STATE OF THE
EDGE
2020

A Market and Ecosystem Report for Edge Computing

www.stateoftheedge.com
# Table of Contents

- Introductions ........................................ 3
- State of the Edge 2020 ................................ 8
- Defining the Edge .................................... 16
- Sizing the Market ................................... 19
- A Tapestry of Use Cases ............................ 26
- Postcards from the Edge ............................ 41
- Resources .......................................... 56
- Membership ......................................... 72
- Credits ........................................... 74
Foreword

Embracing The Third Act of the Internet
By Arpit Joshipura, The Linux Foundation

We stand on the precipice of a profound re-architecting of the Internet called edge computing, which will impact all areas of society. The 2020 State of the Edge report calls this the Third Act of the Internet, and The Linux Foundation will play a critical role.

Founded almost two decades ago, The Linux Foundation is the largest and most pervasive open source organization, dedicated to building sustainable ecosystems around open source projects to accelerate technology development and industry adoption. Edge computing is a natural extension of the work we have always done.

In January of 2019, we launched LF Edge, an umbrella group within The Linux Foundation tasked with building and organizing open source frameworks for edge computing. As of this writing, LF Edge has seven projects that span all elements of the edge computing ecosystem. Key projects include the Akraino Edge Stack, EdgeXFoundry, and the Open Glossary of Edge Computing.

Many people don’t know that the Open Glossary began its life as an appendix in the original 2018 State of the Edge. The Glossary quickly became a go-to resource for the early edge ecosystem, and its creators began looking for a long-term home for the project. Recognizing the importance of documenting and standardizing the language around new technologies, we at The Linux Foundation invited the Open Glossary to be a founding project of LF Edge. Now in its 2.0 edition, the Open Glossary has created a fertile environment for educating broad audiences about edge computing while also advancing the terminology used by practitioners.

I have little doubt that edge computing will change how we sense, evaluate and control our environment. It will impact cloud, robotics, artificial intelligence, healthcare, manufacturing, data centers, mobile devices, smart cities, and autonomous vehicles—and bring forth a wave of innovation that will dramatically improve our lives.

In this wave of innovation, open source will play a key role, and edge computing has already begun its journey as an inherently open ecosystem. At The Linux Foundation, we are dedicated to partnering with organizations like State of the Edge to amplify collaboration, innovation and thought leadership across the entire edge ecosystem.

Welcome to the Third Act of the Internet.

Arpit Joshipura
General Manager, Networking, Edge, and IOT
The Linux Foundation
From the Editors

Just over a year ago, State of the Edge came to life in a conversation over beers. It was a heady time. Edge computing was just emerging into a mainstream niche. Thought leaders and futurists were staking claims, vendors were slapping “edge” onto their products to stay relevant, and very few people even had agreed-upon definitions for even the most basic terms in the industry. The joke at the time was: If you ask 100 people to define edge, you’ll get 112 different answers.

It was in that environment that we birthed the original State of the Edge as a vendor-neutral white paper on edge computing researched and published by a small but passionate group of companies. The organization had an editorial mission. We sought to align and educate, not generate leads or advance people through a sales funnel. It was an experiment, but it worked.

Companies, non-profits, journalists, analysts, and the community at large applauded our efforts, embraced our neutrality and lent their support. Encouraged by the reception, we redoubled our efforts this year, as we prepared this 2020 report.

Since last year:

- Membership in State of the Edge has tripled.
- We published our first single-topic report Data at the Edge.
- The Open Glossary of Edge Computing has become the standard-bearer of a shared edge lexicon. It is now in its v2.x version as an open source project to anchor The Linux Foundation’s LF Edge.
- The Edge Landscape has also become an open source project under the auspices of the LF Edge’s Open Glossary

A Forecast for the Edge

One of the most consistent questions posed by our peers has been a very simple one: how big will the edge be? When will it explode? These are important questions. The answers impact our business strategies. How fast we should move and how much we should invest depends in large part on what we expect to get in return.

Business planners responsible for edge computing turn to expensive market research reports to help make decisions, hoping that someone else has divined the future. Some of these reports are of dubious quality, but you can’t really determine that without dropping a few grand to get a copy. And when you examine these reports, frustration often ensues as you realize that each analyst reaches wildly different conclusions, derived from different models, using different definitions of edge.

State of the Edge does not seek to replace the dozens of third party market reports, many of which are surely worth the fees they request. Nonetheless, we do see an opportunity for a community-supported research model, especially to fill gaps in the available research.
A Focus on Data Centers and IT Infrastructure

We see this year’s report and the model we have constructed as the continuation of a robust conversation. We’ve studied the market and put our best thinking into this report—but we also welcome feedback, comments, and suggestions. We invite you to join the conversation, improve upon or add to the content we present, and help State of the Edge continue to advance the industry with well-researched, vendor-neutral thought leadership.

There is No Finish Line

Edge computing represents a long-term transformation of the Internet that could take decades to fully materialize. This year’s State of the Edge report and its forecast model do not represent final answers; instead, they represent an early start to a robust conversation. We’ve studied the market and put our best thinking into this report—but we welcome feedback, comments, and suggestions. Please join the conversation, consider building upon or adding to the content we present, and help State of the Edge continue to advance the industry with well-researched, vendor-neutral thought leadership.

We look forward to the year ahead.

Matt Trifiro
CMO, Vapor
Report Co-Chair

Jacob Smith
CMO, Packet
Report Co-Chair
Executive Summary

This 2020 edition of the *State of the Edge* aims to explore the edge computing world with an eye toward tomorrow. The report includes empirical research, market sizing, thoughts from industry leaders, and the latest from our collaborative community.

Among the findings are:

- **Over $700 billion in cumulative CAPEX will be spent within the next decade on edge IT infrastructure and data center facilities.** This includes edge equipment at access network sites (e.g. radio base station tower sites), pre-aggregation sites (e.g. street-side cabinets), and aggregation and central office sites.

- **The edge is the "Third Act of the Internet", building upon previous phases of regionalization and origination.** While there remain significant unknown challenges at the edge, substantial business opportunities have emerged.

- **The Edge today is a solution-specific story.** Equipment and architectures are purpose-built for specific use cases, such as 5G and network function virtualization, next-generation CDNs and cloud, and streaming games.

- **Edge computing is a global phenomenon.** Asia Pacific currently has the largest edge computing equipment footprint of all the global regions, estimated to be 187MW today. Buoyed by markets like Australia, Japan, Korea and China

- **The deployed global power footprint of the edge IT and data center facilities is forecast to reach 102 thousand MW by 2028.** Datacenters, servers, local breakout and middle mile architecture are areas of intense investment.

- **A common definition of the edge is gaining momentum.** Community efforts including the LF Edge foundation, the Glossary of Edge Computing, and the Edge Computing Landscape provide cohesion.
STATE OF THE EDGE 2020
Cast your mind back a decade to circa 2010. The seeds for the cloud revolution have just taken root. The business models and technologies of public clouds have not yet proven themselves, but the claims are being staked out, operational models are being established, and use cases are getting refined. In many respects, the cloud of a decade ago parallels the edge computing of today.

As an extension to the Internet, edge computing promises to fundamentally transform the applications we build. It is the keystone technology that will enable a rich portfolio of innovative services, from wireless augmented reality to autonomous vehicles. It will catalyze billions of dollars in network upgrades, including the 5th generation of cellular wireless infrastructure, or 5G. It will enable new classes of real-time analysis, supporting billions of devices that collectively consume and produce zettabytes of data.

The edge also introduces a great deal of business complexity, as it requires a wide range of stakeholders for site real-estate, connectivity, IT infrastructure, app development, traffic delivery, service management, and orchestration. It will bring about a rich tapestry of hardware and software solutions, networking architectures, and business relationships that seek to address the seemingly endless use cases that are being pursued.

Many opportunities exist for new entrants, as it’s still early days for the edge. The precise configurations of the edge are still evolving, and its implementation and operational models have yet to mature.

The Internet, the Cloud, and the Edge

The vast majority of edge applications will be woven into the fabric of the internet using emerging connectivity systems, including 5G. These edge applications will leverage the cloud for storing and processing data, provisioning resources, and optimizing business models that offload CAPEX spending. Developers will build edge applications using modern principles, such as continuous integration and continuous deployment (CI/CD), and providers will emerge to serve them, making it easy and cost-effective to deploy at the edge. In fact, the first “killer apps” for edge computing may be ease of use and economics.

There will always be some edge applications that are air-gapped or disconnected from the internet, or which run with limited connectivity. These applications, whether in service of the military, a cruise ship, or a remote oil derrick, will represent only a small percentage of edge applications. The vast majority of edge applications will support (nay, depend on) the internet.

^ A zettabyte is one trillion gigabytes.
Why Now?

Since the 2018 State of the Edge Report, edge computing has started to go mainstream. One of the main reasons edge computing has captured the imagination of multiple industries is the excitement over new kinds of applications that experts expect to be delivered. Many of these applications will be impossible or cost-prohibitive to deliver with today’s infrastructure. In particular, the latencies are too high. However, by re-building the Internet to support edge computing we can bring about a revolution in data-rich, low-latency applications.

Unique to this moment in time, powerful forces have conspired to create the need and the opportunity for edge computing. In particular:

- The speed of machines is replacing that of humans;
- Billions of devices are being attached to the Internet;
- Highly automated infrastructure and clouds have become commonplace;
- People increasingly expect more multidimensional experiences.

Extending the capabilities of today’s internet to support tomorrow’s applications will require a significant re-architecting of the infrastructure that has gotten us this far. In particular, we will need new approaches, equipment, and architectures deployed to the middle and to the last mile. As we look to enable applications that require lower and lower latencies, for example, we will need faster networks and servers in closer proximity to the end user or device.

1 A zettabyte is one trillion gigabytes.
From Human Speeds to Machine Speeds
The original internet was designed for humans: humans loading web pages, humans reading email, humans watching movies. This is why today’s internet—while fast enough for most humans—appears glacial when machines talk to machines. Take for example a robotic drone flying at 60 miles-per-hour. At this rate, it will travel the length of a football field in 4 seconds. Avoiding collisions may require decision-support from an edge server and a delay of 100ms could cause it to crash into something 10 feet away.

As more and more machines come online, businesses will seek to apply the power of server-side processing to their behaviors. This will require an edge-enabled internet that operates at machine speeds. As we re-architect the Internet to reduce latency and jitter on the round-trip path, we create a platform to satisfy the demand for machine-speed edge applications.

From Millions of Devices to Billions of Devices
We are living in a world that already has billions of cell phones, each with a network connection. Increasingly, every car, light socket, factory robot, wall switch, and appliance will have a network connection too. Most electronic devices and systems—whether they operate in cars, in homes, in factories, on street lights, on street corners or on our persons—will have a connection to the internet. These devices will collectively generate zettabytes of data, and edge computing will activate that data and put it to work. Edge computing will place processing power close to the data, thereby enabling lightning-fast analysis in near real time. It will also obviate much of the delays, costs, and complexities of sending all the collected data back to a centralized location for processing.

