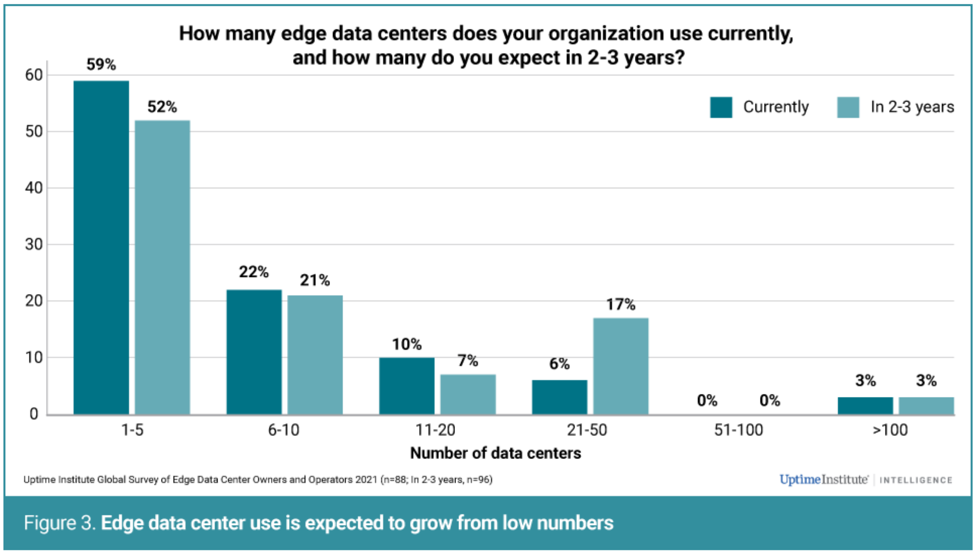
Edge data center demand growth

Demand for edge data centers in recent years has not been as strong as many suppliers had anticipated. Still, there has been steady growth, albeit from initially low volumes. To better gauge demand, Uptime Institute surveyed data center owners and operators, as well as data center product and service suppliers, in early 2021. The findings were consistent: Data center end-user organizations and suppliers alike expect an uptick in edge data center demand in the near term.

Owner/ operator demand

The study suggests that a small majority of owners/operators today use between one and five edge data centers and that this is unlikely to change in the next two to three years. This is likely an indicator of overall demand for edge computing — most organizations have some requirement for edge, but do not expect this to change significantly in the short term.

Many others, however, do expect growth. The share of owners/operators that do not use any edge data centers drops from 31% today to 12% in two to three years' time — indicating a significant increase in owner/ operator uptake (see **Private or shared?**). Furthermore, in two to three years' time, the portion of owners/operators in our study that are using 20 or more edge data centers today doubles (from 9% of respondents today to 20% in two to three years), as shown in Figure 3.


How many edge data centers does your organization use currently, and how many do you expect in 2-3 years?

Uptime Institute Global Survey of Edge Data Center Owners and Operators 2021 (n=88; In 2-3 years, n=96)

Figure 3. **Edge data center use is expected to grow from low numbers**

Over 90% of respondents in North America are planning to use more than five edge data centers in two to three years' time, a far higher proportion than the 30% to 60% of respondents in other regions. The largest portion of owners/operators planning to use more than 20 edge data centers within the next few years is in the US and Canada, closely followed by Asia-Pacific and China.

Large deployments of hundreds or more edge data centers are expected for many applications, including telecom networks, IoT in oil and gas, retail, cloud gaming, video streaming (including expansion into remote regions that will be enabled by rural satellite broadband), public transportation systems (including railways), and the growth of large international industrial companies with multiple local or regional offices. Several large-scale edge data center projects are today at a prototyping stage with one-to-10-site trials. Full-scale deployments involving tens of sites are planned for the coming three years.

