STATE OF THE EDGE 2022

THE LINUX FOUNDATION

Contents

Executive Summary

The three overarching themes of the 2022 State of the Edge report are connectivity, location, and application infrastructure. All three play a crucial role in the development of edge computing. It takes a close look at the broadband access gap in rural areas, an issue that's core to the future of some of the most promising edge computing use cases; it examines the ins and outs of translating cloud native principles of application development and infrastructure management to deploying and running software at the edge, and explores new physical locations where compute infrastructure is being deployed to answer the need for ever more distributed platforms, including both on the ground and in Earth's orbit.

Here are some of the key findings:

- The top use cases for edge computing in rural environments currently are limited to retail needs, such as tracking inventory and in-store security, and 5G infrastructure deployments, which often contain edge computing components.
- The connectivity challenges of rural communities have made the task of bringing edge computing to these communities difficult.
- There is more interest in offering connectivity in rural communities today than ever before, as the pandemic has put into sharp relief the human toll of poor or no broadband access, and as the business case for that connectivity has begun to shift. The limitations of small customer pools for rural broadband are now being surpassed by the prospect of connecting hundreds, thousands, or even millions of devices used in IoT applications.
- Thanks to the emerging private space sector, the costs of both space launches and satellite hardware continue to fall, while constellations of satellites in low earth orbit promise to make satellite internet connectivity faster, cheaper, and more reliable. That connectivity may be an ideal option for otherwise inaccessible edge locations.
- Mobile operators are investigating the use of satellites as cell towers as one way to deliver the larger numbers of base stations needed for 5G and beyond. But over the next decade, satellites could also become compute platforms, with tiny data centers in orbit processing data gathered in space or from workloads sent from terrestrial edge locations.
- Despite the complexity and challenges — and they are massive, even with terrestrial edge infrastructure — data centers in space are likely to be commercially available within five to ten years, at least in a limited form.
- As the internet infrastructure landscape continues to mature, the benefits of centralization now come with certain glaring weaknesses regarding resiliency, redundancy, performance, and regulation. These factors have driven a new wave of investment and development activity outside of the traditional global tier 1 markets.
- After years spent building out centralized infrastructure footprints across a select group of core global hubs, hyperscalers are now looking to move to a more distributed deployment architecture. This is driving the emergence of new data center and network hubs across all geographic regions.

- While Kubernetes (or at least the Kubernetes API) has emerged as the mainstream option for container orchestration at data center scale, orchestration is more complex at the edge. The key is to find a way to abstract processes that aren't OS and architecture specific, because at the edge those are far more heterogeneous than in the cloud.
- As more (and more demanding) workloads have proven suitable for public cloud infrastructure, what's left in many cases are workloads that are best suited for the edge. The future promises an extension of the cloud provider model, one where available compute resources are treated as a fabric that can be utilized as required.
- While some cloud edge infrastructure might be in on-premises data centers, more of it will be in new edge data centers, embedded in edge devices or even built right into the telecom infrastructure.

What Is the Edge?

Researchers and experts from compute and communications industries have been at it for more than a decade now, but this question remains unsolved, not because of its complexity, but because of the edge's mere fluidity and construct. As much as the edge is about compute, it's about communications, and even more about fusion of compute and communications for distribution of intelligence, be it embodied as distributed compute, or distributed communications, or distributed compute and communications.

And for this reason, the edge must be solved collectively by the communications providers (traditionally the telcos) and the compute providers (the cloud), with a laser focus on consumption, treating the underlying tech merely as tools that can be swapped in and out.

And this calls for a new design paradigm: hypercomposed applications, built on demand, with the right type of resources, in the right amount, in the right place, and at the right time. Here the value differentiation must not be benchmarked for a single provider, or its service to the consumer, but across multiple producers and consumers, such that resources are continuously optimized across all producers, as well as across multiple consumers and producers—ideally leading to a circular economy.

And the very first step toward this circular economy is to bridge the plethora of edges distributed across the globe in a circular system that makes consumption of compute and communications as commonplace as any other utility embedded in our daily lives.

This year's State of the Edge report takes a close look at the key elements that must come together—and are coming together, albeit each at its own pace—to enable that new paradigm: ubiquitous network connectivity, new key compute and interconnection hubs, and orchestration tools that work across a diverse set of infrastructure, abstracting it and presenting it as a single platform for developers to build on.

And that's where we stand today with the state of the edge, taking our first step toward underpinning circular economies of the future with Distributed Edge.

Kaniz Mahdi
SVP Technology Architecture and Innovation, Deutsche Telekom

CHAPTER 1

Going to the Country

KENDRA CHAMBERLAIN

The Pandemic and Lessons in Connectivity

The past two years of the pandemic have greatly transformed life around the world. Mask wearing and social distancing have created new norms in our public and private lives, and the shuttering of offices forced businesses to dramatically revise their employment models. And although public health restrictions are beginning to ease, it's clear that the work from home (WFH) model isn't going anywhere anytime soon—and there are big implications for connectivity and edge computing as a result. We'll be using US as an example in this chapter, but the trends and issues highlighted here apply in poorly connected rural areas elsewhere around the world.

The rise of telecommuting during the pandemic taught us two big lessons in broadband connectivity: first, broadband access and connectivity still have substantial gaps among different segments of the US population; and second, residential networks now need the reliability, security, and latency of enterprise networks (or business users need to learn to tolerate consumer-grade connections).

The lack of digital equity in broadband access hit rural communities particularly hard during the pandemic. Employees with shaky to nonexistent broadband connections were unable to continue working from home while offices were closed, and children without access to adequate broadband were left unable to continue their studies from home.

Challenges in Rural Communities

The physical infrastructure of broadband faces many challenges in rural deployments, which explains why service providers have been slow to deploy services in these areas. Rural communities are characterized by smaller populations spread across larger geographies. Simply put, there's little to no business case for internet service providers (ISPs) to deploy physical infrastructure across large expanses of land to reach small customer pools. Consequently, rural broadband access has lagged decades behind urban and suburban counterparts—though that's beginning to change, as we'll see later in this chapter of the report.

The connectivity challenges of rural communities have made the task of bringing edge computing to these communities difficult. On one hand, localized data management can help alleviate some of the latency issues that rural connections may face to support business use cases. But on the other hand, the networks available need to be robust enough to support edge computing from the start. Additionally, the rural edge and IoT use cases (connected tractors, land monitoring, herd feed monitoring, herd pregnancy monitoring, remote utilities and pipeline monitoring and so on) are different enough, and carry enough revenue potential, to justify bespoke connectivity.

Closing the Broadband Gap

Broadband can serve as an economic lifeblood to rural communities. Expanding broadband access in rural areas would bring countless benefits. According to research conducted by Deloitte, robust broadband connections help attract new industries and businesses, increase property values, improve education opportunities, lower unemployment, and create new jobs.

Rural communities will also benefit from edge computing in a multitude of ways, from supporting new business opportunities in rural environments to bringing agriculture—considered one of the economic backbones of these communities—into the next generation, increasing access to healthcare through telehealth applications, improving community resilience to climate change by supporting smart grid or microgrid applications, and integrating more renewables and managing energy use across grids more efficiently.

The WFH model adds a new economic opportunity for rural communities. Now that employees don't need to live within a certain distance of the office, small towns across the US have an opportunity to attract new residents, who can essentially bring their work with them, as long as those employees have access to an adequate broadband connection.

1. Rural Connectivity Today and Looking Forward

Lack of broadband access has plagued rural America for decades. But as business, commerce, education and even healthcare have increasingly become more digital, the impacts of the digital divide between urban and rural residents are growing at an increasing pace.

The FCC estimates some 23 million US residents today do not have access to any broadband connection at home. Where access is available, connections tend to be slower and to cost more.

According to recent research from Omdia, about two-thirds of broadband connections in the US are between 100 Mbps and 500 Mbps, but speeds in rural areas are much slower. Research from the National Association of Counties found that 65% of counties across the US are averaging connection speeds slower than 25 Mbps.

Given these connectivity challenges, rural residents simply use the internet less. Today, rural residents are less likely than urban and suburban residents to have home broadband connections, to own smartphones or tablets, or even to access the internet daily, according to research from Pew Research Center.

The broadband gap is a well-known if not persistent challenge among federal, state, and tribal leaders. Regulators have for years used funding opportunities such as prizes and grants to incentivize ISPs to build out infrastructure in these communities, to varying success.

Once again, the pandemic helped tease out which rural connectivity solutions have been successful and which were not, according to the Institute for Local Self Reliance's DeAnne Cuellar. She points to public Wi-Fi infrastructure, such as connected libraries or community centers, heralded by some as a broadband access solution for rural communities. Public Wi-Fi is a great solution when the business or center is open to the public. But library Wi-Fi isn't accessible when the library is closed.

That's not to say progress hasn't occurred in recent years. The FCC's annual Broadband Coverage Report for 2021 reported that the number of US residents without access to 25/3 Mbps broadband service has been nearly halved since 2016, falling from 30% to 16% by the end of 2019. The report also noted that 83% of rural residents now live in an area with access to broadband services offering 25/3 Mbps.

Research from Omdia estimates broadband access is sitting around 86% in the US, and it expects that number to reach 94% by 2026, a nearly 10% increase in broadband penetration in just five years. While the digital divide is not solely a rural issue—plenty of communities within urban environments have been left behind, too—Omdia expects access to increase somewhat in rural environments during that time period.

Other estimates of the nation's broadband coverage in rural areas paint a much less rosy picture. A 2020 report from the research firm BroadbandNow estimated that around 42 million residents lack broadband access, double the FCC's estimate of about 20 million.

Internet access advocates—and even some of the FCC's commissioners themselves—have long called the FCC's methodology for mapping broadband coverage as inaccurate at best, relying on a census-block approach to mapping coverage that overestimates access to competitive services, and using self-reported data from ISPs that doesn't always align with the on-the-ground realities of broadband access in rural communities.

The FCC's accuracy in accounting for broadband coverage is important because it is the tool by which federal funds for broadband expansion are allocated. Much of the federal government's broadband funding is distributed through the Universal Service Fund, which the FCC established in 1997. Its purpose is to ensure that all Americans across all regions of the US have access to adequate telecommunications services "at just, reasonable, and affordable rates." In 2019, USF's disbursements totaled $8.35 billion.

FEDERAL FUNDING

In 2021, after one year of the COVID-19 pandemic, Congress authorized the Emergency Connectivity Fund (ECF) with a $7.17 billion budget under the American Rescue Plan Act. The ECF was created to help schools and libraries expand their broadband offerings to help support children transitioning to learning from home, with many funding recipients located in rural communities across all 50 states. To date, the FCC has committed $4.69 billion of that funding to applicants.

Even more funding is heading toward closing the rural connectivity gap today. The bipartisan Infrastructure Investment and Jobs Act included a $65 billion investment in broadband, considered a "once-in-a-generation" funding opportunity for states to improve and expand access to broadband.

We're now seeing what the Fiber Broadband Association CEO Gary Bolton calls "the beginning of the largest fiber investment cycle in history."

Tom Ferree, president of 501c3 nonprofit Connected Nation, has dubbed the federal bill a "once in a hundred years" event.

Rural communities are poised to benefit from this windfall, according to the National Association of Counties (NACO)'s Seamus Dowdall. The association considers the federal funding part of a nationwide effort to bring rural counties to the same table as urban communities in terms of broadband access.

2. Technology Challenges for Rural Broadband

Rural broadband penetration greatly mirrors the challenges the US faced in electrifying the country and bringing telephony to rural areas in the early 20th century. Rural communities are composed of small populations spread out over larger geographic areas, while the physical infrastructure of broadband is optimized for large populations packed into dense communities.

CONNECTING RURAL AMERICA

It's no coincidence that one of the earliest and most widely available broadband technologies for rural residents has been DSL, which utilizes the copper wires of telephone lines to deliver internet to rural communities. Similar to the Rural Electrification Act, which ensured every American resident had electricity, the Communications Act of 1934 guaranteed every American access to telephony. That infrastructure has since been upgraded and repurposed to support bringing DSL Internet to rural communities. For the most part, if a rural resident has a telephone line, it's likely their phone provider is also offering an internet service over the same wires.

Over the past decades, cable service providers have slowly expanded their offerings to more rural areas in competition with telephone companies, bringing DOCSIS internet over coaxial cables to more and more homes. But without the support or regulation from the federal government, cable providers have tended to deploy broadband services only in the densest of rural populations—for example, in downtown areas of small towns and villages, leaving the majority of residents within the greater community without access to service.

Lack of competition is an important factor in broadband access in rural communities, according to Omdia's Peter Boyland. Without competition, service providers in rural areas simply have less incentive to upgrade their services, leaving rural residents paying higher prices for lower speeds, outdated data caps, and subpar services.

The main drivers of rural broadband deployments today are regulatory incentives, public-private partnerships, and community-led projects, according to Boyland. But deploying infrastructure is just one piece of the puzzle. Ensuring that rural residents have access to the same speeds and services at similar price points is another important component to the digital divide. On that front, not all broadband technologies are created equal.

WIRELINE VS WIRELESS

Wireline broadband infrastructure, which is often more reliable than wireless offerings, faces a multitude of challenges when being deployed in rural settings. Copper-based DSL technologies have been widely available, but network upgrades have not kept pace with the demands of today's users, and telco ISPs are now favoring more fiber build outs to keep up with speed and latency demands. Cable's DOCSIS technology offers better return on speed and reductions in latency, but cable providers, too, have begun integrating more fiber in their networks, while cable connections have leveled off in recent years. Fiber, which is considered the gold standard in broadband network technology, is also the most expensive to deploy, and it is unfeasible for commercial rural deployments.

In response to the dearth of services, rural communities are now building their own fiber broadband networks to service residents at the city or county level rather than relying on service providers to deploy infrastructure. These community broadband networks, which are often fiber to the home (FTTH) networks, have been widely successful in bringing next-generation broadband services to communities that would otherwise be left using older technologies. According to the Institute for Local Self Reliance, there are currently more than 1,700 community broadband networks across the US. These services tend to offer the higher speed tiers that fiber enables, at more reasonable prices, while delivering ripple-effect benefits to local businesses, farmers, and manufacturers and boosting local economies in the process.

Wireless internet technologies have been heralded for years as a silver-bullet solution to obstacles that rural environments pose for offering broadband, and fixed wireless access (FWA) is currently one of the fastest-growing broadband technologies in rural areas. Research from Omdia estimates FWA penetration will increase nearly 10% in the next five years.

But FWA services have a few major caveats that can prevent rural customers from receiving the same type of connectivity that urban and suburban customers receive.

FWA services are easier to deploy in rural settings because they require less physical infrastructure. A single cell tower can service a wide swath of land, and lower-band frequencies offer good signal propagation that can travel large distances to reach end users. But because FWA services rely on mobile networks to deliver internet to the home, they tend to have narrower data caps and lower throttling thresholds than wireline residential services.

In recent years, 5G has been promoted as being able to offer fiber-like speeds to customers through FWA models. But 5G's network topology is not well-suited to serve rural communities. The fiber-like speeds promised by 5G require the use of high band mmWave spectrum, which doesn't travel very far and cannot penetrate objects such as trees or buildings. This alone makes 5G deployments tricky in urban settings and virtually untenable in rural areas. And because the signals do not travel very far, 5G networks require small cell densification, backed by tons of new fiber backhaul—much more infrastructure than earlier LTE technologies require.

Still, some cellular operators are deploying 5G in rural areas using low band spectrum in the 600—900 MHz and C-band ranges. These signals are much better at traveling further distances, but they cannot support the lightning-fast speeds that generated much of the hype around 5G in the early days. Low band 5G towers have about the same range and speed offerings as 4G LTE, offering speeds between 30 Mbps and 300 Mbps.

Satellite internet is one of the older wireless broadband technologies and has been the main connectivity option for rural communities where no wired infrastructure options exist. It can be deployed literally anywhere a receiver can be installed, and it requires practically zero physical infrastructure to be deployed. But satellite internet is far from perfect: the service is notoriously unreliable, often interrupted by weather; its latency poses a significant challenge for end users; and the receivers are much more expensive than FWA receivers. Traditional satellite internet is provided using satellites positioned in a geosynchronous equatorial orbit

(GEO). SpaceX's Starlink service promises to offer faster speeds and lower latency using low Earth orbit (LEO) satellites instead. But it's unclear whether an LEO satellite service will perform better during inclement weather, or whether Starlink's latency reduction claims will be realized in real-world applications. Overall, Starlink is just beginning to offer commercial services, and the LEO internet model is still largely untested.

Rural Edge Computing

Data and connectivity have become integral to nearly every aspect of life in the modern era, which is why demand for edge computing is expected to grow sharply.

But the technology is just beginning to emerge in rural areas today. According to Omdia's Roy Illsley, the top use cases for edge computing in rural environments currently are limited

Wireline Infrastructure

DSL: Telecom companies have developed an array of DSL technologies to improve speed and performance of high-speed internet over copper wires. Copper technologies have evolved in stages, with the fastest-performing technologies reaching up to 100 Mbps in a laboratory setting. But the fastest speeds reached over copper wires typically require fiber deployments, where the copper wire is transmitting the data over the last mile.

DOCSIS/Coaxial cable: Cable ISPs have gradually evolved the DOCSIS technology to deliver faster speeds over coaxial cables. Most cable ISPs have upgraded their networks to DOCSIS 3.1, which uses Hybrid Fiber-Coax (HFC) technology to deliver 1 Gbps service to customers located in large and mid-sized cities across their footprints. HFC networks use a fiber backbone to connect to a central office or local node and then use coaxial cables for the last-mile connections to the end user.

Fiber technologies (FTTx): Fiber broadband is the gold standard across every metric of broadband. It offers the highest speed potential, the largest capacity, and the lowest latency. It also serves as the backbone of every other type of internet technology, including satellite internet and LTE. The demand for speed and capacity across all sectors of modern life—and the rise of data-hungry applications—has steadily pushed fiber closer and closer to the end user, thanks to the deployment of HFC, fiber to the node (FTTN), and small cell densification with fiber backhaul.

FWA: Fixed wireless access refers to enterprise and residential internet services that use cellular networks to deliver connectivity to the home or business. FWA services tend to have lower latency than satellite internet offerings and can be offered at faster speeds. However, because the service relies on cellular networks, capacity is a huge problem for these networks, and service providers often impose data caps and throttling onto customers to help manage traffic.

Satellite and LEO satellite: Satellite internet has many limitations that make it a difficult solution for rural connectivity. Service quality is frequently interrupted by adverse weather, as thick, precipitation-filled clouds will literally block the signals from reaching receivers installed at the customer's premises. Satellite internet also struggles with latency, as the data is transmitted all the way up into space and then back down to the receiver. SpaceX's Starlink satellites are deployed in LEO, which SpaceX claims will help reduce latency and offer faster speeds.

Private LTE: Cellular operators have recently begun offering private LTE networks to support things like industrial IoT (IIoT) applications, smart grid infrastructure, and precision agriculture. Private LTE networks can handle high volumes of data traffic coming from IoT devices, for example, and can offer more reliability, security, and customization to customers.

to retail needs, such as tracking inventory and in-store security, and 5G infrastructure deployments, which often contain edge computing components.

As retail warehouses, logistics, factories, and even electrical grid applications become more automated, demand for small data centers and edge computing will only increase in rural communities, as will the demand for reliable connectivity.

