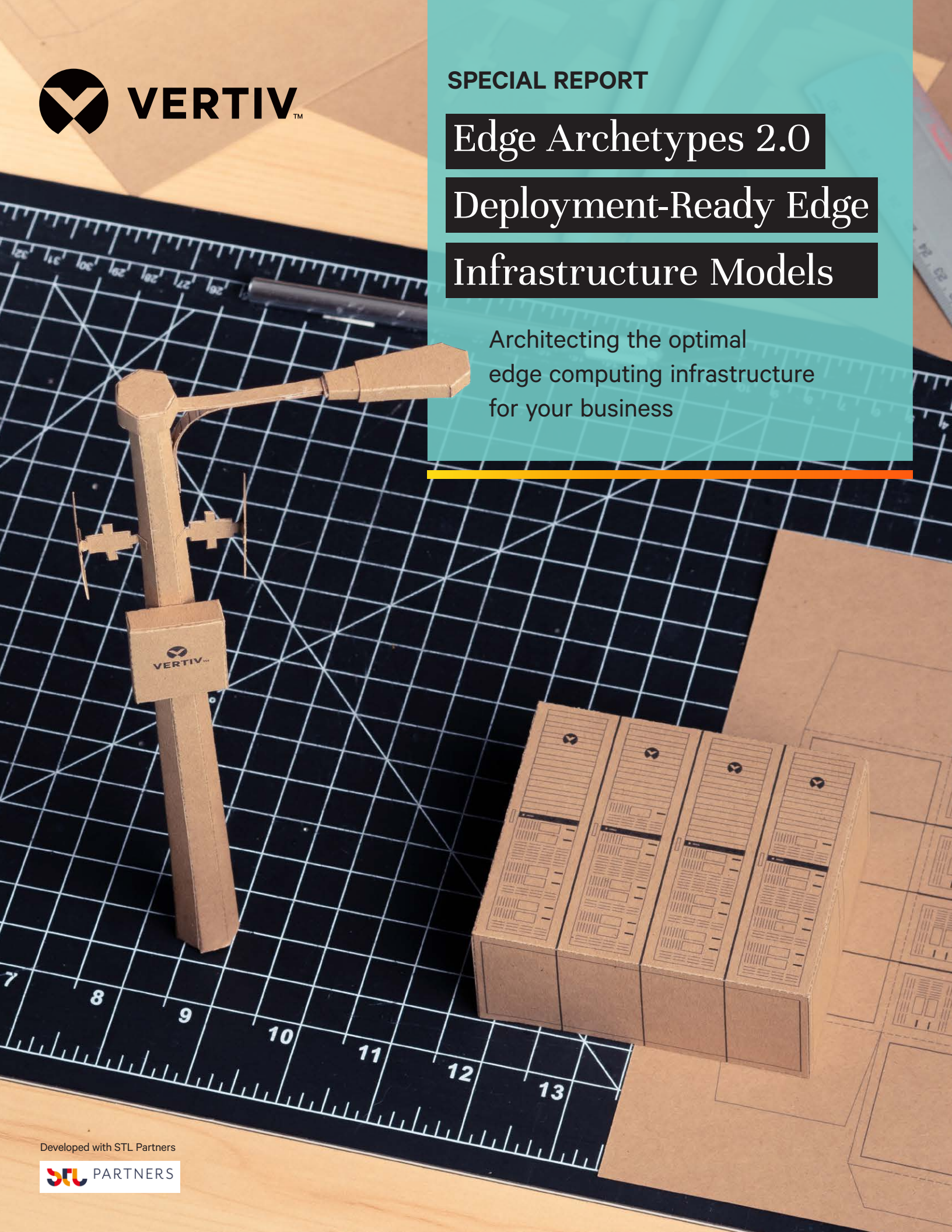
VERTIV™

SPECIAL REPORT

Edge Archetypes 2.0

Deployment-Ready Edge Infrastructure Models

Architecting the optimal edge computing infrastructure for your business



Developed with STL Partners

STL PARTNERS

Executive Summary

Physical infrastructure is key in any edge computing strategy. The power, cooling, and enclosure equipment, as well as the compute it supports, provides the foundation on which applications can run and enables countless edge use cases.

Making the right physical infrastructure choice is even more important at the edge given that many deployments are in locations where additional support and protection is required. Navigating edge infrastructure is also made more complicated with the broad and varied definitions of edge. These factors make it challenging for the 49%[1] of enterprises exploring edge computing deployments. They must make decisions on how best to use existing infrastructure and where to make investments today to support the needs of tomorrow. Fortunately, there is an ecosystem of suppliers, system integrators and other channel partners with experience and expertise in edge deployments to provide support.

Building on Vertiv's work on the Edge Archetypes[2], which provided a taxonomy for categorizing edge use cases, this report takes those archetypes a step further to define four distinct edge infrastructure models. The framework was developed based on interviews with a range of industry practitioners, data center experts, solution providers, and industry bodies across smart city, healthcare, manufacturing, and retail applications. With thorough analysis of the edge computing needs of different industries and use cases, the following edge computing infrastructure models were defined:

[1] STL Partners survey with 699 industry professionals globally from manufacturing, retail, healthcare and transport & logistics industries, May 2021
[2] Defining Four Edge Archetypes and Their Technology Requirements

Device Edge	Micro Edge	Distributed Edge Data Center	Regional Edge Data Center
• On-device • Attachment or built-in • Outside (e.g., street lamp) or inside (e.g., manufacturing equipment)	• Small number of servers or rack • 0-4 racks • At enterprise site (e.g., retail shop floor, factory, IT closet, municipalities)	• Small data center • 5-20 racks • Enterprise site (e.g., warehouse), telecoms network site, parking lot	• Mid-sized data center • 20+ racks • Regional location, (e.g. Tier 2 or 3 city)

Key Findings

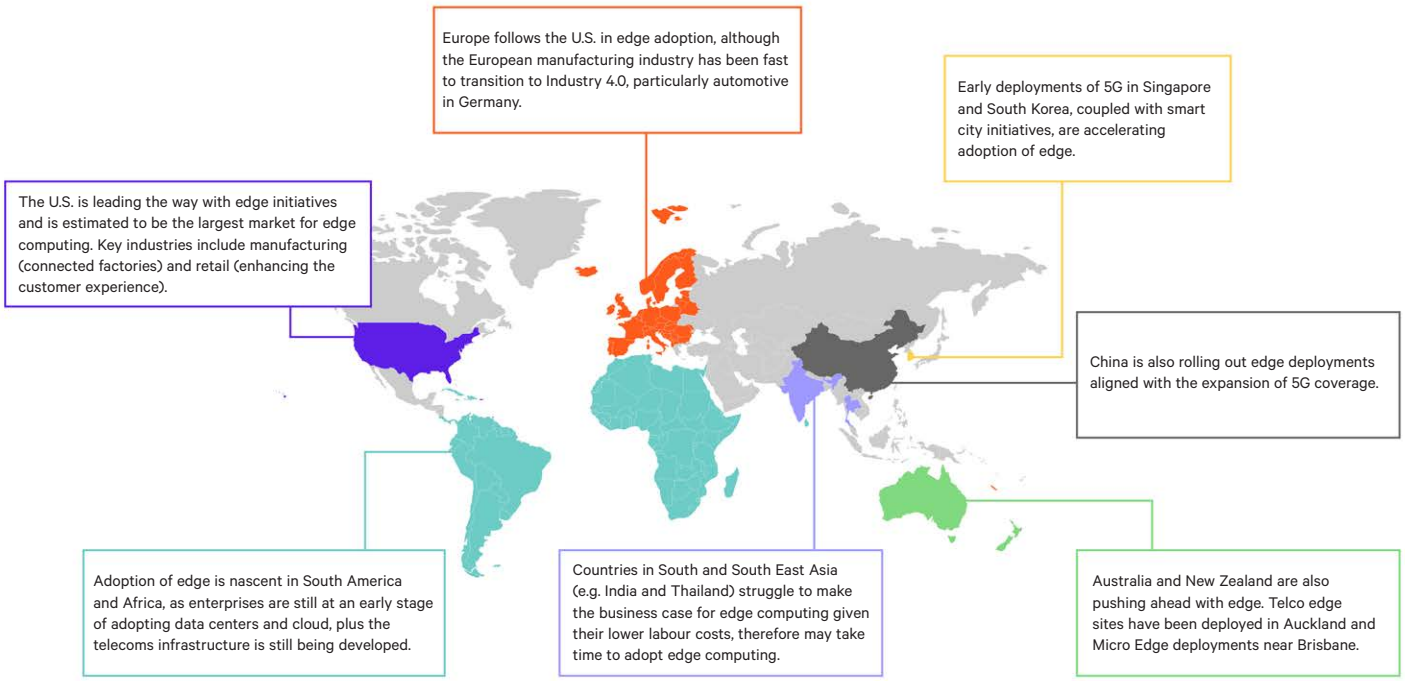
- Edge computing infrastructure will not act as a substitute for cloud. The total number of edge sites is estimated to grow by 226%[3] from 2019 to 2025. Equally, cloud will continue to grow at a CAGR of 10%[4].
- The United States is leading the way with edge initiatives and is estimated to be the largest market for edge computing[5], driven by key industries such as manufacturing.
- The most developed edge computing deployments are those aligned with the Human-Latency Sensitive edge archetype (e.g., cloud gaming) followed by Data Intensive (e.g., video analytics) and Machine-to-Machine Latency Sensitive (e.g., stock trading). Use cases from the Life Critical archetypes (e.g., autonomous cars) are still mainly at an exploration or proof of concept stage.
- Most Life Critical archetype use cases will use the Device Edge infrastructure model in the medium term, whereas Data Intensive, Human-Latency Sensitive, and Machine-to-Machine Latency Sensitive use cases will accelerate the transition from Regional Edge Data Center to Micro Edge and Distributed Edge Data Center infrastructure models in the near term.
- Coordinating the many elements of edge computing (software, hardware, infrastructure, etc.) is challenging and requires an ecosystem of partners to support the 66% of enterprises that prefer to have an entire edge solution coming from a single lead vendor.

