

STATE OF THE  
**EDGE**  
2018

**A Market and Ecosystem  
Report for Edge Computing**



# Prologue

## Building a Community Around Edge Computing

Over the last couple of years, edge computing has emerged from relative obscurity to become one of the most talked about trends in internet infrastructure. Some critics believe that there is no need for edge computing, but I firmly believe it will be one of the most significant technology disruptions in my lifetime—building upon and exceeding the massive transformations of the public cloud.

The disruptive potential of edge computing is fueled by the unprecedented growth of data, the imminent impact of 5G networks, the growing importance of latency and regulation in dealing with data, and the emergence of a distributed computing architecture that favors specialized hardware like GPU's and offloads. As a result, infrastructure is starting to evolve at "software speed" – iterating rapidly and attracting a wide array of contributors.

Like most new ecosystems at early stages of development, the excitement and potential of edge computing is a complex set of definitions by a wide range of participants, adding to the confusion around this topic. This report seeks to rationalize the different factions and help the industry converge on a common definition of edge computing and its related concepts.

Although there is no single edge, nor a single type of edge computing, we all benefit greatly from a shared understanding and a strong dialogue. In this regard, the inaugural *State of the Edge* report (with its diverse group of supporters and authors) is built on a compelling premise: that collaboration and openness can greatly accelerate even the most complex ecosystems.

A fantastic example of this mindset is the [Open Glossary of Edge Computing](#), which was developed for this report but has also been turned into an evergreen open source project where contributors are invited to provide suggestions, corrections and additions. Another example is the [Edge Computing Landscape](#), to which anyone may suggest edits and additions. Already, many standards groups and organizations, including the Telecommunications Industry Association (TIA) and the Cloud Native Computing Foundation (CNCF), are contributing towards these efforts. This is a powerful trend to support, and I encourage you to get involved and add value wherever you can.

The opportunities in edge computing are immense. In a recent talk, Tim Hockin, a Principal Software Engineer at Google and one of the leaders of the Kubernetes project, proclaimed that it's an "exciting time for boring infrastructure." I couldn't agree more. The level of innovation in every aspect of infrastructure down the component level is accelerating and we are witnessing a new renaissance.



**Ihab Tarazi**  
Former CTO, Equinix

# Forward

It's hard not to get caught up in the cacophony about edge computing.

Whether you're exploring the newest open source projects, watching provocative keynotes, perusing the most recent tidal wave of vendor press releases, or merely having casual discussions with customers and peers, one thing is clear: everyone is in a rush to define, understand and conquer the edge. But what is the edge really?

We posit four principles:

- The edge is a location, not a thing;
- There are lots of edges, but the edge we care about today 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.

It is in this context that we, as well as our colleagues at Ericsson (UDN), Arm, and Rafay, created the inaugural *State of the Edge* report to cut through the noise, bring some order to the discussion, and build a community that cares deeply about edge computing and the innovations that will be required to bring its promise to fruition.

When it comes to evolving the internet, there is no finish line. This is an ongoing community effort. We encourage you to join our [Slack group](#) and help shape future editions of the State of the Edge report as well as participate in offshoot projects, such as the Open Glossary of Edge Computing.

Sincerely,



**Matt Trifiro**  
CMO, Vapor  
Report Co-Chair



**Jacob Smith**  
SVP Engagement, Packet  
Report Co-Chair

The State of the Edge 2018 Report was curated and sponsored by the following companies:

**arm**



**packet**



Independently researched and prepared by:



# Executive Summary

This report—the first in an ongoing series—aims to show the state of edge computing today and educate the reader on where this collection of technologies is heading in the near future. Among the findings:

- Today's internet is built around a centralized cloud architecture, which alone cannot feasibly support emerging application and business requirements.
- Edge computing has the potential to improve the performance, scalability, reliability and regulatory compliance options for many critical applications.
- There is a strong historical basis for edge computing, growing out of the efforts by global CDN's in the 1990s to distribute the delivery of content to the network edge.
- Edge computing resources can be located are on the operator side or on the user side of the last mile network. Resources on the operator side of this line are referred to as the infrastructure edge; on the user side, the device edge.
- Device edge resources are often constrained by power and connectivity. At the infrastructure edge, there is the potential for dynamically scalable resources that mimic a centralized cloud experience (although at smaller scale).
- Edge computing and centralized cloud services are not mutually exclusive. Multi-tier, hybrid hierarchical architectures exist that can leverage a wide variety of infrastructure efficiently.
- The recent widespread growth and adoption of cloud native technologies is a primary enabler for a robust edge computing ecosystem.
- Accelerators (Smart NIC's, GPU's, etc) have a central role in edge computing due to an emerging distributed architecture, as well as power constraints and workload requirements.
- Interest in (and use of) edge computing is growing rapidly with significant interest from infrastructure manufacturers, network and cloud operators and application developers.
- These trends have various business model implications, including pricing and purchasing models that account for resource contention, Service Level Agreements (SLA's), and expectations of physical security.
- There is a tremendous opportunity to simplify and grow global infrastructure with edge computing.



# Table of Contents

Foreward	2
Prologue	3
Executive Summary	4
Chapter 1: Setting the Stage	6
Chapter 2: The Edge and its Ecosystem	13
Chapter 3: Software at the Edge	26
Chapter 4: The Genesis of Edge Computing	31
Chapter 5: Key Drivers for the Edge Ecosystem	38
Chapter 6: Applications of Edge Computing	45
Chapter 7: Implications	60
Epilogue: A Call to Community	69
Credits	71
Edge Computing Landscape	75
Open Glossary of Edge Computing	78

CHAPTER 1

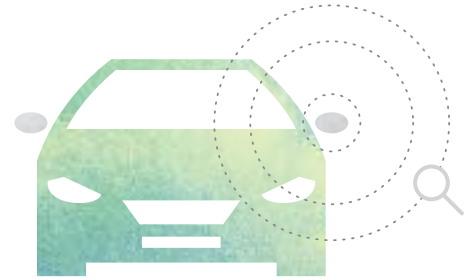
# *SETTING THE STAGE*

*“The edge is a once-in-a-generation  
innovation cycle. The current internet  
doesn't give us what we need...  
We need to rebuild it.”*

*– Yuval Bachar, LinkedIn and Open19 Foundation*

## Why Edge? Why Now?

A car going 60 miles per hour travels half the distance of a football field every two seconds. Moving at that speed, a driver—whether human or machine—has only a fraction of a second to identify a risk and press the brake in order to prevent an accident.



Self-driving cars, the ones you've probably read about or maybe you've seen experimentally on the road, promise to reduce collisions by not getting distracted and by making decisions faster than human drivers. However, in their current configurations, they are impractical except as experiments. They are data centers on wheels—and this is as absurd as it sounds. They contain hundreds of thousands of dollars of specialized IT gear and sensors, have a great deal of complexity, and consume a substantial amount of power that could otherwise go towards other things, such as moving the car.

This is not to say that self-driving cars won't have many, powerful onboard computers—for they surely will. Nor do we claim they won't be able to navigate roads when there is no network connectivity. What we believe is: they are unlikely to ever be solely self-contained systems.

Self-driving cars—in the interest of safety, traffic flow coordination, regulatory compliance and cost reduction—will combine their own capabilities with data and decision-support systems that exist elsewhere. Each self-driving car will have onboard embedded control loops that can operate the vehicle without network connectivity (albeit, most likely in a degraded fashion).

The wide-scale deployment of self-driving vehicles will leverage real-time connections to edge resources in order to incorporate traffic flow, situational awareness and external decision support into its driving algorithms.

In this environment, edge computing becomes the catalyst for the wide-scale deployment of autonomous vehicles. Software for autonomous driving will be deployed to both the vehicle and to nearby edge nodes. The software running on the device (the car) will coordinate with software running on the infrastructure at the edge. These two edge computing configurations (device and infrastructure) connect over the last mile network and work in harmony to reduce collisions, improve traffic, and increase life safety.

We are witnessing the convergence of many forces driving edge computing, from the advent of super-powerful GPUs to the business imperatives driving network upgrades. Bringing these forces together will create the opportunity for new, complex, and unprecedented edge applications.



# *What is Edge Computing?*

Edge computing places high-performance compute, storage and network resources as close as possible to end users and devices.<sup>1</sup> Doing so lowers the cost of data transport, decreases latency, and increases locality.

Edge computing will take a big portion of today's centralized data centers and cloud and put it in everybody's backyard.

Edge computing will unleash a cornucopia of opportunities for every size and shape of institution, from governments to grocery stores. Applications will leap from the pages of science fiction and become practical services.

We will witness a major transformation of today's internet. Billions of IoT devices will generate zettabytes of data, which will be processed at the edge. A vast array of new, exciting applications, ranging from augmented reality (AR) to self-driving cars, will leverage edge services. There will be new regulations, new multi-billion dollar companies, and new discoveries unearthed about ourselves and our world.

This is the realm of edge computing.

<sup>1</sup>One of the biggest challenges of this report has been to document the most accurate and least controversial definitions of edge computing and related terms. Recognizing the challenge, we've created the [Open Glossary of Edge Computing](#), a collaborative open source project that can accommodate an unlimited number of authors contributing toward iterating and improving upon the definitions in our shared lexicon.

# *Extending the Cloud to the Edge*

Today's centralized cloud concentrates its resources in a relatively small number of large data centers, mostly in remote locations where land and electrical power are cheapest. This conventional model of cloud computing has revolutionized how we build and deliver applications, but it's reaching its limit.

A profound side effect of centralized data centers is that they concentrate cloud resources in locations that may be thousands of miles from the end user or device that depends on them. That may not sound like a big deal; after all, modern networks are fast, right?

It turns out, they're not fast enough. When data must travel hundreds or thousands of miles, there are a few issues we encounter. First, data can't go faster than the speed of light, and this is a major component of the total time it takes to transmit data over extended distances. Second, each mile that data travels (and, correspondingly, each network hop that it traverses) makes it more likely that data will be held up or dropped due to other users on the network; there are no guarantees of end-to-end data delivery, let alone in real-time, on the modern internet.

In order to understand why latency is such an important issue for the emerging edge, we must first recognize the explosive growth of machine-to-machine connections on the internet and the reality that machines can "talk" to each other in timescales that are imperceptible to humans. The number of IP-connected devices on the internet, whether it's going to be 20 billion or 50 billion<sup>2</sup>, doesn't matter most. What matters most is how machine-scale latencies change the demands on the internet.

The limits of human perception tend to bottom out in the 10ths of a second range. That's why a movie shown on a screen at 24 frames per second (some argue it is 29.97!) appears perfectly smooth. The explosion of data at the edge makes latency critical. Without ultra-low latency, it is not possible to move great volumes of data across the internet. As the data collected at the edge grows, the need for low latency becomes more critical.

For applications that operate at human-scale, where an acceptable response might take a few seconds, the distance data needs to travel may not make a noticeable difference and, if it does, then your Facebook page may take a bit longer to load or your video may buffer more than you'd like—but it's mostly an inconvenience. Annoying, yes; but not life-threatening.

In comparison, for applications that operate at machine-scale, where latency is measured in microseconds, even small delays in data transmission can have severe and even fatal consequences. More than just inconvenient, delays at this scale can create risk to property, as in a cloud-controlled robotic lathe, or risk to people, as in a self-driving car.

<sup>2</sup> There is controversy over how many IoT devices and machines will be connected to the Internet and how quickly the count will grow. Analyst projections tend to range between 20 and 50 billion IP-connected devices by 2020. While projections are just guesses—and in this new world, could be off by a few billion either way—it's clear there will be an explosive number of machines at the edge internet, all generating a flood of data, looking to communicate with each other and the cloud at very low latencies.

# *Edge Data Center vs. Edge Cloud*

You can have edge computing without an edge cloud, but you can't have an edge cloud without edge computing.

In reality, data centers are just buildings or structures that house IT equipment. Resources in a data center become part of a cloud only when a private enterprise or a public cloud provider incorporates them into their cloud via a software-driven control plane—which can be as privately-controlled as Amazon Web Services or as publicly-available as OpenStack.

What's unique about the edge cloud is not just the presence of edge attributes, such as location and latency, but also the possibility of cloud services running on the devices themselves. Imagine controlling workloads on an edge device in the field with the same tools you use to control workloads in the centralized cloud. By extending the cloud control plane across the last mile network, edge devices may be presented as part of the extended cloud infrastructure, making it possible to program the devices using cloud principles and in ways that are tightly coupled to the cloud proper.

Cloud resources on the device edge won't be as flexible, fungible and scalable as those on the infrastructure edge, though they will have the advantage of being the preferred option for some workloads. For example, certain applications will benefit from a "zero hops" relationship to the end user, for performance, security, or reliability reasons.

Developers and operations teams often won't care where a workload runs, provided it meets stated SLAs or policies. In these instances, automated schedulers will map workload requests onto target environments on both sides of the last mile network. These schedulers will place workloads on the device edge and the infrastructure edge, depending on many criteria.

What's important to note is this if you intend to build an edge cloud, it must be built atop data centers and devices at the edge.

# *The Next Generation Internet*

We're already seeing conventional services, from content delivery to Infrastructure as a Service (IaaS), being enhanced with edge computing. Today's internet and cloud services will all benefit from edge computing and the advantages it brings in terms of latency, inexpensive data transport and locality.

Edge computing is more than just flashy, futuristic applications. Smart cities, IoT and autonomous vehicles are all coming, but they will often be preceded by more conventional uses of edge computing referred to as edge-enhanced applications, such as extending today's content delivery networks.

The remainder of this report, as its title suggests, shows the state of edge computing today and aims to educate readers on where we see this innovative collection of technologies heading in the near future. Use cases, both immanent and future-facing, are covered in detail throughout the "Chapter 6: Applications of Edge Computing" and "Chapter 7: Implications" sections of this report.

Edge computing is here, it's real—and every future-facing network, cloud and application operator must look closely at the unique benefits this technology offers.

CHAPTER 2

# *THE EDGE AND ITS ECOSYSTEM*





*“Computing will become an increasingly movable feast... Processing will occur wherever it is best placed for any given application.”*

*– The Economist "Life on the Edge"*



# Defining Edge Computing

Like many other promising new technologies, edge computing has generated a large amount of discussion throughout industry and academia, and amongst a wide variety of organizations. Although earlier attempts have been made at defining edge computing, they typically lack a collaborative approach which incorporates viewpoints from many different perspectives within the community to create a solid, agreed-upon definition that can be used as the basis for practical systems.