As retail warehouses, logistics, factories, and even electrical grid applications become more automated, demand for small data centers and edge computing will only increase in rural communities, as will the demand for reliable connectivity.

Service providers will need to build out intelligent edge platforms and leverage cloud-native solutions in rural areas to help address the challenges presented by rural environments, and they will likely rely on 5G wireless networks to do so. But it may take a few more years before we see widespread deployments of 5G infrastructure for the purposes of serving rural communities.

In the interim, edge computing may be able to tackle some of the connectivity issues in rural communities today. For example, localized computing of data can cut down on latency issues and increase efficiency of data traffic moving across rural networks, allowing companies to utilize performance and latency-specific applications more effectively in areas where connectivity may be limited. Micro data centers, which can be as small as a utility cabinet, will be integral for rural edge computing. They can be deployed directly on sites such as offices, retail locations, healthcare facilities, farms, banks, and schools.

Despite its slow adoption rate, precision agriculture is a great example of this. Farmers using precision agriculture tools and applications will generate thousands if not millions of data points from things like measuring soil moisture and nutrient content to plant growth and stress, to temperatures and precipitation—to even operating autonomous combines. Other promising use cases are security via drone or camera across large agriculture sites, tracking herds and detecting imminent births or injuries (animal not moving for a long period), optimized feeding for dairy herds and worker safety. All of that data needs to be managed locally at the farm, for example, but it doesn't necessarily need to travel off the farm. Such applications will require connectivity, of course, but because the needs are localized, that connectivity can be served by an array of network technologies.

Ultimately, the intended use case will determine whether edge computing is feasible in a rural setting. More connectivity infrastructure needs to be deployed in rural areas before these communities will be able to realize the full benefits of edge computing.

Telehealth is another great example: rural communities typically lack adequate access to healthcare—another lesson the pandemic taught us. Edge computing plays an important role in telehealth applications: patient data, ranging from medical notes during visits to measuring vitals through wearables or health sensors installed in the home, will be better collected and managed

at the edge in real time, for example. But both healthcare facilities and the residents accessing care will need some form of reliable connectivity to take advantage of telehealth systems.

The good news is that there is more interest in offering connectivity in rural communities today than ever before, as the business case for that connectivity has begun to shift. The limitations of small customer pools for rural broadband are now being surpassed by the prospect of connecting hundreds, thousands, or even millions of devices used in IoT applications. So far, at least one wireless company is zeroing in on this opportunity and has launched a rural edge computing initiative focused on supporting data-heavy applications around logistics, energy, industrial IoT, and farming in rural environments.

Edge Computing Use Cases

In rural settings, access to connectivity and the use cases being supported by that connectivity will determine whether edge computing is feasible in the short term. But there is plenty of demand today for enhanced connectivity and edge computing in rural environments. Here are some of the use cases that rural edge computing will enable:

Remote work: The WFH trend could be an economic boon for rural communities as the importance of location for a career is reduced. But if rural communities are going to support employees moving out to the country to work from home, edge computing services will need to expand in rural settings to support business IT applications.

Smart city applications and autonomous vehicles: The widespread adoption of autonomous vehicles will require a proliferation of edge computing services along roadways to ensure the cars can navigate properly. This will become particularly important in rural areas, where infrastructure will need to be purposefully built out to support autonomous cars in areas that may have no connectivity at all. Similarly, small rural towns and villages will need enhanced edge computing capabilities to implement smart city applications such as turning street lights on or off or monitoring traffic cameras.

Industrial IoT/energy use cases: Plenty of industry verticals are deploying IoT-type devices and sensors to improve business operations in rural settings, particularly in the energy sector. Edge computing can enable renewable energy centers as well as conventional oil and gas facilities and grid stations to monitor and optimize energy production and distribution. Utilities will require edge computing to manage and analyze data generated from smart grid applications to track energy production and distribution in real time.

Precision agriculture: Precision agriculture is considered an important solution to assuring the US' food security as climate change wreaks havoc on food systems by helping farmers maximize land use and crop yield with less resources. But the agriculture sector has been slow to adopt the technologies. The federal government has recently launched a few initiatives to boost access to precision agriculture solutions, including the USDA's Agriculture Innovation Agenda and the FCC's precision agriculture task force.

Telehealth: Telehealth boomed during the pandemic as healthcare facilities worked to manage patient loads and mitigate the spread of COVID-19. And telehealth has the potential to completely transform rural healthcare access, once the connectivity and edge computing needed to support the services are built out in rural areas. Telehealth appointments can increase efficiencies for healthcare facilities serving large rural footprints—meaning more patients can be seen each day—while also saving patients time, energy, and money by seeing the doctor virtually at home. Telehealth can also increase access to specialists for rural residents by scheduling consultations or even examinations that are held virtually at the local clinic.

OTT content delivery and gaming: One of the consumer-driven use cases for edge computing that became a crucial benefit for many during the pandemic is content delivery and online gaming. Edge computing can help OTT service providers deliver better user experiences for customers and can improve online gaming experiences by reducing latency.

Widespread broadband access played a vital role in helping our communities and economies become more resilient in the face of nationwide shutdowns during the public health emergency. But the pandemic also showed us which communities are being left behind when lacking access to broadband.

Build It and They Will Come

The pandemic showed us exactly how essential connectivity is, and after two years of lockdowns, there's no question that access to broadband is integral to our day-to-day lives. Widespread broadband access played a vital role in helping our communities and economies become more resilient in the face of nationwide shutdowns during the public health emergency. But the pandemic also showed us which communities are being left behind when lacking access to broadband.

The rise of the WFH model presents both a challenge and an opportunity for rural communities. On the one hand, giving employees the freedom to work from wherever they live could offer a boost to rural economies as more people move out of the city and into the country. But these communities will need to have the connectivity required to attract and support people who are now able to take their work with them.

Closing the broadband gap in rural environments will also offer these communities opportunities in accessing and benefiting from the data-heavy and intelligence-driven revolutions occurring across all sectors of life, from increased access to healthcare and education to better management of energy production and distribution, to ensuring food security for the nation in the years to come. Edge computing will be an integral piece of the puzzle for these communities.

ABOUT THE AUTHOR

KENDRA CHAMBERLAIN is a freelance industry analyst and journalist, covering telecommunications, smart grids, and IoT. Her work has appeared in Rider Research, Rethink Research, BroadbandNow, and FierceWireless. Follow @KendraRC976 on Twitter and find her on LinkedIn.

CHAPTER 2

Data Centers in Space

MARY BRANSCOMBE, SIMON BISSON

Thanks to the emerging private space sector, the costs of both space launches and satellite hardware continue to fall, while constellations of satellites in Low Earth Orbit (LEO) promise to make satellite internet connectivity faster, cheaper, and more reliable. That connectivity may be an ideal option for otherwise inaccessible edge locations. Mobile operators are investigating the use of satellites as cell towers as one way to deliver the larger numbers of base stations needed for 5G and beyond. But over the next decade, satellites could also become compute platforms, with tiny data centers in orbit processing data gathered in space or from workloads sent from terrestrial edge locations.

In many edge cases, the old real estate mantra "location, location, location" is appropriate, because it's about getting the compute in exactly the right place. But tradeoffs must often be made, with the difficulty of providing connectivity in remote locations (for remote deployment and management and collection of data for further analysis). Large geographical areas in many countries still have no mobile data coverage. Even developed countries with strong coverage have poor network performance in rural areas, because of terrain as well as the distance from the cell mast.

Satellite connectivity is a key technology for expanding coverage of wireless communications networks to more remote areas, including oceans (oil rigs, for example, or cruise ships, which nowadays are effectively floating data centers, with huge connectivity needs), for temporary installations for sporting and entertainment events, and for emergency services. (Imagine firefighters equipped with real-time satellite imagery and accurately directing fire suppressant or water, or live video to help a remote surgeon guide a paramedic in a trauma center.)

	ORBIT	NUMBER OF SATELLITES FOR GLOBAL COVERAGE	TIME PER ORBIT	TIME IN SITE/ GATEWAY	RTT LATENCY	MASS	LIFETIME
GEO	35,000 km (22,000 miles)	3	24 hours	continuous	600-700ms	~3500kg	15 years
MEO	5,000-12,000 km (3,100-7,500 miles)	8-30	6-12 hours	2-8 hours	<150ms	~700kg	12 years
LEO	500-1,600 km (300-1,000 miles)	100+	40-90 minutes	15 minutes	<50ms	5-1,000kg	4-7 years

CHART: Satellite performance and operational characteristics in different orbits

Understanding Orbits

Low-cost commercial space launches make possible constellations of much smaller satellites in LEO that deliver higher bandwidth and lower latency than the traditional satellite broadband from geosynchronous satellites. In fact these LEO constellations can offer lower latency than fiber in some circumstances.

Geosynchronous satellites, as the name suggests, stay in one position and in effect orbit in sync with Earth's rotation around 35,000 km (22,000 miles) above the Equator. A fixed antenna always points at the satellite. While it takes only three satellites to provide global coverage, and while they can relay information to satellites in other orbits, these are the most expensive satellites to launch and build. They offer no coverage above or below the 70-degree latitudes.

Medium Earth Orbit (MEO) satellites are much smaller and orbit at lower altitudes, with lower latency and data speeds of around 1.6 Gbps. They don't have to orbit the Equator—polar orbits enable them to cover the highest and lowest latitudes, which is why they're used for global positioning systems such as GPS, GLONASS, and Galileo—but more satellites are required for full global coverage. Russia and Canada have used highly elliptical orbits to provide coverage to high latitudes, with the slowest and lowest part of the orbit used to minimize the need for tracking antennas.

LEO satellites don't have to be smaller than MEO satellites—the International Space Station is in LEO—but they often are. CubeSats start at 10cm on each side. LEO satellites are cheaper to build and launch, with a cluster of more than 50 smallsats in a single rack. They have the lowest orbit (making them ideal for high-resolution satellite imaging) and move faster, completing an orbit in as little as 40 minutes.

But the low orbit that gives low-latency connections also means that many more are needed to deliver global coverage. There are additional issues; space weather, for example. Coronal mass ejections heating and expanding the upper atmosphere have more effect on low orbit smallsats, and so their constellations require a relatively high percentage of on-station backups to ensure coverage is not affected.

To deliver enough connectivity for large numbers of users, providers plan to create mega-constellations, with tens of thousands of satellites in multiple orbits, using lasers for mesh connectivity and multi-hop communications between them.

At the speed of LEO orbits, the topology of a constellation changes frequently and the mesh network enables connectivity handovers as the satellites move. Instead of needing sufficient (expensive) ground gateways for an entire constellation, satellites can transmit data to another unit that's within the range of a gateway for downlink, using direct satellite-to-satellite optical links. As these operate in a vacuum there is less distortion, so links can offer higher bandwidth than free-space optical connections on Earth.

The most sophisticated approaches to satellite edge communications networks will use a mix of geosynchronous, LEO, MEO, terrestrial networks, and possibly even unmanned high-altitude

platforms, or HAPs, (airships, balloons, or solar-powered drones in unregulated airspace) to provide the best coverage, matching dynamic traffic demands to network capacity supply. Antennas that can connect to multiple satellites on different frequencies, in different constellations, or in different orbits, will make these networks more flexible.

In the short term, mobile operators such as NTT and Vodafone are looking to use satellites as "towers in space" for 5G (perhaps using the way in which Open RAN deconstructs and standardizes the usually proprietary and vertically integrated radio access network). Satellite links will carry end-to-end slices that provide network services with the specified Quality of Service (QoS), such as offering live 4k video channels with an MEC platform integrating with a CDN, or as backhaul for 5G service.

Public cloud offerings that deliver ground station control and satellite downlink "as a service" also make this kind of network deployment accessible to parties other than traditional telecommunications providers. Services such as AWS Ground Station and Azure Orbital simplify the management of space assets without investing in a dedicated network of ground stations. This allows satellite operators and customers to acquire data from communications and imaging satellites more effectively, or to use them for connectivity.

Compute Above the Clouds

While satellite connectivity from LEO constellations will be a big boost for MEC, the next step is to extend the edge into space by adding compute capacity to satellites, to process data created in space or pre-process data transferred by satellite from edge locations to the cloud.

Even the larger satellites in geosynchronous orbit are no longer the single-purpose systems they used to be. Payloads can be shared between customers, so adding computing power to a satellite already scheduled to launch would make it more valuable, getting more out of the high cost of launch.

Payloads can be shared between customers, so adding computing power to a satellite already scheduled to launch would make it more valuable, getting more out of the high cost of launch

Some existing satellites can be repurposed with new payloads by uploading new software, although this is limited by the communication frequencies and the capabilities of the satellite hardware. With a shared payload, performance may degrade because the satellite platform must accommodate all the different payloads. All the familiar problems of multi-tenant hosting are transferred to a far more hostile environment.

LEO hardware lifespans can be surprisingly long. While SpaceX's Starlink smallsats have a design life of around five years, there are constellations like Planet's Dove imaging smallsats, where the oldest devices have been on station since 2015. This imposes additional constraints. Older units, while still usable, have older compute and memory technology. Any workload scheduling tools used to manage a cluster in orbit must take this into account.

LEO payloads are increasingly built with standardized hardware and, in some cases, commercial off-the-shelf processors. NASA's PhiSat-1 uses Intel's Movidius Myriad 2 chip (frequently used in AI-powered IoT cameras at the edge) to identify and discard satellite images obscured by clouds. That saves about one-third of the bandwidth on an expensive downlink.

AI processing could also be used for computational photography (capturing better images by detecting how far away objects are and how fast they're moving) or to identify features of interest, such as forest fires or floods, that should be photographed and prioritized for downlink. NTT and SKY Perfect plan to offer on-satellite image processing by 2025 to similarly reduce the cost and time of satellite downlink as one of the capabilities of their multi-orbit (GEO, LEO, and HAPs) computing network.

Loft Orbital plans to put a customer's own payload on its satellites and give them a web portal to control it, or build the payload for them so they can concentrate on developing applications for the payload. Currently this "space infrastructure as a service" market is focusing on more traditional satellite applications such as imaging, sensing, and satellite communications for IoT devices, with software-defined radios that can be switched to use different antennas for multiple customers on one satellite. The company plans to offer compute on satellites for processing and analyzing imagery and IoT data.

Some startups, such as OrbitsEdge, plan to put off-the-shelf 5U 19-inch rackmount servers in the standard satellite payload frame that can be used to process, cleanse, aggregate, and analyze data. Others, including Lacuna Space and OQ Technology, have started by running tasks on existing satellites, downloading data from IoT devices over LoRaWAN and Narrowband, but are moving toward launching their own dedicated satellites to gain more control.

More ambitious (and currently theoretical) approaches promise secure data storage or distributed compute in LEO constellations, using the in-orbit laser links used for the switching and routing that enables multi-hop communications between satellites. Moving workloads or, more likely, results (due to the limited inter-satellite bandwidth) between satellites will require sophisticated orchestration for several reasons: the available compute and storage resources in different satellites, power and cooling costs, latency, and communications bandwidth. The cost of moving a workload to another satellite must be less than simply sending the data to an edge compute node in the satellite gateway on Earth. Load balancing will necessitate efficient discovery of the satellites that are capable of running a workload and available to do so, based on what workloads are already running and when they are expected to complete.

The calculations for moving workloads or pooling resources between satellites to enable them to handle more demanding workloads will be complex, given the costs of evicting and migrating a workload. Doing that on resource-constrained satellites will require a very efficient, ultra-light container management system.

At least initially, satellite edge data centers will be suitable for tasks such as handling imaging and sensor data from IoT devices and other satellites, where the latency of the connection

or the ability to process data gathered in space is critical and the tradeoff between energy consumption, time delay, and computation cost make sense, given the restrictions on processing power and other resources (including operating power) in the harsh environment of space. Augmented reality and offloading compute for UAVs, drones, and self-driving vehicles might require workload scheduling so complex as to require techniques such as Reinforcement Learning.

Hardware Constraints in Space

Even the lower launch costs of LEO satellites quickly add up. The weight and size of even microdata centers and edge compute nodes and hardware that will operate outside the Earth's protective atmosphere has traditionally required hardening to deal with cosmic radiation, especially solar flares, and extreme temperatures.

The Myriad 2 processor used on PhiSat-1 is usually found in terrestrial devices, but it is based on SPARC V8 LEON4 controllers developed jointly by the European Space Agency and CAES for rugged and high availability domains, including space. It was tested at CERN for performance under high levels of radiation. CAES is also developing space-capable RISC-V based processors.

Tests of commodity 64-bit data center servers on the ISS have shown that standard commercial hardware can operate successfully in space without extensive hardening or shielding, although with some limitations. Hard drives have higher failure rates at extremes of hot and cold temperatures, even on Earth. Even with the more stable temperatures on the ISS (kept around 65-80F to avoid cold spots that could lead to condensation and corrosion), SSDs fail more frequently in space and CPUs make more errors. They are correctable errors, but the system must be designed to check for and correct them. The servers on the ISS were shut down when high levels of cosmic radiation were expected; satellite compute must feature similar precautions. Distributed computing in a constellation of satellites might allow you to move a workload out of harm's way, for example when solar flares or a more severe coronal mass ejection is predicted.

The Earth currently experiences over 2,000 of the lowest-category solar storms (G1 and G2, on a scale that goes up to G5) in a decade, and NASA studies suggest that we're entering a period of more active solar weather as solar activity tamps up. Disruptions may become more common.

Space weather predictions won't always be reliable, either. SpaceX lost 40 satellites soon after launch to a solar storm that NASA originally believed to be traveling too slowly to cause damage when it reached Earth. When scheduling and resourcing allocations for satellite compute, satellite operators may also have to take into account changes to the thermosphere caused by solar weather. This is the layer of the ionosphere where solar activity changes the temperature of the atmosphere. Ultraviolet radiation causes atmospheric particles to become electrically charged and to refract radio waves. Solar weather can affect downlink speeds, which might affect calculations about where to host a workload or when to move it.

Moving to silicon photonics for on-board electronics can extend component life and reduce the amount of shielding required, which reduces size, weight, and launch costs. That's because the metal wires currently used for interlinks pick up charge as a satellite moves through the Earth's magnetic charge, and intermittent voltages reduce processor life. Optical interlinks don't have the same problem, but they are currently less common and more expensive.

The familiar data center issues of power and cooling are present in space too. Solar power is sufficient to power compute nodes on a satellite, but it's available only when the satellite is in sunlight, not shadow. So satellites will need batteries to power compute for part of the orbit, and when allocating workloads the scheduler must take into account satellite position and battery capacity.

When a satellite is in shadow, the external temperature will drop to around 250F below zero. In sunlight temperatures rise to about 250F, so the compute node will require cooling. The ISS used water cooling for the test servers (tapping into a heat exchange system that uses cold plates, heat exchangers, an internal water loop, and an external ammonia loop that circulates through large radiators).

On satellites it's usually done with a combination of shade (often provided by the solar panels), insulation, heat pipes, and radiating fins. That can be effective but it can also fail. The recent GOES-18 launch was done specifically to replace a satellite with an inadequate heat pipe that wasn't transferring enough heat away from the electronics, causing the imaging system to fail more frequently than expected. Adding servers to a satellite may require more advanced cooling techniques.