[3] Data Center 2025: Closer to the edge
[4] Technavio, 2021
[5] Edge Computing Market - Global Forecast to 2025

Introduction: The State of Edge Infrastructure Today

Twenty years ago, the data center market pendulum swung toward centralized computing to improve efficiencies in data processing. Now, the pendulum is swinging the other way toward edge computing. Edge computing refers to compute and storage that sits between centralized data centers and the end-users, device, or source of data. On one hand, edge computing can be considered an alternative to cloud and central data centers, when those options are unable to meet latency requirements, or it is too costly to transfer high volumes of data over long distances. On the other hand, edge computing is also a driver for cloud adoption. An edge site can act as a staging-post for data that is ultimately sent to the cloud for processing, storage, or long-term analysis.

Over the last two years, edge computing adoption has significantly increased, in parallel with the continued growth of cloud. According to a recent survey by STL Partners, 49% of enterprises in specific industries are actively exploring edge computing[6] and it is estimated that the total number of edge sites will grow by 226% from 2019 to 2025[7]. However, adoption varies across geographies. This is due to the level of maturity of adjacent technologies (e.g., artificial intelligence), the existing telecommunications infrastructure, government policy, and the size of certain industries in the country. For example, manufacturing is driving edge computing adoption in the U.S. and Germany and is predicted to account for the largest share of European enterprise edge spending in 2021[8].



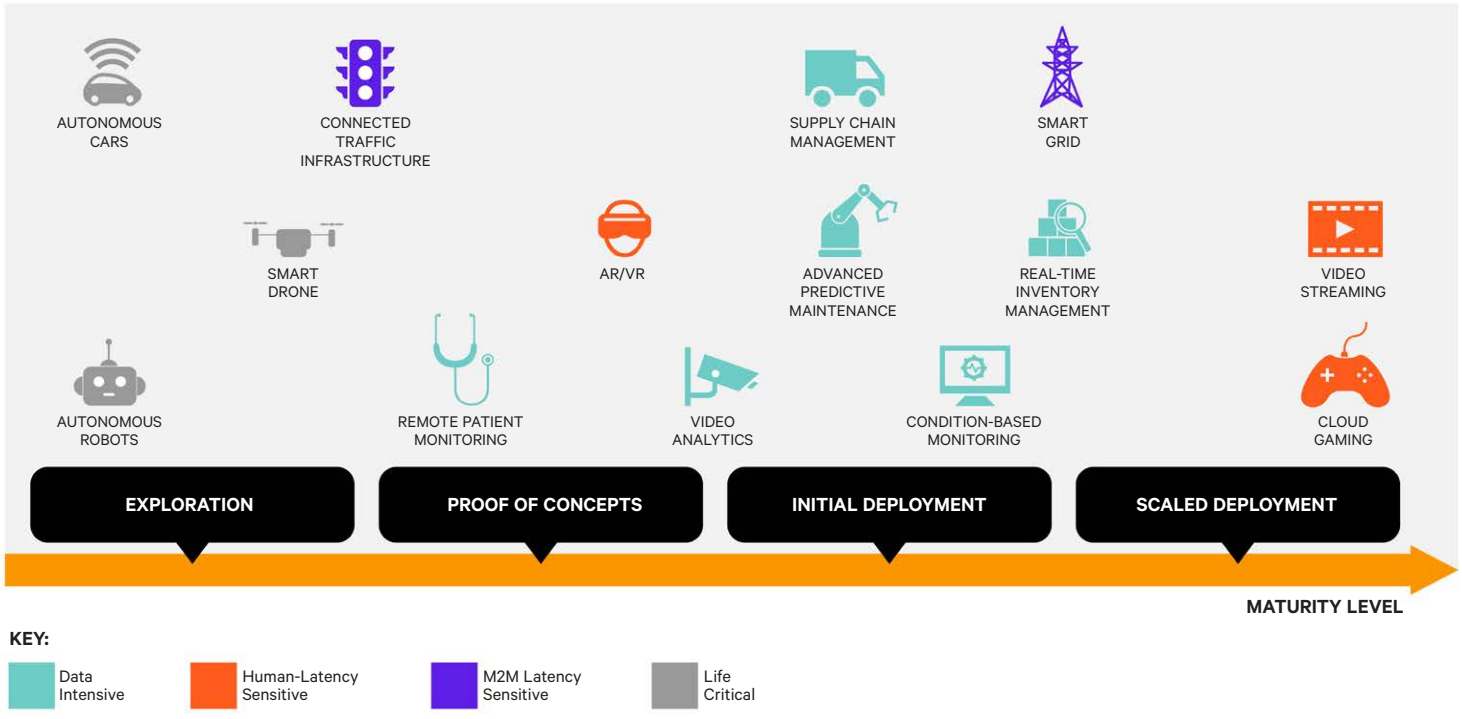
Enterprises see edge computing as a key enabler to overcome challenges related to data security and reliability, in addition to improving application performance. For example, large clusters of data centers could become prime targets for attack. Splitting up the core into multiple edge sites may cost more per kW but eliminates the threat of simultaneous denial of service. Edge Computing also promises to benefit a wide range of industries across a diverse set of use cases. From cloud gaming to smart grids for electricity distribution networks to autonomous robots in industrial settings, all these use cases have something to gain from processing data closer to the end-device. Early adopters are already implementing innovative solutions, moving past proof-of-concepts and initial pilots to multi-site deployments at scale. One example of this is Lloyds Register, a maritime services company, which has deployed edge computing across fleets of ships[9] to optimize fuel consumption through data insights. The adoption of edge computing will also be supported by a growing ecosystem of suppliers, system integrators and other channel players. The distributed nature of edge computing requires a network of edge players with the reach and capability to deploy, service and support edge infrastructure.