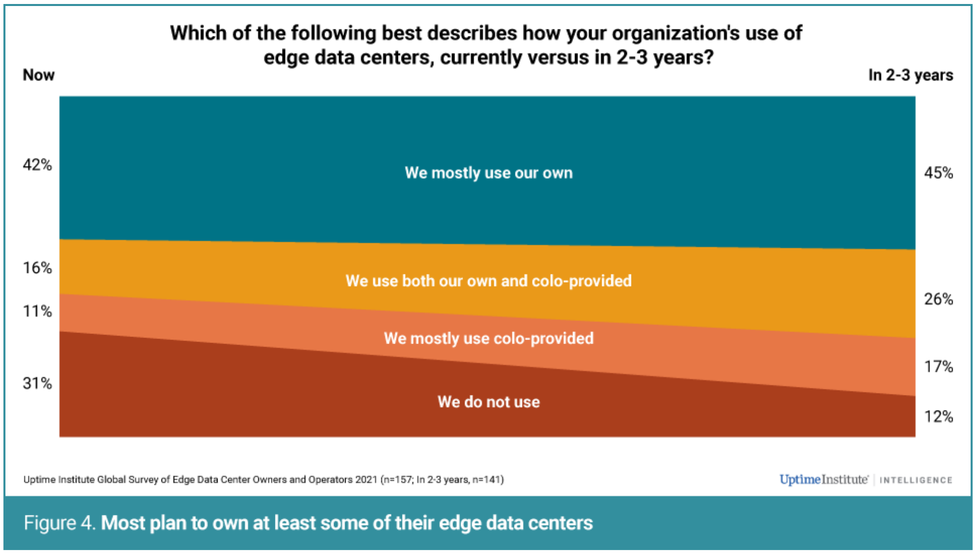
Edge IT densities

Our research shows that edge data centers are being used for a variety of IT loads, ranging from less than 10 kW per data center up to a few megawatts for colocation. Edge data center rack densities today are typically at either 5 kW/rack or so (for example, for telecom applications) or at around 10 kW/rack for more computationally intense workloads.

As demand builds during the coming few years, is it unlikely that this wide range of IT capacities will change. One reason is that edge centers serve a wide array of different edge computing use cases. At some future point, when edge data centers are deployed in high volumes, it is possible that suppliers of edge data centers will standardize on a few IT loads, to speed up production and delivery logistics.

Private or shared?

Almost half of the data center owners and operators (45%) in our study expect to mostly be using their own, private edge data centers in two to three years' time (see Figure 4).


Which of the following best describes how your organization's use of edge data centers, currently versus in 2-3 years?

Uptime Institute Global Survey of Edge Data Center Owners and Operators 2021 (n=157; In 2-3 years, n=141)

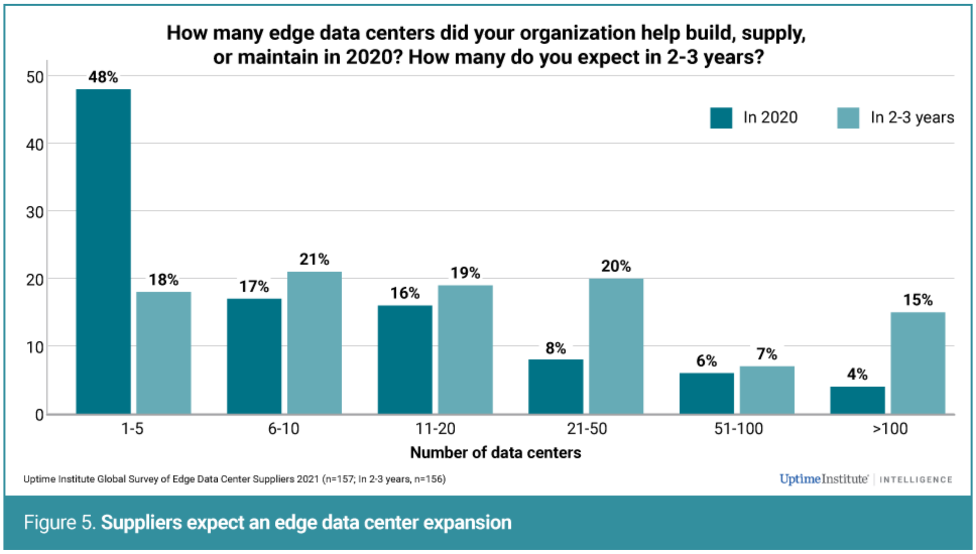
Figure 4. **Most plan to own at least some of their edge data centers**

Clearly, the use of shared edge facilities is anticipated by many. The portion of survey respondents that are mostly or partly using colocation edge data centers today (27%) is expected to grow significantly (to 43%) when looking two to three years ahead. However, it remains to be seen how the benefits of using shared edge data centers might influence the strategies of organizations going forward, particularly as more shared edge facilities become available.

Reasons for using shared edge data centers include greater business flexibility/reduced complexity when managing multiple, geographically dispersed sites and a preference to keep capital investment low.