One possible architecture for satellite-based computing is to replicate terrestrial edge IoT networks, treating most satellites in a constellation as sensors or network hardware, but with larger, more compute-capable hardware used to provide in-orbit processing.

With compute in orbit, maintenance must obviously be conducted remotely. LEO satellites in particular should be considered semi-disposable, with a lifespan not dissimilar to public cloud hardware (but much shorter than most industrial systems).

One possible architecture for satellite-based computing is to replicate terrestrial edge IoT networks, treating most satellites in a constellation as sensors or network hardware, but with larger, more compute-capable hardware used to provide in-orbit processing. Here there is an opportunity to move compute hardware out of the increasingly congested low-orbit constellations and into MEO. Laser links from LEO satellites would send data to MEO for processing, with data returned to ground either via LEO or direct links to ground stations.

Using MEO satellites for processing reduces risks from space weather events, allows the use of better shielded hardware and supports better thermal protection, giving hardware a longer life. There is also the opportunity to use larger solar arrays and batteries, supporting more

compute and memory. In practice there will likely be a trade-off between using smallsats and larger satellites. Mixed fleets of devices will become increasingly important.

Satellites are probably safe from physical tampering (although not from damage, whether accidental or deliberate). But the security of data and workloads relies on the security of the network, satellite gateway, and ground station they connect to, which should be treated like any other cloud network. As we've seen since the invasion of Ukraine, internet connections to satellites are as vulnerable to disruption due to attacks as are terrestrial networks. Fixed ground stations will remain a weak point, so moveable antennas that can connect to multiple satellites in different orbits or on different frequencies may be important for security as well as flexibility.

The nature of satellite compute makes it ideal for imagery and sensing in remote areas. There are some compelling humanitarian applications: detecting floods and wildfires, tracking crop health and signs of drought, or monitoring endangered animals at risk of poaching. But it will also be useful to authoritarian regimes, with applications such as tracking refugees trying to cross a remote border. Organizations collecting satellite data or considering uses for satellite compute should put in place an ethical framework, to guide their choices and decisions. This is something that questions about the use of Russian facilities for satellite launches have made more of a priority.

Despite the complexity and challenges, data centers in space are likely to be commercially available within five to ten years, at least in a limited form. Initially, developing applications for them will be more like working with IoT devices than building for on-premises, access, or regional edge. But for the right applications, LEO compute could be the next frontier.

ABOUT THE AUTHORS

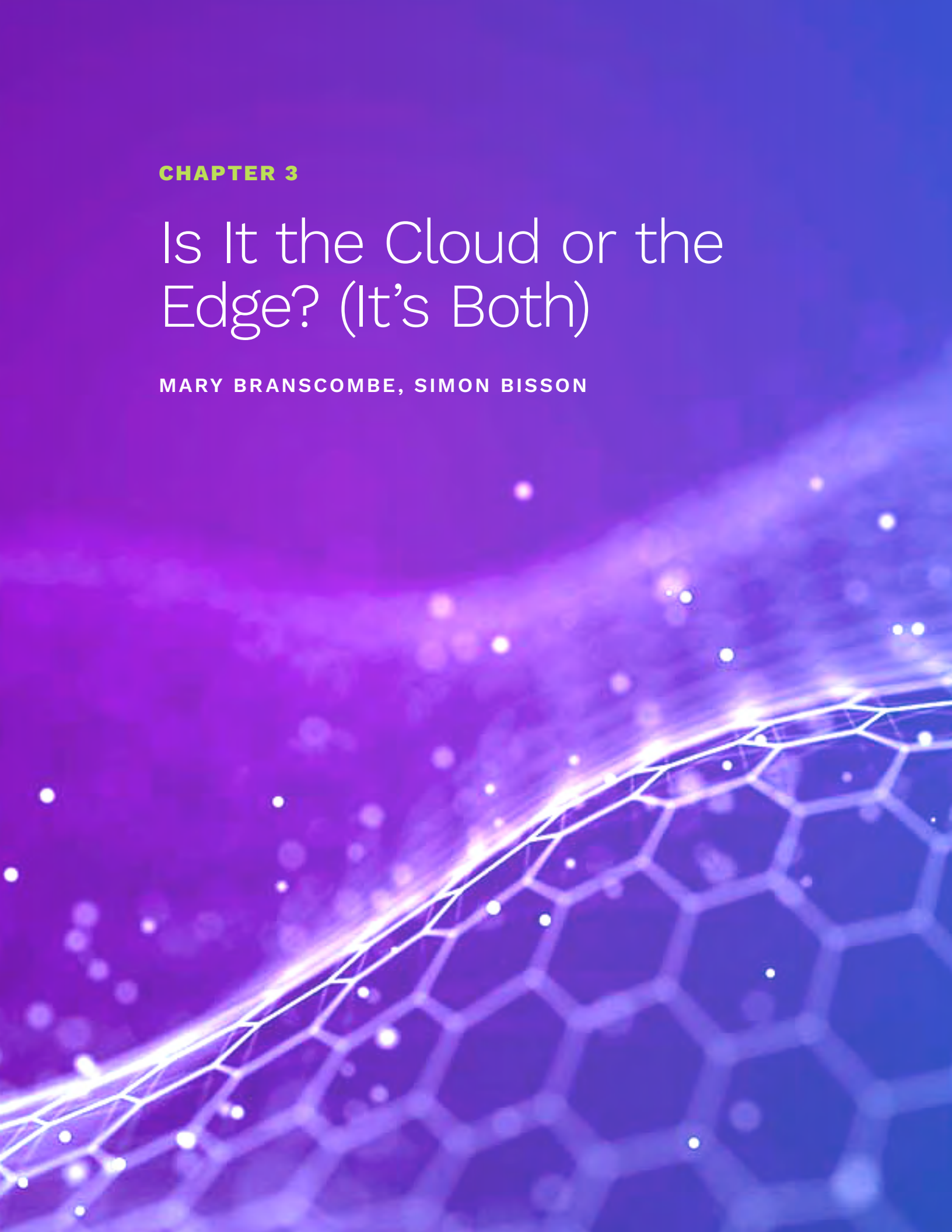
MARY BRANSCOMBE is a freelance technology journalist who writes for a wide range of titles. She has been a technology writer for nearly three decades, covering everything from early versions of Windows and Office to the first smartphones, the arrival of the Web, and most things in between: from consumer and small business technology to enterprise architecture, cloud services, and mainstream AI. She also dabbles in mystery fiction about the world of technology and startups.

SIMON BISSON has been a freelance technology writer since the '90s, covering everything from startups and consumer gadgets to small-business technology, enterprise architectures, developer tools, and application development. In what he sometimes describes as "career as verb rather than noun," he moved into IT from academic, military, and telecoms research (sometimes all three at once), running the technology side of one of the UK's first national ISPs and working as a consultant on many early ecommerce sites, with a natural progression into journalism to explain to other people how to do similar things. He still writes code (currently aircraft tracking for Raspberry Pi).

CHAPTER 3

Is It the Cloud or the Edge? (It's Both)

MARY BRANSCOMBE, SIMON BISSON



Cloud and edge are too often seen as competing alternatives. In fact, they're part of the same continuum of putting colocation, compute and storage in the most effective place, done in the most efficient way.

Edge adoption will ultimately be driven by business value, whether that's cost savings from consolidated hardware that's easier to manage or the availability of applications and services that solve business problems better at the edge because that's where the users or the data (or both) are located. If the edge is more complicated to manage or develop than the cloud, cheaper hardware or lower latency won't be enough to attract users. Many organizations want their familiar cloud services brought to the edge where the data they want to process is created, on their choice of infrastructure — and with the flexibility of cloud consumption models. They want the cloud experience.

The original hybrid cloud promise was compelling: the experience of a public cloud but running in a private cloud on your own infrastructure. As the concept matured, it became clear that it should also be connected to public cloud infrastructure for any workloads that don't need to be on your own infrastructure, because a private cloud will rarely offer the economies of scale and innovation that a hyperscale public cloud can offer. As more (and more demanding) workloads are proven suitable for public cloud, what's left in many cases are workloads that are best suited for the edge.

That's not redefining edge to be the same as on-premises computing, because this "cloud edge" is deployed and managed very differently. And while some cloud edge infrastructure might be in on-premises data centers, more of it will be in new edge data centers, embedded in edge devices, or even built right into the telecom infrastructure. Regardless of where it is, however, users will be able to consume services on cloud edge infrastructure the same way they consume services on traditional cloud infrastructure.

It's Still the Cloud, Only at the Edge

The future promises an extension of the hyperscale cloud provider model, one where available compute resources are treated as a fabric that can be utilized as required. However, when considering the role of edge devices, we can think of it as a two-dimensional fabric, one where latency (and therefore proximity) is a resource that needs to be considered when deploying workloads.

Organizations and, especially, developers want to have access to familiar cloud services and patterns at the edge, rather than having to deliver and use new systems for edge workloads. This is more than just using cloud services to configure and manage IoT devices (although more sophisticated IoT cloud services cover everything from creating images for devices that turn them into cloud service endpoints to applying regular security updates). Instead, this infrastructure extends the cloud experience to the edge with similar primitives and services.

By extending the compute, network, and storage fabrics to the edge, there's no need to use new development models. Orchestration tools will deliver application elements where they're needed, taking advantage of the design-by-contract model at the heart of microservice-based cloud native applications.

Although the same components may run in the core and at the edge, they don't need to offer the same services or the same level of detail. Edge devices with privacy-preserving shopper analytics can let retailers, for example, measure the effectiveness of store layout and displays, manage queues, and have all that shown in the same cloud CRM service as analytics about which marketing emails perform best.

If you're using image recognition to spot defects on your manufacturing line or to make sure parcels are routed onto the conveyor belt for the right delivery truck, the real-time nature of the problem means you'll want to run the inferencing locally, so you can divert substandard components or incorrectly routed packages before they go any further.

But you might choose to build those machine learning models in a cloud service before deploying them to the edge, and before uploading a subset of the data to run analytics to understand trends. If you're using predictive analytics to avoid costly failures to manufacturing equipment, you can run that in a cloud-based machine learning service without the overhead of managing your own infrastructure, an approach similar to the one hardware developers take, leveraging digital twins to monitor and predict device performance. Here, though, digital twins are a deployment model, with code developed using cloud native tools and then delivered on infrastructure where it's needed, when it's needed.

If the data you upload to the cloud is potentially sensitive or regulated, you may want dedicated edge hardware to run sensitive machine learning models. Vice versa, you may want to use cloud services to process data that needs to be kept in a particular region, like healthcare information or smart city data.

You can also configure, manage, monitor and update this cloud-connected edge infrastructure through the same cloud services you use for IaaS and PaaS: tools and processes to manage access and identity, set security policies, build and deploy applications using CI/CD and DevOps, apply patches, and perform data governance. Keeping the same security architecture for core and edge reduces risk, ensuring that the same rules are used everywhere and that no errors occur when those rules are copied and implemented across different platforms.

If the data you upload to the cloud is potentially sensitive or regulated, you may want dedicated edge hardware to run sensitive machine learning models. Vice versa, you may want to use cloud services to process data that needs to be kept in a particular region, like healthcare information or smart city data.

Cloud Infrastructure Merges With Telco Infrastructure

Another cloud edge infrastructure pattern extends hyperscale data centers into much smaller telco facilities, with dedicated cloud hardware in the same racks and cages as the infrastructure used to provide connectivity. This approach has the added advantage of using existing physical infrastructure; newer generation cellular gateways have a much smaller footprint than older models, leaving space in the cabinet for other purposes, like edge servers.

That's particularly suitable when telco operators want to support workloads like augmented reality and autonomous navigation, where the latency of even a fast cloud network connection is too great (and the data would put a significant load on their Internet backhaul). The use of pico-cell 5G networks helps here, as they offer very low latency connections between edge servers in the cabinet and devices.

Although some telcos have experimented with providing their own cloud services, they don't have the same experience reaching and supporting developers as the hyperscale clouds. By partnering with existing "adopted at scale" cloud providers, telcos can offer customers the near-real-time, low latency cloud services they want without developer teams needing to build substantially new workflows.

New Cloud Services With Edge Needs

The range of what can be provided as a cloud service continues to grow, and these new services are sometimes best suited to run on cloud edge infrastructure.

Azure, for instance, offers private 5G "as a service." The company is working with AT&T to provide a service that uses Azure private MEC with Azure Private 5G Core to deploy easy-to-use private wireless networks. AT&T Private 5G Edge might enable a car dealership to have a private cellular network that customers can use to configure their connected car before they drive off the lot (or perform vehicle diagnostics securely). A hospital might track equipment like ventilators or wheelchairs on a private network and could roam the devices onto a public mobile network to track those assets if they're sent off site for maintenance or loaned to another healthcare provider.

Or, operators can run their public wireless networks on cloud infrastructure located at the telco edge, which includes cloud services for security, monitoring, app development and deployment, IaaS, PaaS and virtual network functions, while maintaining full control of their own customer data. With such a deployment, operators can offer customers edge infrastructure for high bandwidth, low latency applications like analyzing IoT data, which they can manage alongside the apps and services they consume in the cloud.

A Spectrum of Connectivity and Cloud Coupling

Cloud edge infrastructure, the computing infrastructure deployed at the edge as an extension of a traditional cloud and consumed in the same way, needs some connectivity to the cloud, at least to set the system up (and for billing), but can often run disconnected for long periods

of time. Sensors at a remote oil field might only batch and send data to the cloud once or twice a day; a cargo or cruise ship might only be connected to enterprise cloud services running in a data center on land when it's in port. There's a spectrum of options for how much connectivity is required and how tightly coupled the edge infrastructure is to the cloud that delivers services and identities that vary between different solutions and providers. However, this infrastructure is effectively part of the cloud and doesn't function independently.

Minimal connectivity options are particularly suitable for agriculture, when used in conjunction with white space networking to give low power coverage over large areas. The same technologies can be used for disaster response, dropping in preconfigured ruggedized edge systems paired with local wireless networking and with BGAN or similar satellite uplinks for access to machine learning systems running in far-away hyperscale clouds.

This integrated approach puts hardware at the edge that delivers a platform designed to support cloud services. This close coupling may also restrict what hardware is supported. Although hyperscale cloud infrastructure is no longer as homogenous as it once was, testing and debugging updates on a wide range of hardware is time-consuming and prone to failure. Minimizing variability is key to being able to update cloud services on edge infrastructure as frequently as they're updated in the cloud (where there can be hundreds of updates to different microservices each month).

In many cases, the hardware will be provided by the cloud provider — possibly in a "hardware as a service" model, where customers pay by usage the way they would for a cloud service, with no upfront hardware or license costs. When more heterogeneous hardware is supported, customers will also have to do more work to deploy and manage this edge infrastructure, and fewer cloud services may be available.

The edge infrastructure typically uses the same security, virtualization and encryption approach as the public cloud, with similar shared models in which customers remain responsible for the security of their applications while the cloud provider secures the infrastructure. However, edge customers might also be responsible for the physical security of the infrastructure and for providing reliable, consistent network connections.

Tightly Coupled Cloud Edge

With standardized, validated hardware that's effectively a cut-down version of the setups deployed in public cloud data centers (bought from familiar IT and telecoms OEMs, delivered and installed locally as servers, racks or scale units but managed through a cloud service), edge infrastructure like Azure Stack Hub and AWS Outposts creates a mini cloud data center.

This infrastructure can still run disconnected once it is set up. Some, like Azure Stack Hub, can even be deployed without an Internet connection (although some features won't work as well because they rely on access to cloud services, and you'll have to pay a fixed annual subscription rather than paying just for the services consumed). You can use them with even intermittent connectivity to the cloud and pay by usage for the cloud services you choose rather than having upfront costs.

These integrated systems with preinstalled software are the most prescriptive in terms of hardware (although there is usually a choice of form factors), but that means they offer the widest range of cloud services: IaaS and PaaS services like storage, database-as-a-service, containers, caching, load balancing, serverless and event-based app models, as well as cloud marketplaces for third-party software and services.

Developers write applications to run on this edge infrastructure using the same tools and APIs as they would for the same services in the public cloud, including hybrid applications that can be deployed in the public cloud or at the edge by treating the edge infrastructure as a region or VPC subnet. Admins use the same cloud services as in the public cloud to deploy and monitor workloads (including third-party tools like Terraform). Meanwhile, system administration and management are handled automatically by the cloud provider or a partner (though operations teams have flexibility in scheduling update windows if they would affect availability).

This is the edge infrastructure that cloud providers are putting directly into the infrastructure footprints of existing telecoms. Telco providers are placing Azure Stack Hub and AWS Wavelength hardware in switching centers and other locations at their network edge so that developers can build ultra-low latency applications like live 4K streaming or near-real-time object detection at the edge — but they can scale them as required. Size, weight and power constraints limit what compute you can put inside the smallest of edge devices, like drones or smart cameras. But with the cloud services running at the network edge, the latency is low enough to offload demanding AI and analytics tasks to make those devices far more powerful.

Smaller, rugged fully coupled cloud edge devices are also available, which are designed for IoT and machine learning scenarios in remote environments that double as ways to transfer data to the cloud by physically shipping it (on a device that can process the data while in transit), like AWS Snowball Edge and Azure Stack Edge. These devices can be powered from inverters in standard off-road vehicles for rapid deployments.

Service-Coupled Cloud Edge

With plenty of edge infrastructure already deployed, not everyone wants to roll out new (and often pricey) hardware or outsource all their infrastructure management responsibilities to get public cloud services at the edge.

There are a wide range of hardware-decoupled approaches, including multiple options from individual cloud providers to cover all the different use cases. These aren't edge-specific; they cover hybrid and multi-cloud options, giving you a unified, cloud-managed approach for a wide range of infrastructure that includes the edge but isn't always optimized for it.

Azure Arc takes VMs, bare metal, databases or Kubernetes clusters — locally or in any public cloud — and turns them into Azure-managed infrastructure where you can deploy IaaS, evergreen cloud database services, serverless and application services and machine learning services, while taking advantage of cloud security, policy and governance. Google Anthos takes a similar approach. However, because it's Kubernetes-specific, it may be less suitable for more constrained edge environment where Kubernetes (at least as we've seen it in the public cloud) isn't appropriate.

Amazon ECS Anywhere lets you run Amazon's container management on your own infrastructure, which turns that into managed instances and allows you to burst into AWS when you need additional capacity. (Amazon EKS Anywhere is a similar offering for running a local Kubernetes cluster using a cloud control plane.)

Azure Stack HCI is a hyperconverged offering that you buy as a cloud service (so there are no upfront hardware or license costs), with the option of managing, monitoring, securing and backing up workloads using cloud services. It can run both Azure Kubernetes Service and the Azure services available on Azure Arc, but the admin experience is more like working with on-premises servers than Azure Stack Hub and the minimum footprint is two servers. Hardware administrators get access to their own console, while application admins use the equivalent of the cloud Azure Portal to build and deploy resources onto managed hardware.

IBM Cloud Satellite combines cloud, data center and edge infrastructure into a single location where you can run a small number of IBM Cloud services and software from IBM and RedHat marketplaces. (You can also run that on a prepackaged IBM Pak System instead of on your own infrastructure to use it as a full-coupled edge system.)

VMware's Tanzu also offers edge capabilities, building on top of vSphere using its remote office/branch office architecture to host cloud applications. VMware's approach to cloud at the edge provides managed infrastructure with a cloud-hosted control plane that lets you migrate workloads from other VSphere systems as well as other Kubernetes hosts.