[6] STL Partners survey with 699 industry professionals globally from manufacturing, retail, healthcare and transport & logistics industries, May 2021
[7] Data Center 2025: Closer to the edge
[8] International Data Corporation's (IDC) Worldwide Edge Spending Guide

In 2018, Vertiv released a report, Defining Four Edge Archetypes and Their Technology Requirements, which provided an industry-first framework for categorizing use cases. These four archetypes helped enterprises and edge data center operators better understand the common underlying requirements across similar use cases. The four archetypes are:

- **Data Intensive:** Use cases where the amount of data makes it impractical to transfer over the network directly to the cloud, or from the cloud to the point of use, due to data volume, cost, or bandwidth issues.
- **Human-Latency Sensitive:** Use cases where services are optimized for human consumption or to improve the human experience with technology-enabled services. Speed is the defining characteristic of this use case as delays in delivering data directly impact a user's experience.
- **Machine-To-Machine (M2M) Latency Sensitive:** Use cases where services are optimized for machine-to-machine consumption. Because machines can process data much faster than humans, speed is the defining characteristic here (and the consequences for failing to deliver data within the required time "budget" can be higher than for Human-Latency Sensitive use cases).
- **Life Critical:** Use cases that directly impact human health and safety. Speed and reliability are therefore paramount.

Interviews with experts in vertical industries and the data center space revealed that the archetypes vary in their levels of maturity. **Human-Latency Sensitive** edge use cases (e.g., cloud gaming) are the most mature and already reaching scaled deployments. The growth of 5G and increasing fiber deployments will further accelerate this maturity. Conversely, **Life Critical** use cases will take much longer to adopt edge computing. This is because they have stringent requirements for latency and reliability and often need regulation changes to be implemented at scale. Smart drones are an example. Governments need to be confident that autonomous drones will not impose any threat to human life before relaxing airspace restrictions. Similarly, connected traffic infrastructure is still at an early stage. In the U.S. alone, only 7%[10] of traffic lights are smart.



[9] WWT, 2020: Three real-world case studies for how manufacturers can maximize edge computing
[10] Vertiv interview program – quote from interviewee (Director – Experience AI, automotive manufacturer).

Moving From Use Cases to Infrastructure

Three years after the release of the original archetypes report, the edge computing market is still evolving, and companies are continuing to develop their edge computing solutions. Use cases have progressed from concept to real applications deployed in the field. These software applications need adequate infrastructure that can support high-bandwidth, low-latency data processing at the edge.

The term "edge infrastructure" refers to the physical compute infrastructure (servers, power, cooling, enclosures) that is deliberately positioned anywhere between the end-device and central data centers. This also includes hosting compute capabilities on premise, something that is obviously not new for many enterprises. In fact, some are re-investing in existing on-site infrastructure (e.g., servers, network closets, or data centers) to optimize applications and implement new use cases. For example, a multinational pulp and paper manufacturer[11] is enabling data intensive applications such as advanced predictive maintenance by leveraging data centers at its larger mills.

Working to a strict definition, true edge infrastructure should use standard off-the-shelf IT infrastructure and be set up on cloud principles to host cloud-native applications and workloads. Legacy on-premise infrastructure that is monolithic or based on proprietary hardware is not considered "edge compute" under this definition.

To date, the market has not been clear about what constitutes edge infrastructure.[12] Enterprise customers want to adopt edge solutions today with a level of certainty that these solutions will meet future needs. Similarly, edge data center operators must invest in infrastructure today that will support tomorrow's applications. Both sides need answers to key questions on edge computing infrastructure:

- What does the edge look like in terms of physical infrastructure?
- What will the measurable benefits be of deploying IT closer to the applications?
- Who will own and operate the edge computing infrastructure?
- How can we implement it effectively and at scale?

In this paper, we will explore the key factors that influence edge infrastructure, including the use case, industry, and external environment. As part of this research, we conducted 22 interviews with a range of industry practitioners including enterprises, data center experts, solution providers and industry bodies.

Building Your Edge: Four Edge Infrastructure Models Provide the Foundations

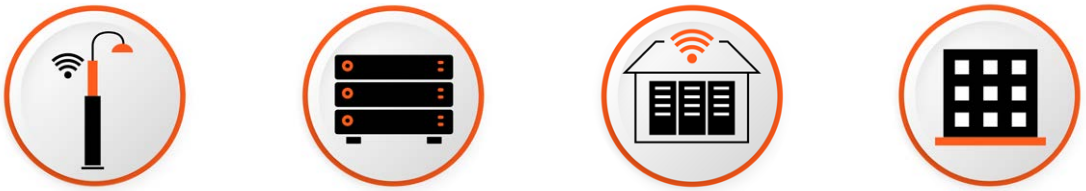
Vertiv has developed an innovative framework for categorizing edge infrastructure into specific models to help organizations make practical decisions around deploying physical infrastructure and compute at the edge. The term "infrastructure" is used instead of data center as not every edge deployment can be described as a data center form-factor per se.[13] The models help align the terminology that is used when discussing edge computing. They encompass the variety of edge deployments seen today, as well as the evolution in deployments expected over the coming years.

The four edge infrastructure models are the following:

- **Device Edge:** The compute is at the end-device. It is either built into the device (e.g., a smart video camera with artificial intelligence capabilities) or is an "add-on edge," stand-alone form factor that directly attaches to the device (e.g., a Raspberry Pi computer attached to an automated guided vehicle). When the compute is built in, the IT hardware is fully enclosed within the device, so it does not need to be designed to endure harsh environments. For example, when the compute is attached to the outside of a camera it must be ruggedized, but if it is built into the camera it is within a controlled environment so ruggedization is not necessary.
- **Micro Edge:** A small, standalone solution that ranges in size from one or two servers up to four racks. It is often deployed at an enterprise's own site (e.g., for a manufacturer it could sit on the shop floor of the factory, or in a back office). It can also be situated at a telco site (e.g., a rack of servers located at a telco base station). The Micro Edge can be deployed in both conditioned and unconditioned environments. In conditioned environments (e.g., IT closet), the Micro Edge does not require advanced cooling and filtration as external factors such as temperature and air quality are stable. In unconditioned environments (e.g. a factory shop floor), the compute is ruggedized and the Micro Edge requires specialized cooling and filtration to account for the harsher external factors (e.g., high temperatures and dust).

[11] Interviewee from Vertiv research program, 2021
[12] Edge computing infrastructure refers to the edge IT stack as well as the physical facilities that support it (e.g. power, cooling, security, enclosures).
[13] A typical data center environment would normally include: fiber connection, uninterruptible power supply, cooling, security, cabling, raised floor.

- **Distributed Edge Data Center:** A small, sub-20 rack data center that is situated at the enterprise site, telco network facilities, or at a regional site (e.g., in modern factories or large commercial properties).
- **Regional Edge Data Center:** A data center facility located outside core data center hubs. As this is typically a facility that is purpose-built to host compute infrastructure, it shares many features of hyperscale data centers (e.g., is conditioned and controlled, has high security, and high reliability).