Supplier demand

Most suppliers of edge data centers in our study are today helping to build, supply or maintain between six and 10 edge data centers per year. Many expect this will grow to 21-50 annually in two to three years' time, as shown in Figure 5. The portion of suppliers that built, supplied or maintained more than 100 edge data centers in the year 2020 is minimal today; looking ahead two to three years, 15% of suppliers in our study expect they will be handling 100+ edge data center projects (a threefold increase from today).



Uptime Institute Global Survey of Edge Data Center Suppliers 2021 (n=157; In 2-3 years, n=156)

Figure 5. **Suppliers expect an edge data center expansion**

Very few of the suppliers surveyed said that they deliver yearly volumes exceeding 1,000 edge data centers; this is unlikely to change in the short-term future. Notably, almost half of the suppliers (48%) helped build, supply or maintain as few as 1-5 edge data centers in 2020. This is expected to fall to 18% of suppliers in two to three years, which indicates meaningful expected supply-side volume growth (from the low number of edge data centers supplied today).

On a regional level, suppliers in the US and Canada are particularly bullish; almost one in three suppliers (32%) expect a yearly volume of more than 100 edge data centers in two to three years' time. Those in Europe are also optimistic; roughly one-fifth of suppliers (18%) expect an annual volume of more than 100 edge data centers in the coming two to three years. Some suppliers report a recent change in market demand, with orders and tenders going from being sporadic and for a single-digit number of sites, to more requests for larger projects with tens of sites. The change may be attributed to a solidification of clients' corporate strategy for edge computing.

Both data center owners/operators and suppliers expect a wide range of IT loads will be deployed in edge data centers in the coming two to three years. However, a higher portion of suppliers expect there will be more edge data centers with an IT load of 100 kW or more (compared to the expectations of owners and operators).

Some suppliers view services playing a large role in their future edge data center business. Almost 70% of respondents — including data center owners, operators and suppliers — foresee suppliers providing edge data center services.

Edge data center services can include scheduled activities (maintenance), nonscheduled activities (break/fix) and outage recovery. Data center infrastructure management (DCIM) software and services, including those that exploit artificial intelligence, will play a key role in supporting these services. Remote monitoring, predictive analysis and site-performance optimization are already being used to enable more efficient, reliable edge data centers. In the future, these software and services will likely be integrated with other systems to enable additional functions, such as WAN connectivity management and optimization.

Factors that impact demand

What will drive or hold back demand for edge data centers? This section discusses various factors that could accelerate or hinder the future of edge data center buildouts, including infrastructure that enables edge data centers, such as public and private 5G networks.

Demand drivers and enablers

Just as there are numerous types of edge data centers (or edge workloads), there are numerous forces driving demand for edge data centers. Some of these drivers are global, such as the COVID-19 pandemic spurring additional growth of cloud services due to increased remote working, streaming entertainment services and online shopping. Others are more regional and can involve the speculative buildout of supporting infrastructure for edge data centers, such as 5G networks and rural broadband.

The COVID-19 pandemic slowed overall business growth in some industries, dampening demand or delaying projects for new edge data centers for some companies. For others, greater use of online services, such as video conferencing, has helped drive additional demand. Looking ahead, the longer-term impact of COVID-19 on edge data center demand is expected to be positive due to its effect in accelerating digitization across multiple areas and industries.

The growth of cloud services is having a significant impact on edge demand. In recent years, large cloud companies have launched edge products that extend their cloud platforms to distributed edge locations. These products include software, such as Google Anthos and Microsoft Azure Stack, as well as software and IT hardware packages, such as AWS Outposts. All serve a similar purpose: they allow organizations to use cloud services, application programming interfaces (APIs) and managed infrastructure services in their own sites with WAN connectivity back to a cloud region (core data centers). A cloud provider's network is separated from an organization's intranet with border gateway protocol (BGP), a protocol commonly used for route advertisement between the networks. Software developers build and deploy applications using local computing power and onboard storage, utilizing the same user interface as for public cloud regions. Organizations' IT operations teams use the same tools as for cloud deployments, including for workload management and software updates.

For the many organizations already developing and managing cloud workloads, using familiar user interfaces and known APIs can make it easier to develop and operate workloads at the edge when compared with using separate (privately developed proprietary) tools and APIs. Cloud providers' edge products are commonly provided as-a-service (on-demand), which can be attractive to organizations seeking to reduce their capital requirements for edge deployments.