From Fixed to Automated Infrastructure and Clouds
Developers now have many abstraction layers for building and deploying applications worldwide, often on machines they never touch or see, such as those owned by the public cloud providers. These abstractions give developers and enterprises the convenience and economics of provisioning compute, storage, networking, and other services—nearly instantaneously—from a web browser, a command line, or even programmatic system. As infrastructure automation and cloud moves to the edge, they will further abstract the complexities of distributed architectures. As more developers become familiar with cloud automated infrastructure, and more enterprises adopt modern continuous-delivery models of development, the edge will become a natural and common way build and enhance applications, much like CDNs are today. These trends will accelerate the edge.

From Flat Experiences to Multidimensional Experiences
Anybody who has used an Oculus headset or played Pokemon Go can see how we are moving from a world dominated by the flat, static, two-dimensional internet into a world filled with rich, multidimensional experiences, such as virtual reality, augmented reality, and tactile applications. These multidimensional experiences will extend far beyond gaming. For example, they may include remote and robotic surgery, which will depend on presenting complex, real-time feedback to the surgeon. Any lag in this experience will drastically impact the surgeon’s ability to precisely control the outcome. These new experiences will demand the low-latency capabilities of edge computing.
The Third Act of the Internet

Much like a classic three-act play, the history of the modern internet can be separated into three key acts: origination, regionalization, and (now) edge computing.

Act I: Origination
In its original form, the internet was mostly a network. There was no “cloud” to speak of, and the most important thing was to get data from point A to point B, making it possible to connect any web browser to any server by routing packets across a shared system of networks. This architecture worked well enough to establish the internet, and for many years provided enough performance and resources for the applications we desired. As the cloud emerged, much of the internet terminated in large, centralized data centers, such as those owned by the major cloud providers, and requests for the applications and data occupying those data centers would often have to traverse thousands of miles and be subject to delays from network hops, best-effort-routing, and the speed of light.

Act II: CDNs and Regionalization
With the advent of the modern web browser and new higher-speed connection technologies, such as DSL, the internet continued to grow exponentially through the 1990s. With this growth came demand for high-fidelity websites, streaming movies, and instant page loads—and with these demands, the internet as it was originally constructed began to fail.

To solve these problems, a second tier of internet infrastructure was deployed, in the form of regional data centers and Content Delivery Networks (CDNs), both of which brought internet infrastructure closer to the users. The performance advantage of this infrastructure evolution was clear: By positioning vital internet infrastructure closer to its users, large volumes of requests were able to remain within a single region and often did not need to transit across the country—let alone oceans—in order to retrieve essential content. This saved time and expense while improving the user experience.
This regionalization of the internet created a second tier of infrastructure, and brought about significant performance and scaling enhancements, and also helped drive the expansion of 3G and 4G cellular data networks. However, as the internet and its users continue to evolve towards very high bandwidth, low latency use cases, a further step is, which brings us to the third act.

Act III: Edge Computing

Where we stand today is at the edge. The third act of the internet seeks to resolve the problems with our current infrastructure and the challenges that come with supporting the applications we desire. To meet the current and future needs of a diverse set of users, from consumer to commercial to connected public safety and other life-critical use cases, our model for deploying infrastructure needs to go beyond regional deployment and operation, and also be deployed more locally.

How Will the Edge be Deployed?

No amount of simulations and laboratory tests can identify all the challenges in operationalizing the edge at scale. Many of the industry pioneers who are delivering commercial solutions today have experienced the hard knocks of being early to market. Many have pivoted their edge computing strategies from technology trials into commercial operations as the market realities have emerged. This is a testament to the agility of industry pioneers and the fluid market dynamics that stakeholders must navigate to succeed with edge computing. These same pioneers believe that they have created sustainable commercial differentiation through early market learnings.

It would be naïve to assume the industry has edge computing figured out. A quick survey suggests ample room for innovation, both technical and commercial. For example, software for workload management and orchestration may very well become a significant category for edge computing, particularly as edge computing implementations become more distributed and complicated.

Many edge business opportunities will emerge once the infrastructure has been deployed. In the interim, though, there are substantial business opportunities in providing enabling technologies that will form the substrate of global edge infrastructure. A few worthy of note include data centers, servers, local breakout and middle mile architecture.

Data Centers

During the Third Act of the Internet, vast numbers of edge data centers will be deployed in diverse locations. Often, they will take the form of factory-built micro modular data centers, such as those resembling shipping containers. These micro modular systems can be trucked to a location, craned onto concrete pads or piers, and be set up and made operational in hours or days, whether in the parking lot of a factory or at the base of a cell tower. In other cases, infrastructure owners will repurpose existing structures. For example, the Open Networking Foundation and its partner companies are funding the CORD project, which stands for Central Office Re-architected as a data center, and which seeks to convert existing central office structures into edge data centers.
In contrast to hyperscale data centers used for the public cloud, the physical locations of edge data centers are important and depends on a variety of factors including the connectivity bandwidth and latency, and the security demands of the services that the edge infrastructure supports. To achieve this, edge data centers are physically distributed, as opposed to hyper-data centers, which a highly centralized. The power ratings of hyperscale data centers are 10’s to 100’s of MW, while the power ratings for edge data centers are 10’s to 100’s of kW.

For large amounts of traffic, the circuitous trip back to the core can be avoided by building edge data centers and deploying the topology of the network and its functions at a fundamental level, through local breakout and other technologies. One data center serving a large region can be augmented with hundreds of edge data centers—which now become disaggregated network nodes performing key network functions at the very edge of the network, closer to their users.

**Servers at the Edge**

Most edge applications have devices in the field that connect to application servers, where much of the processing occurs. These servers will be located somewhere, most likely in a facility that approximates a data center, with robust Internet and device connectivity.

As practitioners deploy edge applications, the decisions they face about servers will be similar to the decisions they face today: who owns them, where do they live, and how will they connect to the devices?

At the outset, the servers that support edge applications will often be owned by end-users, deployed on premises, and connected to devices over a private network. They will require the end users to purchase and operate all of the equipment—including the servers, the devices, and the networks. This infrastructure to support edge is nascent and innovative enterprises may have to roll their own until it matures. Until there are public “edge clouds” and ubiquitous high-speed networks (e.g., 5G), custom edge installations may be the only way to reliably implement an edge application.

However, as the demand for edge applications grow, the cloud will drift to the edge. Edge computing will become part of the standard internet topology. Aggressive companies are already deploying modern edge infrastructure. Designed to be rented, leased, and sold to end users, this edge infrastructure will become the commons substrate upon which we build modern applications. Cloud-like business models will emerge as developers begin to create (or migrate) edge applications onto shared infrastructure, compensating service providers on a pay-as-you-go basis.

In the not too distant future, a developer might provision a cloud instance in Chicago East or Chicago North in the same way that they, today, can provision one in US West or US East. Much of the edge will become part of a cloud-like experience, delivered on a much more fine-grained basis.

**Local Breakout**

It would be an onerous task to completely re-architect today’s networks to reduce network round-trip delays. For example, in today’s 4G LTE networks, it’s common for all data to route through a centralized facility before being handed off for processing. This legacy network topology would defeat any attempt to do processing locally, as the data would need to “trombone” out to the centralized facility and then be returned for processing, which will typically add dozens to hundreds of milliseconds to the round-trip time.
But incremental steps can enable edge computing in data centers that are within a few milliseconds of the access edge of the network by doing what's called “local breakout” of traffic at the network edge. That is, some traffic at the network edge in the same Access Point Name (APN) is broken out at the network edge while the rest goes back to a central EPC (Evolved Packet Core). Historically, this could not be done—you could either offload all the APN traffic or exactly none. But we are now seeing new techniques, such as SGW-LBO (part of the ETSI MEC architecture), which effectively replicates the 5G user plane function in the 4G EPC.

**Middle Mile Architecture**

One less-visible factor that also impacts edge computing is the middle-mile network infrastructure. This is the system of network infrastructure that connects the last-mile network itself upstream to an aggregation point, such as regional or national data center. To get there, large volumes of traffic must pass through many intermediate network nodes and paths across the middle-mile, which do not provide the key network functions required to make data routable.

This means that these nodes (many of which are left over from voice and earlier generations of data networks) have to be maintained, upgraded, and fixed to keep the network operational. If they happen to crash, or if they can’t handle the amount of traffic being sent, the network is down or, at best, vexingly slow for many users. Both outcomes are to be avoided by any operator.

This middle-mile infrastructure is expensive to maintain and, compared to upgrades in the core or last-mile networks, attracts comparatively little investment in many cases. To completely revamp this infrastructure at a time when network operators are being pressed to invest in the early waves of 5G deployment—especially in dense urban areas—is a tall order.
Towards A Common Language

One of the most confounding aspects of the edge computing market is its lack of an agreed-upon definition for even its most basic term: edge. The definition for edge computing lies in the eye of the beholder, offering many view through a single prism that gives rise to multiple definitions. Each is influenced by incumbent systems, powerful stakeholders, and business investments.

Each particular definition of “edge” often reflects the definer’s role in the value chains for digital services. Those engineers who build embedded systems for automobiles might argue the edge is the bumper of a car. The real estate investors who own cell towers and buildings might call those the edge. The owners of regional data centers and internet exchanges might argue their facilities are the edge.

When we talk about edge computing, we typically emphasize proximity. By positioning compute, network, and storage resources as close as possible to users—or so the story goes—latency can be reduced and can thereby make applications practical that would otherwise be impossible. But physical proximity is only part of the challenge. Much of the challenge also resides in the topology and capabilities of the networks employed. It does no good to have a server next to the data if the data must take a circuitous route to get there, which is the case in many of today’s networks.

Edge computing distinguishes itself from centralized computing in several other ways.

In particular: edge computing enables the accurate geographical location of its workloads, with managed quality of service and the network topology used for connectivity. The same cannot be said for centralized computing. Since centralized computing is increasingly deployed in hyperscale data centers supported by cloud abstractions, it creates the perception of having seemingly infinite resources that are provisioned with imperative processes.

In contrast, edge computing platforms generally have scarce resources and are often supported by declarative provisioning processes. This requires a different model of resource allocation and the inherent complexity it adds. To address the scarcity of resources and declarative provisioning processes required, many of the early edge computing platforms have been purpose-built for the specific use cases they support. Increasingly, however, standardized platforms will become easier to deploy as edge computing matures, solutions are densified, and advanced service management and orchestration solutions are developed.
The Edge We Care About

The 2018 State of the Edge report developed a definition for the edge, presented as four assertions:

- The edge is a location, not a thing.
- There are lots of edges, but the edge we care about is the edge of the last-mile network.
- This edge has two sides: an infrastructure edge and a device edge.
- Compute will exist on both sides, working in coordination with the centralized cloud.

This definition of edge proved instructive. It provided a framework for the original State of the Edge 2018 report and became an anchor definition in spinout open source project, the Open Glossary of Edge Computing (now part of the Linux Foundation’s LF Edge).

These assets express the edge in terms of the last-mile network, namely the Device Edge and the Infrastructure Edge, which operate in unison with one another and with the cloud.

Device Edge
The Device Edge consists of devices and servers on the “downstream” side of the last-mile network, such as:

- Self-contained end-point devices, such as smart-phones, wearables and automobiles.
- Gateway devices such as IoT aggregators, switching and routing devices.
- On-premise server platforms.

To connect to the internet, equipment on the Device Edge must utilize some form of last-mile network, such as that provided by a wireless cellular or cable system operator.