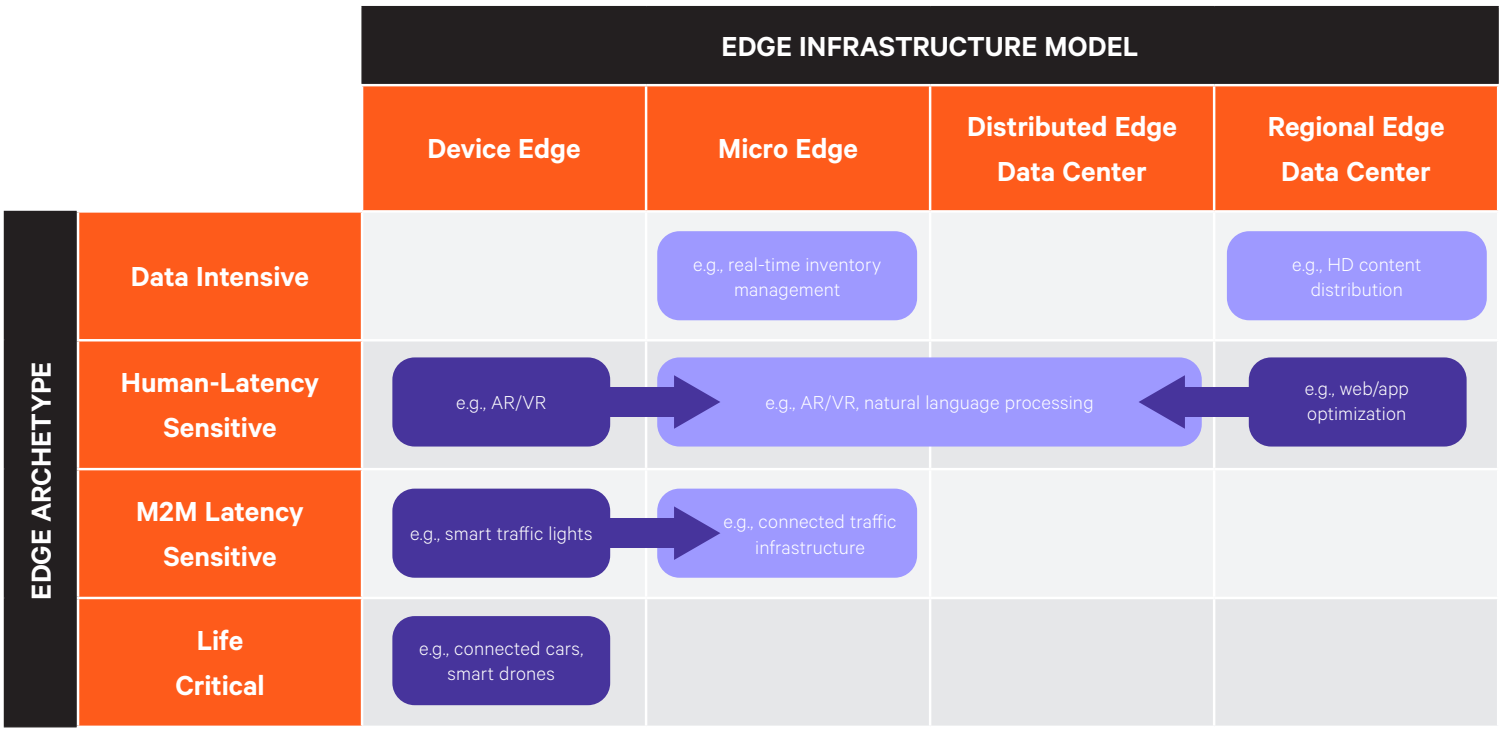


		EDGE INFRASTRUCTURE MODEL			
		Device Edge	**Micro Edge**	**Distributed Edge Data Center**	**Regional Edge Data Center**
CHARACTERISTICS	**Location**	Smart devices (e.g. in vehicle, street lamp, IoT)	Enterprise site (e.g. retail, factory floor, IT closet, municipalities)	Enterprise site (e.g. warehouse, office), telecoms site, parking lot, tier 2/3 city	Tier 2/3 city[14]
	Number of Racks	0	0-4 racks	5-20 racks	20+ racks
	Power	Up to 1 kW	Up to 20 kW	Up to 200 kW	Up to 4000 kW
	Tenancy	Single Tenant	Single Tenant	Single Tenant / Multi-Tenant	Multi-Tenant
	External Environment	Controlled (within Device), Harsh & Rugged	IT Closet, Commercial & Office, Harsh & Rugged	Harsh & Rugged, Commercial & Office, Conditioned & Controlled	Conditioned & Controlled
	Passive Infrastructure	May or may not have power and filtration, no cooling, etc.	Has power with limited cooling and filtration, etc.	Tier 1+	Tier 3+
	Edge Infrastructure Provider	Device manufacturer or in-house solution within enterprise / government	Hardware OEM, data center provider, telecoms operator or in-house solution within enterprise / government	Colocation provider, hyperscale cloud provider (public cloud), telecoms operator	Colocation provider, hyperscale cloud provider (public cloud)
	Expected Deployments	Millions	Hundreds of thousands	Thousands	Hundreds

** by 2030 per major region*

Identifying the appropriate edge infrastructure model depends on the use case being deployed. Since similar use cases often have similar requirements it can be helpful to start by identifying the edge archetype.

Typically, the lower the required latency, the closer the edge infrastructure must be to the end device. For this reason, **Life Critical** use cases often need to be hosted at the **Device Edge**, while **Data Intensive** use cases are often hosted on-premise at a **Micro Edge**.



KEY:

Infrastructure model that is typically deployed today

Infrastructure model that we see being most deployed going forward

- **Data Intensive:** As Data Intensive use cases require the edge to be close to the source of data to prevent high bandwidth costs, on-premise deployments are desirable. A Micro Edge provides a good balance of short data transmission distance (thus limiting bandwidth costs) and greater compute capabilities than a Device Edge.
- **Human-Latency Sensitive:** The Human-Latency Sensitive archetype is dominated by consumer applications (e.g., website speed optimization[15]) for which an on-premise edge solution isn't an option. For this reason, most Human-Latency Sensitive use cases today are hosted at Regional Edge Data Centers. However, as latency needs move to the sub-10 millisecond range and edge data centers become more available at access[16] locations, Distributed Edge Data Centers will be a favorable option. Human-Latency Sensitive applications for business (e.g., AR/VR) are typically hosted on the Device Edge today to meet latency requirements but will move to on-premise Micro Edge as these are increasingly deployed by enterprises.
- **M2M Latency Sensitive:** Machines can process data much faster than humans, therefore speed is the defining requirement of Machine-to-Machine Latency Sensitive applications. The Device Edge meets these latency requirements, but there will be a move to the Micro Edge as enterprise edge adoption becomes more widespread, particularly for machine-to-machine devices that are too small or low cost to justify a Device Edge. For example, in manufacturing, vendors are putting compute on the factory floor itself. A small edge device in a self-contained enclosure with built-in power and cooling.
- **Life Critical:** Low latency is critical for these use cases since they directly impact human health and safety. The Device Edge provides the lowest latency; therefore, many Life Critical use cases rely on this model.

[14] Tier 2 and 3 cities often have a population below 1 million and will rarely have an internet exchange/peering point within the city yet, nor a hyperscale data center. Examples include Austin in the US or Berlin and Milan in Europe.
[15] Website speed optimization uses edge computing to decrease the load time of a page. Many ecommerce providers have experienced negative impacts on revenue when sites are slower, and Google observed that a 500 millisecond delay to page response resulted in a 20% decrease in traffic.
[16] When the edge sits at access locations, it is at sites or points of presence owned by a telco operator (e.g. cell towers, central offices, or an ISP's node). LF Edge established the access edge in their edge continum.

In practice, enterprises consider other factors alongside their use case requirements when making infrastructure decisions. These important considerations include:

- **Environment:** Temperature, pollution, and presence of particulates all have an impact on the infrastructure that is required (e.g., the degree of cooling and filtration). The noise produced, including electrical noise, must also be considered especially if the space doubles as an office. For example, communications cables cannot be run near elevator shafts.
- **Use case:** The quantity and speed at which data must be processed influences how close to the end-device the compute must be. The type of workload (i.e., compute intensive versus storage intensive) also impacts the edge infrastructure, as more compute intensive workloads (e.g., high-definition video) require more power and therefore more cooling.

5G will accelerate edge adoption

5G will be a significant factor in determining edge adoption, since the deployment of 5G acts as a catalyst for the move to the edge. The regions that are further ahead with their rollout of 5G (North America, Europe and East Asia) will therefore be at the forefront of edge adoption. To learn more about how edge use cases will benefit from 5G, see Vertiv's previous research.