State of the Edge sponsors and contributors – as well as the wider community of organizations, academics and other industry figures who are deeply interested in edge computing – have come together to generate the definitions not just of edge computing itself, but also of its component parts and layers that are used throughout this report. These definitions are the result of thousands of discussions, hundreds of research projects and many hours of focused collaboration on the key terminology that must be established for the industry as a whole to move forward with edge computing.

To that end, the first question that must be answered in a discussion on edge computing is – what is edge computing? This is a complex question in itself, and one with many potential answers. Our definition of edge computing simplifies the question of what is and is not edge computing into simple terms, making this new area of technology simple to understand.

**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 side and the device side.*

This definition can be understood by dissecting its two parts.

1. Edge computing in a theoretical sense
2. The more practical details of what a real-world edge computing architecture looks like.

As the second paragraph of our definition of edge computing states, the edge location we are primarily concerned with for the purposes of this report is the last mile network. This is because the last mile network forms the demarcation point between the network operator (e.g. the service provider) and the user and devices who connect to that network.

As our definition of edge computing suggests, edge computing is not limited to existing on just one or the other side of the last mile network. On the contrary, edge computing can and does exist on both sides of the last mile network concurrently; but the ways in which this is achieved and what this means for everyone from the user to the cloud provider or application developer vary significantly. Although edge computing resources on either side of this line of demarcation may interoperate, it is useful for our understanding to categorize(d) them separately.

# Device Edge

The *device edge* refers to edge computing resources on the downstream or device side of the last mile network. These include conventional devices, such as laptops, tablets and smartphones, but also objects we don't normally think of as internet devices, such as connected automobiles, environmental sensors and traffic lights.

Some devices will be single function, such as embedded sensors, designed to perform very specific tasks and deliver streams of data to the network. Other edge devices will act as specialized gateways, aggregating and analyzing data and providing some control functions. And yet other edge devices will be fully-programmable compute nodes, capable of running complex applications in containers, VMs, or on bare metal.

Resources on the device edge are as close as you can possibly get to the end user, both physically and logically. This makes them well-suited for tasks which require as close to zero latency as possible, such as user interface processing, activating the brakes on an autonomous car, or other use cases which require low latency but which do not require complex processing. Although many edge devices such as the modern smartphone are more powerful than the desktop computers of years past, the complex processing required for advanced applications such as Artificial Intelligence (AI) and machine vision have outpaced the resources available on these edge devices.

While there is significant debate as to whether resources on the device edge should be thought of as part of the edge cloud proper, it's already clear that many device edge resources will be connected to the cloud and be managed as extensions of the cloud. To some extent, this discussion is moot, as major cloud providers already give their customers the ability to manage device edge resources as if they were a natural extension of core cloud services, including deploying cloud-like workloads to those devices.

Regardless of how resources on the device edge are managed, it is clear they will largely be connected to the infrastructure edge over wired and wireless networks and that workloads running on the device edge will be coordinated with workloads running on the infrastructure edge. Modern tools may even give developers a choice as to which side of the last mile they'd like to run. In many cases it will be both more reliable and less expensive to run workloads on the infrastructure edge rather than entirely on the edge devices. Additionally, there are many infrastructure functions that are not performed on the device and are much better supported with edge computing.

The device edge will be the basis of many useful applications which require the lowest possible latency, as device edge resources are as close as it is possible to be to the end user. However, most applications will be built from services that span both the infrastructure edge and the device edge, leveraging the unique capabilities of each as part of the edge cloud.



# Infrastructure Edge

The *infrastructure edge* refers to IT resources which are positioned on the network operator or service provider side of the last mile network, such as at a cable headend or at the base of a cell tower. While “last mile network” is a high-level term which has many nuances and exceptions when you dig into the details, the infrastructure edge can generally be thought of as large-scale facilities owned and operated by a service provider.

The primary building blocks of the infrastructure edge are edge data centers, which are typically placed at 5-10 mile intervals in urban and suburban environments. While resources in edge data centers on the infrastructure edge are further away, physically and logically, from users than resources on the device edge, they are still close enough to provide low round-trip latencies (5-10ms) to most devices and they are also capable of housing equipment that is orders of magnitude more powerful than what exists (or is easily accessed) on the device edge.

The infrastructure edge can be viewed as a mid-point between the device edge and the traditional centralized cloud, aiming to combine the advantages of both. Constructed from a potentially large number of edge data centers which are located within a few miles of their intended end users, the infrastructure edge will provide a “cloud-like” experience, but with the locality (and thus many of the latency advantages) of the device edge. While not as scalable as a centralized cloud data center, the typical infrastructure edge deployment will provide enough compute, data storage and network capacity to support elastic resource allocation for workload within a local operational area, such as a city.

The infrastructure edge may not match the latency performance to end users that the device edge is able to provide, but it will likely be no more than one network hop away and within only a few milliseconds of total latency. In addition to being capable of supporting far more demanding applications than the typical device edge (such as AI, due to more powerful resources) the infrastructure edge provides a high performance springboard to centralized cloud resources.

For many applications, the infrastructure edge will provide the ideal balance between latency and resource density. Importantly, edge data centers will be capable of supporting the wide array of computing hardware that is quickly defining distributed architectures, including specialized accelerators such as FPGAs, GPUs and smart NIC’s. There are, of course, many cases where the lowest possible latency of the device edge or the larger resources of the centralized cloud are a better fit for the application workload at hand, and so they should be used instead.

# Visualizing the Edge Topology

When discussing the comparison between the device edge and infrastructure edge, it is helpful to visualize differences in the scale of compute, storage and network resources available at each of these locations.

The resources provided by a single edge data center deployed at the infrastructure edge dwarf those of a 100 edge device heterogeneous network, which in itself may often be challenging to establish and dynamically assign a specific application workload to.

With a sufficient infrastructure edge deployment where multiple edge data centers are deployed in a given area (such as a city), the resources available at the infrastructure edge are significantly larger than those which can be provided by the device edge, in cases where the infrastructure edge is not congested.

## RELATIVE SCALE OF RESOURCES IN KW

Edge Device Operating  
*Operating Standalone*



100 Edge Device Hetnet  
*Using Available Resources*



Infrastructure Edge  
*Micro Data Center*



# Infrastructure Edge Sublayers

The infrastructure edge itself can be split into two sublayers: the *access edge* and the *aggregation edge*.

The access edge is the part of the infrastructure edge closest to the end user and their devices. Edge data centers deployed at or very near this sublayer are typically directly connected to a radio or other front-line network infrastructure, and they are used to operate application workloads for complex tasks such as machine vision and automated decision support for large-scale IoT. Edge data centers deployed at the access edge, a sublayer within the infrastructure edge, may also connect to other edge data centers which are deployed above them in a hierarchical architecture at the aggregation edge sublayer.

The aggregation edge refers to a second sublayer within the infrastructure edge which functions as a point of aggregation for multiple edge data centers deployed at the access edge sublayer. The purpose of this layer is to provide a reduced number of contact points to and from other entities, such as a centralized cloud data center and the infrastructure edge and to facilitate the collaborative processing of data from multiple access edge sublayer edge data centers. The aggregation edge is typically two network hops from its intended users but is still much closer to them than the centralized cloud data center, and it is thus able to achieve far lower latencies.

Consider the example of a city-wide HD video surveillance deployment. The vast majority of the video footage which is recorded by the system is of no interest to those outside the local area. Therefore, it is wasteful, difficult and expensive to transmit this large volume of video data to a centralized cloud data center which may be hundreds or thousands of miles away. Instead, this data can be sent to the infrastructure edge.

First, this data is received by edge data centers at the access edge sublayer. These edge data centers process the video data using technologies such as machine vision and then pass data of interest to the edge data center above them at the aggregation edge sublayer. This aggregation edge data center analyzes video data from a wide variety of sources and can use cross-referencing between them to produce valuable insights such as improved object or person detection. In this example, both the access and aggregation edge sublayers are critical; the access edge performs the bulk frontline processing and passes only that which is of interest to the aggregation edge data center for more analysis.

# Edge Data Centers

As mentioned previously, the edge data center is the key building block of the infrastructure edge and of its access and aggregation sublayers. The term edge data center itself does not define the scale of the data center, either alone or as part of an infrastructure edge deployment.

Individual edge data centers are typically 50 to 150 kW in capacity. These infrastructure edge nodes are capable of supporting the same type of computing hardware as large, centralized cloud data centers, from x86 or ARM general-purpose CPUs to specialized accelerators such as FPGAs, GPUs and TPUs, as well as data storage and network equipment. In this way, edge data centers are capable of providing the same type of resources and supporting the same type of services as the centralized cloud data center, albeit at a smaller scale individually than those hyperscale installations.

The edge data center is required to be deployed in a range of non-traditional locations such as at the base of cellular network towers, in structures which were not originally designed to support data center equipment due to their flooring or cooling systems, and even in some cases directly outdoors with no sheltering structure available. These requirements have given rise to innovative edge data center designs which aim to support the maximum density of IT equipment within the smallest possible footprint, with under 10 feet in diameter considered a magic number in terms of the range of structures these edge data centers can be deployed in.

Interoperation between multiple edge data centers within the local area is key to ensure that the infrastructure edge, and the edge cloud as a whole, is capable of supporting cloud-like services as users move across a local coverage area such as a city in a high-performance, reliable and redundant fashion. By interconnecting multiple edge data centers with a mesh network of fiber, the network or cloud operator can create an infrastructure edge that meets these requirements.

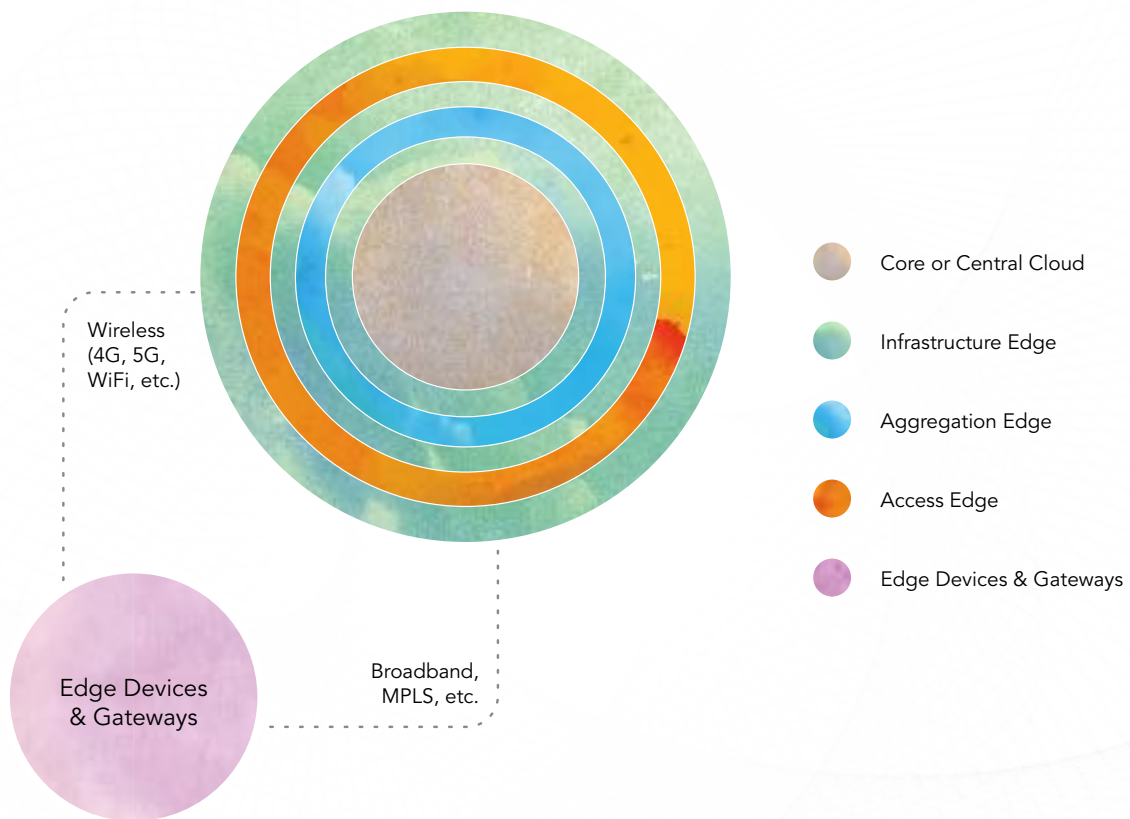
When interconnected in this way, multiple edge data centers can be represented to an external workload orchestration or management system as a single, cohesive unit where each edge data center is similar to a rack within a large centralized data center, with the advantage of multiple points of locality to choose from in order to achieve the best performance for latency-sensitive application workloads depending on which edge data center the user is closest to at the time.

# Cloud Interoperation

Edge computing does not exist by itself. Despite the level of computing power and performance that is achievable between the combination of the device edge and infrastructure edge, both of these entities benefit immensely from tight, cohesive interoperation with the centralized cloud.

As can be seen in the diagram below, both the device and infrastructure edge can be viewed as complementary to, and even as extensions of, the existing centralized cloud. By connecting these distributed resources together and creating an edge cloud which spans from the current centralized data center, through the infrastructure edge and its sublayers through to the device edge, the cloud operator is able to dynamically allocate resources and direct any application workloads to the optimal location for them at a given point in time, regardless of whether that is in the device edge, infrastructure edge or the centralized cloud. This process can be automated for maximum efficiency, as if all of the distributed edge computing resources between the device edge and infrastructure edge were specialized racks within a single, large cloud data center.

EDGE CLOUD LAYERS





# What is Fog?

*“One of the most common misconceptions about fog computing is that it’s just another term for edge computing. Fog computing is an end-to-end horizontal architecture that distributes computing storage, control, and networking functions closer to users along the cloud-to-thing continuum.”* – Helmut Antunes, Chairman, OpenFog Consortium

In 2012, Flavio Bonomi, a researcher at Cisco, coined the term “fog computing.” Back then, Cisco was advancing what they called the “Internet of Everything,” which, in Cisco’s view, would emerge by pushing cloud capabilities throughout the network, using their network equipment.