Infrastructure Edge
The Infrastructure Edge consists of equipment on the “upstream” side of the last-mile network. In the context of communication networks, the Infrastructure Edge consists of the Access Edge and The Regional Edge. Edge computing at the Infrastructure Edge consists of compute platforms which are collocated with:

- Access sites which house network access equipment, such as cellular radio base stations, xDSL and xPON access sites.
- Aggregation hubs, such as those which house DAS (Distributed Antenna Systems) and serve as an initial aggregation of transmission connections from the access sites.
- Regional data centers and central offices, where access controller, switching equipment and other service gateway functionality is commonly deployed.
SIZING THE MARKET
How Big Will Edge Be?

According to the State of the Edge Forecast, over $700 billion in cumulative CAPEX will be spent within the next decade on edge IT infrastructure and data center facilities.

Regardless of how the edge computing market is measured, it will be massive.

It’s not that edge computing will replace today’s internet. Rather, edge computing is complementary to and will surge alongside hyperscale data centers, as both are propelled by the insatiable digital appetites of consumers and enterprises.

This report distinguishes between the Device Edge and the Infrastructure Edge. The Device Edge includes end-point devices and gateways, and enterprise on-premise servers. The Infrastructure Edge includes edge equipment at access network sites (e.g. radio base station tower sites), pre-aggregation sites (e.g. street-side cabinets), and aggregation and central office sites.

This report forecasts edge data center and IT infrastructure investments in terms of capital expenditures (measured in dollars) and the power footprint of the IT infrastructure deployed (measured in megawatts). The predicted capital expenditures offer insight into the market opportunities for edge IT equipment and infrastructure providers. The megawatt (MW) power footprint of the edge IT equipment estimates the scale of infrastructure needed to support the forecasted use-case demands.

The edge footprint can also be measured by the number of racks deployed. However, this approach is not used in this report since compute densities are likely to change significantly over the forecast period. The forecasts in this report only represent a part of the entire infrastructure investments needed. Not included in the forecasts are public fiber and wireless communications (e.g. 5G small-cells) for edge data center access and backbone connectivity and the investments and service revenues for edge applications. In addition, there is extensive on-premise IT infrastructure being deployed for traditional applications such as CRM and ERP that is not included in the device edge forecasts.

Laying the Foundation

Edge today is very much a solution-specific story with exotic equipment and architectures that are purpose-built for specific use cases, such as 5G and network function virtualization, next-generation CDNs and cloud, and streaming games. This represents the first phase in a journey towards a future where the edge becomes a readily available, viable part of the internet fabric with user-friendly interfaces for developers to exploit.
As today's edge computing use cases seed initial infrastructure deployments, they will provide the building blocks for additional edge computing use cases and further infrastructure investments. Over the coming years, edge computing will become pervasive and in time will transition from use-case specific to platform-centric solutions. With this transition, edge platforms will abstract infrastructure complexity away from developers using sophisticated management and orchestration software, and create user-friendly environments for developers to deploy innovative edge services and applications with relative ease.

Amazon achieved a platform-centric approach with its AWS service. With AWS, Amazon has combined its wares with those of others to enable its user-friendly AWS self-service platform environment. AWS laid the foundation for tremendous experimentation and innovation, and enabled entirely new industries and digital paradigms, such as cloud media streaming services and social networking. The same will occur with edge computing once it transitions from use-case specific to platform-centric solutions.

At the Epicenter of Opportunity

Today the Infrastructure Edge is primarily deployed in the central offices and network aggregation sites of communication networks. The Device Edge is deployed on end-point devices and gateways and on-premise enterprise servers. The Device Edge forecasts in this report focus on server-side architectures and therefore exclude end-point devices and gateways and only account for on-premise enterprise servers.

We cannot predict all the new services, business models, and industries that edge computing will create in the future. But we can be sure these will drive unprecedented demand for computing, storage, and networking capabilities, both at the edge and in the cloud. We can also identify some of the key services that are emerging and forecast their future growth trajectories and edge computing demands.
Forecasting the Edge
The forecasts in this report focus on 11 markets and industries, with several edge computing use cases for each. These include:

- **Communication Network Operators (CNO)** which are deploying edge computing equipment primarily to support their network upgrades. Forecasts were developed on a country-by-country basis to identify upgradeable network infrastructure, and to predict the upgrade rate and edge computing requirement.

- **Enterprise IT**, where the number of workloads migrating to the cloud were estimated over the forecast period. The forecast then estimated the portion of these workloads that require Infrastructure edge functionality. These estimates assessed the needs of enterprise IT applications. They also accounted for the availability and maturity of the Infrastructure edge, which was modeled using profiles comparable to historical growth rates that Amazon saw in its AWS service revenues.

- **Smart grids**, where the Infrastructure edge forecasts included future use cases for generation, distribution, transmission, renewables, operations, and maintenance. Forecasts were derived for each use case based on the maturity of available infrastructure and the service capabilities needed. The forecasts also referenced other infrastructure investment estimates, such as expenditures for communications network infrastructure.

- **Smart cities**, for which smart building, traffic signaling, and lighting use cases were assessed in addition to public safety, public venues, and utilities. The number of connected endpoints were initially estimated for each use case. The edge computing demands were then forecasted for each use case based on a variety of factors including: the service demands, maturity, and availability of edge infrastructure.

- **Manufacturing**, where the number of manufacturing facilities globally were estimated and use case demands for edge computing were assessed in the context of Industry 4.0. These use cases included asset tracking, remote operations, logistics and warehousing, operational automation, diagnostics, maintenance, and security enforcement.

- **Retail** for the establishments of the top 100 global retailers. The retail edge computing use cases that were assessed include, digital signage, in-store experience, proximity marketing, and supply chain optimization.

- **Healthcare**, which focused primarily on hospitals and clinics. Addressable market estimates were derived on a country-by-country basis using key factors such a hospital beds-per-person, GDP-per-capita, and the Gini Index. The healthcare edge computing use cases included, continuous patient monitoring and intervention, remote patient care, intervention and surgery, cognitive assistance, physical therapy, and patient record management.

- **Automotive**, where the addressable market for edge computing technology was forecast for a percentage of new vehicles sold globally. The automotive edge computing use cases were forecast individually, and included infotainment, traffic management, assisted and autonomous driving and operations and maintenance.
• **Residential** consumer services, which included edge computing use cases for infotainment, smart appliances, security, assisted living, and energy management. Forecasts were derived for each use case on a country-by-country basis and accounted for edge technology availability and maturity. In addition, key adoption factors included household size, disposable income, and income distribution (i.e. the Gini index).

• **Mobile consumer services**, for which smartphone penetration estimates were derived on a country-by-country basis. Affordability measures and evolving service expectations were used to estimate the market demand for edge computing use cases. These use cases included gaming, media and entertainment, information, social, health and fitness, messaging and communications and internet.

• **Commercial Unmanned Airborne Vehicles (UAV)**, where the installed base and shipments of commercial UAVs were estimated. Edge computing forecasts were developed for several use cases including: apping and surveying, photogrammetry, and 3D and digital elevation modeling.

• **Other markets**, which were forecasted to range between 15 and 20% of the total edge computing market.

The market forecast estimates for edge computing were applied to reference architectures, with pricing, power consumption, and infrastructure utilization estimates. With these, edge infrastructure capital expenditures and power footprint estimates were forecasted.

### The Edge Reaches Globally

Edge computing is a global phenomenon. Some countries will be more aggressive than others in embracing edge computing, and there are a variety of important macroeconomic factors to consider. Notable factors include: Market scale and competitiveness, industrial activity, population density, communication infrastructure maturity, and GDP and Gini indices. Edge computing use cases will only be adopted by consumers when they are affordable and compelling, and by enterprises and public service organizations when they deliver adequate value and can be easily operationalized. These and other factors underpin the forecasts that are presented in this report.

Asia Pacific includes some of the most affluent and technologically advanced countries in the world—and some of the poorest. In addition, five out of 10 of the most populated countries in the world—including China and India—are in Asia Pacific. Asia Pacific currently has the largest edge computing equipment footprint of all the global regions, estimated to be 187MW today. This footprint is measured by the aggregate IT power rating of the edge infrastructure equipment and the edge data center facilities that are deployed.

In Asia Pacific, the footprint is currently buoyed by markets like Australia, Japan, Korea and China, and is forecast to increase to 25,409MW by 2028, and account for 36.7% of the global edge computing footprint. China and India will drive a significant proportion of growth in Asia Pacific, in addition to Japan and Korea. Edge computing has also seen early market momentum in the United States, where public cloud and regional data center providers, and particularly CNOs, have been relatively aggressive in driving market
solutions. The IT power footprint of edge computing infrastructure in North America is forecast to reach 14508MW by 2028 and account for 21.0 percent of the global edge footprint.

Multinational CNOs in Western European are the largest customers for edge computing equipment to support network upgrade efforts. The IT power footprint of edge infrastructure in Europe is forecast to reach 20403MW by 2028, and account for 29.6% of the global edge footprint. Other regions include Latin America and the Middle East and Africa, which are forecast to account for 6.8 and 5.8 percent, respectively, of the global edge computing footprint by 2028.

The deployed global power footprint of the edge IT and data center facilities is forecast to reach 102 thousand MW by 2028, with 68% of the deployments being on the infrastructure edge and the remaining 32% on the device edge. The value proposition for the infrastructure edge is expected to accelerate once the edge transitions from user-specific to platform-centric architectures, particularly after the 2024 timeframe.

The global annual CAPEX for edge IT and data center facilities is forecast to reach $146 billion in 2028 with a 35 percent CAGR. The annual CAPEX for infrastructure edge equipment is forecast with a CAGR of 42 percent, compared with a CAGR of 25 percent for the device edge. After 2025, the global annual CAPEX for device edge infrastructure is forecast to stabilize and in 2028, to decline moderately, as CAPEX for infrastructure edge capabilities continues to increase.
Deployed Global Footprint

Global Annual CAPEX Spend on Edge

Source: Telaga Research / State of the Edge
CHAPTER 4

A TAPESTRY OF USE CASES
Looking Forward

The edge computing market and its infrastructure demands today are not the same as they will be in the future. Today, edge computing does not support a ‘one-size-fits-all’ approach to infrastructure investment. Edge platforms are distributed with scarce resources relative to the seemingly infinite resources of hyperscale centralized cloud data centers. Until edge computing matures with large scale implementations, deployments will be use case dependent, often with proprietary hardware and software platforms, because of resource scarcity. This is in stark contrast to hyperscale data centers that are built on massive general-purpose platforms with the scale needed to support virtually any use-case.

Sophisticated infrastructure orchestration algorithms and real-time resource trading and auction schemes are being proposed and will enable greater scalability once the edge computing market matures and edge becomes widely available. When this occurs, we expect that generic, as opposed to use case specific edge platforms will become the norm.