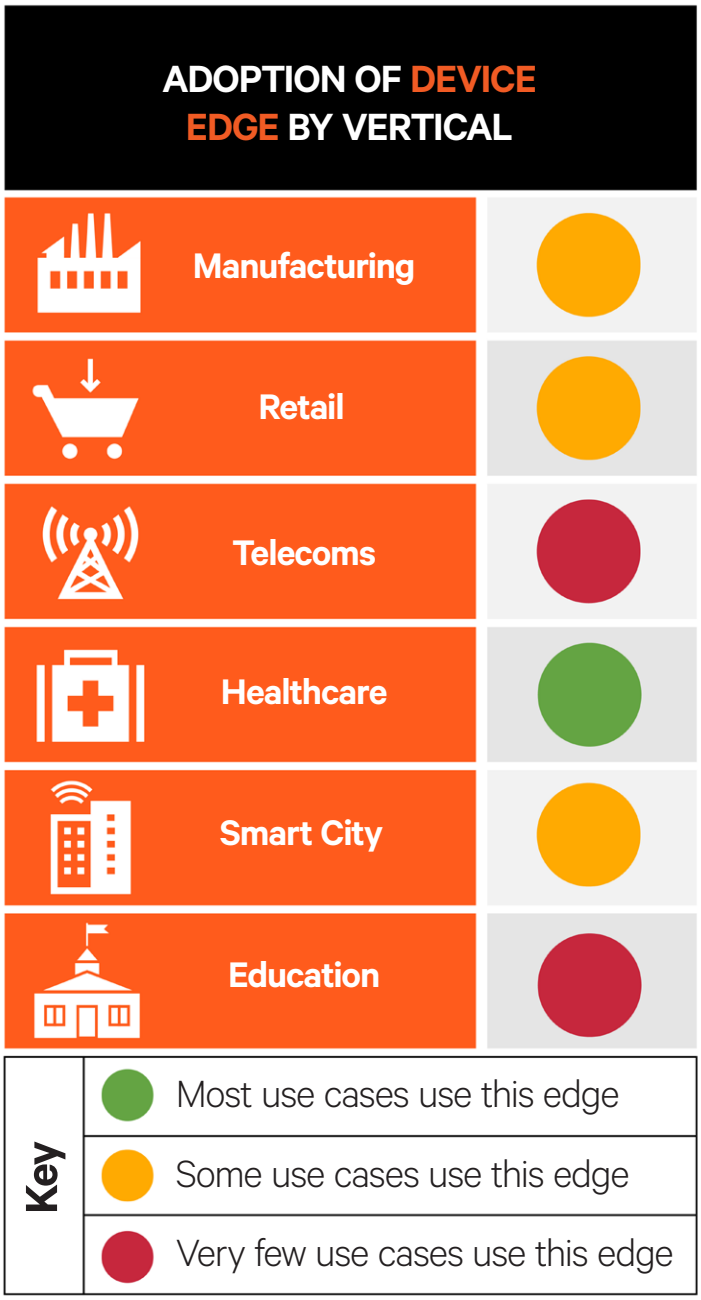
> 5G is starting now and will take 3-5 years in the large developed markets. We think that will accelerate the path to the edge.
>
> **VP Innovation,**
> **Leading Tower Company**

> It's a challenge because these offices were never meant to house IT gear, so we have to go in there and update the electrical. Now we're producing heat in the space, so we have to talk cooling. Especially if it's a space where people are working; we don't want to overheat them, and we don't want to make it too noisy for them either.
>
> **Technical solutions architect,**
> **World Wide Technology**

- **Legacy equipment/infrastructure:** The decision to deploy edge infrastructure in an existing data center versus creating a new, stand-alone deployment ultimately depends on whether an existing legacy data center already exists. For a Micro Edge, the specific shape of the infrastructure is driven by the space into which it must fit (e.g., if there is insufficient floor space, the infrastructure must be wall-mounted).
- **Enterprise operations:** The choice between upgrading an existing on-premise data center and introducing a new stand-alone deployment also depends on if the enterprise can afford the downtime required to upgrade its existing infrastructure. Enterprises for which downtime is costly may benefit from paying a premium for a pre-fabricated data center that can be built off site and then deployed quickly.
- **Security and maintenance:** If edge infrastructure is in an exposed location where people could damage it, the enclosure must be designed with additional security. If employees will need to regularly maintain or interact with the IT gear, it must be easily accessible (e.g., not out of reach on the ceiling).
- **Communications infrastructure:** If the edge is in a remote location and the infrastructure isn't there to transport data over the network (e.g., mining, agriculture), a more robust on-premise solution is needed.

Navigating the Edge Infrastructure Models: Key Recommendations

Device Edge



Use cases that leverage a Device Edge include those in the Life Critical archetype such as drones, autonomous vehicles, robotic surgery, and in-hospital patient monitoring. A Device Edge is suitable since it can meet the mobility requirements for a device, such as a drone, to navigate autonomously within the context of the environment it is travelling through. It also provides ultra-low latency, which is necessary for Life Critical use cases. Finally, it allows some aspects of the use case to function (e.g., navigation, local alarms) even when connectivity is not available due to limited coverage or network failure.

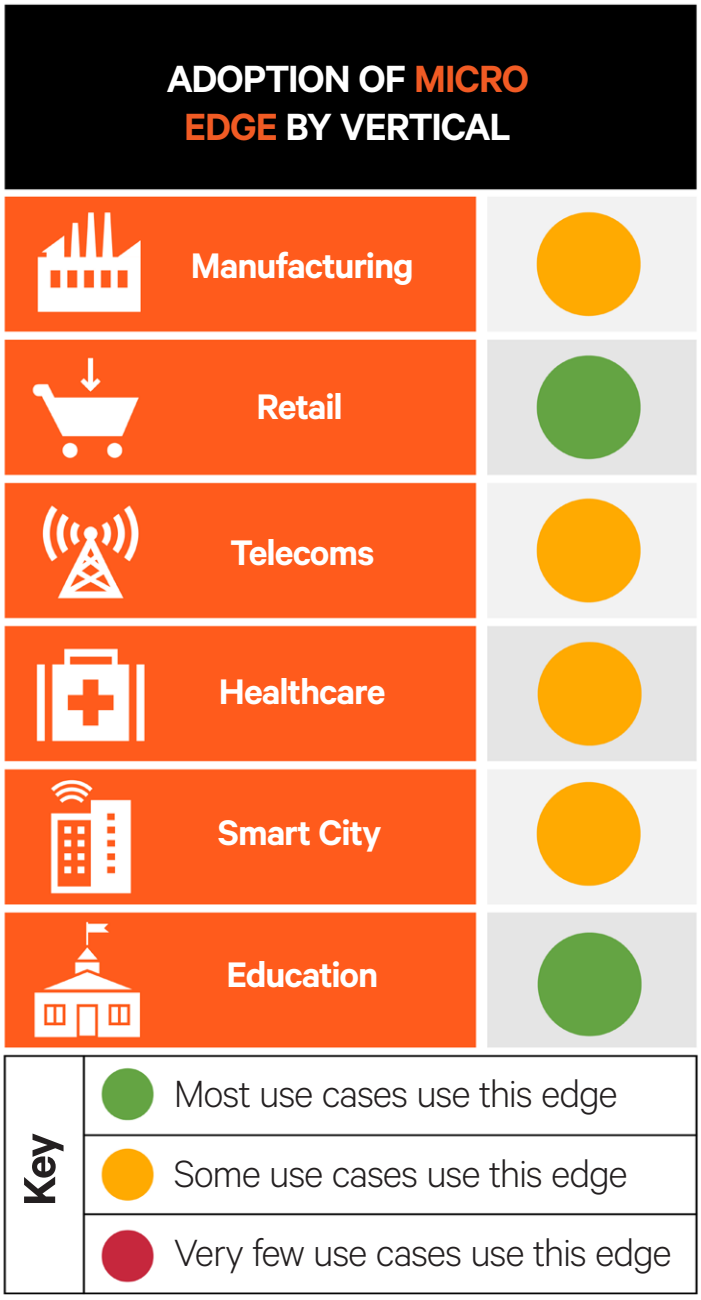
As a result, healthcare is one of the sectors with a high adoption rate of Device Edge since many use cases will need to detect life-threatening situations quickly and reliably, whether the patient is in the hospital or being cared for remotely. The manufacturing sector also has Life Critical use cases, which is why machine control systems mainly run on the equipment itself (a form of Device Edge).

Key recommendations when deploying a Device Edge:

- Add-on Device Edge is more appropriate for retrofitting legacy equipment, but greenfield deployments may want to consider embedding the compute into the device. However, these are often proprietary devices that do not lend themselves to integration with generic edge computing capabilities.
- Device Edge has limited compute capacity. Adding more compute will make end-devices much heavier, so always consider the power/weight trade-off[17] (which is of greater concern in cases where the device is battery operated or has no access to a power supply).
- Be mindful of the data collected by the end-device. Use cases, such as smart security cameras, connected traffic infrastructure, and drones, collect visual or location data about people. It is therefore important to be cognizant of the potential challenges around data privacy and sharing as this could be a contentious issue.