Early on, fog computing struggled to separate itself from Cisco and to differentiate from edge computing. At one point, even its creator jokingly called it a “marketing term for edge computing.” However, in recent years, especially as Cisco transitioned its fog work to the OpenFog Consortium, fog computing has found a solid niche focusing on architectural patterns and technology for distributing workloads across the entire continuum from core to device.

In the context of this evolved definition of fog computing, fog becomes a type of workload that can be run atop edge computing infrastructure. In other words, IT equipment placed in edge environments can be incorporated into the end-to-end architecture that spans from the centralized core to the in-the-field edge device. Fog computing is a type of workload supported by edge computing, not a competitor; and like edge and cloud, the two work best together.

The OpenFog Consortium is the definitive organization for fog computing. They’ve partnered with the IEEE to publish standards for fog computing, and they’ve partnered with ETSI MEC to incorporate MEC-style edge computing into the fog continuum.

For more information, visit [www.openfogconsortium.org](http://www.openfogconsortium.org)

# What is MEC?

Multi-Access Edge Computing (MEC) plays a key role in the edge computing ecosystem. It refers to the official edge computing standards and practices that have emerged from ETSI, the large European standards body. The core of the effort is an open framework for applications and services that are tightly coupled with the Radio Access Network (RAN), providing open interfaces for integrating software services into the wireless cellular networks.

In 2013, Nokia Siemens introduced MEC as a way to augment cellular technologies with edge applications. At the time, Nokia branded this capability “Liquid Applications” and sought to retain it as a proprietary technology. Nokia had recognized an opportunity to open up their cellular network baseband units (BBUs) as a computing platform, and pioneered a new class of edge computing workloads that would benefit from close proximity and direct access to the RAN and its data. These applications sought to combine many cloud-like capabilities with the cellular network, exposing key APIs and general-purpose compute resources to third party developers.

Nokia soon realized that the best way to attract developers to MEC at large scale would be to advocate for an open standard, supported by many manufacturers. In 2015, Nokia— along with many partners that included Intel, IBM, NEC, NTT Docomo, Vodafone, and Orange—created an Industry Specification Group within ETSI. This aim of this group was to further the use of MEC.

In 2017, ETSI repositioned the MEC standards group to expand its aim beyond only cellular networks, recognizing that many MEC capabilities would soon be applied to wired and wireless networks alike. To reflect this change, the name of the group was changed from the original Mobile Edge Computing to Multi-Access Edge Computing. This change illustrated the growing shift in the industry towards edge computing for a variety of use cases, some newer than others.

The core goal of MEC, borne out of its cellular network operator origins, is to create a standard software platform, API and programming model to ease the development of edge applications that interface with network, primarily with the cellular RAN. Although MEC workloads can be run on dedicated network infrastructure devices, the infrastructure edge provides the ideal blend of locality to the RAN or other network infrastructure and processing resources for MEC services.

CHAPTER 3

# *SOFTWARE AT THE EDGE*



*“Kubernetes is a platform for building other platforms. It's a better place to start; not the endgame”*

– Kelsey Hightower, Developer Advocate, Google Cloud Platform

# Edge Applications

In many ways, edge computing is defined by its applications. Although with the cloud-like resources provided by the infrastructure edge in particular, the vast majority of applications can operate from edge computing infrastructure, there are two main categories of application which we can use to better understand the impact of edge computing on our application workloads.

## Edge-Native Applications

The first of these application categories is edge-native applications. Edge-native applications, as their name suggests, are applications which require the unique characteristics provided by edge computing to function satisfactorily, or in some cases to function at all. These applications will typically rely on the low latency, locality information or reduced cost of data transport that edge computing provides in comparison to the centralized cloud. Examples include the operation of autonomous vehicles, which necessarily rely on the processing of large amounts of complex data in real-time to make rapid and accurate decisions, and HD video surveillance systems which require a low cost of data transit from camera to data center to be economically viable. In the majority of cases, these applications are not viable or in some cases, possible at all to operate from the traditional centralized cloud data center, and so rely on edge computing.

Many edge-native applications, such as autonomous vehicles, can also be thought of as latency-critical applications. These are applications which will experience failure if a specific latency range is exceeded. One example of these is an autonomous aircraft; consider the case where an autonomous aircraft is asked to take off on an aerial surveillance flight, but the latency between the flight control application in the data center and the autonomous aircraft itself is measured to be too high for safe operation. In this case, to protect the safety of the aircraft and any people or property that may be endangered by its unsafe flight, the aircraft can refuse to fly.

*“Edge native applications will soon revolutionize the internet. We’re already seeing a new generation of developers building services that bring the cloud to the very edge of the last mile network, creating game-changing value in conjunction with the top wireless operators.” – Jason Hoffman, CEO MobileEdgeX*

# *Edge-Enhanced Applications*

Edge-enhanced applications are applications which do not require the unique characteristics of edge computing to operate, but that benefit from them nonetheless. Rather than innovative new use cases that may capture the imagination such as autonomous vehicles, edge-enhanced applications encompass such tried-and-true areas as complex web hosting, content delivery and Infrastructure as a Service (IaaS). These are applications which in many cases already operate today from the centralized cloud data center, but which can all benefit from the lower latency and reduced cost of data transport that edge computing through the use of edge data centers deployed at the infrastructure edge is capable of providing. Whether they are enhanced in terms of performance or operational cost, these applications benefit from edge computing.

In many cases, edge-enhanced applications are also latency-sensitive applications. Unlike latency-critical applications, latency-sensitive applications will typically not experience failure if their specified latency range is exceeded, but if this occurs it will result in a sub-optimal user experience. An example of this is a simple voice call; the call may not fail if the latency is longer than is ideal, but the user will become frustrated as they keep talking over the other participants.

## *Developer Takeaway*

From the developer perspective, the addition of edge computing may appear to add another layer of complexity beyond that which exists today. However, as will be seen in the discussion on workload orchestration in a later chapter, the majority of this complexity can be abstracted from the developer and from the user by the use of a sophisticated workload orchestrator.

As a developer, you may not need significant work for your existing application to operate from edge computing infrastructure. Many current applications will function from an edge data center just as well as they do today from a centralized cloud data center. Unless your application must be operated from edge computing resources for performance or cost requirements, you may not make a specific choice as to whether it operates in the centralized cloud or at the edge at all.

Every application is different, and as always there will be exceptions to any general rule. But, for many developers today, the hurdle they must jump in order to capitalize on edge computing and the performance and cost advantages it can bring is low, and in some cases will be nonexistent.



# *Edge Orchestration, as with Kubernetes and Apache Mesos*

Running real-time workloads across the highly-distributed infrastructure presented by edge computing introduces many complicated challenges to developers and operators. How do you even begin to decide which workloads should run where?

Fortunately for edge computing enthusiasts, some of the recent work in container orchestration systems, such as Kubernetes and Apache Mesos, have looked to start solving for these types of complex scheduling problems. Because each of these systems allows you to install customer schedulers, they can be extended to take into account increasingly sophisticated levels of edge criteria for workload placement, automating decisions in real-time, abstracting away the complexity from developers and operators, who would prefer to simply specify the SLAs they require.

A custom scheduler for edge computing might contemplate many sophisticated attributes requested by workloads. In addition to the typical scheduling attributes such as requirements around processor, memory, operating system, and occasionally some simple affinity/anti-affinity rules, edge workloads might be interested in specifying some or all of the following:

- Geolocation
- Latency
- Bandwidth
- Resilience and/or risk tolerance (i.e., how many 9s of uptime)
- Data sovereignty
- Cost
- Real-time network congestion
- Requirements or preferences for specialized hardware (e.g., GPUs, FPGAs, etc.)
- And so on...

Modern scheduling algorithms collect all information about globally-available resources and then use best-fit algorithms.



CHAPTER 4

# *THE GENESIS OF EDGE COMPUTING*

*“Information moves, or we move to it.  
Moving to it has rarely been popular and  
is growing unfashionable; nowadays we demand  
that the information come to us. This can be  
accomplished in three basic ways: moving  
physical media around, broadcasting radiation  
through space, and sending signals through wires.  
This article is about what will, for a short time  
anyway, be the biggest and best wire ever made.”*

– Neal Stephenson, “Mother Earth Mother Board,” *Wired*, 4.12

# A Brief History of the Edge

Edge computing has a strong historical basis, drawing on a continuous evolution throughout the preceding decades of computing, network technologies and user needs.

Although many of the technologies which edge computing uses and enables are at the forefront of innovation, it is useful to examine edge computing in the context of this evolution to see its true value and to understand the impact it will have on existing workload orchestration and purchasing models.

One key takeaway from this evolution, and why it is covered in this section, is that each of these eras in computing have been defined as much by their business implications as their technical impact. A technology is useful in so much as it supports a business goal, and the progression from personal computing to edge computing shows how these goals have evolved over time.

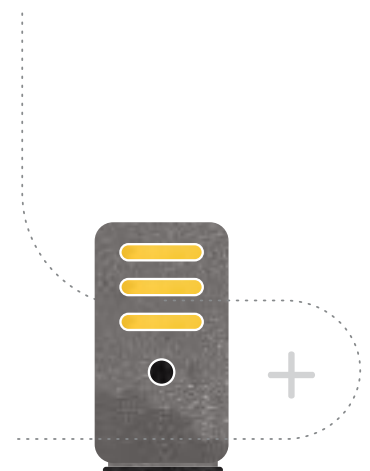
## Mainframe Computing Era

From a hardware perspective, we can broadly chart several eras built around integrated circuits instead of mechanical or electromechanical computing systems such as the Harvard Mark 1 (IBM ASCC).

Computers of this era, epitomized by machines such as the IBM 70941 and its descendants, were large and expensive systems typically used and owned by government agencies and large corporations. The IBM System/360 brought the notion of a general-purpose computer to a broader market by enabling programs to run on all models in the range, rather than building a completely separate model for different markets. Programming shifted from processing card punches in batches to the use of terminals for programming on a time-sharing system.



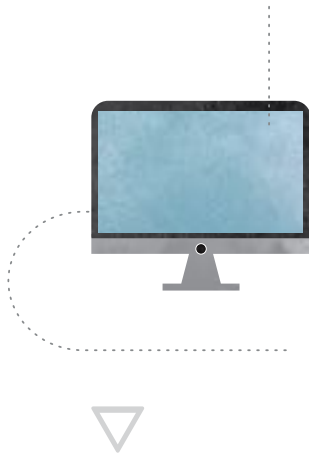
1950s



## Personal Computing Era

During the 1970s, a shift occurred in computing system architecture that was at the time unprecedented: centralized mainframe computing systems began to give way to the personal computer, bringing the compute and data storage resources users needed closer to them; this made computing physically decentralized for the first time, a pattern that continues today with edge computing but for other reasons. The personal computer ran applications and stored data locally, typically with limited network connectivity until common use of the internet in the 1990s.

In this personal computing era, the way an organization obtained access to more compute and data storage resources was simple: more computing hardware was purchased, which was then typically deployed locally in the organization's offices or in an early form of the dedicated data center. Although simple, this approach was expensive and inflexible; even a temporary need for compute and data storage resources could require an organization to purchase vast quantities of hardware, which would need to be accounted for going forward even if they were underused.



Throughout the 1990s and 2000s, the rise of the public internet and other high-speed network connectivity brought a new model of computing to the fore. Cloud computing utilizes a small number of large-scale remote data centers, rather than a large number of locally-deployed computers to provide compute and data storage resources. In contrast to personal computing, an organization no longer had to purchase hardware and deploy it locally to support the needs of their applications; even temporary requirements could be met by resources in these large cloud data centers. Rather than pay to purchase and maintain their own server infrastructure, many organizations were happy to pay a monthly fee for the use of these cloud services.

2000s

## Cloud Computing Era

Cloud computing brought significant benefits to many organizations, and allowed the use of operational, rather than capital expenditure to fund the operation of crucial applications. The temporary, elastic allocation of resources was also a significant benefit for many application operators who would previously have had to purchase the required compute and data storage resources themselves and decide what to do with them after their needs returned to normal. However, as cloud computing physically centralized the compute and data storage resources used by applications, it generated new problems; performance, locality and data sovereignty.



## Edge Computing Era

A model of computing was needed which merged the best of cloud with the best of the personal computer. By combining the density, flexible resource allocation and pay-as-you-grow pricing of cloud computing with the ability of the personal computer, due to its proximity to the user, to support higher performance, locality and meet data sovereignty concerns, the next step in the evolution of computing architecture in both technical and business terms could be achieved.

This next step is edge computing. By combining the best of the cloud and personal computing models, edge computing combines technical and business advantages to create the next great era of computing. Compute and data storage resources at data center densities are positioned locally to their users for the highest performance, even on the infrastructure edge being often within 10 miles; and they are capable of supporting the same flexible resource allocation and pay-as-you-grow pricing as cloud computing. Flexibility, performance and cost are achieved.



2018

# Early Edge Services

Edge services have actually been around for decades; over time, there has been an evolution to our present day meaning. That evolution starts with the creation of the world wide web.

Not long after Tim Berners-Lee and his colleagues helped create the world wide web, Berners-Lee saw that its early popularity was going to pose problems for the networks underpinning the web. In the '70s and '80s, communicating with users on other networks required the networks to be connected at Network Access Points or NAPs (for example, MAE-East). These were initially run by government agencies and nonprofit organizations, and later by telecom providers such as Sprint and MCI (later acquired by Verizon) internet.

The rapid expansion of web traffic meant the original sites for NAPs became too crowded, leading to the creation of commercial Internet exchanges (IXs) run by companies such as Equinix in the US, as well as organizations such as DeCIX and AMS-IX in Europe. Even though exchanges provided more efficient transfer of traffic between entities, they didn't completely solve the problem of capacity or latency. Flash crowds of users could still easily overwhelm a network and servers concentrated in a few locations, and congestion at peering points resulted in packet loss and retransmission. Additionally, users that were far away would experience poor performance for video and audio services due to latency as well as congestion and other issues.

Between 1994 to 1996, there were academic studies appearing on the issue of internet performance, leading to discussion around methods for caching web content via proxy servers<sup>4</sup>. Berners-Lee was challenging people to come up with ways to solve the growing internet performance problem. Math professor Tom Leighton and several grad students formed a company called Akamai, which translated the idea of putting a content cache in the subscriber network into a service. The service bypassed congestion at peering points and improved the ability of web sites to load quickly for the general public.