As well as picking the hardware, you're also responsible for a certain level of infrastructure management with this approach. The public cloud service manages the workload, but typically you manage the platform accessed by the public cloud service. Also, different solutions give you different amounts of choice: IBM Cloud Satellite relies on RedHat Enterprise Linux and OpenShift, while Azure Arc works with any Cloud Native Computing Foundation (CNCF) certified Kubernetes cluster.

App-Coupled Cloud Edge

Many edge hardware devices aren't suitable for running cloud services, but they are ideal for running applications deployed, configured and managed by cloud services. IoT is the most obvious use case, with services like Azure IoT Edge and AWS GreenGrass. But projects like KubeEdge make the cloud the control plane for any containerized application running at the edge. You can also use KubeEdge for service-coupled, hardware-decoupled scenarios.

Azure IoT Edge and AWS Greengrass run cloud analytics and business logic in containers that run on edge devices (individually or as a pipeline for data processing, running cloud services that provide event-driven serverless functions, machine learning and data stream analytics alongside third-party services and your own code on your own hardware). An open source runtime manages the workloads and delivers communications to downstream leaf devices that can't run containers, or to optional cloud services that can build edge workloads using digital twins or configure and monitor multiple edge devices (like making the small tweaks

that improve machine learning model performance or detecting when devices are offline). That gives you real-time services like anomaly detection, industrial image recognition and video analytics or safety solutions like smart wristbands that enforce social distancing — using models and code built to run in the cloud but can now run at the edge with intermittent connectivity or even offline.

The lines between different types of cloud edge infrastructure aren't always clearly drawn. You can run AWS IoT Greengrass on Outposts servers, on Snowball Edge devices or on suitable PCEI infrastructure, depending on whether you want a more turnkey approach or more control. If you want scale and resilience for industrial applications, Azure IoT Edge services and modules can run on private cloud hardware like Nokia's Digital Automation Cloud or on Kubernetes infrastructure using KubeVirt.

If you want to integrate cloud services into your edge infrastructure rather than just adopting the somewhat prescriptive approaches of the hyperscale cloud providers, you can use tools like Akraino (which provides validated, production-ready blueprints including a CI/CD pipeline) and Public Cloud Edge Interface (PCEI) (which offers open APIs and orchestration functionality) to implement your own edge architectures for cloud services without starting from scratch.

Turning CDNs Into App Networks

One of the oldest forms of edge computing is still one of the most popular. While we often think of the edge as a place for compute, it's long been a way to manage content delivery. Content delivery networks like Fastly, Akamai and Cloudflare have provided a consistent cache layer for static content, protecting application servers from load spikes and DDoS attacks.

The rise of web development techniques like JAMstack, using JavaScript, APIs and markup to pre-render web frontends, allows application developers to preload user experiences in CDNs. Instead of hosting code on your own servers, you can push it directly to a CDN from a CI/CD platform, updating the entire edge with a single git pull request. Hyperscale cloud providers are using these techniques to enhance their own web application platforms, with implementations like Azure Static Web Apps taking advantage of Azure's own metro edge CDN.

We're also seeing an increasing level of intelligence in CDNs like Cloudflare, taking advantage of their own metro locations to provide local points of presence for elements of larger scale applications. Technologies like Cloudflare Workers provide programmable networking on the edge, preprocessing packets and delivering them to core applications based on edge rules — for example, handling geographic routing without requiring dedicated load balancers.

The same metro edge hardware is also now hosting WASM and WASI applications, either hosting WASM in JAMstack content or running WASI code directly. Advances like these have changed the role of the CDN significantly. Although it still functions as a way of using the edge to protect core resources, it's now also offloading functionality, both networking and code, adding a new layer to cloud native application architectures.

CHAPTER 4

Born in the Cloud, Works at the Edge

MARY BRANSCOMBE, SIMON BISSON

For many organizations, moving to the cloud is about agility: shifting from capex to opex and getting the opportunity to scale up applications when there's demand and turn them off when there isn't. But the changes you need to make to the processes of development and deployment of your solutions are inherently good practice: integrating development and operations; testing and securing code as you write and deploy it (rather than once you have it running); automating the pipeline that gets it into production and the environment it runs in.

DevOps and GitOps are key to using cloud services efficiently, but the concept of treating infrastructure as code also applies to managing edge infrastructure. Delivering consistent and reliable updates for software and infrastructure configuration as quickly as possible matters for any infrastructure, but it matters even more for remote sites with constrained resources, heterogeneous hardware, and critical workloads. Containers are a key part of delivering workloads at the edge, especially when they've been developed for the cloud, as are policies or feature flags that make sure workloads are placed on devices that have the right resources (compute, storage, or sensors).

While Kubernetes (or at least the Kubernetes API) has emerged as the mainstream option for container orchestration at the cloud and data center scale, orchestration is more complex at the edge. It calls for a lot more abstraction, because processes at the edge are more heterogeneous, with embedded operating systems, custom images, and, commonly, script-driven system configurations.

Kubernetes at the Edge

Cloud-native development models are a good way to think about edge workloads. They build on similar abstractions, using microservices and containers to encapsulate functional modules, which are then managed and deployed by declaratively managed orchestration tools. This code-first approach treats all aspects of an application the same way, whether it's code, networking, security, storage, or even virtual infrastructure.

Edge deployments are hard to manage using traditional tools. By their very nature, they're likely to be geographically distributed, working at a metro or even a cellular base station level. That means using a lot of automation, as the cost of manually managing services at this scale can be prohibitive. With automation being key, we need to consider using tools that integrate with both CI/CD platforms and DevOps environments to deliver applications and services.

Kubernetes' cloud-native heritage offers a lot to edge developers. It provides a sensible level of abstraction, hosting applications in familiar containers and working with Linux and Windows. There's no need for developers to understand what the underlying hardware is, all they need to know is that their code runs and that they have access to tools that enable them to add declarative security and networking rules to their applications.

Taking an automated approach makes Kubernetes a logical choice for managing edge applications, especially where you're working within tight resource constraints. Its default resource-based scaling model can be used to ensure that applications don't overload hardware limitations (as long as the appropriate safeguards are built in), while recent enhancements

have added support for event-based scaling that should help the management of IoT-based edge applications or other message-driven services.

Taking an automated approach makes Kubernetes a logical choice for managing edge applications, especially where you're working within tight resource constraints

It's important to note that Kubernetes isn't and shouldn't be your only edge management solution. It's designed to run on a base OS installation that supports both it and any hosted container application. You will need tooling and staff for remote management of that base OS and of the Kubernetes environment, and so it can perhaps best be thought of as a new layer in a DevOps environment. Meanwhile, the concept of DevOps in a large-scale, geographically distributed environment isn't yet fully operational, with a number of challenges that have yet to be addressed. Your edge installation will need hardware operations, including networking and storage, platform operations, for both the base OS and Kubernetes, and a final application operations layer for your edge workloads.

A typical installation will have a standard Linux to host the Kubernetes control plane, with nodes running containerd and the kubelet service. The controller and nodes, however, don't need to be physical instances, they can be a virtual infrastructure, running on a standard server or a small cluster.

Which Edge?

Your choice of edge platform will depend on what type of edge you're targeting. Systems running in metro edge data centers will have different resource constraints from systems at the telco edge or further out at the industrial IoT edge. If Kubernetes was used on aircraft, ships, and in vehicles, power and/or space could become bottlenecks in addition to other resource constraints, such as CPU, memory, and storage.

As a result, edge Kubernetes instances will need to be implemented using a different set of principles from core cloud-based systems. Those space and power constraints would also affect the available system resources; for instance, a set of x86 edge servers running in a metro data center will have different resources from a small Raspberry Pi-based cluster running on a factory shop floor — even though they will be managed the same way, using the same APIs, even possibly running the same containers.

In one possible scenario, the cloud acts as a central hub, handling application deployment for systems running on the edge. Where metro edge systems are available, these may run software that aggregates data from the device edge, where your applications and workloads run.

Challenges for Kubernetes at the Edge

We shouldn't take Kubernetes at the edge for granted. It's not a one-size-fits-all solution, and there are challenges when it comes to implementing it on edge hardware. An obvious one, of course, is its overhead; it does require its own resources, and this will affect your hardware choices. In practice Kubernetes is likely to require at least a single server rack, and more if you're building out a cluster.

A Variety of Edge-Flavored Kubernetes

While you can run vanilla Kubernetes at the edge, especially if you're using x86 rack servers and platforms like Azure Arc, you also have the option of edge-focused Kubernetes distributions like KubeEdge, K3s, or MicroK8s. These are all certified Kubernetes distributions, with full API support, while offering lower footprints and increased reliability. They can run on both x86 and Arm hardware and scale from metro and telco edge servers to small-form-factor hardware running on factory floors or out in agricultural sites. Supporting a wide range of use cases requires that these Kubernetes systems can work without an uninterrupted network connection, delivering data as and when needed, and offering Kubernetes API endpoints for application deployment and management using familiar tooling and processes.

KUBEEDGE

KubeEdge is a native edge computing framework that's designed to support application orchestration and metadata synchronization between the cloud and the edge. With workload migration a key use case for the telco edge, this model is likely to become increasingly important. However, despite support for synchronization, it's also possible to operate KubeEdge systems independently, allowing them to disconnect from the network where necessary.

While other edge-focused Kubernetes flavors are completely standalone, KubeEdge uses a host cloud Kubernetes system to run its core controllers and hub, which manage the system's components running at the edge: nodes, pods, and an integrated MQTT broker for working with device-generated events. There's also support for service bus connections to remote web APIs, allowing you to offload application UIs to external systems.

This approach will help with deployments to areas with limited bandwidth, allowing systems to be configured where connectivity is available, before being transported to target locations. As users are experimenting with Kubernetes in agriculture, on ships, or for disaster relief, there's likely to be significant demand for features like this. Usefully, KubeEdge supports the MQTT messaging protocol, making it easy to integrate with existing IoT hardware, supporting existing industrial IoT systems.

Like other edge-focused Kubernetes implementations, KubeEdge supports native Kube-API at the edge, so you can bring third-party plugins and services to your edge devices. This approach can help support experimental device support via Akri or Krustlets. With support for x86 and Arm hardware, KubeEdge is ready for most common edge systems.

K3S

Originally developed by Rancher Labs (now part of SUSE) and hosted by the CNCF as a sandbox project, K3s is a lightweight Kubernetes designed for use on edge hardware, especially on Arm-based systems, like those using Raspberry Pi Compute Modules. It supports both Arm7 and Arm64, so it is compatible with most Arm-based Linux releases. Arm7 support allows the use of older 32-bit hardware, allowing you to work with lightweight single-board computers for small deployments.

Unlike larger-scale Kubernetes installations, K3s is a single binary that requires less than 100 MB. It includes its own storage tools, based on SQLite, as well as support for key Kubernetes tooling, like the Helm package deployment environment. It also requires just a minimal set of dependencies, allowing you to tailor a minimal host environment keeping both resource requirements and possible attack surfaces to a minimum. A single binary also means that all the tools needed to control your applications are in one process, thus simplifying resource management.

Having a small-footprint Kubernetes flavor that supports all the familiar tooling, while aiming to be secure on install, is an important step on the road from the cloud to the edge. You get the same experience as with a hyperscale provider but sized appropriately for edge-class devices.

MICROK8S

Canonical's MicroK8s takes a similar approach to K3s, with a small-footprint environment that works as well on a single node as it does in a larger-scale cluster. The same MicroK8s tooling runs on developer systems as on production edge hardware, simplifying development and deployment as developers can, to a certain extent, test applications on their own machines. Of course, there's a wide variety of edge hardware, so their testing needs may still extend beyond their own computers. Setup is automated, requiring only a single command to get a cluster up and running. Once configured, systems can automatically update, systems' administration tasks can be kept to a minimum.

Designed to be fault tolerant, MicroK8s treats every node as a worker node, while at the same time offering a complete set of API services. Lose any node and the system switches to another, keeping your service running. This approach solves many edge reliability problems, allowing you to restart failed nodes while keeping applications running. As nodes restart, they are added to the pool, adding standby API servers. You do need at least three nodes in a cluster, running a high-availability version of SQLite, Dqlite, to handle service storage. Nodes run as a voting cluster, with API support on any node.

Build Once, Push Everywhere: DevOps at the Edge

Using cloud technologies at the edge in this way fits well with modern development methodologies, treating your edge as the build target for a CI/CD process and your artifact repositories as the trusted source of code running on those edge systems. The new generation of application bundles, based around containers, reduces the amount of work needed to adjust application code to the specific edge hardware it will be running on.

These new bundle models build on familiar cloud-native processes and tools, as well as on tried and tested web application deployment models. Technologies like CNAB, the Cloud Native Application Bundle, can package all the elements of a Kubernetes application and its related services, simplifying deployment to remote servers. Bundling resources into a CNAB package using a tool like Duffle lets you store the resulting package in a standard registry and then deploy it programmatically.

While CNAB was designed for multi-container applications, it's easy enough to use for smaller instances, too, giving us a way to manage signed trusted installers for widely distributed edge workloads.

Closely related is the WebAssembly Bindle distribution format. This owes a lot to web packaging tools like Yeoman, providing a model for packaging and storing application resources to be treated as labeled parcels, with simple text-based descriptions. Where a bindle can work well on the edge is its ability to be used to describe different packages for different targets, allowing you to tailor deployments so they can deliver appropriate versions of your code to different classes of edge device.

It's perhaps best to think of working with edge systems using cloud technologies as moving to an application-first operations model. Edge hardware is prepopulated with the appropriate services, and then application containers are delivered as and when needed.

It's perhaps best to think of working with edge systems using cloud technologies as moving to an application-first operations model. Edge hardware is prepopulated with the appropriate services, and then application containers are delivered as and when needed. This model separates application operations from edge operations, using declarative infrastructures and bare-metal provisioning to handle the deployment of new hardware. Network and storage can be managed in a similar manner, using service meshes, while new container management tooling, like podman, link your edge systems with OCI-compliant registries to handle application deployment, with a focus on container lifecycle management.

At a lower level, traditional devops approaches are key, using services like Foundries.io to deliver custom Linux images with versioning and updates, without the need for a Linux expert to be on hand.

Working With Edge Hardware

One advantage of using technologies like Kubernetes to power edge systems is its support for a wide variety of hardware. Small x86 systems are easy to find, with certified systems available for most common edge platforms. Devices can be rack mounted or delivered in ruggedized cases.

Arm hardware is also supported, adding options like rack-mounted systems based on server-class Arm processors, or dedicated systems based around single-board computers, like the Raspberry Pi. One useful form factor is the Raspberry Pi Compute Module, which provides either a socketed or DIMM connection to Pi hardware, thus saving space. Dedicated mounting systems, like the TuringPi, are ideal hosts for small K3s or MicroK8s clusters, with four CM devices and the necessary networking and support hardware.

At the same time, for the device edge, we're seeing the rise of platforms like FreeRTOS, AzureRTOS, and Zephyr. With hardware standardizing on a limited set of platforms, this will reduce the risk of fragmentation of OSes for MCUs in IoT devices, simplifying management and providing support for common driver models for technologies like Akri, allowing the direct integration of microcontroller-class devices in cloud-native environments.

More Than Plain Kubernetes

Kubernetes isn't the only option for running code on the edge. In some cases, it's a tool for orchestrating alternative application hosts, like the WebAssembly-based Krustlets, or working directly with device drivers in Akri.

Krustlets

WebAssembly offers edge platforms with a consistent development target, based on JavaScript engines. While initially intended to run binary apps in the browser, the development of the WebAssembly System Interface (WASI) allows application developers to target a simple virtual processor model that runs WebAssembly binaries. Code can be compiled and run on thin edge hardware without requiring the overhead of a container environment.

Krustlets are a prototype that explores how this approach can be used with Kubernetes, replacing the familiar containerized binaries with Rust code targeting WASI runtimes. Code can be managed and deployed from Kubernetes using its familiar tooling and declarative configurations. By tailoring WASI runtimes to edge hardware, especially single-board computers and other low-power, low-cost solutions, Krustlets allow the same orchestration platform to manage microcontroller applications as well as your hubs and gateways.

The technology is still experimental, but it shows a lot of promise. Embedding WASI runtimes as device firmware gives us a consistent deployment target for centrally developed code. Using memory-safe languages like Rust also reduces risk, ensuring that edge code should be safe to run on a wide range of different hardware platforms. All that's necessary is a WASI runtime and a network connection.

Scheduling Edge Devices with Akri

With industrial IoT being a key edge driver, there's a need to manage standard devices like cameras or sensors with edge services. Microcontroller-class devices like these are considered leaf devices, delivering data into our applications. Akri builds on the existing Kubernetes

device plugin model to take it beyond its basic hardware capabilities to also support common device drivers. As it's built on Kubernetes standards it's compatible with most common edge Kubernetes distributions, including K3s and MicroK8s.

Using Akri, it's possible to determine what nodes have what capabilities, and then to schedule appropriate workloads. A camera-based service will be tasked to run on camera-capable hardware, ensuring that code is allocated appropriately, and helping you design and deploy edge hardware that works with your planned applications. Akri is designed to use open protocols to discover hardware, such as the Open Network Video Interface Forum protocol, so you're not limited to vendor-specific devices.

Event-Based Scaling at the Edge

Choosing to implement an application operations approach to the edge allows you to offer what is, at heart, a serverless approach. Once you have an applications platform in place, simply deploy your containers and your code should run, orchestrated by your Kubernetes or a similar distribution. While Kubernetes scaling is based on a resource model, tools like KEDA add support for event-based scaling, a model that's more appropriate for many edge scenarios. Using KEDA systems can be configured to add new nodes as events are received, then disposing of them once they have run.

Any environment that can host containers and that can be automated is suitable, though some form of workload orchestration is preferred. Linux supports remote bare-metal installations, and thin distributions like Flatcar are ideal for hosting edge applications.

Running Alternative Cloud Services at the Edge

Kubernetes isn't your only option for delivering cloud services to edge platforms. Any environment that can host containers and that can be automated is suitable, though some form of workload orchestration is preferred. Linux supports remote bare-metal installations, and thin distributions like Flatcar are ideal for hosting edge applications.

You can use this approach to work with infrastructure as code tooling to manage edge hardware as well as your code. HashiCorp's Nomad helps manage workloads in both the cloud and on the edge, deploying containers and managing bare-metal hardware. You can use it as a control plane for your edge deployments, working with other HashiCorp tooling; for example, managing networking across the cloud and the edge with Consul and using Terraform to ensure consistent environments across your entire estate. You can also reduce security risks by using Vault to manage your secrets for you, ensuring that they're not hard coded into your edge apps, where it would be relatively easy for malicious actors to get access to your hardware.

You can experiment with Bindles using the currently experimental Hippo service, built on the tip of nomad. This uses the experimental WAGI environment (mixing WebAssembly with the original Common Gateway Interface web application environment). Hippo is a standalone low-impact PaaS that provides a way of deploying and managing lightweight applications; as such, it shows a lot of promise as a way of delivering an extremely easy-to-manage edge environment for WebAssembly-based applications.