[17] The trade-off of power, weight and cost for AR/VR headsets is explored in Apple Glass: An iPhone moment for 5G?

Micro Edge



A Micro Edge can be located close to the source of data due to its small size and relative ease of deployment (compared to a larger data center). It therefore offers low latency and decreases the cost of data transmission, making it a suitable infrastructure model for use cases across the following three archetypes: Data Intensive, Human-Latency Sensitive and Machine-to-Machine Latency Sensitive. In space-constrained industries, such as retail or education, a Micro Edge is an attractive solution as it limits the real estate required, allowing compute to be deployed in a smaller footprint. For example, a large supermarket chain with 16,000 locations in Europe is deploying a Micro Edge in stores for local data collection and processing, and is also adding central data centers for aggregation and general IT management.

Key recommendations when deploying a Micro Edge:

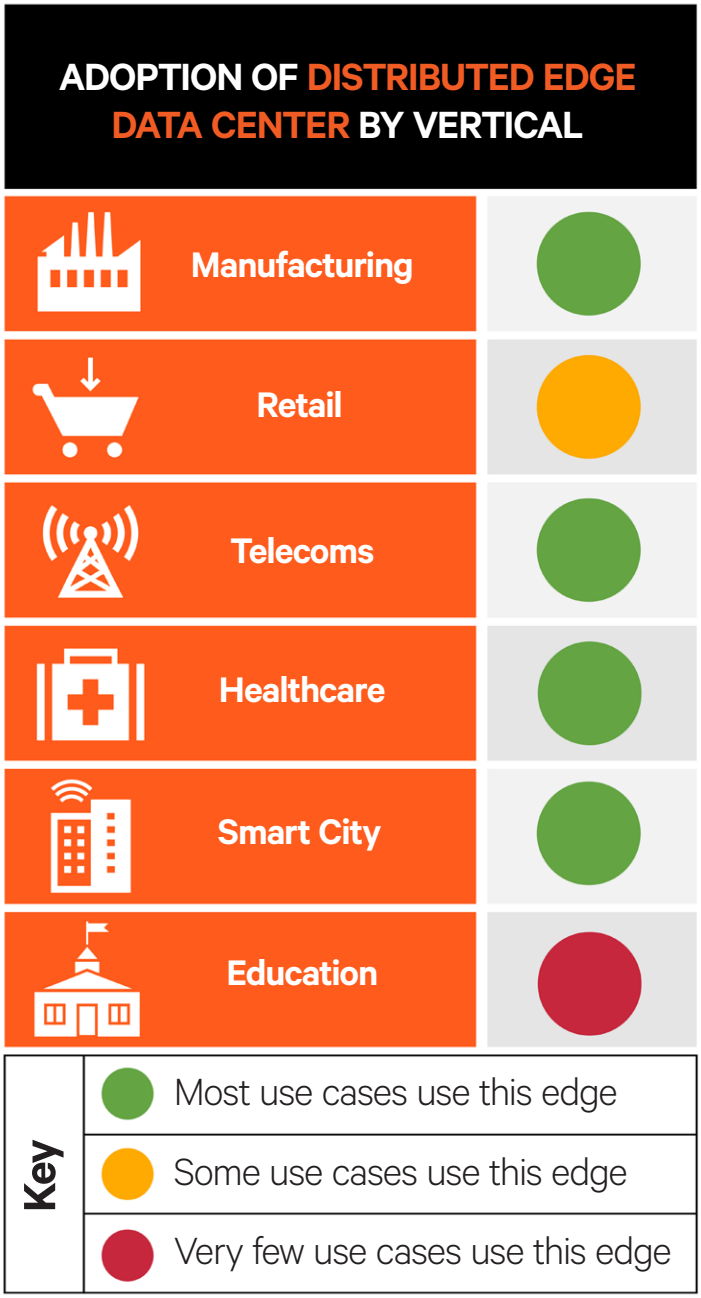
- Consider the space available (it may be necessary to attach to the walls or ceiling), the function of the space (if customers or workers will be present), and the security requirements (when the infrastructure is easily accessible, a physical layer of security is necessary). Micro edge deployments often cover areas with different electrical feeds, regulations, site access (e.g., elevator height), site control (store manager, plant manager) and technical expertise.

> Physical and virtual infrastructure have to be coordinated together, otherwise it simply will not work.
>
> **Jon Abbott, Technologies Director, Vertiv**

- If decisions around software, hardware, and infrastructure are made by different stakeholders, maintain alignment between these stakeholders so the decisions are made in parallel, not sequentially, as this results in a more successful solution.
- Select your equipment type. Hardened equipment is made for less controlled environments, so it can withstand 122 degrees Fahrenheit. Enterprises can use generic, commercial off-the-shelf (COTS[18]) servers instead, which are cheaper, but the shelf life of these servers is greatly reduced when they run above 86 F. While both types of hardware require an enclosure, the supporting infrastructure for COTS servers must offer greater control for temperature, humidity and power. An economic balance of standardization and location-specific tailoring is necessary.

[18] COTS - Commercial off-the-shelf products that readily available for sale and are designed to easily integrate with existing systems (rather than being custom-made or bespoke).

Distributed Edge Data Center

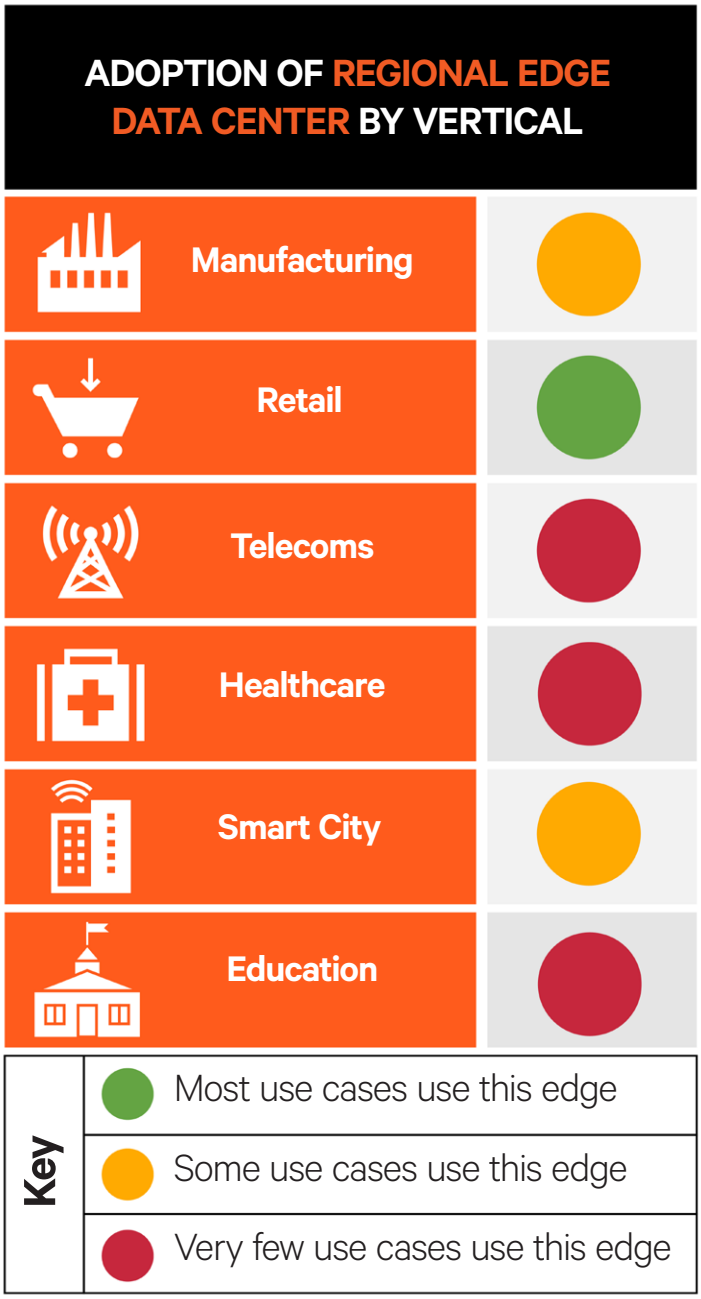


Like a Micro Edge, Distributed Edge Data Centers are located at the enterprise site and are suitable for many industry use cases as they offer low latency and decreased bandwidth costs. The research found that telcos use Distributed Edge Data Centers to host both consumer applications and their own internal network functions, which are Machine-to-Machine Latency Sensitive. Similarly, medium and large manufacturers will use these smaller data centers for their Internet of Things (IoT) use cases. For medium-sized manufacturing facilities, most of the edge infrastructure will reside within an eight-rack data center.