Although not immediately apparent in the diagram in its patent filing<sup>5</sup>, MIT's system (later Akamai) is an early example of an edge service. Indeed, by 2000, Akamai and others were referring to a Content Delivery Network (CDN) as an edge network; by 2001, the W3C standards body recognized (though never formally approved) a method for allowing a content to be combined by a reverse proxy server on behalf of an origin server—a useful idea for combining dynamic content with cached content. Edge Side Includes (ESI) is essentially a markup language that allows the edge server to control centralized services.

Now in 2018, CDNs are explicitly marketing themselves as providers of edge compute services. Customers can programmatically invoke functions that the vendor has developed. But the evolutionary path of edge computing extends beyond content delivery and web applications, into a multi-tenant environment where a customer's own applications can reside and maintain state.

<sup>4</sup><https://www.w3.org/Conferences/WWW4/Papers/155/>

## EARLY HINTS OF AN EDGE SERVICE ARCHITECTURE

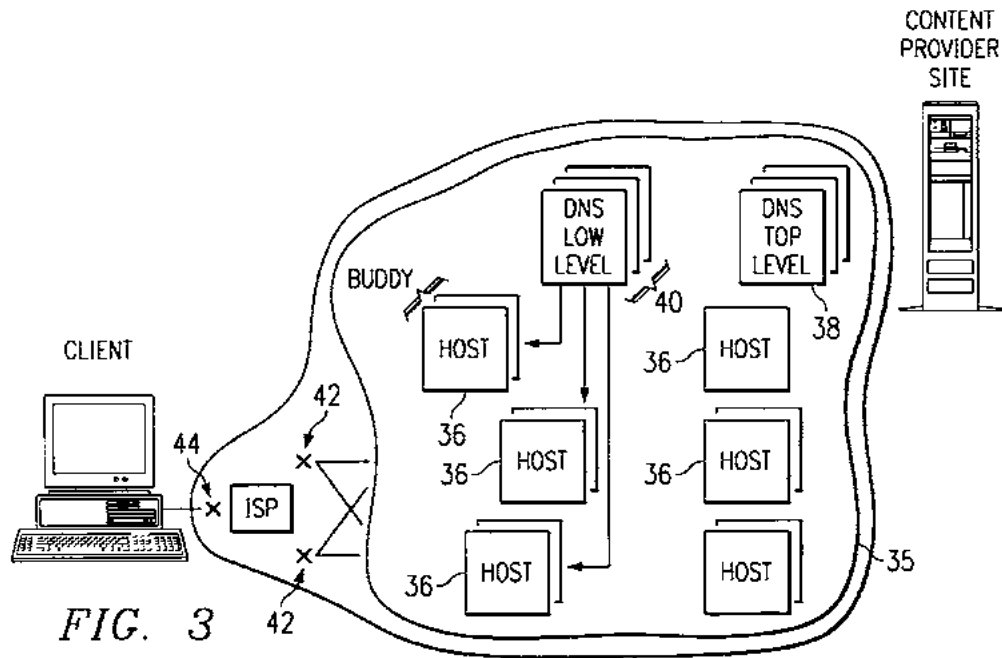


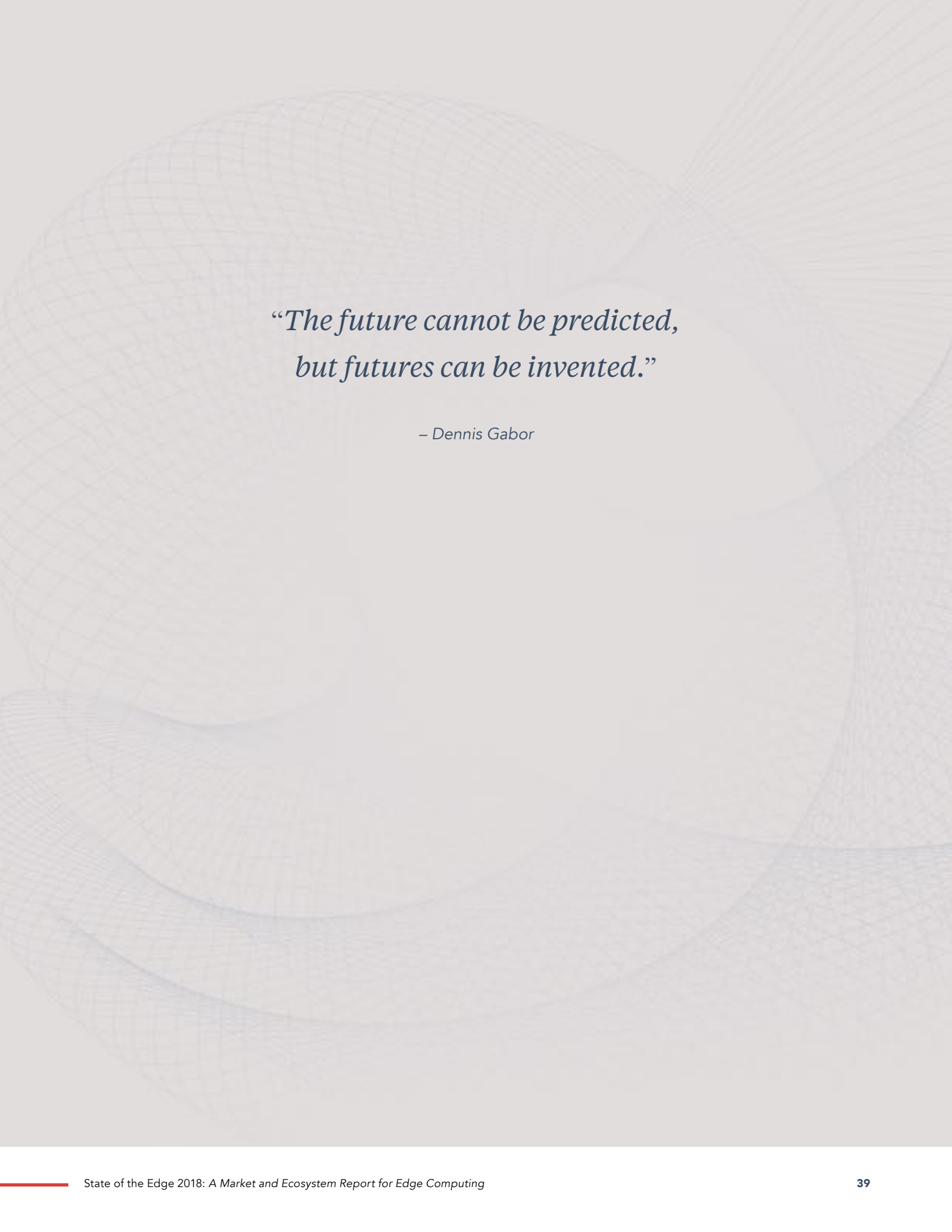
FIG. 3

The named inventors in U.S. Pat. No. 6,108,703 are Professor Tom Leighton of the Massachusetts Institute of Technology (MIT) and his student Danny Lewin (now deceased), a PhD candidate at MIT.



CHAPTER 5

# *KEY DRIVERS FOR THE EDGE ECOSYSTEM*



*“The future cannot be predicted,  
but futures can be invented.”*

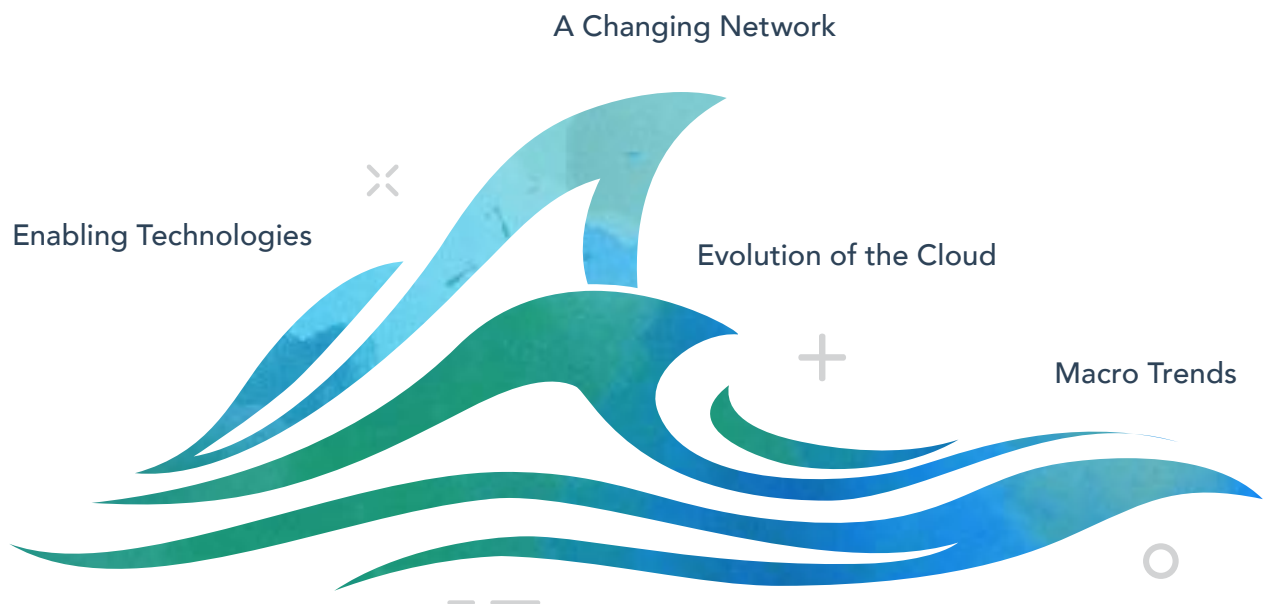
– Dennis Gabor

# Surfing the Giant Wave of Edge

The surge of edge computing has been building momentum in the deep waters of the internet since the 1990s when two MIT researchers left academia to form Akamai,<sup>6</sup> the first CDN. Back in those days, and until fairly recently, the primary challenge at the edge was to quickly deliver content to human-operated devices. CDNs made our websites load faster and our videos buffer less frequently.

Today, the perfect storm of edge computing is upon us. A powerful tide is surging across the internet landscape. If you stand quietly, you can almost feel the swell of edge computing as it rises above the water's surface, preparing to crash against the firmament of centralized clouds. Like the giant waves at Teahupo'o Reef, it will leave both fresh growth and destruction in its wake, creating potent riptides that drag the cloud towards the edge.

Many vigorous forces conspired beneath the surface to create today's giant wave of edge computing. Looking out at a three- to five-year horizon, the trends powering the wave of edge computing can be categorized thusly:



<sup>6</sup>"Akamai Technologies." Wikipedia, Wikimedia Foundation, 15 June 2018, [en.wikipedia.org/wiki/Akamai\\_Technologies](https://en.wikipedia.org/wiki/Akamai_Technologies).

# Enabling Technologies

- **Cloud Native Software** – The cloud native movement and overall application portability are changing how workloads are being developed and deployed. Containerization and microservice architectures are enabling applications to be more mobile and distributed. So far, that has resulted in developers building cloud native applications that leverage various combinations of private and public core cloud services. These approaches are the foundation that enables developers to take advantage of edge computing services.
- **5G** – Wireless networks supported by NFV and C-RAN are key enablers of IoT and edge computing. Companies like AT&T and Verizon spend tens of billions of dollars a year on wireless infrastructure and spectrum – AT&T is expected to spend \$25bn in 2018. The slow but inexorable move towards the virtualization of network and compute resources through standards-based efforts potentially places wireless operators on the inside track to providing edge computing services, for their own networks as well as their users.
- **Distributed Computing Architecture** – Both incumbent and startup firms are investing heavily in specialized accelerators for AI, ML and other high-performance applications, which are charting a course beyond Moore’s Law by using a distributed architecture. Non-traditional hardware companies like Google, Facebook and Amazon are also designing their own chips for AI and ML applications in an attempt to vie for developer mindshare. For more traditional workloads, data center capacity requirements will push the use of more power efficient processor designs such as ARM in edge compute services alongside the standard x86 chips that populate servers in core data centers.
- **Diverse Architectures** – The software community has responded to the increasingly diverse and specialized hardware ecosystem with an increased appetite to support architectures beyond traditional x86, including ARM, Open Power and even RISC-V.
- **Blockchain** – Widespread adoption of blockchain has the potential to drive consumption of edge computing services as well as underpin the use of those services. There are many applications where edge compute services can aid in data storage and performing the computation needed for blockchain services. Conversely, blockchain can underpin systems that record and reward performance of distributed edge compute services.
- **Edge Data Centers** – Projects like Open Compute and Open19, as well as a competitive modular data center ecosystem, are creating a supply chain that is able to effectively deploy and manage infrastructure at the edge.

# *Evolution of the Cloud*

- **Hybrid & Multi-Cloud** – Hybrid and multi-cloud architectures are becoming the norm. Vendors such as RightScale have conducted surveys which have found that cloud users run applications in an average of 1.8 public clouds and 2.3 private clouds.
- **Hyperscaler Investments** – Edge computing services from Cloud Service Providers (CSPs) are areas of heavy investment. Microsoft, for example, plans to invest \$5bn in IoT products, services, and research over the next four years, highlighting major CSPs' focus on the development and deployment of edge services.
- **New Performance Vectors** – Major CSPs are starting to focus on new performance vectors, including bandwidth between VMs and an evolving ML and AI ecosystem. Providers have been boosting bandwidth between compute resources and some are offering developer tools that provide applications direct access to networking hardware.
- **AI and ML** – The development of processing-intensive applications involving AI and ML is critical for many use cases. Data scientists are crunching larger data sets and using new techniques to do massive algorithmic parallelism on models in order to improve business operations. An exponential increase in the compute cycles needed to train AI models is expected to continue in foreseeable future, creating significant demand for additional localized compute, data storage and network resources.
- **Bare Metal Adoption** – With demanding use cases and a focus on network performance and security, many SaaS platforms and technology-enabled enterprises are embracing Bare Metal as a Service (BMaaS), building comfort with non-virtualized IaaS options.
- **Application Delivery** – CDNs are turning into edge compute delivery networks. While best known for caching and delivery of content (images, files and video, for example), CDN vendors are building new applications and APIs on and growing their distributed compute resources. They are aligning marketing messages around edge computing while evolving their distributed systems into full-fledged edge computing platforms.