Similarly, Microsoft's Azure Sphere platform uses its own Linux to deliver applications from cloud repositories to remote single-board devices. By building on Microsoft's Pluton security processor, Sphere hardware has the added advantage of offering an end-to-end secure software supply chain, from development PC to edge device. This approach works surprisingly well, using Sphere to build custom hardware as well as working with its single-board computer partners, and with the option of using Sphere-based gateways to bring older SCADA hardware into modern edge environments.

At a larger scale, Azure Arc and similar services from Google and Amazon allow you to manage edge infrastructure from cloud tools, using familiar resource models to deliver applications to managed edge hardware. This approach is perhaps best for larger systems, where a handful of racks of x86 hardware are used to host cloud-native applications along with a supporting virtual infrastructure.

CHAPTER 5

The New Hubs

JABEZ TAN

The development of the world's internet infrastructure landscape has historically taken its cue from other more established industries, such as financial services and logistics. These industries have been known to generate economies of scale through centralizing their core ecosystem within a select few global markets. For financial services, the London, Singapore, Hong Kong, Tokyo, Frankfurt, and New York markets have long been seen as the world's top financial and economic hubs. Those markets, therefore, have also been natural logistics, import, and export hubs, given how the flow of capital impacts industries that are especially capital intensive.

The data center and network infrastructure sectors are no exception to this trend, and global tier 1 markets, such as London, Singapore, Hong Kong, and Tokyo, have grown to become the internet's first layer of aggregation points where international submarine cable systems and data center infrastructure naturally gravitated. This has created a centralization effect in order to build upon the economies of scale and ecosystem benefits of being able to access a variety of other locations and markets from a single location. To bring this closer to home, think about the compounding effect of retail outlet malls versus single store locations. Would a consumer prefer to drive 20 minutes to access a single retail store? Or would they rather drive a little bit longer to a mall where they can access many different stores and brands all under one roof? In that same vein, centralization is a crucial step toward generating strong ecosystems and developing network effects for the end user.

As the internet infrastructure landscape continues to mature, and has experienced both sustained and accelerating levels of adoption and usage on a global scale, the benefits of centralization now come with certain glaring weaknesses regarding resiliency, redundancy, performance, and regulation. These factors have driven a new wave of investments and development activity outside of the traditional global tier 1 markets that are now experiencing elevated levels of land and power constraints given the high level of population density residing within these economic and infrastructure hubs.

Decentralization of Hyperscale Demand

Hyperscale clouds started off by centralizing their footprints and deployments in a select group of core global hubs and are now looking to move to a more distributed deployment architecture. Singapore and Hong Kong are prime examples of this initial strategy in the Asia Pacific region given their position as connectivity hubs with plenty of submarine cables concentrated in those locations. Structure Research projects that global internet infrastructure will continue to shift to a more distributed and decentralized model. More submarine cables will be constructed to connect emerging markets to core markets. This will lead eventually to a shift in data center builds and deployments to more localized and in-country architectures. Hyperscale clouds will continue to fill gaps in their infrastructure map, especially in markets with large population densities like Indonesia, India, South Korea, and Japan. Hyperscale clouds have also shown an appetite and ability to build their own data centers in APAC and may continue to do so in the future within strategic locations. Although it may initially appear that demand for colocation data center services would decrease if hyperscalers choose to build their own data centers, hyperscalers are notorious for underestimating their data center capacity needs, which typically leads to a meaningful amount of overflow demand where colocation providers can potentially benefit.

New Opportunities for Colocation Companies

Hyperscalers moving more convincingly to the edge will create more opportunity for colocation providers. Hyperscale cloud infrastructure is, by nature, highly centralized to take advantage of economies of scale. That is about to change as public cloud adoption continues to accelerate across both mature and emerging markets, with customers demanding increased performance—especially in markets that do not yet have in-country public cloud infrastructure.

Hyperscale clouds are not likely to self-build in smaller increments at the edge, because that does not scale efficiently. But what hyperscalers have done is build different scenarios through the deployment of converged infrastructure appliances.

Hyperscale clouds are not likely to self-build in smaller increments at the edge, because that does not scale efficiently. But what hyperscalers have done is build different scenarios through the deployment of converged infrastructure appliances. Amazon Web Services (AWS), for example, launched Local Zones in 2019, which is basically an edge data center (via colocation) with Outposts appliances set up and connected back to a core cloud region. Microsoft launched Azure Edge Zones, a small increment of compute located in its edge points of presence locations (PoPs) and is now moving into wireless carrier PoPs. In a matter of just two years, AWS expanded its Local Zone to 14 locations—all in the US. In 2022, AWS confirmed plans to take Local Zones global, and they will add more than 30 locations. Some of the countries confirmed include Argentina, Australia, Austria, Belgium, Brazil, Canada, Chile, Colombia, Czech Republic, Denmark, Finland, Germany, Greece, India, Kenya, Netherlands, Norway, Philippines, Poland, Portugal, and South Africa. Many of these countries do not have in-country cloud regions, and Local Zones will provide a way for hyperscale cloud providers like AWS to deploy local cloud infrastructure to serve local enterprises and compliance-sensitive verticals such as government agencies. These Local Zones deployments will tether back to AWS' core cloud regions.

But some large countries, like India, Australia, Brazil, and Canada, already have core cloud infrastructure regions deployed. In these markets, Local Zones are moving to far-flung edge locations that either have clusters of end users or are significant markets that do not quite have the critical mass to justify a full core cloud region deployment. Canada is an interesting case. AWS chose to build its core region in Western Canada in Calgary, Alberta. It is a good bet that Vancouver will be the Canadian Local Zones location referenced. On the telco edge, AWS referenced Wavelength but did not detail specific expansion plans. That current footprint is still largely concentrated in the US. Finally, AWS sort of relaunched the 1U and 2U Outposts products that were already mentioned a year ago. These are compute increments that will end up at the "far" or "rugged" edge and in places where connectivity is somewhat unstable. The AWS

footprint continues to push out to the edge, and its strategy is becoming clearer. Cloud regions are going to be built strategically where a market has scale and also has proximity to other sizable clusters in neighboring countries. To the extent possible, AWS will try to "dual-serve." For latency-sensitive workloads and data sovereignty requirements in markets with no core, it will set up edge nodes. And to reach an on-premises data center or far edge locations, it will encourage end users to deploy Outposts and connect back to the core. And that is at the heart of the wider strategy. Different infrastructure increments will have varied feature and service sets available. But it will all connect back to the core to access AWS' full range of capabilities.

North America

The North America region is home to its own set of tier 1 markets that include Northern Virginia (NoVa), Silicon Valley, Dallas, New York, and Chicago. And since the US has the most mature data center market in the world, it was also the first region to start the decentralization process. A thriving big tech company community in the Silicon Valley market created supply constraints for data center infrastructure development. This drove demand into the first wave of alternative markets that began around four to five years ago with Phoenix and to some degree in Las Vegas. Through 2021 and into 2022, Structure Research has seen a second wave of alternative markets such as Hillsboro, Oregon, and Salt Lake City, Utah, emerging as viable locations to address the data center infrastructure demands of the US West Coast region that had been historically centered around Seattle, the San Francisco Bay Area, and Southern California.

In the US East Coast, the NoVA market has become the aggregation point for the vast majority of the data center infrastructure demand in the region due to the New York market being inherently space constrained from the very early days.

In the US East Coast, the NoVA market has become the aggregation point for the vast majority of the data center infrastructure demand in the region due to the New York market being inherently space constrained from the very early days. It has taken longer for the US East Coast to decentralize compared to the US West Coast since there was not the same level of land constraints in NoVA relative to the Silicon Valley market. This is, however, beginning to change as land prices in the NoVA market have risen to materially high levels. This has in turn driven a first wave of alternative markets that include Atlanta, Georgia, and Columbus, Ohio, on the US side, as well as the Toronto and Montreal markets in Canada.

Columbus is a unique example of a strategic hyperscale market where AWS deployed its core cloud node known as the US-East-2 region. Google has also recently acquired land in Ohio to do the same for its cloud platform. Both AWS and Google have the capability and scale to develop their own data centers, and Ohio is ideal given its central location in the Midwest and the fact that both land and power are abundant and affordable. But that is not stopping other colocation data center operators from entering the market and is, in fact, creating opportunities for data center providers pursuing the spin-off effect from hyperscale presence to go after overflow scenarios, redundancy requirements, and connectivity ecosystems for network and cloud on-ramps.

Europe

In Europe, most of the data center development and demand activity has historically centered around the FLAP + D markets (Frankfurt, London, Amsterdam, Paris, and Dublin). That is beginning to change as other emerging markets continue to develop well across Europe driven by demand for new cloud regions (especially from Microsoft).

Madrid stood out in 2021. The historically underdeveloped market is seeing a wave of new builds to meet projected hyperscale demand as public cloud adoption takes off. Existing operators and a raft of new entrants (both established operators and newcomers) are vying for a piece of the action, but we don't expect this will lead to oversupply. Structure Research is forecasting that the Madrid market will effectively quadruple in the 2022-2026 period, albeit from a low base. The density of networks and connectivity that is building around Iberia is elevating the strategic importance of both Madrid and Barcelona, and the latter is in a strong position to provide a cable landing alternative to Marseille.

Microsoft, Google, and Oracle Cloud all opened cloud regions in Zurich in 2019, but this did not lead to the exponential growth that has been enjoyed by Madrid. However, activity ramped up significantly in 2021 as existing data center providers were all expanding capacity and attributing the demand to hyperscale activity. AWS plans to launch a Zurich region in 2022. It usually prefers to self-build but could turn to third-party operators for one or more of what we expect will be three availability zones (AZs).

Interest in Berlin is building, especially from hyperscalers, as power constraints in Frankfurt and latency requirements increase in importance. Microsoft opened a cloud region in Berlin in 2019, and Google is following suit, although the timing has not been disclosed.

Some notable data center acquisition and development activity occurred in the Nordics in 2021 as there is no denying the conduciveness of this particular region for data centers given its climate for free air cooling, abundance of land, and access to renewable energy sources.

Middle East

In the Middle East, the United Arab Emirates (UAE) continues to be the region's most vibrant data center market, and activity continued apace in 2021 from both local and international data center operators driven by new submarine cable investments such as the 2Africa cable system. On the hyperscale side, AWS plans to open a UAE region in the first half of 2022, while Oracle Cloud opened one in Abu Dhabi. Both are second cloud regions for the Gulf area: AWS's other region is in Bahrain, while Oracle's is also in UAE, in Dubai. The two UAE regions for Oracle Cloud are to meet local regulatory requirements set by the Dubai Electronic Security Center (DESC). It is following a similar path in Saudi Arabia, where it is set to open a second region in 2H22 to meet Saudi Arabian Monetary Authority (SAMA) rules. Other hyperscalers may well follow suit.

Israel hit the headlines in May 2021 when the government awarded contracts to public cloud providers, which required a minimum of two in-country data centers at least 25 km apart.

Microsoft and Oracle had already announced cloud regions (Microsoft back in early 2020 and Oracle earlier in 2021), but they missed out on the contracts. The bulk (70%) went to AWS, and the remainder went to Google Cloud, with services set to start in 2023, but Microsoft and Oracle are both pressing ahead with regions in Israel. There is a combination of self-building, build-to-suit (with a real estate development angle), and wholesale colocation to serve the hyperscalers.

2021 has been a year of significant activity on the African continent as the region has finally started to expand beyond South Africa and the competitive environment is heating up in markets such as Lagos, Nigeria, and Nairobi, Kenya. South Africa (primarily Johannesburg) will continue to be the largest market in the continent.

Latin America

In the Latin America (LATAM) region, most data center development and hyperscale cloud deployment activity have been centered around São Paulo, Brazil. Multiple operators are starting to build out pan-LATAM footprints, with Santiago, Chile, and Queretaro, Mexico, being top of mind to meet hyperscale demand.

Asia Pacific

Finally, in Asia Pacific (APAC) 2021 marked the shift in the region's tier 1 market rankings in terms of data center investment and development activity. Singapore and Hong Kong have long been considered the top data center markets in APAC (excluding China). This changed in 2020-2021 with the Singapore government's moratorium restricting new data center builds within its borders, while Hong Kong has seen a notable decline in expansion pipeline and leasing activity from US-based hyperscale cloud platforms due to the increased political turbulence. These factors have catapulted the other two tier 1 data center markets in APAC—Japan and Australia—to become the center of a wave of investment and development activity. Structure Research recently published granular data center supply and demand studies on both of these markets, which shows the staggering influx of new builds in the pipeline. Most of the activity has centered around Tokyo and Sydney, although, due to the size of both countries, secondary markets such as Osaka and Melbourne have seen an acceleration in both cloud and data center demand.

The Next APAC Growth Markets

While APAC core hubs like Singapore, Hong Kong, Sydney, Tokyo, Shanghai, Beijing, and Guangzhou will continue to see stable levels of hyperscale demand, another set of emerging markets in APAC will see accelerated hyperscaler-driven growth and move into the tier 1 market status conversation. These markets are India (particularly Mumbai, Hyderabad, Delhi, and Chennai), Jakarta, Indonesia, Manila, Philippines, and Seoul, South Korea. The Southeast Asia region has often been underrated and overlooked for hyperscale demand, although that is about to change with the current moratorium in Singapore driving new supply clusters in adjacent, proximate markets that include Johor, Malaysia, and Batam, Indonesia.

Elsewhere in Southeast Asia, markets such as Kuala Lumpur, Manila, and Bangkok are seeing the early signs of hyperscale demand starting to materialize in a meaningful way. Taiwan continues to be an interesting market that many data center operators evaluated in 2021, though it appears that it is currently not high on most providers' expansion priority list. The bull case for Taiwan continues to be around three main aspects: the perceived safety in Taiwan's geographic location not sharing a terrestrial border with mainland China, coupled with recent sizable investments from US-based tech companies that include Google and Microsoft; Taiwan's central location in the APAC region positioning it as an ideal location for new submarine cable deployments; and Taiwan being a hub for semiconductor manufacturing, and an established data center region, with Google now on its third data center self-build in the country, seen as validation of Taiwan's future growth prospects.

Expect an Explosion of Activity

The global expansion of internet infrastructure is happening at a rapid pace, and the pandemic was at the center of a confluence of events that are accelerating the trends already in pace. The next decade will see an explosion of activity as a result. Core markets will continue to expand, and decentralization will create new infrastructure locations at the edge and in strategic and difficult-to-access places. Moving to the edge is fundamentally about getting infrastructure closer to end users, whether that be in new and emerging markets or wherever critical masses of end users cluster. The underlying infrastructure may be provided by a hyperscale platform or an independent operator, and it will have to come in various increments and form factors. But data center infrastructure, often in the form of colocation, is the foundation upon which everything will be built.

ABOUT THE AUTHOR

JABEZ TAN is the Head of Research at Structure Research, an independent industry research and consulting firm devoted to the cloud and data center infrastructure services markets with a specialization in the hyperscale value chain. He leads Structure's coverage of the Asia Pacific region and building of the firm's proprietary market-share data and deep-dive supply and demand analysis on a growing set of global data center markets. Jabez has served as the lead analyst and subject matter expert for more than $2 billion in data center M&A transactions providing in-depth due diligence and strategic advisory. He is a regular keynote speaker at industry events who gets regularly quoted as an expert in industry press. Jabez has a Bachelors of Science in Aerospace Engineering from the University of Maryland, College Park.

SPECIAL SECTION

Essays

LF EDGE

Essay

STEWART MCGRATH

CEO, Section

What Will Edge Computing Unlock?

Cloud computing was the third major shift in modern compute paradigms following the mainframe and then client-server models. Cloud computing has delivered many benefits to users, including lower operational burdens, greater scalability, access to advanced system resources and capabilities, and standardized deployment processes. Among the biggest impacts of cloud computing, I would argue, is the accelerated pace of innovation resulting from these technological benefits.

The pace of technological innovation in the cloud era has been phenomenal. Although we cannot ascribe all of this to cloud computing, it has clearly been a fundamental contributor to the rate at which new technology is being developed, deployed and operationalized at scale. Small, medium and large organizations can spin up and tear down massive amounts of compute and easily access networking and software systems to support their innovation and development efforts in a way that was unprecedented in the pre-cloud era.

In early 2021, nearly 1,500 major data centers were distributed across the United States; many of these data centers were powering “Cloud Compute” capabilities. With that many data centers distributed across just one country, why does edge represent any improvement in compute possibilities?

Central to the concept of edge compute is the notion of not only running workloads at locations physically closer to a group of end users but running closer to all end users. This requires a distributed computing solution rather than just distributed computers. Cloud computing data centers represent many distributed computers, but they do not present a distributed computing layer. When we add the distributed computers available in Telco, ISPs and the developing 5G infrastructure — and even on-prem compute capacity — we have many possible targets for application logic and data. When we evolve these distributed computers into a cohesive distributed computing (or Edge) solution, we can then change the innovation game once more.

Provided we give developers the same benefits of cloud (ease of operational access and control, scale, standardization of deployment and application dev lifecycle, etc.) along with additional benefits of edge (reduced latency, reduced data backhaul, locality compliance, etc.), then we will have an Internet with an innovation process that is both fast and significantly more flexible.

Cloud computing brought a previously unimagined speed of innovation. Edge is bringing an innovation experience unencumbered by the limitations of a centralized cloud compute model.

Cloud computing brought a previously unimagined speed of innovation. Edge is bringing an innovation experience unencumbered by the limitations of a centralized cloud compute model.

Many use cases have been mapped out already by this unencumbered edge innovation, including autonomous vehicles, smart buildings, homes and cities, remote healthcare, retail and supply chain management, banking, manufacturing and industrial IoT, improved energy monitoring and consumption efficiencies, and smart agriculture. All of these represent opportunities for a more productive, healthier and more sustainable planet.

Although these are undoubtedly incredibly exciting opportunities to improve the planet for the entire human population, they are based on what we have thus far imagined is possible with an unencumbered innovation process powered by edge. What may be more exciting is that which has not yet been imagined.

Cloud critically unlocked many innovation aspects simply by democratizing access to vast amounts of powerful computing resources. As we democratize access for application developers to true distributed computing systems (e.g., edge), rather than simply presenting them with distributed computers, we will be able to help them innovate without being held back by the latency, networking and compliance limitations of the centralized cloud.

What could be more valuable than another impending fundamental shift in our tech industry's ability and scope to innovate?

ABOUT THE AUTHOR

STEWART MCGRATH is the CEO and Co-founder of Section, an edge hosting platform that helps modern engineers deliver better web applications. With extensive experience leading companies in the technology space, Stewart's passion for building teams focused on bringing technologies to market drove him to co-found and lead Section. As applications continue to evolve and end user demands for performance, security, and functionality increase, he envisions a world where developers are unencumbered by infrastructure and a better internet is powered by the edge.

LF EDGE

Essay

ILDIKÓ VÁNCSA

Sr. Manager Community & Ecosystem,
Open Infrastructure Foundation

The Hidden Edge

Edge computing has been in the spotlight for years, circulating in a seemingly endless hype cycle. It is no wonder that, when you ask someone about exciting new use cases, they list things like self-driving cars, drones, augmented and virtual reality experiences and other examples more reminiscent of a sci-fi movie than everyday life.

Digital technology is going through a very fast-paced evolution that has a big impact on not just how we live but on the course of humanity as well. Depending on who you talk to, the edge might not even be on this planet! However, what is most exciting about the innovations that edge computing triggers is that we can find solutions to pressing issues that affect our lives today. Just take a look at healthcare, agriculture or even seemingly simple things like waste management.