Key recommendations when deploying a Distributed Edge Data Center:

- There may be an investment required to upgrade an existing data center or network room, and the time to deploy the changes could have a costly impact on operations. This impact on downtime must be weighed against the cost of buying a new, pre-fabricated data center that can be deployed quickly on site.
- It is recommended to build spare capacity into the data center in order to maintain flexibility in the future, but it should be noted that overbuilding to prepare for all outcomes is costly and may not be necessary. Finding the balance of what is needed today versus what will be required tomorrow requires users to consider the evolution of their edge use case in their given industry.
- When building redundancy into the data center, consider both the value of the applications being run and the stability of the environment (e.g., in some countries the grid is unreliable, so the risk of power loss is significant).
- Sometimes it is not necessary to deploy a Distributed Edge Data Center at the enterprise site, as a “near-premise” deployment meets requirements such as those related to latency or security. This could still be owned by the enterprise or it could be a multi-tenant facility serving multiple enterprises.
- If a Distributed Edge Data Center is used as a colocation facility, it needs to have layers of security and isolation to provide this multi-tenant edge computing experience. This may include gates, locks and cameras.

Regional Edge Data Center



A Regional Edge Data Center acts either as an edge compute site or as an intermediary site where edge data is sent for pre-processing before being sent to the cloud. It satisfies both low-latency and data-intensive use cases, therefore all the edge archetypes leverage Regional Edge Data Centers. Human-Latency Sensitive consumer use cases, in particular, rely on Regional Edge Data Centers since on-premise edge deployments (e.g., low-latency media streaming or immersive gaming) are not an option.

Regional Edge Data Centers are often adopted in retail, as they can reduce the need to deploy compute infrastructure across individual retail stores.[19] In cases where the retailer has invested in individual on-premise deployments, the data center can act as an intermediary data processing site.

Key recommendations when deploying a Regional Edge Data Center:

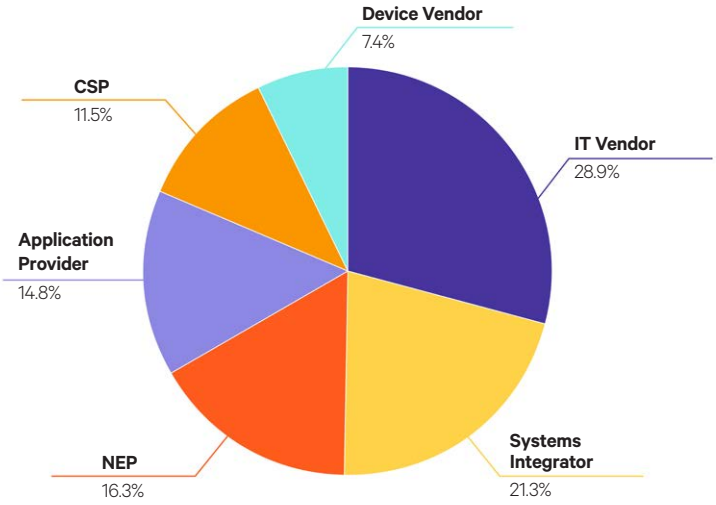
- Security and isolation are necessities (since many regional data centers are multi-tenant facilities). Customers must ensure the data center has adequate mechanisms for securing the tenant's infrastructure and data.
- Consider the specific use case when designing the edge computing infrastructure (e.g., more compute-intensive workloads will likely require more power, and therefore, more cooling).
- Location is a key consideration. If data sovereignty is a factor, data may need to be stored within the jurisdiction of the end customers. However, if the key factor is latency (< 50 milliseconds), target a strategically important location that reduces latency across as many end sites as possible. This will often be a data center that is at or very close to a major internet exchange.
- Major public cloud providers are extending their public cloud to local data centers (e.g., AWS Local Zones), which will allow enterprises to distribute their cloud applications more easily. However, there are two key considerations: public cloud providers are at an early stage of deploying these local clouds, and certain applications (and the data) will not be suitable for storing and processing on a public cloud (partly due to government regulations).

[19] According to the International Data Corporation's (IDC) Worldwide Edge Spending Guide, retail is the second-largest and fastest-growing industry in the European enterprise edge market.

It Takes an Ecosystem to Build the Edge

Infrastructure is only one piece of the puzzle for any organization looking to implement edge-enabled solutions. There are many elements that influence the building of the edge — software, hardware, infrastructure, orchestration, management, etc. — and enterprises will struggle to coordinate these elements on their own.

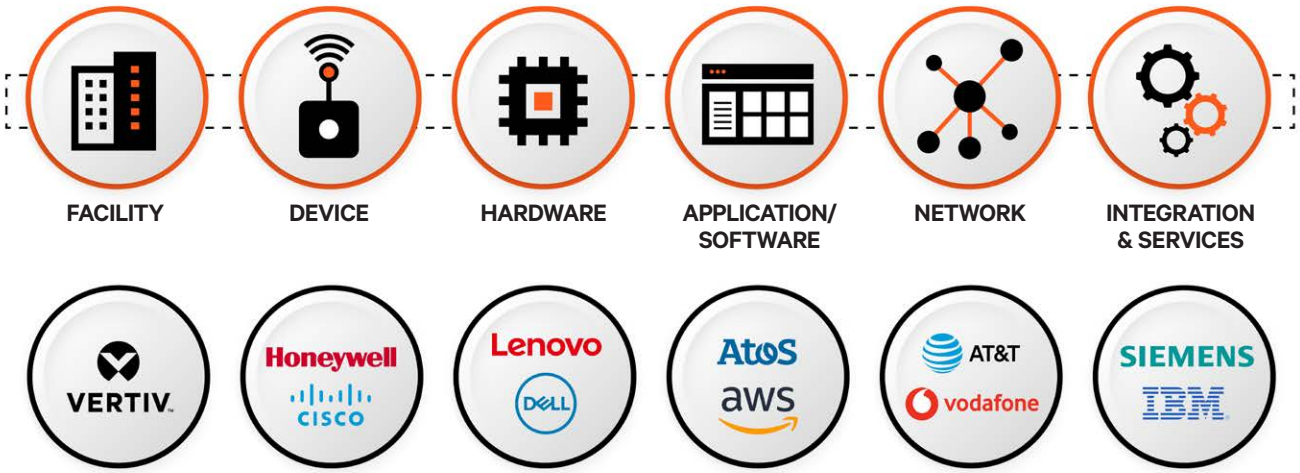
There is no one-size-fits-all with regards to how enterprises want to purchase these next generation information and communication technology (ICT) and edge solutions. Research shows 34% of enterprises prefer a do-it-yourself approach whereby they select different components from each vendor. The 66% who prefer entire solutions to come from one lead vendor, vary in terms of who that primary vendor is: IT vendor versus systems integrator versus network equipment provider, etc.



Source: STL Partners survey with 699 industry professionals globally, May 2021

Whether it is an enterprise building its own edge or a service provider deploying edge infrastructure to run applications or allow others to run workloads, collaboration with other players in the edge computing ecosystem is essential to success. Building strong relationships with industry specialists (e.g., Siemens or Honeywell in manufacturing) ensures solutions meet vertical-specific needs and can integrate successfully with existing systems and infrastructure.