# A Changing Network

- **Undersea Cables** – CSPs are building private global networks and making significant investments in submarine cabling. Google and Microsoft have whole or part ownership in 17 different submarine cable networks between them, highlighting how major CSPs are becoming large network service providers in their own right. They are doing this to help enterprises maintain application performance across different, disparate geographies.
- **Cloud Based Security** – Security functions are becoming increasingly cloud-based and decentralized. The size and variety of DDoS attacks have shown the need to distribute mitigation services among different regions at colocation facilities. Attacks can be more effectively deflected when malicious traffic is intercepted close to its points of origin.
- **Data at the Network Edge** – Industrial IoT services have significant potential to create huge datasets, with scalable processing and network services needed to manage them. Some predictions estimate that smart factories alone could be generating petabyte-scale data sets on a daily basis in the near future. Filtering and processing data in nearby edge data centers makes sense, and flexible interconnection services will be needed.
- **Edge Processing** – Processing power and network connectivity is spreading outward from core metro markets. Driven by a need to process data from devices and in support of activities like in-factory AI and ML, the massive amounts of data coming from remote systems indicates that more data filtering, processing and delivery needs to be managed closer to end users, at the infrastructure edge.
- **Software Defined Networking** – SDN technology allows network and security services to be less constrained by the ownership of dedicated physical assets. SD-WAN services are an example of how software is being used to create overlay networks that bond different transport types (MPLS, cellular, public internet access from DSL, fiber and cable providers) together to enable more resilient, performant and affordable networks.

# Macro Trends

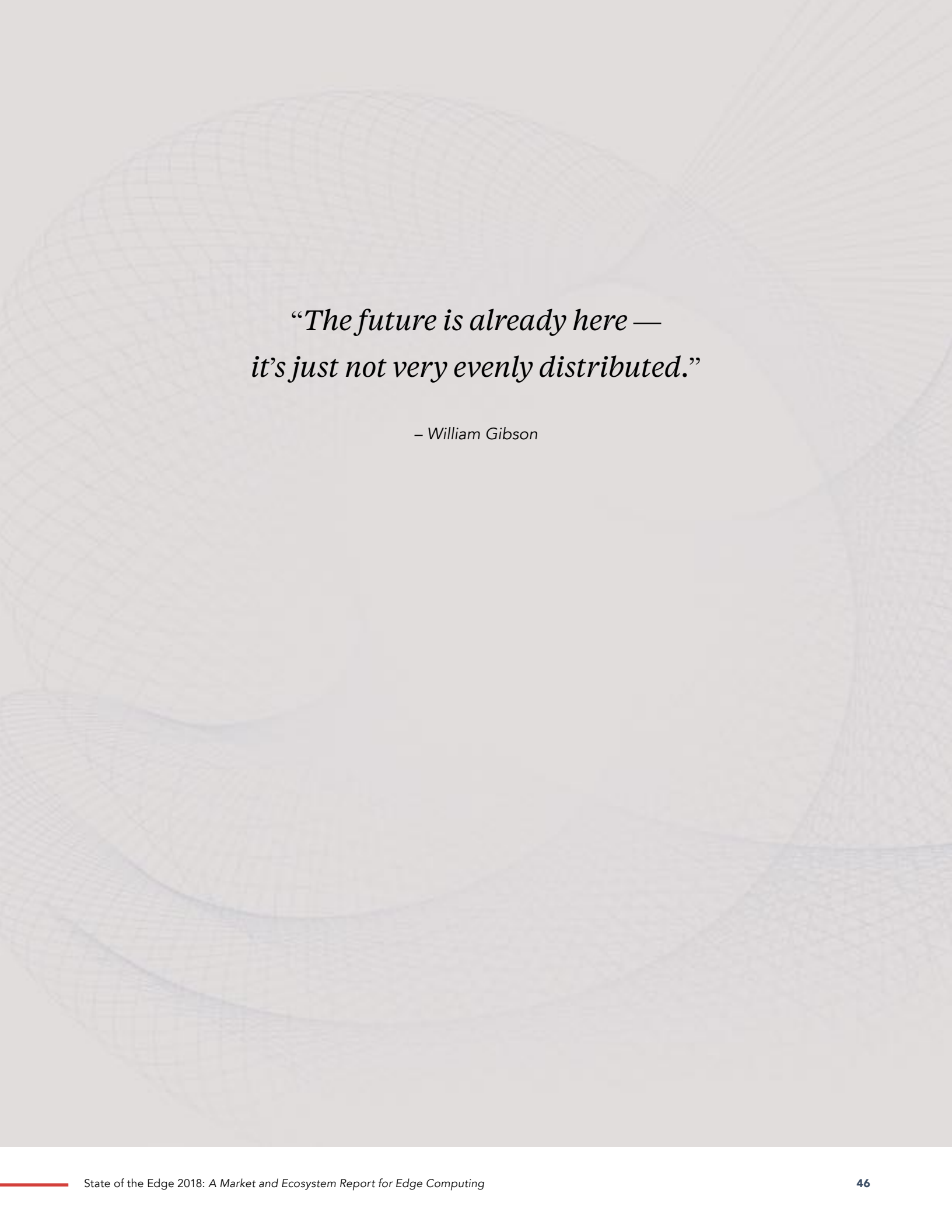
- **GitHub Generation** – An entire group of developers has grown up in the cloud, creating a massive pool of infrastructure and software-enabled innovators who are helping to architect solutions that take advantage of both centralized and distributed resources. Microsoft's recent acquisition of GitHub only reinforces the strategic influence of the developer ecosystem.
- **The Digital Enterprise** – Companies that aim to lead their sectors of technology to enable business transformation can vary widely between different industries, but successful digital transformation efforts start with a customer-centric view of products and services, then re-imagining meeting customer needs with new digital services underpinned by cloud technologies. Transformation efforts are well underway in financial services, retail, travel, media/publishing, and many other industries. With customers increasingly accessing services via mobile devices and generating more data, an edge-first approach to IT will be needed as part of most digital transformation plans.
- **Massive Scale Investment** – From the major CSPs, cellular service providers, data center Real Estate and Investment Trusts (REITs) and media or cable companies, to disruptive VCs like the SoftBank Vision Fund, enormous amounts of capital is being invested in building the next wave of the digital economy, which will be based on data that is best processed, stored and delivered close to its users with edge computing.
- **Regulation** – As IT touches even more aspects of people's business and personal lives, new regulatory constructs are emerging that require data to remain in market. This is encouraging enterprises and other large-scale users to further embrace application and infrastructure portability, and use computing resources in a growing number of locations.
- **Global Business** – Successful digital transformation will make information and services available in new markets on a global basis, but will also require investment in localized network and data center infrastructure. Performance for applications in distant markets, along with regulatory compliance for data locality (such as GDPR in Europe) means a more sophisticated approach to IT buildout will leverage edge computing services.



CHAPTER 6

# *APPLICATIONS OF EDGE COMPUTING*





*“The future is already here —  
it’s just not very evenly distributed.”*

– William Gibson

# Pushing Applications to the Edge

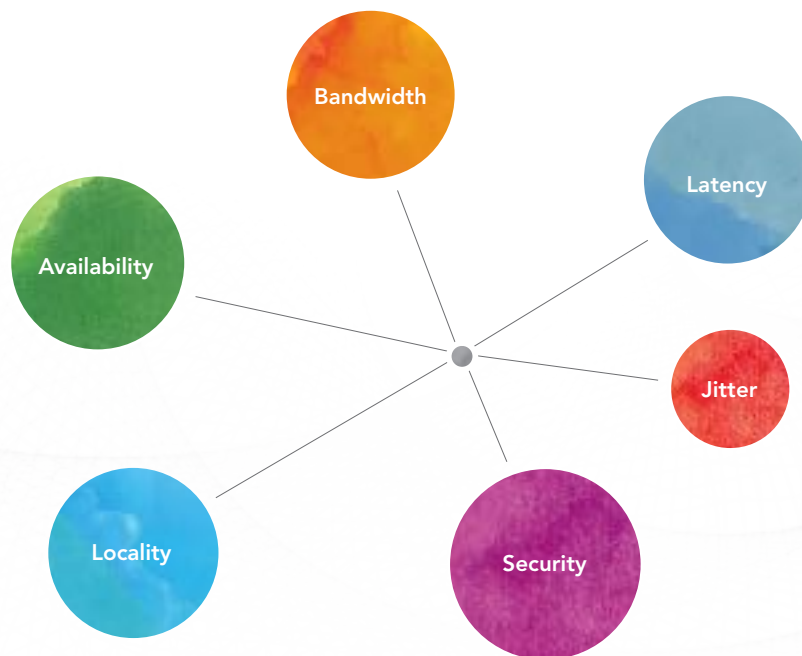
It might be tempting to dismiss edge computing as yet another clever technical “solution” looking for a problem. That would be a mistake because relentless demands of latency and scale make edge computing inevitable.

Edge computing will meet these demands by making it possible for developers to build new and previously impractical classes of applications, such as those that process massive streams of data from internet of things (IoT) devices to new use cases, such as autonomous vehicles.

Developers starting from scratch will build edge native applications—edge applications designed from the ground up to embrace edge computing. Moreover, existing applications will be *edge enhanced*—adapted and extended to utilize edge computing without completely rewriting them—and this will improve their performance and lower their impact on the network.

The applications presented here illustrate how edge computing solves key problems for a range of high-value use cases. The narrative presented for each documents the drivers and challenges for each use case. A quick reference table accompanies each use case and distills requirements into five key characteristics: bandwidth, latency, jitter, availability and security. These tables provide an at-a-glance view of the needs of each application, showing the wide variety of needs that edge computing can fulfill.

## FACTORS IMPACTING EDGE WORKLOADS





## *Large-scale IoT and IIoT*

IoT and Industrial IoT (IIoT) systems have been the subject of considerable interest, not to mention hype, over the past five years. To date, these systems have fulfilled many important real-world roles from smart thermostats to connected control systems. But large-scale IoT deployments such as smart cities have not yet materialized. A number of trends must converge to make IoT ubiquitous. Low-power wireless sensors need to proliferate, the creation of software ecosystems must advance, and the overall deployment of capital to support IoT must accelerate. As these trends drive IoT forward, there must be contemporaneous improvements in the network and data center infrastructure that these systems will rely on.

### With edge computing

Large-scale IoT and IIoT systems challenge the traditional centralized data center architecture in use today, such as in a smart city application, where millions of wireless sensors may be connected to the cellular (or other wireless) network. Depending on the type of device, it may transmit data at intervals, or transmit constantly. The range of possible IoT devices from traffic lights to gas sensors makes it difficult to categorize them in one set of needs. Nonetheless, we find common characteristics across the majority of IoT and IIoT use cases where edge computing solves one or more significant problems.

Though any single IoT device may be relatively low bandwidth, millions of them in aggregate can generate enormous amounts data and bring the network to its knees. Many IoT and IIoT sensors provide critical data to latency-sensitive applications. For example, sensors on the power grid or on oil pipelines will require low-latency analysis to prevent damage to people, property and the environment. Availability is also key for most IoT devices, and the data they generate is local in scope. Although a great deal of sensor data may be repetitious, it still must be processed in order to identify trends and advance machine learning models. For these applications, shipping massive amounts of data to a far-off centralized data center is less practical than processing at the infrastructure edge, with extracted results sent upstream.

Latency sensitivity, the need for high availability and the demands of data locality in IoT and IIoT applications means edge computing on both the device edge and the infrastructure edge is ideal to support these systems. Removing their reliance on long-distance network connectivity and locating dense compute, storage and network resources locally allows a large-scale IoT system to operate effectively, with excellent performance and resource usage.



## Video Games

Online video games have been a prime leisure activity for millions of people. Many video games are considered sports in their own right. Popular video game console services such as Xbox Live and Playstation Network have brought online gaming into the living room and high-powered mobile devices like the Nintendo Switch as well as iOS and Android devices have made them mobile. Today, the releases of popular video games rival the latest blockbuster movie, and their appeal often endures for longer.

In many modern games, reaction times are measured in milliseconds; in a fighting game playing at 60 frames per second today, a user may have 1/60th of a second to input a command that can win or lose the match. Latency, and even worse, jitter, are the bane of gamers and can even stop players from competing altogether. Most of these issues are due to the long-distance network connectivity required to reach a centralized data center where the game is hosted.

### With edge computing

While games usually require low to medium bandwidth, they are extremely sensitive to latency and jitter. By hosting the game for local players on a micro data center at the infrastructure edge, gaming services can offer a better experience for gamers. The short-distance network connectivity offers faster round-trip-times, fewer points of contention and less opportunity for routing failures which increases both the responsiveness and availability of games.

Ameliorating latency and jitter vastly improves the gaming experience for users. A Google search for lag related to any particular game (e.g., "Fortnite lag") will show a plethora of articles and online discussions that compare measurements and offer first-hand experiences of various Wi-Fi routers and internet service providers with players regularly exploring which combination offers the best results. It's not uncommon to see players change service providers just to lower latency, making edge computing a powerful customer retention tool.

In addition, as online gaming has increased in popularity, it has become a popular spectator sport. The finals of tournaments for games such as League of Legends or Street Fighter V can draw live viewerships equivalent to many other sporting events, and the load of this video traffic can be considerable on the network and data center infrastructure that is supporting it. The live video streaming aspect of "eSports" benefits from edge computing by allowing video encoding and storage to happen at the infrastructure edge to ease the load on the backbone network.



## VR and AR

Virtual and Augmented Reality applications (VR and AR, respectively) are seen as the future of human-computer interaction. These technologies open the door to new business tools and immersive leisure applications. Remote surgery where a surgeon is able to fully appreciate depth and hand-to-eye coordination, virtual city tours where points of interest are displayed in real-time to a user based on their location and driving assistance through AR are all powerful applications of these technologies, all of which rely on complex off-device processing.

Many of the AR and VR applications we desire are impractical on conventional networks due to the low latency they require between a user and the data center operating the application. These are real-time applications, highly-sensitive to delay; but they rely on the compute power of the data center to process complex rendering algorithms and use collaboration from multiple data sources to improve their speed and accuracy. This means many VR and AR applications are impractical with the centralized data center model; the achievable latency is too high, as is the cost of transporting large amounts of data.

### With edge computing

VR and AR require low latency to appear natural to a user, but also require large volumes of complex data to be processed using techniques such as 3D rendering and machine vision. This combination makes edge computing the ideal platform for VR and AR; a dense micro data center at the infrastructure edge provides the muscle to support VR and AR, at low latency due to its locality.