By 2028, the edge market will be dominated by consumer applications

The edge computing demands across eleven markets and industries are forecasted in this report. It is expected that by 2028, 4G and 5G mobile consumer and residential consumer (smart home) applications will come to dominate the edge computing footprint, and robust demand will also come from CNOs, enterprise IT and automotive industry verticals. The edge IT power footprint for mobile and residential consumers is forecast to reach 16938 and 10843 MW, respectively by 2028. The edge IT power footprints for CNOs and enterprise IT is expected to increase over the forecast period to reach 7117 and 5800 MW, respectively. The variety and scope of edge use cases are expected to increase tremendously as edge matures with platform-centric offerings, particularly for consumer mobile and residential use cases.
Edge Footprint by Segment in 2028 (MW)

- Communication Service Providers: 7117 MW
- Enterprise IT: 5800 MW
- Residential: 10843 MW
- Mobile Consumer: 16938 MW
- Retail: 4361 MW
- Healthcare: 4195 MW
- Automotive: 5006 MW
- Commercial UAV: 447 MW
- Smart Grid: 1509 MW
- Smart Cities: 3052 MW
- Manufacturing: 2374 MW
- Other: 7397 MW

Source: Tolaga Research / State of the Edge
**Telecom & Wireless**

**Operators are Leading the Charge**

The largest demand for edge computing currently comes from communication network operators (CNOs) as they virtualize their wireline and wireless network infrastructure and accelerate their network upgrades, including 5G. Although wireless, and particularly 5G, tends to capture the limelight for edge computing, most edge investments today are for virtualizing wireline infrastructure, including SD-WAN equipment, core network routing and switching equipment and data gateways.

Communication network operators are upgrading their network and service delivery architectures to capitalize on advancements in cloud and virtualization technologies and increase the agility and efficiency of their services. Network function virtualization (NFV) and software defined networking (SDN) solutions that underpin next generation technologies like 5G and are being implemented on edge platforms that CNOs are deploying across their networks. 5G is expected to accelerate demand for edge solutions since it natively supports service capabilities such as ultra-low latency connectivity, bandwidth management, ultra-reliable connectivity and data geofencing.

![CNO Edge Compute Footprint Distribution](image-url)
The first commercial 5G networks were seen in 2019, a year ahead of schedule. 5G is much more than an a 4G upgrade. Rather, it is virtualized by design and can simultaneously support legacy wireline and 2G, 3G and 4G wireless technologies, as software on the same hardware fabric. Network control and data plane separation is inherent to 5G for local breakout, and 5G enables flexible network architectures for ultra-low latency, broadband and ultra-reliable services.

5G and similar upgrades to the wireline networks will require edge computing for network virtualization and automation, as well as to enable new services. This will cause the system operators to invest significant capital to build and lease edge data centers and deploy IT equipment at the edge. These network upgrades will create the capacity to deliver local and distributed computing capabilities for a growing number of digital services.

The global power footprint of the edge computing equipment for CNOs is forecast to increase from 231 to 7383 MW between 2019 and 2028. In 2019, CNO deployments will represent 22 percent of the global edge footprint. The CNO edge computing footprint will decline to 10 percent of the overall global footprint in 2028, as other use cases come to the fore.

In 2019, 94 percent of this footprint will occupy network central office sites, with the remaining 6 percent in access and aggregation sites. Between 2019 and 2028 the aggregation edge footprint for CNOs is forecast to increase from 5 to 38 percent of the total CNO footprint, as CNOs virtualize their access networks at scale and deploy ultra-low latency services.

History suggests that most CNOs are particularly effective with mass market consumer services, but less effective with specialized consumer and enterprise services. The exception is CNO wholesale services, where CNOs have been successful in selling wholesale connectivity services to large enterprises. Some of the largest wholesale customers for connectivity services today are Tier 1 media and cloud service providers, including AWS, Google, Microsoft and Netflix. These same companies are important ecosystem participants for edge computing.

Some CNOs will opt to market their edge service capabilities to their wholesale customers. Others will adopt retail enterprise strategies for fear that a focus towards wholesale edge computing will render them ‘dumb pipes’ for the foreseeable future. It is still too early to ascertain the winning strategies for CNOs.
Mobile Consumer Services

The wireless edge strengthens

Mobile consumers use smartphone and tablet computing devices with 4G and 5G mobile network connectivity for an ever increasing range of applications. They can be notoriously fickle, but also have a propensity to drive the viral adoption of compelling applications once they become available. To find these viral opportunities, developers depend on user-friendly environments for experimentation.

Today this is achieved with cloud connected client devices that are equipped with advanced features, such as cameras, GPS location, accelerometers and software development kits (SDK). New advanced features are being developed. These include wearable and integrated sensor-based technologies, tactile interfaces and capabilities such as virtual, augmented and mixed reality and real-time artificial intelligence. These advanced features are relatively nascent but benefit from tremendous investments and are expected to transform mobile consumer services in the coming years. In many cases, these transformed services will require edge computing infrastructure.

Media and entertainment, information services, gaming and advanced information and internet services are early candidates for edge infrastructure. The range of use cases are expected to increase in the future to include those related to social networks, health and fitness, and messaging and communications. The infrastructure edge footprint to support mobile consumer services is forecast to increase from 124MW 2019 to 16938MW in 2028.

Source: Tolaga Research / State of the Edge
Residential Consumer Services

The Smart Home
Smart appliances and a growing variety of digital services, such as those for security, infotainment, assisted living and energy management are driving smart-home capabilities. Today these capabilities are primarily implemented directly in end-point devices and residential gateways (i.e. the Device Edge) with support from centralized cloud functionality.

Over the coming years, we expect that digital services for smart-homes will become increasingly sophisticated, with immersive capabilities that aim to significantly enhance residential consumer experiences. As this occurs, we expect that an increasing number of residences will implement smart home and appliance technologies. In some cases, these capabilities will be deployed in the Device Edge. However, there will be other cases where advanced service features and user expectations will drive demand for Infrastructure Edge capabilities, particularly as edge becomes more readily available.

We estimate that households with annual incomes in excess of USD 250 thousand in 2019 and USD 143 thousand in 2028 can afford residential digital services. Each digital service has unique characteristics that will drive demand for Device and Infrastructure edge capabilities. Between 2019 and 2028, the Infrastructure Edge footprint to support residential services globally is forecasted to increase from 26 to 10843MW.
Key digital services include infotainment, which will range between 47 and 29 percent of the global edge footprint between 2019 and 2028, smart appliances, which are forecast to consistently occupy 19 percent of the edge footprint between 2019 and 2028, and security with an edge footprint that is forecast to increase from 10 to 16 percent. In addition, the footprint for energy management is forecast to increase from 10 to 16 percent, and assisted living, 2 to 5 percent, between 2019 and 2028.

Automotive

Finding the Edge for Mobility

The automotive industry has faced decades of disruption. These disruptions are still in full swing and are driving fundamental changes in vehicle designs and their features and functionality. The latest disruptions to the automotive industry are the culmination of more than a decade of innovation. In 2009 Google kicked off its ‘self-driving car project’ to pioneer the subsequent proliferation of assisted and autonomous driving solutions. Tesla launched its Model S in 2012 which revolutionized electric vehicle technology and proved that the tremendous barriers for entering the automotive industry could be overcome. Car makers are responding to heightened competition and market disruptions with innovative vehicle designs, which aim to embrace the digitization of the vehicle experience.
The edge computing requirements for automotive use cases are impacted by several design considerations. In particular, there are already extensive edge computing footprints deployed in vehicles today. These footprints will continue to increase for the foreseeable future to support sophisticated use case demands. We refer to this as the Device Edge in the context of this report. While the Device Edge addresses much of the functionality needed, services like infotainment, traffic management and operations and maintenance have some emerging Infrastructure Edge requirements to localize use case functionality, improve response times and reduce network bandwidth costs. As vehicle use cases become more sophisticated and assisted and autonomous driving grows in popularity, coordination and cooperation amongst vehicles will become increasingly important. Infrastructure Edge solutions, built out in sufficiently close proximity to highways and roadways will be required to achieve this reliably and at scale.

The forecasts in this report focus on the Infrastructure Edge investments needed by the automotive industry. Early demand for the Infrastructure Edge is expected to come from infotainment and traffic management use cases. Additional Infrastructure Edge use cases that are expected to emerge in the future include, autonomous and assisted driving, and proactive operations and maintenance. The global Infrastructure Edge IT power footprint for the automotive industry is forecast to increase from 87MW to 3405 MW between 2019 and 2028.

Enterprise IT

Infrastructure Edge Complements Centralized Cloud

When workloads require edge computing capabilities, enterprises ordinarily deploy the workloads with on-premise (Device Edge) equipment. The Infrastructure Edge provides a means for enterprises to migrate Device Edge workloads to off-premises locations and public clouds and reap comparable benefits to that of traditional workload migrations. There are a variety of reasons that an enterprise might deploy workloads at the Infrastructure Edge as opposed to (or in addition to) the centralized cloud. These include data management, enhanced user experience, security and privacy, and to support latency sensitive and bandwidth intensive services.

Public cloud providers including AWS and Microsoft have cloud products which can either be deployed in the Device or Infrastructure Edge. AWS’ Outpost and Microsoft’s Azure IoT Edge products enable edge implementations of proprietary cloud technologies. For enterprises, the value proposition for products like Outpost and AzureStack is that they integrate seamlessly with the AWS and Azure public clouds. Both Outpost and AzureStack where initially targeted towards on-premise (Device Edge) implementations, but more recently have also been applied to Infrastructure Edge implementations to meet customer demands.

To make their products available for Infrastructure Edge implementations, it is unlikely that AWS and Microsoft will deploy edge infrastructure facilities. Instead we expect that they will seek partnership and collaboration opportunities with CNOs and regional data center providers that have edge infrastructure facilities. For this, AWS and Microsoft might partner directly with facility providers or indirectly through deployments with their enterprise customers.
Edge infrastructure CAPEX to support enterprise IT is forecast to USD 8.7 billion by 2028. By this time, it is estimated that 9.3 percent of enterprise IT cloud workloads will be deployed at the Infrastructure Edge.

**Smart Cities**

**An Exciting Time for Boring Infrastructure**

Cities provide many services that are well suited for Device and Infrastructure Edge capabilities. It has taken many years and numerous technology trials for key stakeholders to become actively engaged in smart cities, with incentives to deploy and operationalize large scale solutions. Notable solutions include smart-lighting, traffic management, smart buildings and public venues and services for public safety and utilities.

The Global IT power footprint for smart city Infrastructure Edge equipment is estimated at 158MW in 2019 and forecast to reach 3052MW by 2028. In 2019, over 50 percent of the power footprint is associated with Smart Building and Public Safety solutions, such as video surveillance and incident first responder capabilities. Smart Buildings solutions are forecast to reach 36.9 percent of the overall smart-city footprint by 2028, 24.8 percent for Public Safety and the remaining associated with Traffic Signaling, Smart Lighting, Public Venues and Utilities.
2019: Smart Cities and Buildings
Edge IT Power Footprint (158MW)

2028: Smart Cities and Buildings
Edge IT Power Footprint (3052MW)

Electrical Utilities: Infrastructure Edge IT Equipment Power Footprint (MW)

Source: Tolaga Research / State of the Edge
Utilities

Smart Grids with an Edge

Today smart grids rely primarily on Device Edge platforms. Increasingly, these platforms will be complemented with Infrastructure Edge solutions as grids become smarter. Environmental sustainability initiatives are driving global efforts to improve the efficiencies of electrical utility services and to drive alternative renewable power sources, such as wind and solar.