By building and deploying smart sensors, you can extend and scale tasks that only humans could do in the past. To go a step further, you can also optimize them. As an example, soon we will be able to fully automate taking care of livestock while ensuring a better environment and living conditions for the animals at the same time. Or, municipalities will be able to optimize the placement of trash cans and garbage collection routes for cleaner cities and rural areas alike.

Use cases like these are practical and impactful, and solutions to the technology challenges are within reach. For instance, 5G networks and services are being rolled out globally while I write this article — a substantial change from a few years ago. At the same time, edge infrastructures still need to overcome certain challenges to achieve long-term sustainability. Chief among them is scale.

The scale of end-to-end infrastructures that serve edge use cases is massive and is expected to change dynamically, including further growth. The combination of scale and flexibility means these systems will be complex, despite the best efforts in this area to reduce that complexity. Still, the real challenge is geographical distribution, as the components of these solutions are not located within the walls of the same data center anymore.

Sustainable products and solutions need to be designed to handle complexity and provide automation during the whole lifecycle of the system, from deployment to operations. Just a simple task, like patching a system, can get very hard when the number of edge sites grows to thousands and more and when they are located at the end of often unreliable network connections. This gets even more pressing when you try to carry out a security patch, which is becoming the norm rather than the exception, as we keep on relying on digital technology for even the simplest things in our lives.

To support the growth of heterogeneous edge environments, interoperability is a requirement that is in the spotlight once again.

To support the growth of heterogeneous edge environments, interoperability is a requirement that is in the spotlight once again. Various hardware and software building blocks have to fit together as the trends point toward multi-vendor solutions to build out the end-to-end infrastructures to run applications, from yet another source, on top.

As more edge computing use cases are deployed in production, we are reaching the phase where open source communities can play an important role in overcoming the aforementioned challenges. Various groups in the ecosystem cover every layer from the underlying hardware through software infrastructure up to the applications. As each component and project reaches a state of maturity, they can all work together to overcome integration and interoperability challenges.

Open source infrastructure platforms and building blocks — such as the Linux kernel, OpenStack, Kubernetes, StarlingX and Ceph, just to mention a few — provide you with a set of de facto standard interfaces that support the organic growth of edge infrastructures by providing a clear set of interfaces between the different layers of the stack.

The next step for these open source edge infrastructure building blocks is to have more companies and organizations get involved in existing projects and also contribute openly where they see gaps. That way, they can share their experiences about production deployments, as well as systems, that they are about to roll out, and they can work together with the communities to identify and fill the gaps to keep up the fast pace of innovation!

ABOUT THE AUTHOR

ILDIKÓ VÁNCSA works for the Open Infrastructure Foundation as Senior Manager, Community & Ecosystem. As part of her role, she is the community manager for the StarlingX open source edge cloud project. Her focus areas within the foundation also include telecommunications, NFV, and edge computing. She is also co-leader of the OpenInfra Edge Computing Group and active participant in open source communities such as Anuket, State of the Edge, and more.

LF EDGE

Essay

JON LIN

EVP & GM, Data Center Services
Equinix

Connecting the Underconnected

Connecting the underconnected is a topic that is near and dear to my heart, in large part because it is a critical component of the conversation around diversity, equity, and inclusion. As we think about the state of the edge, it's important that we look beyond technology and business trends to consider how we can impact the digital divide that persists in our society.

Access to connection and quality and performance of that connection are fundamentally important to how we live our day-to-day lives. It is a base requirement to progress in our society and it is an important civil rights issue. The broadband element of the digital divide disproportionately affects low-income households and rural communities, who often lack the infrastructure and resources needed to connect to the internet.

According to the White House, there are an estimated 30 million Americans living in areas with little to no broadband infrastructure (in 2015 the United States FCC defined broadband as any connection with a download speed of at least 25 Mbit/s and an upload speed of at least 3 Mbit/s), and even more with internet speeds suboptimal for standard internet use. Additionally, according to a Pew Research survey 24 percent of adults who live in rural areas are more likely to say access to high-speed internet is a major problem in their local community: compared with 13 percent of urban adults and 9 percent of suburban adults.

These statistics represent a gap in our society that often gets overlooked.

Most people, as they go about their days—clicking away, ordering pantry staples or using their smart devices—don't think about edge computing or last-mile connectivity and how all the pieces connect to enable them to live their lives in such a way. As the State of the Edge 2021 report discussed, there's an enormous amount of innovation that goes into each of these experiences, from specialty computing devices and cloud native software to "hard" infrastructure like edge data centers and undersea cables.

But the truth is that a significant number of people in the U.S. (and of course the world at large) don't have access to the most basic part of the equation: a quality, affordable internet connection. This prevents people from being able to work from home, do their homework, apply for jobs, and access critical health, financial, and government services. It also puts people in unserved or underserved areas at an incredible disadvantage and is an impediment to uplift.

Despite these headwinds, I am optimistic for the future. Right now is a pivotal time for broadband in the U.S. The bipartisan Infrastructure Investment and

Edge computing is a new frontier in our digital world, and as such it offers an opportunity to push for digital inclusion.

Jobs Act (IIJA) includes $65 billion for expanding broadband internet access and adoption. While this funding offers great opportunity, it is one that we must use wisely. Deployment of IIJA funds should incorporate the following:

- Consider the complete range of options, taking an unbiased approach towards technology choices when building out broadband infrastructure.
- Ensure healthy competition and pricing, which will make broadband access more affordable.
- Encourage and make it more affordable and simpler for people living in areas with broadband gaps to apply for stipends.
- Advocate for more accurate and equitable digital mapping, to guarantee underserved areas are properly identified.
- Continue to drive awareness that the disparity in broadband access is far more nuanced than a geographical issue.

Edge computing is a new frontier in our digital world, and as such it offers an opportunity to push for digital inclusion. By building out edge computing infrastructure, we are setting up for a future with tangible benefits for society, including new job opportunities and innovation across a wide variety of industries. Keeping equity in mind during this buildout is critical.

A Call to Action

So when thinking about what's next in innovation, I implore industry professionals to remember and help spread awareness that access to affordable, high-performing broadband is imperative to digital inclusion in today's society. Having that mindset at the individual level will impact the way we go about our daily work, creating the technologies, infrastructure, and services that can benefit all people in our society.

At Equinix, we are always looking for ways that we can come together as an industry to work on digital inclusion. One of the ways we've done that is by joining the CEO Action for Racial Equity, a business-led initiative focused on bringing business, communities, and policies together to drive change. To find out more about how to get involved with CEO Action for Racial Equity visit here today.

LF EDGE

Essay

MATT TRIFIRO

Co-Founder,
State of the Edge

Building a Grid at the Edge

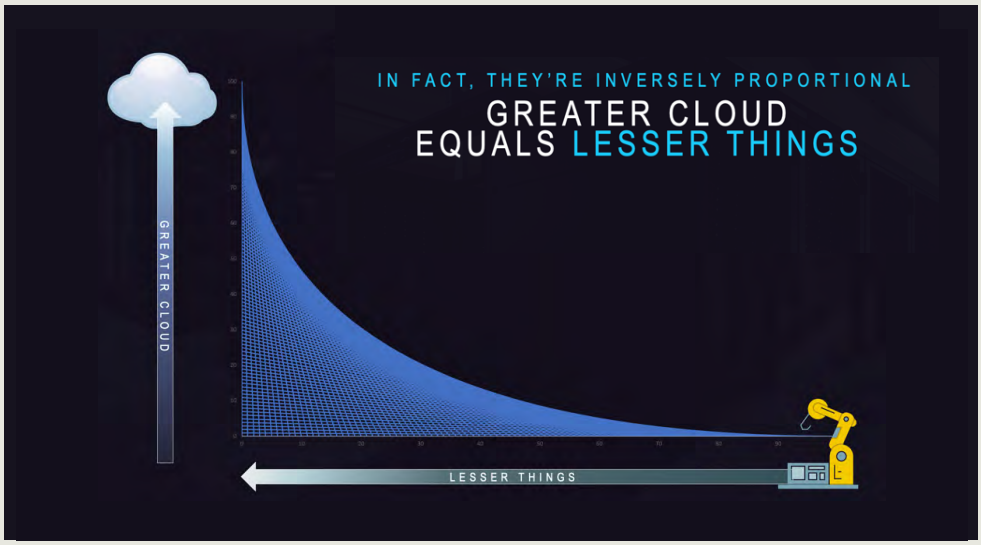
Everything in the World Is a Thing

For all the excitement of the cloud, what we really care about are the "things," those elements that occupy our physical world, like trees, buildings, cars, refrigerators, phones and bridges. The things in our lives have little in common with the cloud's abstracted virtual machines in far-off data centers.

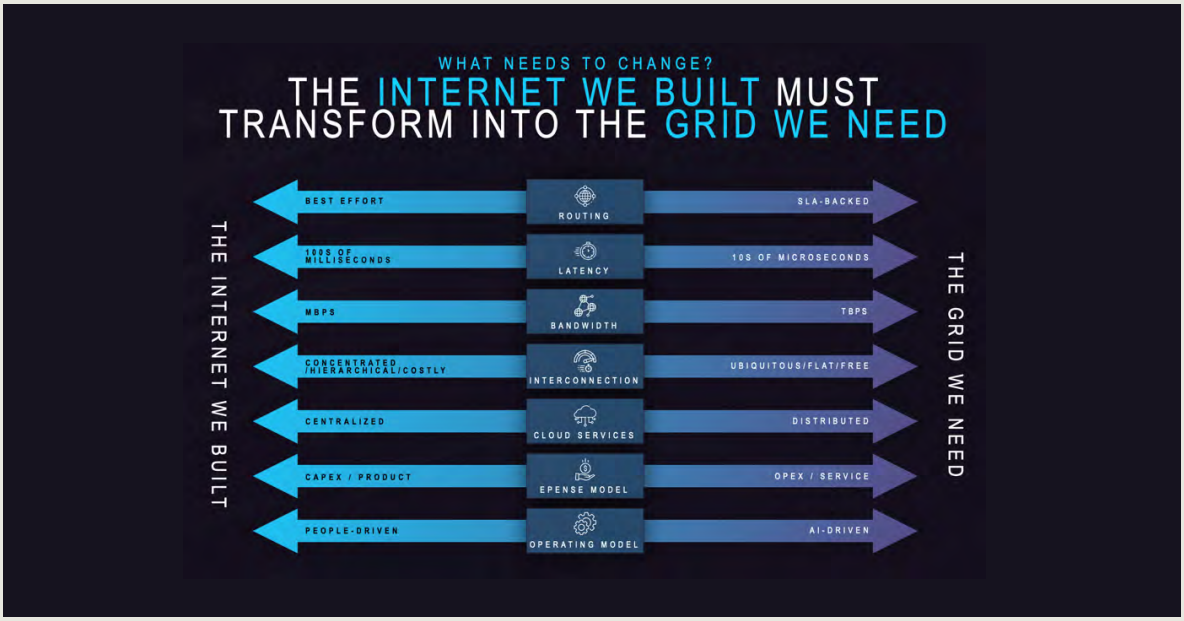
We often toss around phrases like "Internet of Things" and "Cloud Robotics," mistaking the metaphor for reality. The Internet is not the thing, just as the map is not the territory and the cloud is not the robot. If a thing is "here" — like the phone I hold in my hand — then, by definition, the cloud is "somewhere else." The Internet is merely the conduit through which we connect one to the other.

Greater Cloud Means Lesser Things

In today's world, the more computation you place in the cloud, the less computation you have on the thing. It's a mathematical certainty; they're inversely proportional. Greater cloud means lesser things.

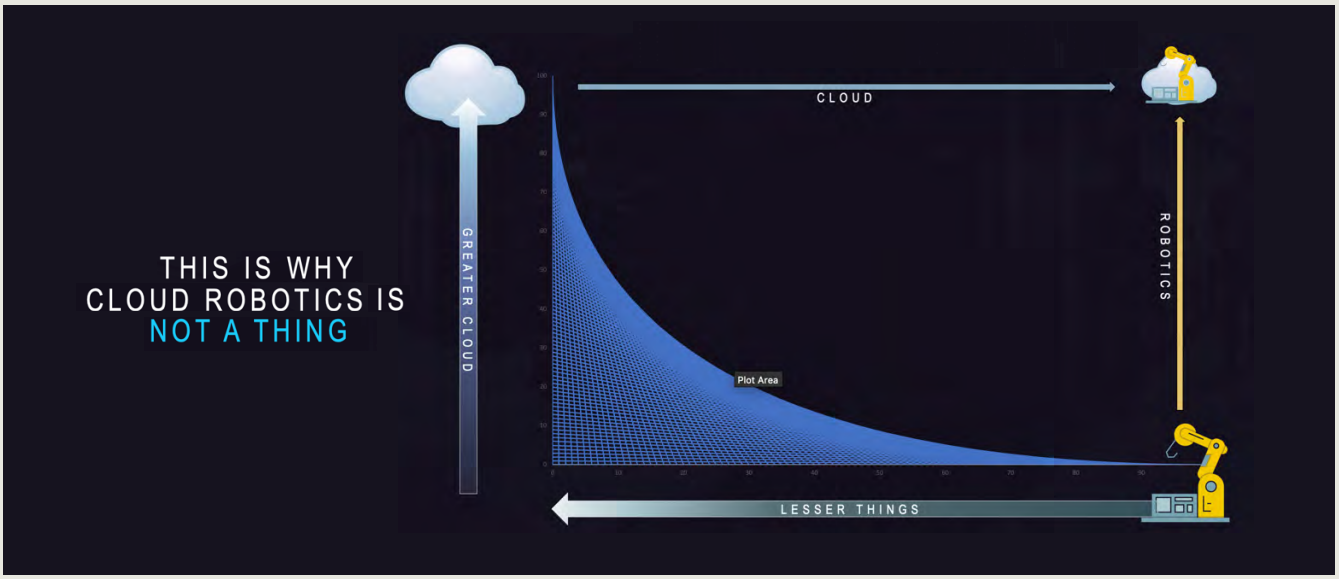


It's easy to get caught up in the excitement of grandiloquent futurists bombasting about subjects like cloud robotics. Surely running a robotic lathe from the Internet can't be that much different than playing a game on my mobile phone. But it is. It's fundamentally different.



Today's Internet was built for humans consuming content, not machines talking to machines. It operates on a "best effort" basis, and transactions can take hundreds of milliseconds, which is orders of magnitude too slow and imprecise for industrial robots.

Existing cloud services are highly centralized, concentrated and hierarchical. This architecture has provided great economies of scale, but at the price of performance and proximity — data must travel hundreds or thousands of miles before it can be processed and acted upon. This is why "cloud robotics" is not a thing. Not Yet.



The Edge Is a Location, Not a Solution

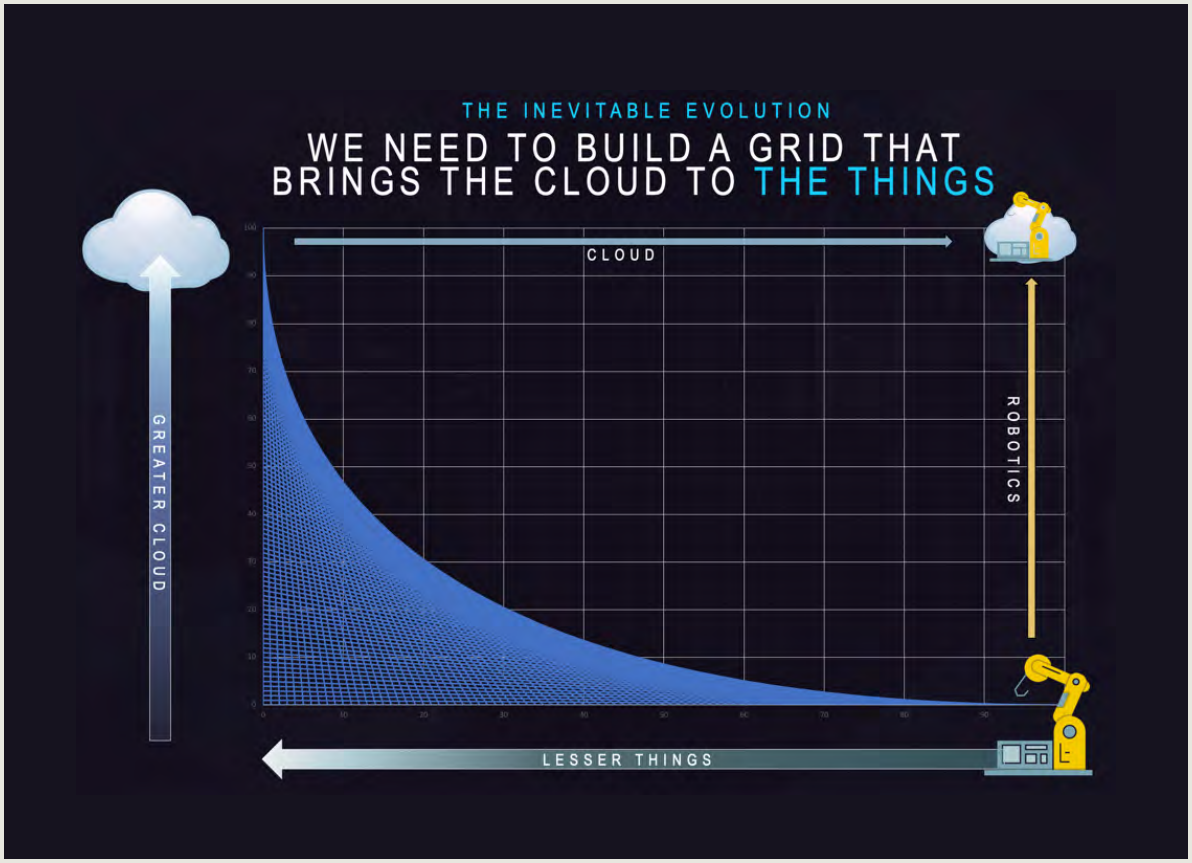
When we talk about edge — when we're not using it as a shibboleth or marketing buzzword — what we really mean is "close to the thing." Edge computing at an oil derrick means computing that is "close to or on the oil derrick." Edge computing at a factory means computing that is "close to or on the factory floor." For a great many edge use cases, "edge" is mostly just a synonym for "on-prem."

Yet, intuitively, we know there must be more to it. We know there must be something important about the concept of edge that isn't being captured by the legacy of on-prem. Edge can't just be a rhetorical device to spruce up what we've already been doing. And it's not.

The difference between edge and on-prem is cloud — and not just any cloud, but every cloud. Public cloud, private cloud, hybrid cloud, multi-cloud. Cloud gives us the ability to scale, the ability to virtualize, the ability to provision and deprovision on-demand and the ability to automate across a shared infrastructure with consumption-based pricing. Most of these attributes don't exist on-prem; or, if they do, they exist only in an ersatz fashion.

Bringing the Cloud to the Edge (and the Things)

The purpose of The Grid in its modern incarnation is to bring the cloud to the things, and because the edge is where we have all the things, we need to build a grid at the edge.



The vision was that computing should be a fifth utility, delivered on-demand like water or electricity. Today, there is an opportunity to build a new grid. This time at the edge.