Although VR and AR span a wide range of use cases, their requirements are fairly uniform. Bandwidth requirements range from medium to high, depending on the application; but latency and jitter must be low as many applications rely on tasks like the identification of a landmark during a virtual city tour appearing rapidly and in a seamless fashion to avoid user frustration. In the case of VR, where tracking of the user's real-time head position and hands occurs, delays in processing these movements can result in motion sickness. To be seamless, availability is key, and for some applications such as remote surgery security is vital.

In AR use cases, the processing required is deceptive. An edge device such as AR-enabled glasses cannot perform all of it; graphics-intensive processing is needed to identify an object, analyze it, and then display relevant information to the user such as in a virtual city tour. In these cases the edge device streams HD or 4K video to the infrastructure edge where this processing occurs. This is a load on the network, making the infrastructure edge vital for it to be successful.



## Autonomous Vehicles

Autonomous vehicles, whether buses, cars or aircraft are one of the most exciting applications being discussed today. The benefits they can bring in terms of improving safety, enhancing efficiency and saving manual labor for millions of people are considerable; but replacing the complex cognitive functions that a human performs hundreds of times every second with a machine is no easy task. Despite the functions the manufacturers of autonomous vehicles are able to integrate into the vehicles, it is both impractical and cost-prohibitive to rely entirely on on-board systems to deliver autonomous vehicles at scale.

The reasons for this are no surprise. Autonomous vehicles rely on high-performance compute, storage and network resource to process the massive volumes of data they produce, and this must be done with low latency to ensure that the vehicle can operate safely at efficient speeds. Using on-device or centralized data center resources would be practical only if autonomous vehicles were speed-limited, making them inefficient for common use. To replace human operators, autonomous vehicles must operate safely at the same speed as human-operated vehicles using data from many sources so a vehicle knows how to react to stimuli it may not be able to directly acquire (e.g., real time data about other cars approaching a blind intersection).

### With edge computing

Autonomous vehicles offer a prime example of how the power of micro data centers positioned at the infrastructure edge can greatly improve the performance of an applications while also reducing the cost. Resources positioned at the edge of the cellular network can communicate with cars wirelessly and provide real time coordination and decision support across an entire metro region. Local resources on the vehicle are used to perform basic data acquisition and decision-making, such as detecting and avoiding a pedestrian directly in the path of a moving car. Other more complex tasks, such as modeling city-wide traffic flow or identifying a badly-damaged road sign can be delegated to the greater resources of the edge data center, which returns the results.

Processing data from multiple autonomous vehicles with the infrastructure edge has another benefit; when data from many sources is analyzed, patterns which were not obvious from a single source appear, and the result can be applied to every vehicle in the area. The needs of autonomous vehicles can be categorize(d) as follows. They require high bandwidth due to the amount of sensor data they collect; some estimate an autonomous car will collect 2 petabytes of data annually, much of which will be analyzed in the infrastructure edge. They also require low latency and jitter, high availability and security to prevent tracking and remote vehicle hijacking.

Without dense resources at the infrastructure edge, many autonomous vehicles will not become practical. Beyond the self-driving car, there are many other autonomous vehicles where the infrastructure edge is the cornerstone of their safe and practical operation. Small aircraft or drones



which operate without human piloting are extremely valuable for use cases including construction, where they can be used to inspect sites for safety and other concerns in a fraction of the time and cost of a human inspector. In these cases the infrastructure edge provides not just the location for offloading video and other sensor data; it also provides the processing capability to control the flight plan of the aircraft in real-time, so that conditions such as urban micro-weather can be avoided before they become a safety problem.



## Edge Content Delivery

Content delivery may appear to be a solved problem. Today we are able to deliver content to millions of users with a centralized data center architecture. However, this ignores the fact that the demand for content continues to grow beyond network resources. Edge content delivery involves innovative use cases such as real-time distributed video encoding, to make the live coverage of major sporting events scalable. This can best be done by augmenting a modern CDN with the ability to run compute workloads in edge locations, as processing at the edge is required to process large volumes of video data and re-encode it rapidly as necessary.

The possibilities for edge content delivery are many, but they cannot be done today with the combination of centralized data centers and current CDN systems. In some cases the latency of a centralised data center is too high, and to be economical the cost of content data transport must be cheap. It is clear that the infrastructure edge enables scalable edge content delivery.

Moreover, high-bandwidth media delivery is switching from a downstream-only activity to a bidirectional one. The advent of live streaming services (SnapChat, Facebook Live, YouTube Live) and the increasingly common 4K (and soon 8K) video cameras on smartphones, will create immense pressure on the upstream network that can be alleviated with capabilities on the infrastructure edge to assist with video encoding, compression and upstream caching.

### With edge computing

Infrastructure edge computing, and the dense compute, data storage and network resources of its edge data centers are the key to making edge content delivery practical. As the resolution and frame rate of video content continue to grow, the demands it places on the network begin to multiply rapidly. In addition, the resources required to re-encode and distribute this video data also increase significantly. The solution to both of these problems is to process video at and distribute it to users from the infrastructure edge, removing strain on the long-distance network.

Personalized content creation, where video or other types of content can be adapted to appeal to a specific user, is also complex. Once the required data is assembled from various sources, content must be created, rendered and encoded into the correct format. Dedicated fast GPU hardware will be used along with CPU resources, to perform processing in faster than real-time; personalization must be instant to the user to avoid frustration from delays. The capacity of the micro data center at the edge is what makes personalized and edge content delivery possible.



# AI and Machine Learning

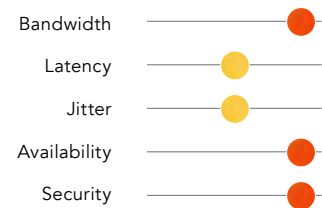
Artificial Intelligence and Machine Learning (AI and ML, respectively) will be the cornerstone of many applications in the near future, from digital assistants to decision support systems which use data to propose a solution to a human operator or an agent on a remote device, or even to initiate an action themselves in an automated system.

Most of us have experienced the delay that occurs between speaking a command and getting a response from Amazon's Alexa or Apple's Siri. During this delay, the user's voice is transmitted to a centralized data center where complex speech recognition and inference algorithms operate on dense compute resources to allow the digital assistant to understand what was said, formulate a response, and then return that result to the user's device. The time required to relay the command to a centralized data center, not to mention the collaborative data sharing between many sources of data required to make these technologies work to the satisfaction of a typical user, are considerable. In addition, the amount of data that is transmitted is large and growing. The centralized data center model is being strained by AI and ML.

## With edge computing

A micro data center deployed within the infrastructure edge can contain hundreds or even thousands of specialised accelerator devices which are dedicated to fast AI and ML application processing, such as GPUs, TPUs and FPGAs. Many micro data centers positioned at the infrastructure edge create the high-performance, low latency foundation required to make AI and ML applications ubiquitous, natural and seamless to millions of users, at higher levels of performance and with a lower cost of data transmission than using a centralized data center alone.

As briefly touched on above, the requirements that AI and ML place on the network are many. AI and ML applications will often rely on transmitting large amounts of data from a device to a data center, and so high bandwidth is required; low latency is also an important consideration. This low latency requirement will become more pronounced as AI and ML systems move more towards delay-sensitive tasks, such as a digital assistant conducting lifelike conversations with humans in real-time, in a non-obtrusive way, such as to interactively schedule appointments or to provide other information. Unlike many other real-time applications, these AI and ML systems are partly tolerant to jitter, though not excessively. Availability and security are both key considerations to ensure that AI and ML systems are frequently used, and to protect them and their users as these systems increasingly handle personal information and are responsible for user safety.



## Video Surveillance

With the growing demand worldwide for video surveillance, new technologies are bringing what was once an antiquated use case supported by analog cameras and tape back into the forefront of innovation. AI and ML technologies, described above, can be used extensively with HD video surveillance to provide machine vision capabilities where a modern video surveillance system can automatically identify people, faces and situations to alert a human operator. The live video feeds from thousands of autonomous vehicles and stationary cameras positioned throughout a city can now generate gigabytes of data per second, all of which must be stored and processed.

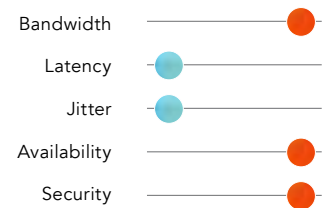
The movement of these vast quantities of data is the problem, however. A large-scale HD or 4K video surveillance system may generate gigabytes or terabytes of data per second, all of which today must be transported over long-distance network connectivity to a centralized data center. This is expensive, difficult and wasteful; the vast majority of data produced by a modern video surveillance system is of no interest to people or systems outside of the local area in which it is generated, and so transporting it hundreds or thousands of miles for processing and storage is a poor use of network resources. The need for these systems continues to grow worldwide, and to accommodate their increased usage the systems used to support them must adapt as well.

### With edge computing

Besides the sheer amount of data that a modern HD or even 4K video surveillance system on a city-wide scale can generate, one important aspect of this data is its locality. The vast majority of video data collected by such a system is relevant only to that local area. Therefore it makes little sense to transmit these huge and growing masses of data hundreds or thousands of miles to a centralized data center, when it can be processed and stored at the infrastructure edge instead. In cases where a person of interest to law enforcement or the national government is identified from video surveillance footage in the local area, either through machine vision provided by AI and ML or a human operator, this finding and the relevant footage can then be sent up to the centralized data center in the cloud. In this way the infrastructure edge serves as the foundation for high-performance, modern video surveillance systems which greatly enhance public safety.

Video surveillance systems typically impose a common set of requirements on the network and data center architecture, but there is some variation depending on the type of camera and the location in which it is deployed. All video surveillance systems, especially with full HD being the current standard and 4K emerging in some newer systems, require high bandwidth. Latency and jitter are typically not as big of a concern, but an exception to this statement are systems in high-security areas where detection of an event by the video surveillance system may trigger an automated response. Availability and security are very high priorities for any video surveillance system, as the effectiveness of the system both as a deterrent and a means to identify crime, accidents of persons

of interest are directly linked to keeping both of these characteristics as high as possible. The infrastructure edge is the ideal place to support modern video surveillance systems as it can function as a local point of bulk data ingestion, as well as support the dense compute resource required to perform AI and ML operations to provide complex machine vision.



## *NFV and C-RAN*

The resources provided by the infrastructure edge are of great interest to network operators as they undergo network upgrades to leverage new technologies while mitigating budgetary pressures.

Today, cellular network operators are at a crossroads. 2G, 3G and 4G networks were built on proprietary hardware; when a new site or an upgrade was required, network operators deployed dedicated infrastructure such as a 2G base station. This was a fixed-function device; it used specialized hardware to perform a function, and could not be repurposed. They are expensive, and as the performance of x86 and ARM processors has grown, they are difficult to justify as single-function devices.

Network Function Virtualization (NFV) transfers network capabilities from proprietary devices and dedicated hardware to virtualized software running on general-purpose (“white box”) servers. This gives operators more flexible and less costly ways to expand their core networks, moving functions off of dedicated appliances and onto general-purpose, open platforms that are much easier to manage and have a much lower TCO. To scale up a network site, simply add more general-purpose resources. This concept can also be applied to the cellular radio networks using Cloud RAN (C-RAN) architectures, which virtualize radio processing previously performed by dedicated hardware devices.

Although these technologies are powerful, to be effective the processing for these functions must be performed as close to the edge of the network as possible; ideally in the same location as the dedicated hardware devices are today, at the cell sites themselves or neighborhood aggregation hubs for example. Using a centralized data center thousands of miles away will not work due to excessive latency and the cost of data transport. To make NFV and C-RAN practical, the infrastructure edge is needed.

### With edge computing

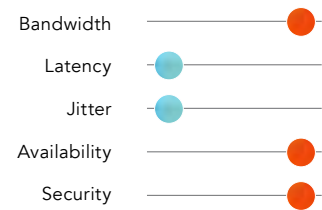
Micro data centers deployed at the infrastructure edge, directly at the cell site, an aggregation hub, a carrier hotel, or another location near the network edge, become the ideal place to host the server, networking and storage devices required to operate NFV and C-RAN to their full potential. The micro data center offers a dense, high-performance collection of compute, storage and network resources, which can provide general-purpose x86 or ARM-based processing which is ideally-suited to processing data from a Remote Radio Head (RRH) in the case of C-RAN, or operating many different types of VNFs in the case of NFV. The performance of these general-purpose compute resources is considerable when it is considered that a micro data center at the infrastructure edge may provide a capacity up to 150 kW in a single unit, with the capability to interoperate with a larger system of multiple edge data centers within the local area for greater distributed processing.

The cost and performance of these resources make the infrastructure edge ideal for the network operator who is seeking to migrate their network architecture to NFV and C-RAN. As these resources are general-purpose, they can be repurposed over time as the requirements of the network operator change. For example, a network operator may begin by virtualizing the key functions of their routers and radio baseband processing equipment. In time, they may also seek to virtualize their firewalls; rather than requiring a new dedicated device to do so, they simply deploy a firewall VNF at the infrastructure edge and route traffic through it as required.

Multi-tenancy, where more than one organization shares the usage of a single location or piece of infrastructure, is becoming increasingly common in wireless networks. With 5G expecting a greater number of smaller cells, the number of sites is set to increase and with it the investment required to construct them. Network operators are looking at innovative ways to host NFV and C-RAN functions on common infrastructure, where multiple operators would host workloads in the same racks (or, even on the same devices) as their competitors, so that they are able to share costs. The infrastructure edge is ideal as it provides enough resources at a low enough deployment cost to make multi-tenant and Network as a Service (NaaS) models a reality.

The requirements that NFV and C-RAN place on the network are stringent. As such, they are one of the most difficult of all the use cases discussed in this section, requiring high bandwidth, low latency and jitter, all with very high levels of availability and security to protect the network from unintentional operation and the leakage of sensitive information. The only practical way to meet these needs is to perform NFV and C-RAN at the infrastructure edge, relying on its dense resources to allow the network operator to meet their challenges that will only increase with 5G.