Smart-grids are being implemented worldwide to improve operational efficiencies and enable capabilities such as real-time consumption management, integration with smart appliances, and micro-grids to support generation from distributed renewable sources. Electrical grids are still in the early stages of being upgraded to enable smart-grid capabilities.

The global Infrastructure Edge power footprint for smart grids is forecast to increase from 28.4MW to 1512MW between 2019 in 2028. By 2028, renewables are forecast as the largest smart-grid Infrastructure Edge use-case, followed by general Operations and Maintenance and Distribution.
Smart Retail

Bricks and Mortar Goes Digital

In recent years, traditional retailers have been disrupted by the onslaught of competition from online alternatives. Many traditional retailers have responded with online solutions of their own, and online retail continues to see robust growth. Even as the online retail market continues to grow, the ‘bricks-and-mortar’ establishments are starting to capitalize on digital services. Amazon’s acquisition of Whole Foods in the United States is a testament to Amazon’s belief it can disrupt and transform retail shopping by aggressively adopting digital technologies. As ‘bricks-and-mortar’ establishments implement digital services, they require the support of edge computing capabilities. Notable use cases include smart digital signage and instore advertising, proximity marketing and other instant pricing and marketing solutions that benefit from real-time analytics and supply chain and inventory optimization solutions.

Both the Device and Infrastructure Edge will be important for retail digital services. It is forecast that by 2028, 20.3% of the retail establishments run by the top 100 retailers globally will have implemented digital services that require Infrastructure Edge capabilities. The global Infrastructure Edge CAPEX for retail is forecast to reach $6.5 billion by 2028.
Both the Device and Infrastructure Edge will be important for retail digital services. It is forecasted that by 2028, 20.3% percent of the retail establishments run by the top 100 retailers globally will have implemented digital services that require Infrastructure Edge capabilities. The global Infrastructure Edge CAPEX for retail is forecast to reach $USD 6.5 billion by 2028.

**Industrial IoT**

**Manufacturing Embraces Digital Services with Industry 4.0**
Manufacturing industries face tremendous disruptions with increased market competition and unprecedented end-user demands for personalized and customized products. These disruptions traverse all facets of manufacturing including, logistics and warehousing, asset tracking, operational automation, security enforcement, diagnostics and maintenance, and remote operations. The Industry 4.0 initiative provides a framework for transforming manufacturing to address industry disruptions. Industry 4.0 emphasizes operational agility using technologies that bring convergence between cyber and physical systems. Industry 4.0 requires horizontal integration across operational technology functions and vertical integration between operational and information technologies.
This horizontal and vertical integration depends on unprecedented plant instrumentation (e.g. Industrial IoT), data orchestration and analytics. Edge computing plays a critical role in allowing manufacturers to transform their operational plants towards Industry 4.0. Although the lion’s share of edge computing for manufacturing will be implemented at the Device Edge, demand for Infrastructure Edge capabilities is forecast to grow as service complexity increases and the Infrastructure Edge matures and becomes more readily available.

The global Infrastructure Edge CAPEX for manufacturing services is forecast to increase from $360m to $3.55b between 2019 and 2028. It is estimated that over a third of the global CAPEX in 2019 will be associated with asset tracking and that operational automation will emerge during the forecast period to reach almost 27.2% of the global CAPEX by 2028.

**Healthcare**

**An Impactful Driver of the Edge**

The healthcare industry is characteristically conservative with digital technologies. It is an industry that is closely regulated, particularly in markets like the United States, and where innovation tends to be driven by consensus rather than outright disruption. Even with this conservativism, hospitals and clinics have digital health strategies with varying degrees of maturity and success.

### Healthcare (Hospitals/Clinics) Infrastructure Edge IT Power Footprint

Source: Tolaga Research / State of the Edge
To support these strategies, hospitals and clinics have implemented edge computing solutions for key use cases including patient record management, remote patient care and intervention and continuous patient monitoring and intervention. The global IT power footprint of the Infrastructure Edge support for Healthcare services is forecast to increase from 2.1 to 4195 MW between 2019 and 2028.

**What the Future Holds**

The market for edge computing is global with the largest footprint forecasted in Asia Pacific, followed by Europe and North America. Both the Device and Infrastructure edge will grow tremendously in the coming decade, with greater growth in the Infrastructure edge to support the technical and commercial demands of current and emerging edge use cases.

In 2019, Communication Network Operators (CNO) were the largest customers of IT equipment and data center facilities for the Infrastructure Edge. Significant demand is also coming with other use cases, such as CDNs for consumer media and entertainment, and public safety and surveillance applications for smart cities. By 2028, mobile and residential consumer use cases will occupy the largest Infrastructure edge footprint, followed by CNOs and enterprise IT applications. This reflects the changing market dynamics that will occur once the Infrastructure Edge transitions from use case-specific to platform-centric architectures.

Edge computing is crucial for many industries that currently find themselves in the midst of the digital revolution. This revolution has been playing out for the last decade in the Second Act of the Internet with the proliferation of cloud and will play out for the next decade with the Third Act of the Internet, which requires distributed computing capabilities delivered with edge computing. As automobiles, cities, factories, and retailers become smarter, they will place tremendous demands on the Device Edge and the Infrastructure Edge. It is crucial that industry players respond to these demands—if they don’t they will be substituted for players who can and will.
POSTCARDS FROM THE EDGE
We asked leading experts from around the world what they thought—and what they were learning—about the edge as we enter 2020. The following “Postcards from the Edge” are excerpts from their responses. Each of the authors below will have an article on the State of the Edge blog that expands upon their thoughts here. A new article will be released every week. Follow @StateoftheEdge on Twitter to get notified when a new post goes live.

**The Edgeless Cloud and Flatnets**

The cloud is going edgeless at the rate of cloud inflation, but carrier wireless service revenues are well into long-term decline. The solution will be highly capital-efficient flatnets that both inflate with the cloud and generate new revenue sources.

Francis McInerney
Managing Director of North River Ventures

**Increasing BMS IQ via Edge Computing**

For those of us who pioneered edge applications early on, it is encouraging to see edge computing architected into 5G. For the CRE market, the ability to loosely connect mobility to both BMS and enterprise systems will positively transform the tenant experience and increase the sustainability of the building itself. Good times!

Art King
Director of Enterprise Services
Connected Real Estate Magazine (Rich Berliner, Publisher)
How Much Will the Edge Cost?

Edge computing emerged on the wireless industry stage several years ago, albeit with several different versions and approaches. No matter which form it eventually takes, edge computing has the potential to be a disruptive technology. In fact, edge computing is quite likely to help realize the promise of 5G, particularly since the new 5G system architecture is designed to capitalize on virtualization. But how much will it cost?

Iain Gillott
Founder and President of iGR

If You Want to Understand the Edge, Just Look at Your Phone

What is edge computing? As one might guess, the answer depends on the context in which you ask the question. Should you think of the edge from perspective of cloud computing, looking out (as most experts do, quite understandably), or should we think of it from the outside in, more from the perspective of the user or the device? If, like me, you’re trying to understand where and how the edge adds the greatest commercial value, then the outside-in approach has some important and very tangible benefits. And if you need more convincing, talk to a Millenial on the street – I bet the don’t care about the Internet or the Cloud unless they can get to it from their phone.

Peter Christy
Independent Analyst
Monolith to Microservices in the Physical World

The promise of Edge is at the forefront of the imminent Fourth Industrial Revolution. This dramatic transformation in app building necessitates equally extensive adaptations to accompanying network architectures and hardware landscapes. On the software side, microservices architecture is emerging to provide flexibility and efficiency to the IoT ecosystem.

Antonio “Pelle” Pellegrino
Founder & CEO, Mutable

The Event-Driven Edge is the Most Important New Idea in Cloud Computing

Modern, centralized cloud computing relies on a request-response architecture. The event-driven architecture emerging at the edge is fundamentally different — and potentially transformative. By putting event-driven applications and web services at the edge, we can take advantage of the infinite streams of event data to create new digital services that don’t just record what humans and their devices are doing, but can also anticipate and predict what might happen next.

Chetan Venkatesh
CEO & Co-founder, Macrometa Corp.
**The Data Layer Challenge at the Rapidly Evolving Edge**

In the early days of the cloud, visionaries predicted that enterprise infrastructure would quickly migrate to few large cloud providers. Unfortunately, the early cloud hype got ahead of reality—because public cloud faces a huge issue it can’t solve on its own: latency. The edge is developing quickly, but the initial emphasis has been on providing compute capabilities, with little thought given to storage.

---

**And Now, The Time Has Come: 5G and the Edge**

We’re at a watershed moment in the evolution 5G, the “put-up or shut-up” moment when operators are assessing the commercial value to justify their investment in a 5G buildout. This is what always happens in the marketing and investment dance of a new cellular generation.

Recent research concludes that the full value of 5G is still being discovered, with select operators like SK Telecom aggressively exploring edge services that can deliver outsized value and demand a premium.

---

Ellen Rubin
CEO, Clear Sky Data

Geoff Hollingworth
CMO, MobiledgeX
What the Cutting Edge Looks Like Today

Most enterprises are still in the early stages of implementing an edge computing strategy, often as part of a digital transformation initiative or, in some cases, a consequence of such an initiative. In talking to these enterprises, we’ve identified common challenges and opportunities. AI, in particular, has emerged as a key leading “edge” use case that can bring significant upside to enterprises while creating tough challenges in terms of power consumption and cooling for facilities providers.

Jim Davis
Principal Analyst,
Edge Research Group

In the Clouds: The Times They Are a Changin’

There is no doubt the cloud—the technical backbone of our current modern world—is changing.

We should stop looking at edge computing as a successor to the cloud, but rather as an essential gateway to what will come next on a global level: a stepping stone for the old world of telecommunications to merge with the new world of software-defined cloud, giving birth to a new layer of infrastructure that can power a technology-driven world for the next 20 years.

Mahdi Yahya
CEO & Co-founder, Ori
**Edge is the New Cloud**

As the founder and CEO of SoftLayer (now IBM Cloud) I had the opportunity to be at the forefront of the cloud era since before anyone really knew what the “cloud” was. The development of the “edge” is leading to as dramatic a revolution.

So who will win the revolution and own the edge? I believe it has got to be a platform of secure edge services that enables developers to protect, accelerate, and innovate cloud properties ranging from websites to media delivery and IoT services. A platform that is origin agnostic, inherently secure, able to hyperscale, and allows users to easily connect to a fully secure SaaS world.

**Lance Crosby**  
Principal Analyst, Founder & CEO at Stackpath

---

**The Enterprise and the Edge**

The shape of the internet is starting to change. 5G wireless deployments, coupled with moves by hyperscalers, eyeball networks, content delivery networks, and new kinds of connectivity (CBRS! balloons! satellites!) promise that whatever the internet looks like in 10 years, it will be different from the one we know today.

What does that imply for today’s broader “edge” ecosystem? In my view, it means that the first chapter of edge won’t be necessarily about latency, but will be driven by innovative requirements of enterprises and the use cases that matter them today.