The growth of colocation data centers is another factor at play. Colos are today present in most major cities globally and providers are likely to expand into more geographies in the future, including smaller cities and remote regions. Oftentimes, they will market their presence in smaller cities as being in edge data center locations. An important enabler for the growth of colos in new regions is the availability of new undersea cables and long-haul terrestrial fiber networks — new fiber rollouts bring new data centers. One example of this is in Australia, where several colocation operators (including DCI Data Centers, Edge Centres and Leading Edge Data Centres) have deployed edge data centers enabled by new fiber rollouts in remote areas.

The advancement of certain technologies, and increased demand for them, is also driving demand for edge data centers. Examples include the increased use of high-definition video cameras for security (including face recognition) and IoT whereby data from sensors is collected for numerous applications, ranging from manufacturing factory quality assurance to digital agriculture and soil moisture measurement. High-resolution video generates large data volumes; edge data centers can be used to process and store data locally to keep down networking costs.

Industrial IoT applications in manufacturing, mining, and more are increasingly running in private edge data centers (often on-site) to optimize production processes and to ensure reliable availability.

Public and private 5G

In some cases, the networking infrastructure that can enable edge data center proliferation will play a role in driving demand for edge sites.

For example, 5G provides a combination of high bandwidth and low latency, making it suitable for large, latency-sensitive workloads, such as cloud gaming. 5G is also expected to become the connectivity standard for many IoT applications, as it supports very large numbers of connected, low-power devices.

Another reason for high expectations for 5G is the massive — largely speculative — investment being made in 5G network infrastructure. Aggressive associated marketing and market education campaigns by 5G suppliers and carriers are helping to attract increasing investments in this area. While many use cases for 5G are proposed, it is not yet clear which, if any, will drive edge data center deployments in large numbers in the near term. Virtual reality and augmented reality workloads are among those that would clearly benefit from 5G networking, but today these workloads are relatively scarce.

Most public 5G networks available today are designed for use by many and are targeted toward high-volume workloads such as online gaming and online content download. Less latency-sensitive workloads that are often described as being "5G use cases" can typically be deployed using more established 4G long term evolution (LTE).

Network slicing is a part of the 5G standard and can play a key role in helping to drive demand for more edge data centers at, or adjacent to, 5G network sites. Network slicing is a way to virtualize physical 5G networks, allowing individual organizations to establish private virtual 5G networks within a public 5G network. Organizations can customize a private virtual 5G network to meet specific requirements (such as throughput and latency requirements) for their workloads.

Another emerging approach is private 5G networks, whereby individual organizations obtain a license for a portion of shared 5G spectrum for their own use. (In some jurisdictions, use of private 5G spectrum may be limited to indoor use only.) A private 5G network allows an enterprise to customize the performance of the network. They can configure communication specifications to and from devices and set policies according to workload demands and priorities. Notably, private 5G networks can be deployed by companies that do not possess public spectrum (public spectrum is owned by telecom operators). Private 5G networks may widen competition for enterprise 5G services to include companies that are not telecom operators. The result could be the development of new services or competitive pricing — both of which will attract larger enterprises to invest in edge data centers.

Private 5G networks allow organizations to set policies according to workload demands.

Some pioneering organizations, among them German auto manufacturers BMW and Volkswagen, deploy industrial IoT using private 5G networks on-premises (on their factory sites), together with cloud provider products on-site (that extend the cloud platform to each factory location). Sensors and devices in the factory are directly connected to the cloud with very low latency. Enabling 5G network optimization is an active area for all large cloud providers; all have recent joint announcements with 5G telecom equipment suppliers to enable private 5G networks on-premises that are linked to their cloud instances.

For some edge workloads, the location accuracy that can be achieved with information supplied by GPS today is insufficient. These include applications that are used to operate "things" such as autonomous aerial drones and robots for industrial use and for consumer services in building reception areas, retail stores, etc. Millimeter wave 5G (mmWave), which provides greater (millimeter-scale) accuracy, may be used in such cases.

When compared with cellphones, an emerging number of roaming "things" require much higher bandwidth between sites in a 5G radio network to process data in real time. Examples include remotely controlled vehicles, vessels and aerial drones and driverless cars in an automated valet system. In some cases, it will require new buildouts of high-capacity fiber and edge data centers to support these types of workloads.

Wireless operators investing in 5G often expect to maximize their financial returns from the network investment by investing more in the upcoming edge compute workloads themselves. Some operators seek to partner with colos to operate edge data centers at their hub sites and central offices. Certain wireless operators now host large cloud companies at their 5G radio access network hub sites in some metropolitan cities — enabling applications with very low latency to run over their public 5G networks.