## *Compute and Network Offloading*

Next-generation mobile applications are typically resource-hungry, demanding intensive computation and high energy consumption. Due to limitations in battery power, computation capacity and physical size, some processor-intensive applications such as natural language processing, virtual reality and interactive gaming cannot be performed smoothly on many mobile devices. Data center capacity at the infrastructure edge can be a means to overcome the resource-constraints of edge devices. Applications can be designed to run partially on the device, and partially on the infrastructure edge, overcoming any device-specific constraints.

To illustrate compute offloading, consider a mobile device operating an AR application. In the eyes of many users, the first duty of a mobile device is to be available; to have enough power remaining to let them go about their day before charging it at night. With the use of ride-sharing services, many people rely on their smartphones to get around, demonstrating how important the ability of these increasingly thin mobile devices to limit their power consumption is.

### With edge computing

The AR application requires complex processing in a number of areas, from 3D graphics to machine vision to function properly. To stay within its desired power consumption level, much of this processing is out of reach of the mobile device, such as a smartphone. Instead of boosting its power usage to cope with the workload, the mobile device can offload the majority of it to the infrastructure edge, which returns the result in real-time, keeping device power usage minimal.

This describes compute offloading. Another concept, network offloading, is also achievable using the infrastructure edge. In certain cases, an alternative path can be used for network data which improves performance or resource utilization. An example of this is local breakout, where the compute, data storage and network resources of the infrastructure edge are used to route network data in transit onto other networks, such as the internet, at an earlier stage than would have otherwise been possible. In traditional cellular network design, this same network data may follow a trombone-shaped path through an aggregation point before this can occur.

Both of these types of offloading require high bandwidth, alongside low latency and jitter. As the workload being offloaded to the infrastructure edge could be from a wide range of use cases, availability and security have to be high, especially for network offloading to avoid disruption.

In both cases of offloading, the result is improved performance and resource utilization from the mobile edge device to the core network infrastructure. Through the use of compute and network offloading, the infrastructure edge provides significant, widespread performance improvements.

CHAPTER 7

# *IMPLICATIONS*



*“I try to look in the future and think backwards. That is when I am at my most comfortable. Because then I am just like the movie, 'Back to the Future.' ”*

– Masayoshi Son, Founder and CEO, SoftBank

## *Moving Beyond US-East*

This report has primarily focused on the historical basis, technical underpinnings and the drivers and use cases for edge computing at the infrastructure edge. But as with any major technology shift, beyond these technical considerations there are a whole host of practical and commercial realities that must be understood in order for edge computing to make a significant impact in the real world.

Most applications today are deployed in a single location, or a handful at most. Sophisticated applications with global reach may deploy in a dozen or more. Amazon Web Services – the largest public cloud - has 16 regions worldwide, with 42 availability zones.

As the edge computing infrastructure ecosystem evolves, everyday users (or more likely their software or applications) will look to take advantage of compute and network resources in hundreds of locations. Currently only the largest CDNs in the world stretch to this kind of footprint – and only running a single, fairly simplistic application.

As we look at building and taking advantage of a robust edge computing ecosystem, defined by a much more diverse and distributed architecture, what are the implications? From the software we use, the way in which we purchase computing resources, and the expectations for service levels, physical security, hardware lifecycling, refresh cycles and more.

The implications are myriad.

*“Just as the electrical grid made it possible to distribute energy everywhere it’s needed, a new edge computing grid will deliver cloud capabilities to every square mile of the planet, transforming the economy in the process.”*

– Cole Crawford, founder & CEO, Vapor IO

# *Edge Will Not Eat the Cloud*

Industry discussions often portray edge computing and our existing centralized cloud as antagonistic towards one another. Many articles—some with provocative titles, such as Peter Levine’s [The end of Cloud Computing](#)—have focused on how edge computing will eat the cloud, but this line of inquiry misses the important point: edge computing and centralized cloud are stronger together than apart. Combining edge and cloud creates a powerful n-tier compute architecture, with resources spanning the entire continuum from core to device.

Many would agree that the key aspect of the cloud experience is automation; in other words, programmatic consumption. Building upon this experience, the dense arrays of micro data centers which will be deployed at the very edge of the operator network in the infrastructure edge will be capable of supporting the same type of programmatic resource allocation as the centralized cloud, although individually at a smaller scale than the hyperscale data center.

Viewing edge cloud resources as an extension of the public cloud experience opens up a powerful collaborative model between the edge and centralized models. In this new and more complex model, workloads may be migrated dynamically in several directions depending on many factors such as the available resources at the edge and centralized data centers, the performance characteristics of the workload, and the maximum acceptable cost to process that specific workload at that time. Workloads can be migrated from an edge device to a micro data center, from a centralized data center to a micro data center, and vice-versa in near real-time to ensure that Service Level Agreements (SLAs) and costs for that workload are met.

What breaks down is the idea of infinitely-scalable resources and services, which has been the hallmark of hyperscale public cloud. While the public cloud has become essentially unlimited to most users, no such construct exists at the edge, where both physical space and certainly power can be constrained. This, along with a physical distribution footprint that may include thousands of locations, reduces the amount of actual infrastructure that can be available in each location. In short: the edge introduces scarcity, and resource contention will be a fact of life.

As with any market, scarcity creates pricing implications. How much are you willing to pay for a specific resource on a given day? As this report details, the kinds of workloads that find value at the edge will often be specialized. Service providers – which may not directly own infrastructure at the edge - will need to become agile at dealing with resource limitations, outages, and other factors that may impact pricing, workload requirements and SLAs further up the stack.

# *Software as the Customer*

With the growth of cloud native and web-scale applications, modern infrastructure automation tools have been stepping up to embrace enormously complex requirements. This is a critical underpinning for edge computing, and one that infrastructure service providers need to intuitively understand to be successful at the edge.

Due to number and placement of locations, diverse business rules, and need for workload portability, nearly all edge workloads will be deployed and managed through sophisticated automated systems. As such, everything about an edge data center will need to be controlled and managed through automation. From provisioning and network management, to telemetry.

We believe that a more nuanced marketplace for compute, storage and network resources will evolve, based upon exacting data to support modern applications that intelligently adds the centralized cloud and infrastructure edge together, pooling their individual resources and at the same time retaining the specializations of each. There are many different aspects to this, beginning with differentiating the types of application workloads each is best suited to operate.

For this type of dynamic resource allocation between centralized and edge data centers to be practical and effective, the real-time resource availability and performance of every data center and their interconnecting networks must be available programmatically to a dynamic workload orchestrator which is able to use this data to determine where a specific application workload is best operated for the performance and cost targets that workload is tagged with, as well as to meet any other SLAs that have been agreed, and paid for, specific to that application workload.

Details such as the current and historical utilisation of the compute resources in each data center, their currently available storage space, and the type of compute resources which are available at the present time must all be considered. Beyond this, network factors are also crucial such as the latency and jitter between the data center and its intended application users and between other data centers, not to mention the bandwidth available on these connections, are key factors in determining whether a specific data center is the optimal place for a workload.

There will be far too many workloads operating dynamically on the gradient of data center resources from the centralized cloud to the infrastructure edge for these workload allocation decisions to be made manually. Although a human application operator who is paying for the use of these resources will initially tag their workload with specific requirements that it has, such as performance, cost or data sovereignty, the task of actually allocating where this and other workloads operate in real-time will be ideally performed by an intelligent orchestrator. By automating this process through orchestration software, the combination of the centralized and infrastructure edge data centers becomes far more valuable together than they would be apart.

## *Special Things in Many Places*

We've all heard "Moore's Law is Dead!" And with the adoption of accelerators, offloaders, GPUs and other specialized hardware, a new distributed computing architecture is quickly emerging. This architecture is thriving on specialized hardware, paired with specialty software, which is often leapfrogged by new versions every 12 or 18 months.

An additional layer of complexity in this area is that between data centers, the equipment that is deployed is often not uniform and may change frequently. For example, one micro data center at the infrastructure edge may provide general-purpose x86-based compute resources; another may be ARM-based, and yet another may be concentrated on providing dense GPU resources for graphics tasks such as real-time machine vision.

The current cloud ecosystem and supply chain is quite adept at making a lot of a few rather generic hardware platforms appear in a few locations, such as the ubiquitous dual-socket x86 server platforms built on Open Compute standards and shipped by the pre-built rack. In fact, tens of thousands of racks are delivered to hyperscale data centers by the truckload each year.

Edge computing will require a very different supply and service chain, one that is facile at deploying and managing a wide variety of unique resources in thousands of locations. Instead of a generic dual socket x86 server, an edge workload may benefit from a custom low-power System-on-Chip (SoC), a programmable smart NIC, and the latest memory technology from Intel. Anyone that has ever tried to deploy a rack of servers in one location, let alone a hundred, understands the logistical and physical complexities that must be addressed.

The ecosystem is already hard at work on solving this "last 100 feet" problem, as well as the automation challenge. On the physical side, projects like Open19 are challenging design standards to remove complex cabling, and allow racks to be deployed in advance of expensive individual compute "bricks." On the automation side, efforts like [WorksOnArm.com](http://WorksOnArm.com), [AccelerateWithOptane.com](http://AccelerateWithOptane.com) and the [CNCF Community Infrastructure Lab](http://CNCFCommunityInfrastructureLab) are helping the software ecosystem build and test against new hardware.

*"People who are really serious about software should make their own hardware."*

– Alan Kay, 1982



## *Physical vs Network Locality*

Infrastructure edge computing and today's centralized cloud can both provide elastic compute, storage and network resource to application workloads, but as has been seen throughout this report these technologies provide different levels of performance, scalability and overall cost.

Also, in the centralized public cloud, the network has been almost completely abstracted. While one can deploy overlays with ease, there is very little insight into the layers below. By and large, this has worked well for most cloud users, but moving to the edge, with its focus on networking use cases and benefits, will require much greater access and control.

In a centralized model, a user might want to answer the question: did my packet arrive? At the edge, the question is likely to evolve: how did it arrive, which networks is the compute I'm using attached to, what is the past performance of those networks, and what is the latency to various points on various networks?

Applications will need to become much more network aware, and service providers are likely going to need to evolve to share, or sell, more of this valuable data with their own users.

## *The Ingredients for a Marketplace*

The move from a centralized-only to a multi-tier marketplace of compute, storage and network resources spanning from the infrastructure edge to the centralized cloud creates many new possibilities in how the use of these resources is purchased. Using our earlier distinction between the infrastructure edge and centralized data centers, examples become clear to illustrate how cloud resource purchasing decisions will look for many in the near future.

An application workload can be tagged with many different characteristics. Perhaps the most stringent of these would be the combination of real-time, low-latency operation, the need for a low cost of data transport and data sovereignty complications. The general location to operate this application workload at is obvious - the infrastructure edge - as it is the only way to satisfy these difficult requirements. Although this case may seem simple, there is an added dimension to the purchasing decision for this type of application operating at the infrastructure edge: cost.

Users of cloud resources are of course no strangers to the cost of using those resources, but this particular example bears some analysis. Within a local area the amount of suitable edge compute resources for a particular workload may be limited. Although the cumulative power of many micro data centers combined is significant, there is a limit to the number of sites at which they can be deployed and the density to which each individual site can scale that may be reached in many areas over time.

In just a few years, the World Cup will be played at various cities in the US, Mexico and Canada – does it not seem likely that these localities will see a massive spike in the need for nearby infrastructure resources? It can be seen that in even moderately extreme cases, there may develop a mismatch in supply and demand between the available infrastructure edge resources in the local area and the number of application workloads which require them.

Ultimately, if the competition for resources is between application workloads which all require the same SLA, the determinant as to which workload will be able to operate must be cost. The highest bidder for those infrastructure edge resources will receive what they need to operate their application workload, at the exclusion of lower bidders. In this way traditional market dynamics are applied to the problem of allocating data center resources; a classic example of scarcity has occurred, and the problem is no longer technical in nature but is a business issue.

This extreme case shows one aspect of this resource allocation process, which in many cases may not occur. In an ideal situation there are spare compute, storage and network resources available at both the infrastructure edge and centralized data centers which allow for their own levels of improved flexibility. Workloads may be operated using cost-based scheduling, where non time-sensitive applications operate on the first slot of resources which becomes available and meets their specific cost targets. Performance-based scheduling will also occur, which may force the dynamic migration of workloads with less-sensitive SLAs from the infrastructure edge to the centralized data center in real-time, as required by the available data center resources.

As the edge computing landscape evolves, as well as its integration and interaction with the traditional centralized cloud, many more nuanced options will develop in regards to how these combined sets of resources can be allocated, purchased and monitored.

Purchasing models such as spot and futures-based approaches will be the focus of much discussion, as will how software and existing automation will adapt to make the most optimal usage of the gradient of compute, storage and network resources spanning from the centralized data center to the edge.

In order to support a robust, dynamic and evolutionary marketplace, the edge computing ecosystem will need to get very good at exposing all of the value available in the market. At the base layer, a simple marketing abstraction such as a VM size or application type will likely not provide enough detail for accurate costing. Exposing exacting details of hardware and network specifications in a standardized fashion will enable all players in the ecosystem above to create the experiences and models that will allow this resource-constrained marketplace to thrive.

## *Epilogue: A Call to Community*

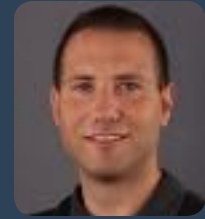
The explosion of smart devices and computing at the network edge has been an incredible trend to witness over the last few years. This unprecedented boom is causing a shift in how data processing is handled in a typical cloud infrastructure. As bandwidth to mobile devices improves with the advent of 5G networks, we'll see a new set of use cases involving localized data processing via IoT devices, faster rendering for games and immersive AR/VR environments that will take advantage of micro data centers on the infrastructure side of the edge as well as more powerful devices and gateways on the device side of the edge.

The [Cloud Native Computing Foundation \(CNCF\)](#) is the home of [Kubernetes](#) and other open source cloud native projects that empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. These techniques enable loosely coupled systems that are resilient, manageable, and observable. Combined with robust automation, they allow engineers to make high-impact changes frequently and predictably with minimal toil.

The CNCF community sees edge computing as a natural extension of cloud native practices, both on the infrastructure side as well as on the device side. Our role is to help the community ensure containers and other Kubernetes components operate well at the edge, taking into account the advantages and limitations of edge environments compared to traditional data centers and cloud. The CNCF community also seeks to pioneer ways to use the Kubernetes control plane to seamlessly connect cloud native applications with the edge devices and infrastructure that they control and depend on.