**Zac Smith**  
CEO and Co-founder, Packet
IoT, AI and Networking at the Edge

In the fast-emerging world of 5G-enabled IoT, edge computing and AI technologies will each play a crucial role in the digital transformation of homes, businesses, workplaces, educational institutions—even entire cities. New solutions for edge network automation are needed to compel this revolution forward. Ideally these new edge environments will be devised with a smaller footprint in mind, requiring no additional hardware and software, and resulting in sustainable, cost-efficient solutions that also optimize space and conserve power usage.

Mike Capuano
CMO at Pluribus Networks

Who Will Build and Pay for the Edge?

There is broad industry consensus on the need and value of the edge, but it is still uncertain how the move to distributed networks will unfold, especially in the enterprise, alongside the rise of IoT and private networks.

Both the enterprise and service providers may try to position themselves to dominate the edge infrastructure, but it is more likely, more efficient and cost effective for them to share the helm of the edge and build new and deeper working and financial relationships than they have been able to do until now.

Monica Paolini
Principal at Senza Fili
Applications at the Industrial Edge

Augmented Reality, Virtual Reality and Mixed Reality are being adopted in manufacturing, heavy industries, retail, utilities, automotive and construction. However, there are barriers blocking AR/MR adoption, including access to highly responsive compute infrastructure, 24/7 availability, and the ability to scale.

To deliver a low-latency experience, services and instances of the applications need to be deployed near the end-users where it is most relevant—near the edge of the network.

Vineet Anshuman  
Co-founder, Cloudlyte

Vikram Balimidi  
Director Of Product and Marketing (SD-WAN Services)  
Cloudlyte

Bridging the Last Mile: Convergence at the Infrastructure Edge

For as long as communication networks have existed, the last mile has presented unique infrastructural challenges and opportunities. By being the most distributed portion of the internet, the last mile is often the hardest to build and operate, but by being the closest to the end user or device, it’s also the most important part of the network for enabling next-generation applications. Converging network and compute at the edge becomes the foundation for a next-generation internet—and the way to fix the last mile.

Cole Crawford  
Founder & CEO at VaporIO
Edge Computing Landscape

Launched with great fanfare in 2018, the original Edge Computing Landscape quickly became a historical artifact as the ecosystem exploded. Today, the Edge Computing Landscape is maintained by a working group at The Linux Foundation under the auspices of the Open Glossary of Edge Computing.

The new Edge Computing Landscape is a work in progress. It is a living document that gets compiled from the contents of a Github repo and is governed as an open source project. Developers, investors, vendors, researchers and others can use it as a resource on the landscape of edge computing.

Make sure your own projects and organizations are on the Landscape. If not listed, first collect the following information:

- Crunchbase URL
- Website URL
- Twitter URL
- SVG Version of the logo
- Linkedin URL (Optional)
- Github Repo URL (Optional)

Then, do one of the following:

1. Follow the instructions for adding a new landscape entry using the web interface.
2. Open a pull request (PR) in the Github repo using these instructions.
3. Create A Github Issue in the Github repo. (Be sure to note which category under which to place your addition include links mentioned above and an SVG format logo.

If you have questions and would like to learn more about LF Edge and its programs, please visit www.lfedge.org.
State of the Edge Blog

The State of the Edge blog publishes diverse opinions from industry practitioners, analysts, and researchers, highlighting thought leadership in all areas of edge computing and adjacent technologies. The State of the Edge team will collaborate on topics and provide developmental support. They will also be flexible on timing. With a goal of publishing a post every 10-15 days, the team will work with authors to decide the ideal publishing sequence.

We are looking for unique, journalistically top-tier content on edge computing to be published via the State of the Edge blog. Content that we’re looking for includes:

- Deep dives on a use cases or vertical: e.g., AR/VR, IoT, smart city, etc.
- Unique perspectives on edge computing: e.g., from the perspective of a telco operator, developer, car maker, cloud provider, etc.
- Broad views on how other technology trends intersect edge: e.g., AI and edge, cloud native and edge, autonomous driving and edge, etc.
- Op-ed style opinion pieces, labeled as such.
- Summaries of first or second party research.

Simple Ground Rules

- No vendor fluff;
- Avoid mentioning specific companies or products; be product- and company- independent;
- Make it well-written, thoughtfully argued and in the active voice;
- Be generally supportive and optimistic about edge computing.

It’s OK to write about what your company does, but please do so generically. In other words, you may describe the technologies, architectures and viewpoints you prefer, but when you do so, please avoid making mentions of companies or products. This does not apply to open source technologies, such as Kubernetes or the Akraino stack. You describe your company and your products in the author bio, which will appear at the end of the article and may link offsite, such as to your website.

How to Submit a Blog Post

If you have an idea for a post that you believe would fit on the State of the Edge blog, please email info@stateoftheedge.com.
Open Glossary of Edge Computing v2

3G, 4G, 5G
3rd, 4th, and 5th generation cellular technologies, respectively. In simple terms, 3G represents the introduction of the smartphone along with their mobile web browsers; 4G, the current generation cellular technology, delivers true broadband internet access to mobile devices; the coming 5G cellular technologies will deliver massive bandwidth and reduced latency to cellular systems, supporting a range of devices from smartphones to autonomous vehicles and large-scale IoT. Edge computing at the infrastructure edge is considered a key building block for 5G. See also: Infrastructure Edge

Access Edge Layer
The sub-layer of infrastructure edge closest to the end user or device, zero or one hops from the last mile network. For example, an edge data center deployed at a cellular network site. The Access Edge Layer functions as the front line of the infrastructure edge and may connect to an aggregation edge layer higher in the hierarchy. See also: Aggregation Edge Layer

Access Network
A network that connects subscribers and devices to their local service provider. It is contrasted with the core network which connects service providers to one another. The access network connects directly to the infrastructure edge. See also: Infrastructure Edge

Aggregation Edge Layer
The layer of infrastructure edge one hop away from the access edge layer. Can exist as either a medium-scale data center in a single location or may be formed from multiple interconnected micro data centers to form a hierarchical topology with the access edge to allow for greater collaboration, workload failover and scalability than access edge alone. See also: Access Layer Edge

Base Station
A network element in the RAN which is responsible for the transmission and reception of radio signals in one or more cells to or from user equipment. A base station can have an integrated antenna or may be connected to an antenna array by feeder cables. Uses specialized digital signal processing and network function hardware. In modern RAN architectures, the base station may be split into multiple functional blocks operating in software for flexibility, cost and performance. See also: Cloud RAN (C-RAN)

Baseband Unit (BBU)
A component of the Base Station which is responsible for baseband radio signal processing. Uses specialized hardware for digital signal processing. In a C-RAN architecture, the functions of the BBU may be operated in software as a VNF. See also: Cloud RAN (C-RAN)

Central Office (CO)
An aggregation point for telecommunications infrastructure within a defined geographical area where telephone companies historically located their switching equipment. Physically designed to house telecommunications infrastructure equipment but typically not suitable to house compute, data storage and network resources on the scale of an edge data center due to their inadequate flooring, as well as
their heating, cooling, ventilation, fire suppression and power delivery systems. In the case when the hardware is specifically designed for edge cases it can cope with the physical constraints of Central Offices. See also: Central Office Re-architected as Data Center (CORD)

Central Office Re-architected as Data Center (CORD)
An initiative to deploy data center-level compute and data storage capability within the CO. Although this is often logical topologically, CO facilities are typically not physically suited to house compute, data storage and network resources on the scale of an edge data center due to their inadequate flooring, as well as their heating, cooling, ventilation, fire suppression and power delivery systems. See also: Central Office (CO)

Centralized Data Center
A large, often hyperscale physical structure and logical entity which houses large compute, data storage and network resources which are typically used by many tenants concurrently due to their scale. Located a significant geographical distance from the majority of their users and often used for cloud computing. See also: Cloud Computing

Cloud Computing
A system to provide on-demand access to a shared pool of computing resources, including network, storage, and computation services. Typically utilizes a small number of large centralized data centers and regional data centers today. See also: Centralized Data Center

Cloud Native Network Function (CNF)
A Virtualized Network Function (VNF) built and deployed using cloud native technologies. These technologies include containers, service meshes, microservices, immutable infrastructure and declarative APIs that allow deployment in public, private and hybrid cloud environments through loosely coupled and automated systems. See also: Virtualized Network Function (VNF)

Cloud Node
A compute node, such as an individual server or other set of computing resources, operated as part of a cloud computing infrastructure. Typically resides within a centralized data center. See also: Edge Node

Cloud RAN (C-RAN)
An evolution of the RAN that allows the functionality of the wireless base station to be split into two components: A Remote Radio Head (RRH) and a centralized BBU. Rather than requiring a BBU to be located with each cellular radio antenna, C-RAN allows the BBUs to operate at some distance from the tower, at an aggregation point, often referred to as a Distributed Antenna System (DAS) Hub (#distributed-antenna-system-das-hub). Co-locating multiple BBUs in an aggregation facility creates infrastructure efficiencies and allows for a more graceful evolution to Cloud RAN. In a C-RAN architecture, tasks performed by a legacy base station are often performed as VNFs operating on infrastructure edge micro data centers on general-purpose compute hardware. These tasks must be performed at high levels of performance and with as little latency as possible, requiring the use of infrastructure edge computing at the cellular network site to support them. See also: Infrastructure Edge, Distributed Antenna System (DAS) Hub
**Cloud Service Provider (CSP)**
An organization which operates typically large-scale cloud resources comprised of centralized and regional data centers. Most frequently used in the context of the public cloud. May also be referred to as a Cloud Service Operator (CSO). See also: Cloud Computing

**Cloudlet**
In academic circles, this term refers to a mobility-enhanced public or private cloud at the infrastructure edge, as popularized by Mahadev Satyanarayanan of Carnegie Mellon University. It is synonymous with the term Edge Cloud as defined in this glossary. It has also been used interchangeably with Edge Data Center and Edge Node in the literature. In a 3-tier computing architecture, the term “cloudlet” refers to the middle tier (Tier 2), with Tier 1 being the cloud and Tier 3 being a smartphone, wearable device, smart sensor or other such weight/size/energy-constrained entity. In the context of CDNs such as Akamai, cloudlet refers to the practice of deploying self-serviceable applications at CDN nodes. See also: Edge Cloud, Edge Data Center, Edge Node

**Co-Location**
The process of deploying compute, data storage and network infrastructure owned or operated by different parties in the same physical location, such as within the same physical structure. Distinct from Shared Infrastructure as co-location does not require infrastructure such as an edge data center to have multiple tenants or users. See also: Shared Infrastructure

**Computational Offloading**
An edge computing use case where tasks are offloaded from an edge device to the infrastructure edge for remote processing. Computational offloading seeks, for example, performance improvements and energy savings for mobile devices by offloading computation to the infrastructure edge with the goal of minimizing task execution latency and mobile device energy consumption. Computational offloading also enables new classes of mobile applications that would require computational power and storage capacity that exceeds what the device alone is capable of employing (e.g., untethered Virtual Reality). In other cases, workloads may be offloaded from a centralized to an edge data center for performance. The term is also referred to as cloud offload and cyber foraging in the literature. See also: Traffic Offloading

**Content Delivery Network (CDN)**
A distributed system positioned throughout the network that positions popular content such as streaming video at locations closer to the user than are possible with a traditional centralized data center. Unlike a data center, a CDN node will typically contain data storage without dense compute resources. When using infrastructure edge computing CDN nodes operate in software at edge data centers. See also: Edge Data Center