5G — and later, 6G — will most likely become large demand drivers for edge data centers because of these newer capabilities, such as network slicing and mmWave accuracy.

Demand inhibitors

There are numerous, often complex, parameters underpinning the decision to deploy an edge data center. They include decisions around owning and operating the data center (private or leased/shared facility); how to optimize networking resources and costs (using as-a-service software or by developing proprietary software); and various practical operational aspects. Multiple stages of prototyping are common, which means planned edge rollouts can take time. Such complexity also makes return-on-investment estimations difficult.

Some edge computing providers deploy modern fiber optical technology, with very low fiber losses, to provide high capacity and low latency over longer distances than previous fiber networks. This can enable reliable low-latency connectivity down to 5 ms from a single data center serving multiple metropolitan areas. If this approach proves successful for a wide range of edge workloads and if it proliferates, it could significantly reduce the number of edge data centers needed.

In large projects where multiple edge data centers are involved (including those in mesh networked configurations), it can be a complex task to configure and provision edge data center infrastructure, networking, IT hardware, software, and remote monitoring. To simplify large deployments, suppliers can provide edge configurator tools and actively promote standard configurations. To ensure positive deployments and continued supplier demand, edge data center products and services must ensure a positive user experience in remote locations, including for users with unskilled and low-skilled staff.

In populated areas, engaging with local municipalities and jurisdictions can be complex and time consuming. Permitting processes can take many months. Some North American city municipalities can take years to grant zoning permits for the three-phase power supply needed to support edge data centers. Certain large-scale edge applications — such as autonomous vehicles that use roadside data center infrastructure — require policies/permits across multiple jurisdictions (states or countries), which can significantly delay deployments.

Other factors that impact demand

Environmental sustainability is increasingly important for all types of data centers. As edge data centers proliferate over time, their sustainability profile is likely to be scrutinized. Negative publicity may slow demand. To avoid this scenario, edge data center owners/operators and suppliers will need to make a focus on sustainability of designs and approaches standard practice, including:

- Adopting energy-efficient IT and the active use of power-save modes.
- Sourcing green grid power, supplemented by on-site solar or wind power.
- Using energy-efficient power and cooling technologies.
- Promoting data center heat reuse, including in partnership with district municipality heating systems.

For more on data center sustainability, see our recent report **Renewable energy for data centers: Renewable energy certificates, power purchase agreements and beyond**.

The supply of edge data centers and related services is sufficient to meet short-term demand, with many suppliers ready to serve a market that currently has low-volume orders. However, this situation may quickly change if multiple, large-volume projects ramp up simultaneously on a broader scale. Edge data center suppliers have an important role in helping to drive edge data center demand by investing in production capacity to avoid long delivery lead times and by driving supply chain efficiencies, resulting in cost improvements over time.

Standardization of edge data centers is another factor at play. While most edge data center clients have historically sought tailor-made data center specifications, more are now favoring standardized edge data center floor plans and other features enabling serial production.

Logistics is an important factor for large projects of many tens or hundreds of edge data centers. Simultaneously deploying many facilities at different locations is a new experience for most organizations and edge data center suppliers. It requires a different kind of project organization from deploying a single large data center. Access to installation partners capable of handling many projects at different locations and with adequate quality assurance will be key.

These project characteristics in many ways resemble telecom rollouts, and some data center providers (including colos) are actively recruiting staff with telecom backgrounds to meet their needs.

Certain large-volume edge data center projects, including deployments with rural broadband networks, depend on government grant money for their rollout. Demand tied to government-backed projects can be susceptible to delays based on a change in government or new government policies.

Summary

Data center capacity is beginning to be reshaped by edge computing. Distributed sites are being upgraded, and new small edge facilities are being deployed. Edge activity is also driving demand for large core data centers, including for regional colocation and hyperscale data centers.

Edge data center demand is growing across many different industry verticals, particularly those that are adopting IoT approaches, as well as from large cloud providers extending their cloud platforms to the edge. There is complexity involved in developing edge business cases and it is not yet clear that any single use case will drive edge data centers in high volumes. Growth is likely to be the result of a combination of factors, including the demand profile of various edge workloads and the extent of the speculative buildout of supporting networks and shared edge data centers.