The Grid is a paradigm where cloud computing resources are available when needed, precisely where they are needed, just as water is available at the faucet and power is available at the outlet. You use as much or as little as you need, in the location where you need it — and when you turn it off, it's available to others in the same way that water is available to others when you turn off the faucet.

The Grid will need many of the following attributes:

- **Flexible, fungible resources,** competitively auctioned across multiple clouds and automatically provisioned.
- **Pay-as-you go consumption** with simple financial settlement across a multiplicity of suppliers.
- **Declarative provisioning,** allowing developers to express intent around such attributes as location, latency, performance, cost, carbon footprint and SLAs, to be autonomously configured in real-time by AI-assisted schedulers and orchestrators.
- **Distributed resilience,** delivered via autonomic systems that ingest billions of data points to create digital twins of the network, compute and storage systems, and the ability to run predictive models to optimize placement.
- **Machine-speed interconnection** that directs local east/west traffic across a city or region to create complex service chains; Grid Interconnection that creates novel new business models for delivering aggregated real-time services.

Before Jeff Bezos sold his first book online, Internet pioneers that became part of Sun Microsystems created the underpinnings of the first Infrastructure-as-a-Service cloud, and they called it "The Grid" (later to become Sun Grid). The vision was that computing should be a fifth utility, delivered on-demand like water or electricity. Today, there is an opportunity to build a new grid. This time at the edge.

Building The Grid at the edge is why organizations like the Open Grid Alliance have emerged to assemble complex supply chains of hardware, software and services. By building grid integrations that span the entire stack, from dirt to cloud, we can empower the edge to deliver the applications it has promised.

ABOUT THE AUTHOR

MATT TRIFIRO is CMO of the edge infrastructure company Vapor IO. He is a co-founder and former co-chair of State of the Edge and is also the host of the award-winning Over the Edge Podcast.

LF EDGE

Essay

TERRY FU

CEO and Co-Founder,
Spectro Cloud

Kubernetes Unlocks Innovation at the Edge at Scale

I'm writing this in May 2022, in the middle of our industry's conference season and a string of customer meetings with big retailers, banks, telcos and healthcare companies. With every conversation, I'm left with the same takeaway: in 2022, the edge is where Kubernetes is really making a difference for customers, and where business model innovation is burning white hot.

We've talked to retailers about using edge devices in thousands of stores and restaurants to gather and analyze customer purchasing habits to optimize stock, as well as to run point of sale systems, CCTV analytics, digital signage, environment monitoring, equipment health and more.

We've talked to healthcare device companies about bringing powerful analytical tools closer to the edge, to the clinicians as they diagnose and treat patients — even enabling an app store-like experience for health providers to access new clinical features, opening up a new application innovation ecosystem and business model for the device maker.

And we've even worked with a startup that's putting lightweight Kubernetes worker nodes directly on drones to autonomously pick fruit, with the control plane at a nearby ground station. They have plans to scale to over a thousand clusters soon. Kubernetes doesn't get any more edge than that.

These use cases are new, they're fascinating, and they have huge potential to improve the customer experience and ultimately drive bottom-line growth for the business. This is exactly the stuff that IT teams want to be involved in and help drive!

But making it happen means deploying code to and managing potentially hundreds of thousands of edge devices. Indeed, even with the portability of containers and the orchestration features of Kubernetes, edge computing is really a perfect storm for IT and DevOps teams. They somehow have to deal with diverse, resource-constrained devices, distributed at mind-boggling scale in non-traditional environments, without access to on-site IT staffing, and a list of requirements for performance, security, resilience, compliance.

Clearing these infrastructural and operational roadblocks is not easy, but I've watched customers' eyes light up when you show them a clever architectural approach to sidestepping a seemingly intractable obstacle.

For example, take the challenge of pushing software updates to running edge devices in unsupervised locations: how can you perform rolling updates without

I'm so bullish on edge in 2022 because I see the excitement in the eyes of our customers when we show them a path that's free of these kinds of roadblocks.

risking application availability, even in single-server edge configurations? This is one of the edge problems we are solving for, in this instance addressing it with an A/B OS partition and multi-node Kubernetes deployment for the edge device.

Another challenge we often face is the ability to easily scale to thousands of edge K8s locations. Conventional edge architectures that have no separation between the management plane and control plane — or even worse, those that depend on a management server — are not able to scale beyond a few hundreds of K8s clusters. The way to address this is to let the local edge K8s cluster enforce policies so the management plane does not become a bottleneck as new edge locations are added into the mix.

I'm so bullish on edge in 2022 because I see the excitement in the eyes of our customers when we show them a path that's free of these kinds of roadblocks.

And the great news is, there are so, so many open source projects and commercial providers working every day to make edge computing easier — not just on PowerPoint slides but in the real world, through integrations and collaborative, community effort. Take the CNCF's Cluster API, for example. We have always been advocates of declarative management fueled by the open source community, and today Cluster API is the only way for modern K8s management to scale across multiple clusters and locations. Last summer, we extended Cluster API to support bare metal data center environments with our open-sourced Cluster API provider for Canonical MAAS. For edge, we now further extend Cluster API through integration with Docker Engine to fully support containerized multi-node K8s on single-server or multi-server configurations.

The road to the edge may still be winding, but thanks to herculean community efforts like Cluster API, it's getting more and more passable. That's important progress because, fundamentally, IT teams (whether ops, platform, DevOps or somewhere in between) don't want to be spending time on infrastructure care and feeding, nor do they want to be saying no to their business partners' next big idea. IT teams want to be innovators — and that's what the edge has the opportunity for in spades.

ABOUT THE AUTHOR

TENRY FU has more than 20 years of experience in the tech industry and software development. Prior to co-founding Spectro Cloud, he led the architecture for Cisco's multicloud management and private cloud portfolio after Cisco acquired his previous venture, CliQr Technologies. Past experience includes VMware and McAfee. He has more than 18 patents in the fields of scalable distributed systems, enterprise system management, and security. He is a hardcore audiophile and likes hiking with his family.

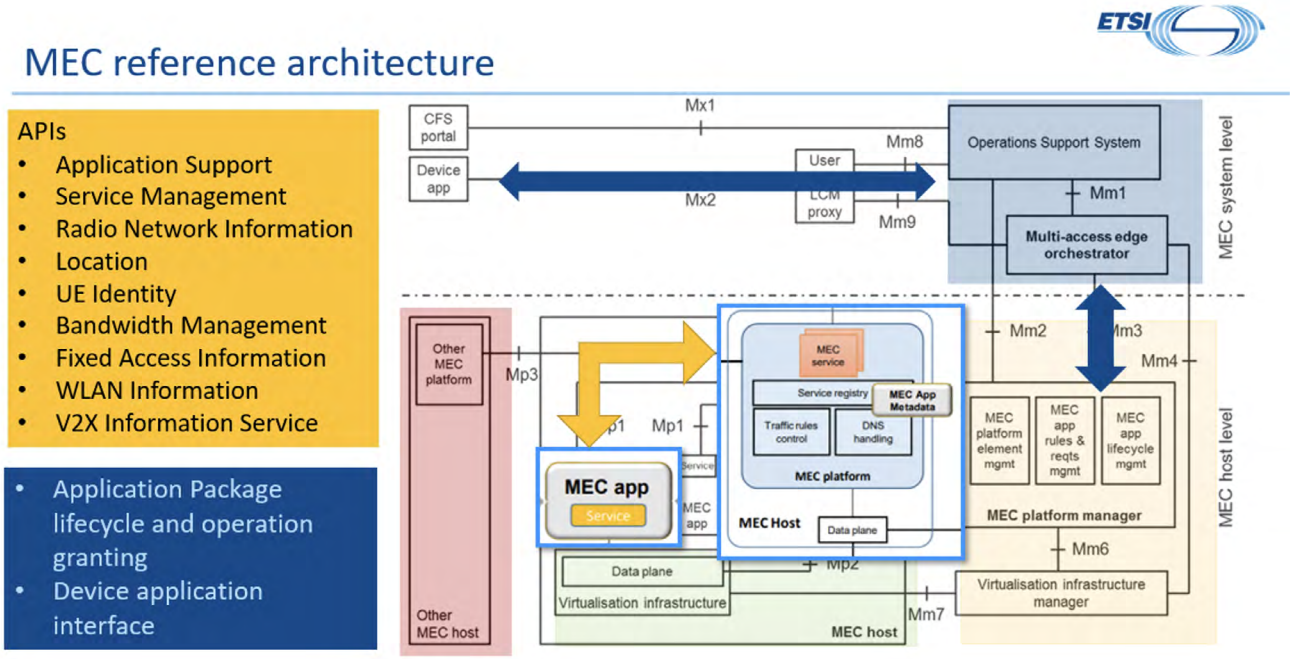
LF EDGE

News & Project Updates

Introducing MEC Federation

DARIO SABELLA, *CHAIRMAN, ETSI MEC*

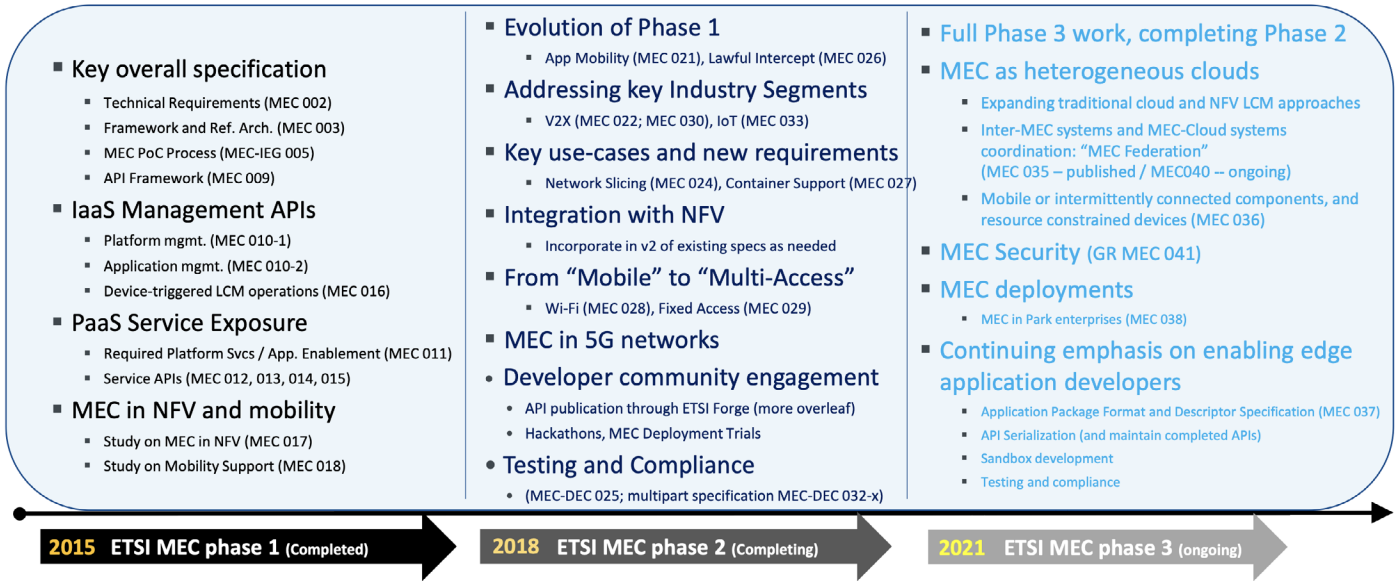
The Multi-access Edge Computing (MEC) initiative is an Industry Specification Group (ISG) within ETSI. The purpose of the ISG is to create a standardized, open environment that will allow the efficient and seamless integration of applications from vendors, service providers and third parties across multi-vendor platforms. ETSI MEC offers cloud-computing capabilities and an IT service environment to application developers and content providers at the edge of the network. This environment is characterized by ultra-low latency and high bandwidth as well as real-time access to radio network information that can be leveraged by applications.



ETSI MEC architecture supports multiple access technologies, such as 5G, Wi-Fi and fixed networks. Regarding MEC in 5G deployments, the standardization work is also aligned with the 3GPP SA6 EDGEAPP, as highlighted in an ETSI white paper co-authored by officials from both standard bodies. The common goal of this harmonized architecture is in fact to offer an interoperable environment, where consistent standards can open the edge computing market, avoiding duplication and market fragmentation.

It exposes a set of RESTful APIs to edge applications, to support multiple use cases: for this purpose, the group is publishing a set of MEC service APIs (also made available in OpenAPI format under ETSI forge website) that can be exposed in the MEC platform and consumed by authorized MEC applications. Also, New APIs (compliant with the MEC API principles) can be added and exposed in the MEC Platform. MEC also enables applications and services to be hosted "on top" of the mobile network elements, i.e., above the network layer. These

applications and services can benefit from being in close proximity to the customer and from receiving local radio network contextual information.



In 2021, the group started its Phase 3 work, where (apart from completing the outstanding Phase 2 work) the focus is on studying MEC security aspects, special MEC deployments (e.g., in Park enterprises) and most importantly MEC as heterogeneous clouds, thus expanding traditional cloud and NFV LCM approaches, i.e., introducing Inter-MEC systems and MEC-Cloud systems coordination, to support MEC Federation. This concept was initially defined in the Group Report MEC 035, as "a federated model of MEC systems enabling shared usage of MEC services and applications." The report analyzes eight use cases that require inter-system coordination, including those in multi-mobile network operator (MNO) environments. Recommendations, evaluations and possible technical solutions to solve key issues are issued for each use case. In particular, two use cases make recommendations to achieve V2X service continuity considering a typical MEC federation scenario of V2X services, in a MEC-system environment with multiple operators and multiple equipment manufacturers. These are key use cases for 5GAA (5G Automotive Association) that in 2021 joined MEC to strengthen the collaboration that started in 2019. Another use case describes a location-based immersive Augmented Reality game where a MEC federation can serve as a solution to limitations in providing an interactive AR application with users connected via different mobile operators.

Additional use cases include:

- An application instance transfer between MEC and Cloud systems
- Inter-system communication involving a MEC system within an MNO's network
- A MEC federation scenario for connecting different services
- A MEC federation scenario for edge service availability on visited (another's operator) networks
- A MEC federation scenario for edge-node sharing

In all these scenarios, interoperability is a key aspect, and stakeholders critically need a harmonized standard. For this purpose, the report served as a basis to derive recommendations for the normative work for MEC in Phase 3. In fact, the concept of MEC federation, initially introduced by GR MEC 035, is further exploited in a recently updated version of the MEC architecture (GS MEC 003), which introduces an architecture variant for MEC Federation, as a key enabler for supporting the requirements coming from GSMA OPG (Operator Platform Group): it enables inter-MEC system communication and allows 5G operators to collaborate among them and with service cloud providers and other stakeholders. In particular, multi-operator environments are key scenarios for automotive use cases (e.g., as required by 5GAA), and the standard support of GSMA OPG requirements is critical for the interoperability in this heterogeneous scenario.

Another relevant normative work started in the area of MEC Federation is **GS MEC 040**, which is introducing key functionalities for MEC Phase 3 work related to "MEC Federation Enablement APIs." These APIs are relevant for the Operator Platform (OP) architecture defined by GSMA OPG; the deliverable is still at draft stage since the normative work is ongoing. The group has already introduced some preliminary functionalities, e.g., registration of MEC system(s) to the federation, MEC Service discovery, application package management, application instance lifecycle management and also data-type definitions related to the information provided by the MEC orchestrator as a part of the "Registration of MEC system to the federation." The new deliverable versions are continuously made available also at their early stages in the MEC Open Area folder, in order to facilitate the information sharing with other organizations and foster collaboration.

LF Edge Project Updates

Akraino is a set of open infrastructures and application blueprints (BPs) for the edge, spanning a broad variety of use cases including 5G, AI, Edge IaaS/PaaS, IoT, Multi-Domain Orchestration and Automotive, as well as emerging Metaverse use cases, for both provider and enterprise edge domains. The Akraino community has created these blueprints, focusing exclusively on the edge in all of its different forms. The blueprints are ready for adoption as-is or can be used as a starting point for customizing a new edge blueprint. Currently, over 20 Akraino blueprints have been tested and validated in real hardware labs supported by users and community members. Akraino collaborates with other open-source communities such as CNCF and LF Networking, as well as Standard Definition Organizations such as ETSI MEC, focusing on edge infrastructure, networking, applications and services.

Akraino Releases 4 and 5 were made available in 2021, including K8-ready blueprints and multi-cloud deployments such as Public Cloud Edge Interface, AI Edge, 5G MEC System, Integrated Edge Cloud, Integrated Cloud Native blueprint families, Automotive, IoT, Metaverse Areas and more. All released blueprints have passed a vulnerability scanning process implemented by Akraino's security subcommittee.

In addition, Akraino has enhanced its API map, integrated upstream components, explored downstream labs and approved new incubation blueprints, including buffer at the edge, smart data transition for CPS and CPS robotics. Two blueprints entered the maturity stage in 2021: Connected Vehicle Blueprint and IEC Type 4: AR/VR oriented Edge Stack for Integrated Edge Cloud (IEC) Blueprint Family.

In November 2021, Akraino cautiously restarted a live meeting schedule with a hybrid event in Palo Alto. A more extensive technical meeting took place in March 2022.

- In 2021, Akraino delivered two releases (R4 in Feb and R5 in Sep) with over 20 blueprints participating, including 5+ new ones.
- We held two Akraino technical meetings with a wide representation from service providers, public cloud providers, technology providers, research institutions and developer communities.
- Akraino held Automotive and IoT Area workshops presented to the Korean Institute on Communications and Information Sciences.
- We developed and agreed on a collaboration plan with ETSI MEC ISG, participated in the MEC Technical Meetings and took part in the ETSI MEC Plugtests 2021. In 2022, we are planning a joint ETSI MEC — Akraino hackathon targeting development of innovative solutions utilizing Akraino blueprints and ETSI MEC APIs and services in the focus areas of Edge Computing and 5G, Automotive and Metaverse. In addition, in 2022, Akraino plans to participate and contribute to other ETSI MEC initiatives such as the ETSI MEC Tech podcast series.
- We held our first Akraino community awards, recognizing individual achievements, remarkable women of Akraino and blueprints of the year.

- Our security subcommittee developed new automated security vulnerability identification features and increased efficiency in the blueprint security certification process.
- We held our first in-person Akraino Reunion meeting since the onset of the COVID-19 pandemic.
- We participated, by providing presentations and technical demonstrations, in industry events, including the LF Networking Developer and Testing Forum, Open Networking & Edge Summit and Open Edge Computing workshop.
- Akraino TSC welcomed multiple guest speakers who discussed a variety of edge technologies during our weekly meetings.

EDGE X FOUNDRY™

EdgeX Foundry is an industry-leading edge IoT plug-and-play, ecosystem-enabled open software platform.

EdgeX is a highly flexible and scalable open-source software framework that facilitates interoperability between devices and applications at the IoT Edge. It accelerates the digital transformation for IoT use cases and businesses in many vertical markets by providing replaceable reference services for device-data ingestion, normalization and analysis. EdgeX Foundry also supports new edge data services and advanced edge computing applications, including enabling autonomous operations and AI at the edge.

The EdgeX IoT middleware platform acts as a dual transformation engine collecting data from sensors (i.e., “things”) at the edge and sending/receiving data to/from enterprise, cloud and on-premises applications. With 8+ million container downloads and as a stage 3 LF Edge project, EdgeX Foundry has broad industry support. It is available under a vendor-neutral Apache 2.0 open-source licensing model under the Linux Foundation.