**Core Network**
The layer of the service provider network which connects the access network and the devices connected to it to other network operators and service providers, such that data can be transmitted to and from the internet or to and from other networks. May be multiple hops away from infrastructure edge computing resources. See also: Access Network
**Customer-Premises Equipment (CPE)**
The local piece of equipment such as a cable network modem which allows the subscriber to a network service to connect to the access network of the service provider. Typically one hop away towards the end users from infrastructure edge computing resources. See also: *Access Network*

**Data Center**
A purpose-designed structure that is intended to house multiple high-performance compute and data storage nodes such that a large amount of compute, data storage and network resources are present at a single location. This often entails specialized rack and enclosure systems, purpose-built flooring, as well as suitable heating, cooling, ventilation, security, fire suppression and power delivery systems. May also refer to a compute and data storage node in some contexts. Varies in scale between a centralized data center, regional data center and edge data center. See also: *Centralized Data Center*

**Data Gravity**
The concept that data is not free to move over a network and that the cost and difficulty of doing so increases as both the volume of data and the distance between network endpoints grows, and that applications will gravitate to where their data is located. Observed with applications requiring large-scale data ingest. See also: *Edge-Native Application*

**Data Ingest**
The process of taking in a large amount of data for storage and subsequent processing. An example is an edge data center storing much footage for a video surveillance network which it must then process to identify persons of interest. See also: *Edge-Native Application*

**Data Reduction**
The process of using an intermediate point between the producer and the ultimate recipient of data to intelligently reduce the volume of data transmitted, without losing the meaning of the data. An example is a smart data de-duplication system. See also: *Edge-Native Application*

**Data Sovereignty**
The concept that data is subject to the laws and regulations of the country, state, industry it is in, or the applicable legal framework governing its use and movement. See also: *Edge-Native Application*

**Decision Support**
The use of intelligent analysis of raw data to produce a recommendation which is meaningful to a human operator. An example is processing masses of sensor data from IoT devices within the infrastructure edge to produce a single statement that is interpreted by and meaningful to a human operator or higher automated system. See also: *Edge-Native Application*

**Device Edge**
Edge computing capabilities on the device or user side of the last mile network. Often depends on a gateway or similar device in the field to collect and process data from devices. May also use limited spare compute and data storage capability from user devices such as smartphones, laptops and sensors to process edge computing workloads. Distinct from infrastructure edge as it uses device resources. See also: *Infrastructure Edge*
Device Edge Cloud
An extension of the edge cloud concept where certain workloads can be operated on resources available at the device edge. Typically does not provide cloud-like elastically-allocated resources, but may be optimal for zero-latency workloads. See also: Edge Cloud

Distributed Antenna System (DAS) Hub
A location which serves as an aggregation point for many pieces of radio communications equipment, typically in support of cellular networks. May contain or be directly attached to an edge data center deployed at the infrastructure edge. See also: Edge Data Center

Edge Cloud
Cloud-like capabilities located at the infrastructure edge, including from the user perspective access to elastically-allocated compute, data storage and network resources. Often operated as a seamless extension of a centralized public or private cloud, constructed from micro data centers deployed at the infrastructure edge. Sometimes referred to as distributed edge cloud. See also: Cloud Computing

Edge Computing
The delivery of computing capabilities to the logical extremes of a network in order to improve the performance, operating cost and reliability of applications and services. By shortening the distance between devices and the cloud resources that serve them, and also reducing network hops, edge computing mitigates the latency and bandwidth constraints of today’s Internet, ushering in new classes of applications. In practical terms, this means distributing new resources and software stacks along the path between today’s centralized data centers and the increasingly large number of devices in the field, concentrated, in particular, but not exclusively, in close proximity to the last mile network, on both the infrastructure and device sides. See also: Infrastructure Edge

Edge Data Center
A data center which is capable of being deployed as close as possible to the edge of the network, in comparison to traditional centralized data centers. Capable of performing the same functions as centralized data centers although at smaller scale individually. Because of the unique constraints created by highly-distributed physical locations, edge data centers often adopt autonomic operation, multi-tenancy, distributed and local resiliency and open standards. Edge refers to the location at which these data centers are typically deployed. Their scale can be defined as micro, ranging from 50 to 150 kW+ of capacity. Multiple edge data centers may interconnect to provide capacity enhancement, failure mitigation and workload migration within the local area, operating as a virtual data center. See also: Virtual Data Center

Edge Exchange
Pre-internet traffic exchange occurring at an infrastructure edge data center. This function will typically be performed in the edge meet me room of an infrastructure edge data center, and may operate in a supplemental or hierarchical fashion with traditional centralized internet exchange points if a destination location is not present at the edge exchange, as is the case with internet-bound traffic. An edge exchange may be used in an attempt to improve end-to-end application latency compared with a centralized internet exchange. See also: Internet Exchange Point (IXP)
**Edge Meet Me Room**
An area within an edge data center where tenants and telecommunications providers can interconnect with each other and other edge data centers in the same fashion as they would in a traditional meet me room environment, except at the edge. See also: *Interconnection*

**Edge Network Fabric**
The system of network interconnections, typically dark or lit fiber, providing connectivity between infrastructure edge data centers and potentially other local infrastructure in an area. These networks due to their scale and most frequent location of operation can be considered metropolitan area networks, spanning a distinct geographical area typically located in an urban center. See also: *Edge Exchange*

**Edge Node**
A compute node, such as an individual server or other set of computing resources, operated as part of an edge computing infrastructure. Typically resides within an edge data center operating at the infrastructure edge, and is therefore physically closer to its intended users than a cloud node in a centralized data center. See also: *Cloud Node*

**Edge-Enhanced Application**
An application which is capable of operating in a centralized data center, but which gains performance, typically in terms of latency, or functionality advantages when operated using edge computing. These applications may be adapted from existing applications which operate in a centralized data center, or may require no changes. See also: *Edge-Native Application*

**Edge-Native Application**
An application which is impractical or undesirable to operate in a centralized data center. This can be due to a range of factors from a requirement for low latency and the movement of large volumes of data, the local creation and consumption of data, regulatory constraints, and other factors. These applications are typically developed for and operate on the edge data centers at the infrastructure edge. May use the infrastructure edge to provide large-scale data ingest, data reduction, real-time decision support, or to solve data sovereignty issues. See also: *Edge-Enhanced Application*

**Fog Computing**
A distributed computing concept where compute and data storage resource, as well as applications and their data, are positioned in the most optimal place between the user and Cloud with the goal of improving performance and redundancy. Fog computing workloads may be run across the gradient of compute and data storage resource from Cloud to the infrastructure edge. The term fog computing was originally coined by Cisco. Can utilize centralized, regional and edge data centers. See also: *Workload Orchestration*

**Gateway Device**
A subcategory of the device edge, referring to devices on the device edge side of the last mile network which operate as gateways for other local devices, with the goal of aggregating and facilitating data transference between local devices, many of which are battery-operated and may operate for extended periods in a low-power state, and external entities such as a data analysis application operating inside an edge data center at the infrastructure edge. See also: *Resource Constrained Device*
Infrastructure Edge
Edge computing capability, typically in the form of one or more edge data centers, which is deployed on the operator side of the last mile network. Compute, data storage and network resources positioned at the infrastructure edge allow for cloud-like capabilities similar to those found in centralized data centers such as the elastic allocation of resources, but with lower latency and lower data transport costs due to a higher degree of locality to user than with a centralized or regional data center. See also: Device Edge

Local Breakout
The capability to put internet-bound traffic onto the internet at an edge network node, such as an edge data center, without requiring the traffic to take a longer path back to an aggregated and more centralized facility.

Interconnection
The linkage, often via fiber optic cable, that connects one party’s network to another, such as at an internet peering point, in a meet-me room or in a carrier hotel. The term may also refer to connectivity between two data centers or between tenants within a data center, such as at an edge meet me room. See also: Edge Meet Me Room

Internet Edge
A sub-layer within the infrastructure edge where the interconnection between the infrastructure edge and the internet occurs. Contains the edge meet me room and other equipment used to provide this high-performance level of interconnectivity. See also: Interconnection

Internet Exchange Point (IXP)
Places in which large network providers, among other entities, converge for the direct exchange of traffic. A typical service provider will access tier 1 global providers and their networks via IXPs, though they also serve as meet points for like networks. IXPs are sometimes referred to as Carrier Hotels because of the many different organizations available for traffic exchange and peering. The internet edge may often connect to an IXP. See also: Internet Edge

IP Aggregation
The use of compute, data storage and network resources at the infrastructure edge to separate and route network data received from the cellular network RAN at the earliest point possible. If IP aggregation is not used, this data may be required to take a longer path to a local CO or other aggregation point before it can be routed on to the internet or another network. Improves cellular network QoS for the user. See also: Quality of Service (QoS)

Jitter
The variation in network data transmission latency observed over a period of time. Measured in terms of milliseconds as a range from the lowest to highest observed latency values which are recorded over the measurement period. A key metric for real-time applications such as VoIP, autonomous driving and online gaming which assume little latency variation is present and are sensitive to changes in this metric. See also: Quality of Service (QoS)
**Last Mile**
The segment of a telecommunications network that connects the service provider to the customer. The type of connection and distance between the customer and the infrastructure determines the performance and services available to the customer. The last mile is part of the access network, and is also the network segment closest to the user that is within the control of the service provider. Examples of this include cabling from a DOCSIS headend site to a cable modem, or the wireless connection between a customer’s mobile device and a cellular network site. See also: Access Network

**Latency**
In the context of network data transmission, the time taken by a unit of data (typically a frame or packet) to travel from its originating device to its intended destination. Measured in terms of milliseconds at single or repeated points in time between two or more endpoints. A key metric of optimizing the modern application user experience. Distinct from jitter which refers to the variation of latency over time. Sometimes expressed as Round Trip Time (RTT). See also: Quality of Service (QoS)

**Latency Critical Application**
An application that will fail to function or will function destructively if latency exceeds certain thresholds. Latency critical applications are typically responsible for real-time tasks such as supporting an autonomous vehicle or controlling a machine-to-machine process. Unlike Latency Sensitive Applications, exceeding latency requirements will often result in application failure. See also: Edge-Native Application

**Latency Sensitive Application**
An application in which reduced latency improves performance, but which can still function if latency is higher than desired. Unlike a Latency Critical Application, exceeding latency targets will typically not result in application failure, though may result in a diminished user experience. Examples include image processing and bulk data transfers. See also: Edge-Enhanced Application

**Location Awareness**
The use of RAN data and other available data sources to determine with a high level of accuracy where a user is and where they may be located in the near future, for the purposes of workload migration to ensure optimum application performance. See also: Location-Based Node Selection

**Location-Based Node Selection**
A method of selecting an optimal edge node on which to run a workload based on the node’s physical location in relation to the device’s physical location with the aim of improving application workload performance. A part of workload orchestration. See also: Workload Orchestration