The edge computing value chain:



Conclusions and Recommendations

Edge infrastructure remains a complicated topic (as interviews with industry professionals demonstrate). Nonetheless, the edge infrastructure model framework defined in this report can help enterprises navigate the array of edge solutions available and provide guidance on appropriate infrastructure choices.

Looking beyond the edge infrastructure models, Vertiv recognizes that there are complexities associated with the practical task of building edge infrastructure that are unique to each enterprise. An interactive web tool has been developed to enable enterprises and other data center operators to explore key use cases in depth. Organizations will be able to better understand associated workload and infrastructure characteristics, and inform decisions on infrastructure design, build and deployments.

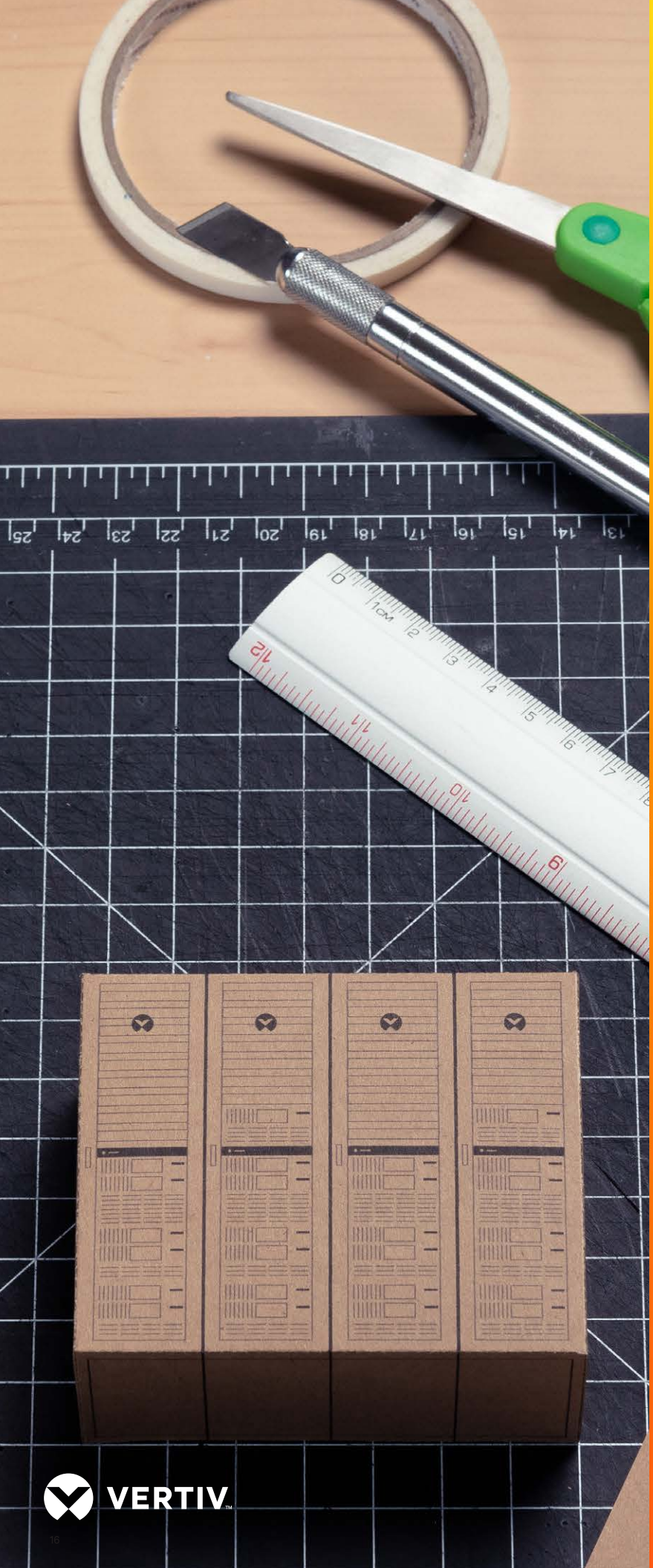
Other key recommendations include:

Enterprises

- **Identify an anchor use case.** There is still uncertainty on the nature of future use cases and their edge computing needs. The first use case must justify the business case for an initial build, so ensure you have a good understanding of why your use case needs edge computing. Understanding which of the workload characteristics is the key driver for edge deployment (e.g., latency, bandwidth, security) will also help inform decisions around infrastructure.
- **Be prepared to manage a variety of edge infrastructure models**. For example, many retailers opt for a Micro Edge in their stores then supplement with a Distributed Edge Data Center close to the stores that can filter and aggregate data from across locations, sending only the necessary information to the cloud.
- **Try not to define a single blueprint for all scenarios.** Even within model types, there will be variation given that different locations have different legacy environments. Companies with multi-national coverage will face geographical differences in climate, pollution, power supply, regulation, etc. (e.g., the EU regulates the number of decibels permitted which could limit infrastructure with fans or switching).

Solution Providers

- **Futureproof edge infrastructure.** Understand the use cases that customers are adopting now and plan to adopt in the future, and build in spare capacity (storage, compute, etc.) as appropriate. Adopting more flexible deployment models will reduce risk.
- **Work with the ecosystem.** Edge is not a single product to be sold by a single vendor, but a solution that multiple ecosystem players build together. Solutions should therefore be standardized, so it is easy for customers to use a solution as a component. Partnering is also important, particularly when looking to meet highly industry-specific needs.
- **Consider new economic models.** Replicating what was done with cloud is not possible. Edge infrastructure has specific needs, so it is important to consider the economic models that ensure power, cooling, security, and space are optimized with new ways of guaranteeing economies of scale.



Appendix: Glossary

ACCESS EDGE	An edge location within the telco network that connects subscribers to the main carrier's backbone network, then onto other networks, the Internet, and hyperscaler clouds.
COLOCATION FACILITY OR SERVICE	A colocation facility, or "colo," is a data center facility in which a business can rent space for servers and other computing hardware. Typically, a colo provides the building, cooling, power, connectivity to others or the internet, and physical security, while the customer provides servers and storage.
CONDITIONED, CONTROLLED ENVIRONMENT	Environments with dedicated systems in place to control for various factors, including temperature and humidity, dust particulates, pollution, etc.
DATA CENTER	A physical facility that organizations use to house their critical applications and data. A data center's design is based on a network of computing and storage resources that enable the delivery of shared applications and data. The key components of a data center design include routers, switches, firewalls, storage systems, servers, and application-delivery controllers.
EDGE COMPUTING	This physical compute infrastructure is positioned between the device and the hyperscale cloud, supporting various workloads. Edge computing brings processing capabilities closer to the end user/device/source of data, which eliminates the journey to cloud providers' data centers and reduces latency.
FORM FACTOR	Overall design and functionality of hardware systems.
HYPERSCALE	In computing, hyperscale is the ability to achieve massive scale, especially for big data and cloud computing. Today, AWS, Azure, and Google Cloud are considered "hyperscalers."
IT/NETWORK CLOSET	A closet or a small room where electrical wiring and computer networking hardware is installed.
MULTI-ACCESS EDGE COMPUTING (MEC)	Type of network architecture that provides cloud computing capabilities and an IT service environment at the edge of the network.
ON PREMISE	Also known as "on premises" or "on-prem," this refers to technology that is hosted within the physical confines of an enterprise's own site.
PREDICTIVE MAINTENANCE	Process of monitoring data from equipment sensors to ensure it is in good condition and to pre-emptively flag if there is a need to repair it, possibly eliminating the need for scheduled maintenance.
RUGGEDIZED HARDWARE	Hardware that is designed specifically to endure challenging environments such as outside pollution, high or low temperatures, humidity, etc.
STAND-ALONE	Able to operate independently of other hardware or software.
TELCO BASE STATION	Transmission and reception station in a fixed location, consisting of one or more receive/transmit antennas, microwave dish, and electronic circuitry, used to handle cellular traffic.

PARTNERS This research report was developed with the support of STL Partners

Vertiv.com | Vertiv Headquarters, 1050 Dearborn Drive, Columbus, OH, 43085, USA

SL-70825 (R10/21)