LF Edge members and EdgeX Foundry contributors have created a range of complementary products and services, including commercial support, training, customer pilot programs and plug-in enhancements for device connectivity, applications, data and system management and security.

Additionally, EdgeX works closely with several of the other LF Edge projects such as Akraino, eKuiper, Home Edge and Open Horizon. EdgeX is part of the Akraino Edge Lightweight IoT (ELIOT) Blueprint and is tested under the Akraino Community Lab. eKuiper is used as the reference implementation rules engine provided with EdgeX. With fellow LF Edge Projects Open Horizons and SDO, EdgeX is part of the Open Retail Reference Architecture project to create a base foundation for edge-cloud retail-centric solutions. Finally, EdgeX is an integrated core to the HomeEdge to drive and enable a robust, reliable and intelligent home edge computing open-source framework.

In 2021, EdgeX delivered its second major release (EdgeX 2.0) — codenamed Ireland. In the Ireland release, the community added all new APIs and removed four years of technical debt accumulation. Also added were the use of a message bus for service communications (as an alternative to REST communications), a new GUI and new device/sensor connectors for GPIO, LLRP and CoAP. EdgeX also released its first-ever long-term support (LTS) version — codenamed

Jakarta — in the fall of 2021. LTS in open-source is rare. With its LTS release, the community announced two years of support for addressing critical bugs or security issues with the product. 2021 also saw massive expansion in adoption of the product in China, which spurred the creation of the Chinese language version of the project's website, as well as new project liaisons with AgStack and the Digital Twin Consortium, and the launch of a Developer Badge recognition program to honor code contributors and bug fixers on the project.

April 2022 marked EdgeX's fifth birthday. In May 2022, EdgeX will have its 10th release (version 2.2), codenamed Kamakura. The Kamakura release will include a new beta feature for capturing system metrics/telemetry, new camera connectors for ONVIF and simple USB/webcams, delayed start options for its microservices, dynamic device profiles and a second version of its command line interface (CLI). EdgeX has consistently released twice a year since its founding in 2017. In that light, the second release of 2022 — codenamed Levski — will come out around November 2022. At this time, the community expects the fall release to include a north-south messaging subsystem and standardization of units of measure along with a complete/full implementation of the metrics/telemetry collection. The community is also exploring the return to live events in the fall, which possibly will include hosting a hackathon event.

eKuiper is an edge lightweight IoT data analytics/streaming software implemented by Golang, and it can be run at all kinds of resource-constrained edge devices. One goal of eKuiper is to migrate the cloud streaming software frameworks (such as Apache Spark, Apache Storm and Apache Flink) to the edge side. eKuiper references these cloud streaming frameworks, and it also has considered special requirements of edge analytics and introduced a rule engine, based on Source, SQL (business logic) and Sink. Rule engine is used for developing streaming applications at edge side.

eKuiper joined LF Edge in mid-2021 as a Stage 1 "At-Large" project and has released three major releases since then. It provides the ability to extend the connection to external systems and the processing function by Python in addition to Go language. More than 10 other key features enrich the stream processing ability and simplify the usages.

eKuiper has been consistently collaborating with other LF Edge projects even before joining the umbrella. eKuiper works closely with EdgeX Foundry. It has been the referenced rule engine microservice of EdgeX Foundry since early 2020. Additionally, eKuiper can be deployed and run in Beatyl.

In 2022, the project will continue to enrich stream processing abilities such as more SQL clause and function support. Additionally, the project will improve the usability to lower the threshold to adopt and contribute. Lastly, we will seek to collaborate with other LF Edge projects and open-source projects under the Linux Foundation.

FLEDGE

Fledge is a mature IIoT open source platform that has been deployed in process and discrete manufacturing since 2018. Three open source communities contribute to the project: LF Edge, LF Energy, and OSDU. Contributor diversity includes industrial users (RTE, Alliander, JEA, Honda Racing, Neuman Aluminum, BRP), industrial suppliers (AVEVA/OSIsoft, FLIR WAGO, Nexcom, Advantech) and machine learning innovators (Google,, Dianomic, BIBA Research).

Fledge offers more than 100 industrial protocols, data mappings and sensor plugins, as well as more than 20 integration solutions connecting to OEE, ERP, logistics, MES, historians, databases, and cloud provider systems.

Fledge's pluggable microservice-based architecture and UI address the OT user community requirements for no-code, low-code and source-code development and provisioning of data pipelines and workloads. Supporting time-series, image, radio-metric, array, vibration, and transactional data, Fledge unifies the OT edge, enabling scale, manageability, and security.

Developers can leverage Fledge quick-start guides and its growing community support for rapid isolated development of new protocols and data mappings for any industrial asset or integration. It is easy to contribute and collaborate by building edge applications using pluggable filters, rules, ML runtimes, or scripting.

Fledge works closely with other LF Edge projects such as EVE and Akraino. EVE provides system and orchestration services and a container runtime for Fledge applications and services. Together, industrial operators can build, manage, secure, and support both Supervisory Control and Data Acquisition (SCADA) as well as non-SCADA connected machines, IIoT, and sensors as they scale. Fledge is also integrated with Akraino, as both projects support the roll-out of 5G and private LTE networks.

In 2022, Fledge is focused on:

- Set point control enabling both internal and external control paths that may be used to alter the behavior of a machine or system being monitored via the Fledge platform. This is not intended to be a replacement for time-critical process control systems, but as a way to impact the high-level functioning of a machine.
- Machine learning operations at the edge, focusing on those functionalities required to enable the creation of better, more accurate machine learning modeling and to improve the accuracy of the inference at the edge. Much of this will be related to data cleansing and preparation for the model building and inference stages.

Home Edge emphasizes driving and enabling a robust, reliable and intelligent home edge computing open-source framework, and an ecosystem running on a variety of smart home devices. To accelerate the deployment of the edge computing services ecosystem successfully, the Home Edge Project will provide users with an interoperable, flexible and scalable edge computing services platform with a set of APIs that can also run with libraries and runtimes. In a home setup with a quantifiable amount of smart devices with less computing and memory power, the orchestration service of Home Edge adds a major advantage. Home Edge has done four major releases since December 2021, vD (Dewberries) being the latest. With the latest release, Home Edge has provided data synchronization with cloud endpoints and APIs for saving/retrieving data from EdgeX containers along with quality improvements.

Home Edge has been the only project under the umbrella of LF Edge to cater to the smart home domain. Developers can leverage quick start guide support for rapid development of new use cases for smart home scenario. Home Edge has successfully collaborated with EdgeX for the development of a data storage feature and is open to work on such collaborations in the future also.

In 2022, Home Edge will work toward the development of the following features:

- MQTT-based independent cloud synchronization module to synchronize data to cloud endpoint (AWS/Azure/Google).
- Android flavor of Home Edge with minimum feature for execution on Android-based devices.
- Repository separation of data storage for modular and easy development.

Starting this year, we have been conducting weekly tagged releases, which makes stable code available regularly to developers. Also docker images (lfedge/edge-home-orchestration-go) are being released in the docker hub for easy download of container images.

We plan to develop sample demonstrations going forward and are open to work with developers toward the same.

Open Horizon is a platform for managing the service software lifecycle of containerized workloads and related machine learning assets, enabling the autonomous management of applications deployed to devices as well as distributed web-scale fleets of edge computing clusters, all from a central management hub.

Open Horizon joined LF Edge in mid-2020 as a Stage 1 "At-Large" project. The two-year-old project has now been at Stage 2 for a year and is actively working toward Stage 3. The project software provides the ability for a single human administrator to enable the autonomous management of more than 40,000 edge devices simultaneously. The software's management hub also enables the ability to handle multi-tenancy for up to 1,000 organizations (clients) per instance.

Open Horizon collaborates with other LF Edge projects such as EdgeX Foundry and Secure Device Onboard (SDO), an automated "Zero-Touch" onboarding service, in order to more securely and automatically onboard and provision a device on edge hardware. Used with Open Horizon, it provides a zero-touch model that simplifies the installer's role, reduces costs, and mitigates poor security practices.

The opportunities Open Horizon will pursue in 2022 include:

- Expanding the edge beyond traditional containerized host devices
- Enabling smart retail
- Promoting smart agriculture and sustainability
- Supporting security best practices on the edge
- Fostering human/cobot interactions
- Exploring application-directed networking

In 2022, the Open Horizon project will focus on:

- Expanding their popular mentorship opportunities and encouraging the involvement of early STEM learners;
- Attracting additional project partners to join IBM and mimik Technology in project leadership;
- Continued collaboration with projects, both within LF Edge and other open-source foundations and organizations;
- Hosting a public instance of Open Horizon in the LF Edge Shared Community Lab;
- Stage 3 maturity.

STATE OF THE EDGE

State of the Edge is a vendor-neutral platform for open research on edge computing that is dedicated to accelerating innovation by crowdsourcing a shared vocabulary for edge. The project develops free, shareable research that is widely adopted and used to discuss compelling solutions offered by edge computing and the next generation internet.

State of the Edge believes in four principles:

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

The State of the Edge project manages and produces the following assets under the LF Edge umbrella:

- State of the Edge reports, such as this one;
- Open Glossary of Edge Computing, a freely-licensed, open source lexicon of terms related to edge computing;
- Edge Computing Landscape, a dynamic, data-driven tool that categorizes LF Edge projects alongside edge-related organizations and technologies to provide a comprehensive overview of the edge ecosystem.

Other LF Edge Projects

BAETYL

SECURE DEVICE ONBOARD

STATE OF THE

EDGE

2022

WWW.STATEOFTHEEDGE.COM

THE LINUX FOUNDATION

Addendum

VERSION
2.0

STATE OF THE
EDGE
2022

Open Glossary of Edge Computing

LF EDGE

Open Glossary of Edge Computing, Version 2.0

The Open Glossary of Edge Computing is an official project of The Linux Foundation and a founding project of LF Edge. LF Edge is an umbrella organization within the Linux Foundation that aims to establish an open, interoperable framework for edge computing independent of hardware, silicon, cloud, or operating system.

For the past year, the Open Glossary team has been quietly updating the glossary in the Github repo, incorporating feedback and addressing suggestions and concerns from within LF Edge, as well as the large community. The team has brokered relationships with other organizations, including the TIA and OpenStack Foundation, and now has working groups developing a taxonomy of edge computing and an edge computing landscape map.

As part of LF Edge, The Open Glossary leverages a diverse community to develop and improve upon this shared lexicon, offering an organization-and vendor-neutral platform for advancing a common understanding of edge computing and the next generation internet ecosystems. The project seeks to curate, define and harmonize terms related to the field of edge computing. Project participants submit common and accepted definitions into an openly licensed repository.

The Open Glossary is governed using a transparent and meritocratic process. Anybody can make additions, clarifications and suggestions by raising a GitHub issue or editing a branch and issuing a pull request. Each issue or pull request is evaluated by the community for inclusion.

- For information on how to contribute to the glossary: See the Contributing Guide
- To view a markdown version of the glossary: See edge-glossary.md
- To send a private email to the project Chair: state-of-the-edge@gmail.com

3G, 4G, 5G

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

Access Edge Layer

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

Access Network

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

Aggregation Edge Layer

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

Base Station

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

Baseband Unit (BBU)

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

Central Office (CO)

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

Central Office Re-architected as Data Center (CORD)

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

Centralized Data Center

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

Cloud Computing

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

Cloud Native Network Function (CNF)

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

Cloud Node

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

Cloud RAN (C-RAN)

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

Cloud Service Provider (CSP)

An organization which operates typically large-scale cloud resources comprised of centralized and regional data centers. Most frequently used in the context of the public cloud. May also be referred to as a Cloud Service Operator (CSO).
See also: *Cloud Computing*

Cloudlet

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

Co-Location

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

Computational Offloading

An edge computing use case where tasks are offloaded from an edge device to the infrastructure edge for remote processing. Computational offloading seeks, for example, performance improvements and energy savings for mobile devices by offloading computation to the infrastructure edge with the goal of minimizing task execution latency and mobile device energy consumption. Computational offloading also enables new classes of mobile applications that would require computational power and storage

capacity that exceeds what the device alone is capable of employing (e.g., untethered Virtual Reality). In other cases, workloads may be offloaded from a centralized to an edge data center for performance. The term is also referred to as cloud offload and cyber foraging in the literature.
See also: *Traffi Offl ading*

Content Delivery Network (CDN)

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

Core Network

The layer of the service provider network which connects the access network and the devices connected to it to other network operators and service providers, such that data can be transmitted to and from the internet or to and from other networks. May be multiple hops away from infrastructure edge computing resources.
See also: *Access Network*

Customer-Premises Equipment (CPE)

The local piece of equipment such as a cable network modem which allows the subscriber to a network service to connect to the access network of the service provider. Typically one hop away towards the end users from infrastructure edge computing resources.
See also: *Access Network*

Data Center

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

Data Gravity

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

Data Ingest

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

Data Reduction

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

Data Sovereignty

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

Decision Support

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

Device Edge

Edge computing capabilities on the device or user side of the last mile network. Often depends on a gateway or similar device in the field to collect and process data from devices. May also use limited spare compute and data storage capability from user devices such as smartphones, laptops and sensors to process edge computing workloads. Distinct from infrastructure edge as it uses device resources.
See also: *Infrastructure Edge*

Device Edge Cloud

An extension of the edge cloud concept where certain workloads can be operated on resources available at the device edge. Typically does not provide cloud-like elastically-allocated resources, but may be optimal for zero-latency workloads.
See also: *Edge Cloud*

Distributed Antenna System (DAS) Hub

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

Edge Cloud

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

Edge Computing

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

Edge Data Center

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

Edge Exchange

Pre-internet traffic exchange occuring at an infrastructure edge data center. This function will typically be performed in the edge meet me room of an infrastructure edge data center, and may operate in a supplemental or hierarchical fashion with traditional centralized internet exchange points if a destination location is not present at the edge exchange, as is the case with internet-bound traffic. An edge exchange may be used in an attempt to improve end-to-end application latency compared with a centralized internet exchange.
See also: *Internet Exchange Point (IXP)*

Edge Meet Me Room

An area within an edge data center where tenants and telecommunications providers can interconnect with each other and other edge data centers in the same fashion as they would in a traditional meet me room environment, except at the edge.
See also: *Interconnection*

Edge Network Fabric

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

Edge Node

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

Edge-Enhanced Application

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

Edge-Native Application

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

Fog Computing

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

Gateway Device

A subcategory of the device edge, referring to devices on the device edge side of the last mile network which operate as gateways for other local devices, with the goal of aggregating and facilitating data transference between local devices, many of which are battery-operated and may operate for extended periods in a low-power state, and external entities such as a data analysis application operating inside an edge data center at the infrastructure edge.
See also: *Resource Constrained Device*

Infrastructure Edge

Edge computing capability, typically in the form of one or more edge data centers, which is deployed on the operator side of the last mile network. Compute, data storage and network resources positioned at the infrastructure edge allow for cloud-like capabilities similar to those found in centralized data centers such as the elastic allocation of resources, but with lower latency and lower data transport costs due to a higher degree of locality to user than with a centralized or regional data center.
See also: *Device Edge*

Local Breakout

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

Interconnection

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

Internet Edge

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

Internet Exchange Point (IXP)

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

IP Aggregation

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

Jitter

The variation in network data transmission latency observed over a period of time. Measured in terms of milliseconds as a range from the lowest to highest observed latency values which are recorded over the measurement period. A key metric for real-time applications such as VoIP, autonomous driving and online gaming which assume little latency variation is present and are sensitive to changes in this metric.
See also: *Quality of Service (QoS)*

Last Mile

The segment of a telecommunications network that connects the service provider to the customer. The type of connection and distance between the customer and the infrastructure determines the performance and services available to the customer. The last mile is part of the access network, and is also the network segment closest to the user that is within the control of the service provider. Examples of this include cabling from a DOCSIS headend site to a cable modem, or the wireless connection between a customer's mobile device and a cellular network site.
See also: *Access Network*

Latency

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

Latency Critical Application

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

Latency Sensitive Application

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

Location Awareness

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

Location-Based Node Selection

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

Micro Modular Data Center (MMDC)

A data center which applies the modular data center concept at a smaller scale, typically from 50 to 150 kW in capacity. Takes a number of possible forms including a rackmount cabinet which may be deployed indoors or outdoors as required. Like larger modular data centers, micro modular data centers are capable of being combined with other data centers to increase available resource in an area.
See also: *Edge Data Center*

Mobile Edge

A combination of infrastructure edge, device edge and network slicing capabilities which are tuned to support specific use cases, such as real-time autonomous vehicle control, autonomous vehicle pathfinding and in-car entertainment. Such applications often combine the need for high-bandwidth, low-latency and seamless reliability.
See also: *Infrastructure Edge*

Mobile Network Operator (MNO)

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

Mobile Virtual Network Operator (MVNO)

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

Modular Data Center (MDC)

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

Multi-access Edge Computing (MEC)

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

Network Function Virtualization (NFV)

The migration of network functions from embedded services inside proprietary hardware appliances to software-based VNFs running on standard x86 and ARM servers using industry standard virtualization and cloud computing technologies. In many cases NFV processing and data storage will occur at the edge data centers that are connected directly to the local cellular site, within the infrastructure edge.
See also: *Virtualized Network Function (VNF)*

Network Hop

A point at which the routing or switching of data in transit across a network occurs; a decision point, typically at an aggregating device such as a router, as to the next immediate destination of that data. Reducing the number of network hops between user and application is one of the primary performance goals of edge computing.
See also: *Edge Computing*

Northbound vs Southbound (and east/west)

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

Over-the-Top Service Provider (OTT)

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

Point of Presence (PoP)

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

Quality of Experience (QoE)

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

Quality of Service (QoS)

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

Radio Access Network (RAN)

A wireless variant of the access network, typically referring to a cellular network such as 3G, 4G or 5G. The 5G RAN will be supported by compute, data storage and network resources at the infrastructure edge as it utilizes NFV and C-RAN.
See also: *Cloud RAN (C-RAN)*

Regional Data Center

A data center positioned in scale between a centralized data center and an edge data center. Significantly physically further away from end users than an edge data center, but closer to them than a centralized data center. Also referred to as a metropolitan data center in some contexts. Part of traditional cloud computing.
See also: *Cloud Computing*

Resource Constrained Device

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

Service Provider

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

Shared Infrastructure

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

Software Edge

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

Throughput

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

Traffic Offloading

The process of re-routing data that would normally be delivered inefficiently, such as over long distance, congested, or high cost networks, to an alternative, more local destination (e.g., a CDN cache) or on to a lower-cost or more efficient network. Local Breakout is an example of using edge computing for traffic offloading.
See also: *Local Breakout*

Truck Roll

In the context of edge computing, the act of sending personnel to an edge computing location, such as to an edge data center, typically to resolve or troubleshoot a detected issue. Such locations are often remote and operate for the majority of the time remotely, without onsite personnel. This makes the cost other practical considerations of truck rolls a potential concern for edge computing operators.

Vehicle 2 Infrastructure (V2I)

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

Virtual Data Center

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

Virtualized Network Function (VNF)

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

Workload Orchestration

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

xHaul ("crosshaul")

The high-speed interconnection of two or more pieces of network or data center infrastructure. Backhaul and fronthaul are examples of xhaul.
See also: *Interconnection*