Edge computing will drive a shift in how all applications are designed and managed. We expect novel solutions to emerge as the communities that design applications for web-scale data centers and cloud collaborate with their peers that operate in non-traditional computing environments at the edge.

By creating cloud native environments in edge locations, it becomes possible to spin up workloads that optimize the delivery of applications. For example, a cloud native augmented reality (AR) application could deliver low-latency services to nearby devices by running Kubernetes pods in precise edge locations, such as at the base of cell towers. Another promising development integrates edge devices, including IoT sensors and gateways, directly into a Kubernetes cluster, just like another federated cluster. CNCF community



Chris Aniszczyk  
CNCF

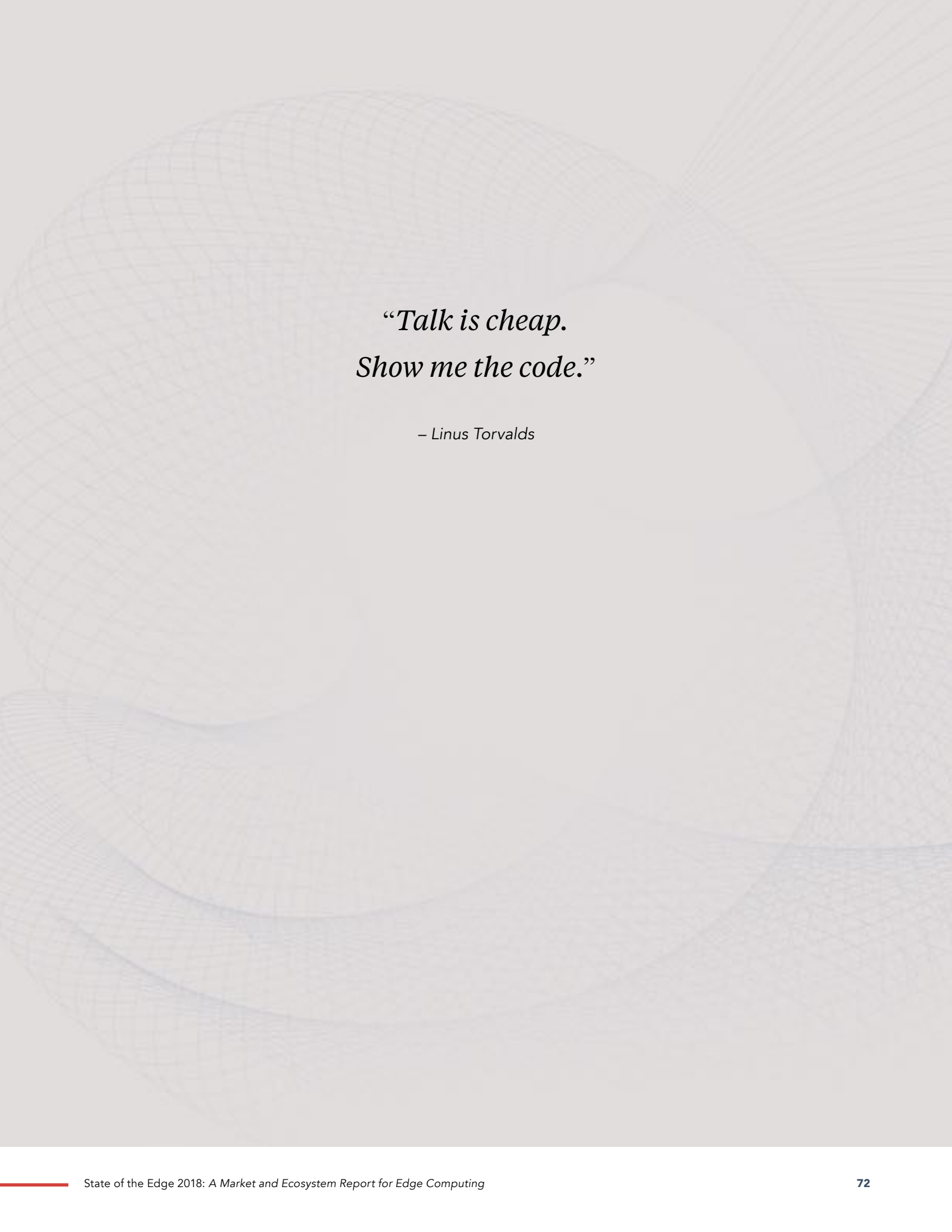
members like Microsoft have pioneered a version of this approach with the [Virtual Kubelet](#) project and their Azure IoT Hub service.

The CNCF is excited to participate in cross-community collaborations, such as the *Open Source Glossary of Edge Computing* which defines the key terminology involved with edge computing, as well as the [Kubernetes Edge/IoT Working Group](#). The working group, in particular, is where the CNCF community will converge to design and document ways to use Kubernetes in edge and IoT environments. We hope you can join us in this Kubernetes Working Group collaboration to see where the intersection of cloud native and edge computing will be.



# *CREDITS*





*“Talk is cheap.  
Show me the code.”*

– Linus Torvalds



# Authors

## **Jim Davis, Principal Analyst, Edge Research Group**

Jim Davis is the founder and principal analyst of Edge Research Group, as well as a contributing analyst for Structure Research. He has a track record of providing insights into the technologies and companies that are influencing markets. Prior to founding the Edge Research Group, Jim was a Principal Analyst with 451 Research for 17 years, covering cloud networking and security services, including content delivery networks, interconnection, peering, and cloud exchange services.

Learn more at [www.edgeresearch.group](http://www.edgeresearch.group).

## **Philbert Shih, Managing Director, Structure Research**

Phil is the Managing Director and Founder of Structure Research, an independent research firm focused on the cloud, hosting and data centre infrastructure service provider markets. Based out of Toronto, Structure Research is squarely focused on the SMB to mid-tier segments with an emphasis on international markets, particularly the Asia-pacific region. The firm provides publication of syndicated and customized research and analysis, strategic consulting, advisory services and public speaking. Clients include service providers, technology vendors supplying service providers and investors. Learn more at [www.structureresearch.net](http://www.structureresearch.net).

## **Alex Marcham, Network Architecture 2020**

Alex is a technologist, writer and researcher focused primarily on network technology, wireless networks, and edge computing. His 10 year career in engineering, marketing and product management has allowed him to translate his market insights into innovative products and solutions. He holds an MSc in network engineering and his analysis and research of industry technologies, trends and tribulations can be found at [www.networkarchitecture2020.com](http://www.networkarchitecture2020.com).

# Special Contributor

## **CNCF**

The Cloud Native Computing Foundation (CNCF) is an open source software foundation dedicated to making cloud native computing universal and sustainable. Cloud native computing uses an open source software stack to deploy applications as microservices, packaging each part into its own container, and dynamically orchestrating those containers to optimize resource utilization. Cloud native technologies enable software developers to build great products faster.

Learn more at [www.cncf.io](http://www.cncf.io).

# Sponsors

**Arm** – Arm technology is at the heart of a computing and connectivity revolution that is transforming the way people live and businesses operate. Our advanced, energy-efficient processor designs have enabled the intelligent computing in more than 125 billion chips. Over 70% of the world's population are using Arm technology, which is securely powering products from the sensor to the smartphone to the supercomputer. This technology combined with our IoT software and device management platform enable customers to derive real business value from their connected devices. Together with our 1,000+ technology partners we are at the forefront of designing, securing and managing all areas of compute from the chip to the cloud. For more information, visit [www.arm.com](http://www.arm.com).

**Ericsson UDN** – Ericsson's Unified Delivery Network (UDN) is a global strategic partnership platform that creates turnkey, value-added services. Ericsson UDN is building the world's first true edge delivery network at webscale, driving performance benefits and cost efficiencies. Together, we create a global edge network that offers the highest performance to consumers. For more information, visit [www.ericsson.com/en/tech-innovation/offerings/udn](http://www.ericsson.com/en/tech-innovation/offerings/udn).

**Packet** – Packet is the leading bare metal cloud for developers. Its proprietary technology automates physical servers and networks without the use of virtualization or multi-tenancy – powering over 60k deployments each month in its 18+ global datacenters. Founded in 2014 and based in New York City, Packet has quickly become the provider of choice for leading enterprises, SaaS companies, and software innovators. In addition to its public cloud, Packet's unique "Private Deployment" model enables companies to automate their own infrastructure in facilities all over the world. For more information, visit [www.packet.net](http://www.packet.net).

**Rafay Systems** – Rafay Systems enables next generation performance improvements for SaaS applications delivered over the Internet. Rafay's Programmable Edge™ platform equips developers with a disruptive set of tools to automatically deploy performance and geography sensitive applications, or micro-services, closer to endpoints. With presence at the infrastructure edge, Rafay's platform enables organizations to deliver a new set of experiences to their end customers. For more information, visit [www.rafay.co](http://www.rafay.co) and join the conversation on Twitter @RafaySystemsInc.

**Vapor IO** – Vapor IO is building the cloud of the future by delivering a suite of hardware and software for edge computing and operating the fastest-growing edge colocation business known as Project Volutus. The company's technology enables highly-distributed micro data centers to be embedded in the wireless and wireline infrastructure, colocated with the last mile or Radio Access Network (RAN), and meshed together with software and high-speed fiber as part of the company's Kinetic Edge, a technical architecture for city-scale edge computing. For more information, visit [www.vapor.io](http://www.vapor.io) and follow the company on Twitter at @VaporIO.



*EDGE COMPUTING  
LANDSCAPE*

# Edge Computing Landscape

One of the most challenging tasks in a fast-evolving market is to understand how technologies and vendors fit together. Simply keeping track of the market players can turn a passionate innovator into a part-time analyst.

Market maps are nothing new, but the open, collaborative approach taken by the CNCF and its partners (including Red Point Ventures) has been particularly effective in helping people make sense of things as the cloud native ecosystem expands and evolves.

Similar to the CNCF, we have embraced the tenets of openness and participation that allow dynamic ecosystems to be inclusive and meaningful. The intent is for the Landscape to be a living document that developers, investors, vendors, researchers and others can use as a resource. See a gap in our landscape map? Hear of a new company or project relevant to the space? We encourage members of the community to [open a GitHub issue](#) and contribute to the landscape map.

# Edge Computing Landscape

The intent for this Landscape is to be a living document that developers, investors, vendors, researchers and others can use as a resource. See a gap in our landscape map? Hear of a new company or project relevant to the space? We encourage members of the community to open a GitHub issue and contribute to the landscape map. Visit <https://github.com/edge-computing/edge-computing-landscape> for more information.



STATE OF THE  
**EDGE**  
2018



# *OPEN GLOSSARY OF EDGE COMPUTING*

## *Toward a Shared Lexicon*

The Open Glossary of Edge Computing seeks to provide a concise collection of terms related to the field of edge computing. The purpose of the glossary is to improve communication and accelerate innovation through a shared vocabulary, offering a vendor-neutral platform with which to discuss compelling solutions offered by edge computing and the next generation Internet.

As an official project under the stewardship of [The Linux Foundation](#), the goal is to help implement a community-driven process to develop and improve upon this shared lexicon.

You can find the official version of the Glossary via this [GitHub repository](#). Proposed edits, clarifications and suggestions are made by filing GitHub issues or creating "pull requests." Each issue, addition or suggested change will be evaluated by the community for inclusion. To contribute to the glossary, refer to our [Contributing Guide](#).

The Open Glossary is presented under the [Creative Commons Attribution-ShareAlike 4.0 International license \(CC-BY-SA-4.0\)](#) in order to encourage use and adoption. Code contributions to the project are licensed under the [Apache License, version 2.0 \(Apache-2.0\)](#).



## 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. Infrastructure edge computing is considered a key building block for 5G.

## Access Edge Layer

The sublayer 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 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: *Baseband Unit (BBU)*, *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: *Base Station*

## 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.

See also: *Central Office Re-architected as a 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 servers, storage, and computation services. Typically utilises 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 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*

## 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. 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*

## 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.

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, Traffic Offloading*

## 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 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 deduplication 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 Availability Zones

Isolated locations within the infrastructure edge, from which application workloads and other edge cloud services operate. An application may intend to be run on a specific edge availability zone for performance, reliability and data sovereignty reasons, or be dynamically allocated to it by a workload orchestration system. Edge availability zones are linked to others for failover.

See also: *Location Awareness*

## 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.

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, Device Edge, Last Mile*

## 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 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*

## 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*

## 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: *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 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 a layer beyond the infrastructure edge to separate and route network data received from the cellular network RAN. Although it does not provide the improved user experience of local breakout, IP aggregation can improve performance and network utilization when compared to traditional cellular network architectures.

See also: *Local Breakout*

## 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*

## Local Breakout

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 local breakout 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 and network utilization for the network operator.

See also: *Traffic Offloading, IP Aggregation*

## Location Awareness

The ability of an application workload to know where it is operating, in terms of its physical and logical location as well as the type and quantity of compute, data storage and network resources which are available to it. May also refer to the ability of an application workload to determine the location of its user, allowing the workload to move to the nearest point of the infrastructure edge.

See also: *Workload Orchestration*

### 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*

### 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*

### 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 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*

## Offload Processing

The use of compute, data storage and network resources at the infrastructure edge to process workloads which have been handed off to the infrastructure edge by other layers of the edge cloud system. An example is an edge device offloading a complex processing task to the edge data centers within the infrastructure edge to conserve its own battery life and limited resources.

See also: *Workload Orchestration*

## 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 realtime 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 utilises 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*

## 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: *Infrastructure Edge*

## 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 use of compute, data storage and network resources at the infrastructure edge to route network data in preference of another network path. This may be seen when the infrastructure edge is providing local breakout, and therefore provides a superior performance network path.

See also: *Local Breakout*

## 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 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, workloads 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 networking 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. **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: *Network Functions Virtualization (NFV)*

## Workload Orchestration

The process of determining where and how an application workload should be processed on the n-tier gradient of compute, data storage and network resources provided by the edge cloud. In most cases, workload orchestration will be performed by an automated system which takes into account the performance, time and cost requirements of an application operator at that point.me layer.

## xHaul ("crosshaul")

The high-speed interconnection of two or more pieces of network or data center infrastructure. Backhaul and fronthaul are the most common examples of xhaul today. Other examples include the connectivity between infrastructure edge sites, and between the infrastructure edge and any other significant network or data center infrastructure within the local area at the same layer.