Appendix A: Key companies

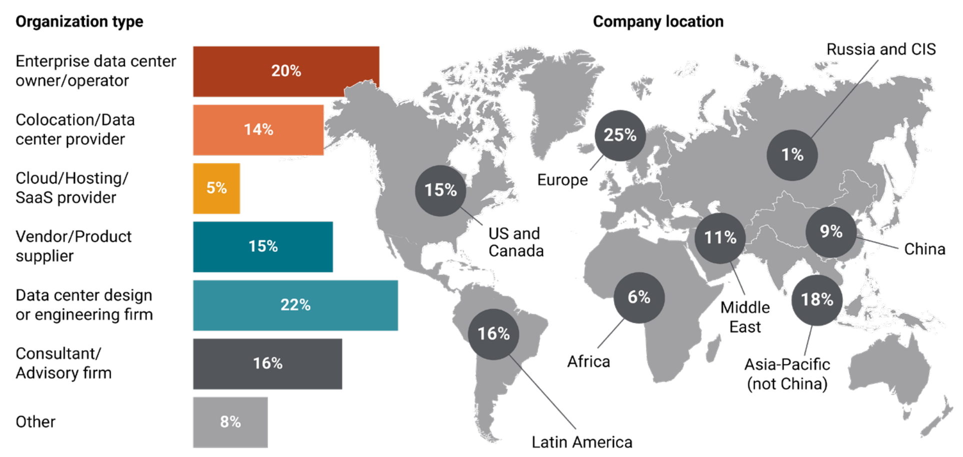
The following companies are examples of notable edge data center suppliers, owners, operators and innovators. They are among the players driving or supporting edge data center demand.

Amazon Web Services
Cannon Technologies
CloudFrame
Compass Datacenters
Delta Electronics
DCI Data Centers
Edge Centres
EdgeMicro
Google
Huawei
Leading Edge Data Centres
Microsoft
Schneider Electric
Silent-Aire (Johnson Controls)
Vapor IO
Vertiv
Zella DC

Appendix B: Methodology

This report draws on roughly two dozen in-depth interviews with selected companies that either use, supply or provide supporting infrastructure for edge data centers, including cloud providers, colocation providers, telecoms and edge data center suppliers.

The report also cites findings from Uptime Institute's Global Survey of Edge Data Center Owners, Operators and Suppliers 2021, which was conducted during the first quarter of 2021. More than 430 decision makers participated in the survey. Figure B1 provides respondent demographics.



Uptime Institute Global Survey of Edge Data Center Owners, Operators and Suppliers 2021 (n=434)

Figure B1. **Survey demographics**

Uptime Institute® | INTELLIGENCE

ABOUT THE AUTHORS

Tomas Rahkonen is Research Director of Distributed Data Centers at Uptime Institute. Dr. Rahkonen has spent over 25 years in global positions in the telecommunications, mobile communications and data center sectors. He most recently served over 10 years as chief technical officer of Flexenclosure, where he managed the design and delivery of prefab data centers to four continents. Contact: trahkonen@uptimeinstitute.com

Rhonda Ascierto is the Vice President of Research at Uptime Institute and is a founding member of Uptime Institute Intelligence. She has spent two decades at the crossroads of IT and business as an industry leader, keynote speaker and executive adviser. Her focus is on innovation and disruptive technologies in data centers and critical infrastructure, including those that enable the efficient use of all resources. Contact: rascierto@uptimeinstitute.com

ABOUT UPTIME INSTITUTE

Uptime Institute is an advisory organization focused on improving the performance, efficiency and reliability of business critical infrastructure through innovation, collaboration and independent certifications. Uptime Institute serves all stakeholders responsible for IT service availability through industry leading standards, education, peer-to-peer networking, consulting and award programs delivered to enterprise organizations and third-party operators, manufacturers and providers. Uptime Institute is recognized globally for the creation and administration of the Tier Standards and Certifications for Data Center Design, Construction and Operations, along with its Management & Operations (M&O) Stamp of Approval, FORCSS® methodology and Efficient IT Stamp of Approval.

Uptime Institute – The Global Data Center Authority®, a division of The 451 Group, has office locations in the US, Mexico, Costa Rica, Brazil, UK, Spain, UAE, Russia, Taiwan, Singapore and Malaysia. Visit uptimeinstitute.com for more information.

All general queries:
Uptime Institute
5470 Shilshole Avenue NW, Suite 500
Seattle, WA 98107 USA
+1 206 783 0510
info@uptimeinstitute.com