**Micro Modular Data Center (MMDC)**
A data center which applies the modular data center concept at a smaller scale, typically from 50 to 150 kW in capacity. Takes a number of possible forms including a rackmount cabinet which may be deployed indoors or outdoors as required. Like larger modular data centers, micro modular data centers are capable of being combined with other data centers to increase available resource in an area. See also: Edge Data Center
Mobile Edge
A combination of infrastructure edge, device edge and network slicing capabilities which are tuned to support specific use cases, such as real-time autonomous vehicle control, autonomous vehicle pathfinding and in-car entertainment. Such applications often combine the need for high-bandwidth, low-latency and seamless reliability. See also: Infrastructure Edge

Mobile Network Operator (MNO)
The operator of a cellular network, who is typically responsible for the physical assets such as RAN equipment and network sites required for the network to be deployed and operate effectively. Distinct from MVNO as the MNO is responsible for physical network assets. May include those edge data centers deployed at the infrastructure edge positioned at or connected to their cell sites under these assets. Typically also a service provider providing access to other networks and the internet. See also: Mobile Virtual Network Operator (MVNO)

Mobile Virtual Network Operator (MVNO)
A service provider similar to an MNO with the distinction that the MVNO does not own or often operate their own cellular network infrastructure. Although they will not own an edge data center deployed at the infrastructure edge connected to a cell site they may be using, the MVNO may be a tenant within that edge data center. See also: Mobile Network Operator (MNO)

Modular Data Center (MDC)
A method of data center deployment which is designed for portability. High-performance compute, data storage and network capability is installed within a portable structure such as a shipping container which can then be transported to where it is required. These data centers can be combined with existing data centers or other modular data centers to increase the local resources available as required. See also: Micro Modular Data Center (MMDC)

Multi-access Edge Computing (MEC)
An open application framework sponsored by ETSI to support the development of services tightly coupled with the Radio Access Network (RAN). Formalized in 2014, MEC seeks to augment 4G and 5G wireless base stations with a standardized software platform, API and programming model for building and deploying applications at the edge of the wireless networks. MEC allows for the deployment of services such as radio-aware video optimization, which utilizes caching, buffering and real-time transcoding to reduce congestion of the cellular network and improve the user experience. Originally known as Mobile Edge Computing, the ETSI working group renamed itself to Multi-Access Edge Computing in 2016 in order to acknowledge their ambition to expand MEC beyond cellular to include other access technologies. Utilizes edge data centers deployed at the infrastructure edge. See also: Infrastructure Edge

Network Function Virtualization (NFV)
The migration of network functions from embedded services inside proprietary hardware appliances to software-based VNFs running on standard x86 and ARM servers using industry standard virtualization and cloud computing technologies. In many cases NFV processing and data storage will occur at the edge data centers that are connected directly to the local cellular site, within the infrastructure edge. See also: Virtualized Network Function (VNF)
Network Hop
A point at which the routing or switching of data in transit across a network occurs; a decision point, typically at an aggregating device such as a router, as to the next immediate destination of that data. Reducing the number of network hops between user and application is one of the primary performance goals of edge computing. See also: Edge Computing

Northbound vs Southbound (and east/west)
The direction in which data is transmitted when viewed in the context of a hierarchy where the cloud is at the top, the infrastructure edge is in the middle, and the device edge is at the bottom. Northbound and southbound data transmission is defined as flowing to and from the cloud or edge data center accordingly. Eastbound and westbound data transmission is defined as occurring between data centers at the same hierarchical layer, for purposes such as workload migration or data replication. This may occur between centralized or between edge data centers. See also: Virtual Data Center

Over-the-Top Service Provider (OTT)
An application or service provider who does not own or operate the underlying network, and in some cases data center, infrastructure required to deliver their application or service to users. Streaming video services and MVNOs are examples of OTT service providers that are very common today. Often data center tenants. See also: Mobile Virtual Network Operator (MVNO)

Point of Presence (PoP)
A point in their network infrastructure where a service provider allows connectivity to their network by users or partners. In the context of edge computing, in many cases a PoP will be within an edge meet me room if an IXP is not within the local area. The edge data center would connect to a PoP which then connects to an IXP. See also: Interconnection

Quality of Experience (QoE)
The advanced use of QoS principles to perform more detailed and nuanced measurements of application and network performance with the goal of further improving the user experience of the application and network. Also refers to systems which will proactively measure performance and adjust configuration or load balancing as required. Can therefore be considered a component of workload orchestration, operating as a high-fidelity data source for an intelligent orchestrator. See also: Workload Orchestration

Quality of Service (QoS)
A measure of how well the network and data center infrastructure is serving a particular application, often to a specific user. Throughput, latency and jitter are all key QoS measurement metrics which edge computing seeks to improve for many different types of application, from real-time to bulk data transfer use cases. See also: Edge Computing

Radio Access Network (RAN)
A wireless variant of the access network, typically referring to a cellular network such as 3G, 4G or 5G. The 5G RAN will be supported by compute, data storage and network resources at the infrastructure edge as it utilizes NFV and C-RAN. See also: Cloud RAN (C-RAN)
Regional Data Center
A data center positioned in scale between a centralized data center and an edge data center. Significantly physically further away from end users than an edge data center, but closer to them than a centralized data center. Also referred to as a metropolitan data center in some contexts. Part of traditional cloud computing.  See also: *Cloud Computing*

Resource Constrained Device
A subcategory of the device edge, referring to devices on the device edge side of the last mile network which are often battery-powered and may operate for extended periods of time in a power-saving mode. These devices are typically connected locally to a gateway device, which in turn transmits and receives data generated by and directed to them from sources outside of the local network, such as a data analysis application operating in an edge data center at the infrastructure edge.
See also: *Gateway Device*

Service Provider
An organization which provides customers with access to its network, typically with the goal of providing that customer access to the internet. A customer will usually connect to the access network of the service provider from their side of the last mile.  See also: *Access Network*

Shared Infrastructure
The use of a single piece of compute, data storage and network resources by multiple parties, for example two organizations each using half of a single edge data center, unlike co-location where each party possesses their own infrastructure.  See also: *Co-Location*

Software Edge
From a software development and application deployment perspective, the point physically closest to the end user where application workloads can be deployed. Depending on the application workload and the current availability of computing resources, this point may be at the device edge, but will typically be within the infrastructure edge due to its cloud-like capability to provide elastic resources.
See also: *Network Function Virtualization (NFV)*

Throughput
In the context of network data transmission, the amount of data per second that is able to be transmitted between two or more endpoints. Measured in terms of bits per second typically at megabit or gigabit scales as required. Although a minimum level of throughput is often required for applications to function, after this latency typically becomes the application-limiting and user experience-damaging factor.
See also: *Quality of Service (QoS)*

Traffic Offloading
The process of re-routing data that would normally be delivered inefficiently, such as over long distance, congested, or high cost networks, to an alternative, more local destination (e.g., a CDN cache) or on to a lower-cost or more efficient network. Local Breakout is an example of using edge computing for traffic offloading.  See also: *Local Breakout*
Truck Roll
In the context of edge computing, the act of sending personnel to an edge computing location, such as to an edge data center, typically to resolve or troubleshoot a detected issue. Such locations are often remote and operate for the majority of the time remotely, without onsite personnel. This makes the cost other practical considerations of truck rolls a potential concern for edge computing operators.

Vehicle 2 Infrastructure (V2I)
The collection of technologies used to allow a connected or autonomous vehicle to connect to its supporting infrastructure such as an machine vision and route finding application operating in an edge data center at the infrastructure edge. Typically uses newer cellular communications technologies such as 5G or Wi-Fi 6 as its access network. See also: Access Network

Virtual Data Center
A virtual entity constructed from multiple physical edge data centers such that they can be considered externally as one. Within the virtual data center, work-loads can be intelligently placed within specific edge data centers or availability zones as required based on load balancing, failover or operator preference. In such a configuration, edge data centers are interconnected by low-latency net-working and are designed to create a redundant and resilient edge computing infrastructure.
See also: Edge Data Center

Virtualized Network Function (VNF)
A software-based network function operating on general-purpose compute resources which is used by NFV in place of dedicated physical equipment. In many cases, several VNFs will operate on an edge data center at the infrastructure edge. See also: Network Function Virtualization (NFV)

Workload Orchestration
An intelligent system which dynamically determines the optimal location, time and priority for application workloads to be processed on the range of compute, data storage and network resources from the centralized and regional data centers to the resources available at both the infrastructure edge and device edge. Workloads may be tagged with specific performance and cost requirements which determines where they are to be operated as resources that meet them are available for use.
See also: Software Edge

xHaul ("crosshaul")
The high-speed interconnection of two or more pieces of network or data center infrastructure. Backhaul and fronthaul are examples of xhaul.
MEMBERSHIP
State of the Edge reports are produced and funded collaboratively by a growing coalition of edge computing companies, with the explicit goal of involving a diverse set of stakeholders to produce original research without vendor bias. Supported by member funds and a community-driven philosophy, the State of the Edge mission is to accelerate the edge computing industry by developing free, shareable research that can be used by all.

Please get in touch to learn how you can be a part of the community: www.stateoftheedge.com/get-involved.

Founding Members

![Arm](image1)
![Edge Gravity](image2)
![Packet](image3)
![RAFAY](image4)
![Vapor](image5)

General Members

![Catchpoint](image6)
![DETECON](image7)
![Edgeconnex](image8)
![MACROMETA](image9)

![Mutable](image10)
![Orl](image11)
![Pluribus Networks](image12)
![Seagate](image13)

![SBA](image14)
![Section](image15)
![StackPath](image16)

Media & In Kind Partners

![Mission](image17)
![Edge Research Group](image18)
![Connected Magazine](image19)
![IGR](image20)

![Computing World](image21)
CREDITS
The State of the Edge is a community-led effort, and draws much of its expertise from the talented staff of its members as well as outside experts.

Financial Modeling and Research
The financial model and significant portions of the supporting narrative were developed by Dr. Phil Marshall of Tolaga Research.

Editors
Jacob Smith and Matt Trifiro led the planning and execution of the 2020 report, and also served as its primary editors.

Special thanks to Jim Davis, Phillip Marangella, Alex Marcham, Stacey Norwood and Rebecca Hurd for their tireless proofing. Additional thanks to Rob Hegblom, Thomas Switala, Joseph Noronha and Peter Christy for evaluating and providing suggestions to improve the forecast model.

Stacey Norwood and Zoe Allen coordinated the Postcards from the Edge. Additional thanks to each of the authors who submitted an essay for publication.

Illustrations & Design
Illustrations are by Mo Moussa. The report was designed by Megan Fentzloff.

Industry Experts
In addition to those listed above, we would also like to thank Monica Paolini of Senza Fili, Iain Gillot of iGR, Harry Smeenk of TIA, Arpit Joshipura of The Linux Foundation, Chris Aniszczyk of the Cloud Native Computing Foundation, Ildiko Vancsa of the OpenStack Foundation, Clay Moran of SBA Communications, Hugh Carspecken of Dartpoints, Lance Crosby of StackPath, Zac Smith of Packet, Haseeb Budhani from Rafay, Ihab Tarazi from Dell, Iyad Tarazi from Federated Wireless, Jason Hoffman from MobiEdgeX, Yves Beaudreau from EdgeGravity by Ericsson, Ivo Rook from Sprint, Mahadev Satyanarayanan of Carnegie-Mellon University. All of these experts offered their time and expertise to our researchers to help prepare the report and validate its accuracy.