THE LINUX FOUNDATION

STATE OF THE EDGE

2021

A Market and Ecosystem
Report for Edge Computing

stateoftheedge.com

Contents

Introductions

Executive Summary

This 2021 State of the Edge report explores today's edge computing ecosystem with a focus on the increasingly interconnected domains of critical infrastructure, networks, software, and hardware. In line with previous reports, the insights are provided by independent experts with the goal of educating and advancing the field. The report also builds upon previous editions with updated market sizing data, as well as thoughts from industry leaders.

Some of the most important findings include:

- Despite (and in some cases driven by) the impact of COVID-19, the deployment of new edge infrastructure and applications continued in 2020. Seven out of ten areas saw increased forecasts compared to last year's report.
- Tremendous infrastructure investments are needed to support the growing device and infrastructure edge demand. We estimate that between 2019 and 2028, cumulative capital expenditures of up to $800 billion USD will be spent on new and replacement IT server equipment and edge computing facilities. These expenditures will be relatively evenly split between equipment for the device and infrastructure edges.
- As a natural extension of cloud computing, the edge cloud construct is increasingly viewed as a key enabler for the "Fourth Industrial Revolution" in which the widespread deployment of the Internet of Things (IoT), the global sharing economy and the increase of zero marginal cost manufacturing deliver unprecedented communication-driven opportunities with massive economies of scale.
- The foundational options for edge infrastructure are expanding, as hyperscale data center providers and telecom service providers jostle for position with suppliers of regional data centers, traditional colocation solutions, micro edge data centers, wireless towers and networking equipment. To a large extent, the Return on Investment for all these players will be determined by the effectiveness of the partnerships that they establish.
- Architecture decisions become significant factors in CAPEX and OPEX as edge deployments scale up and new silicon options emerge. Intel, AMD and Nvidia are in fierce competition, while acceleration options such as SmartNICs (Smart Network Interface Cards), FPGAs (Field-Programmable Gate Arrays) and DPUs (Data Processing Units) are aiming to boost the cost-performance of Artificial Intelligence (AI) and Machine Learning (ML) algorithms.
- SD-WAN (Software-Defined Wide Area Networking) and SASE (Secure Access Service Edge) represent critical virtual networking technologies for edge applications. The security, resiliency and session-awareness that these technologies bring to enterprise connectivity are equally applicable to use cases hosted at the network edge.
- While a segment of edge use cases will emerge that require 5G connectivity, 5G is not a requirement for mainstream edge applications. Many enterprise edge applications leverage wireline networks, while 4G-LTE and its variants LTE-M (Long Term Evolution category M1) and NB-IoT (Narrowband IoT) are adequate for the majority of current wireless connections. CBRS (Citizens Broadband Radio Service) and shared spectrum solutions are also appealing.

- Hybrid and multi-cloud solutions from major players like Amazon Web Services, Google Cloud, Microsoft Azure, VMware and IBM (Red Hat) are extending the cloud experience to far flung locations with the promise of a consistent application and operations experience

- Off-the-shelf edge applications and marketplaces are becoming available, thanks to initiatives such as Google's December 2020 launch of an ecosystem that comprises more than 200 applications from over 30 Independent Software Vendors (ISVs). Similar efforts by IBM and Amazon Web Services, as well as "all in one" pricing models, indicate how cloud-style consumption is influencing the edge.

- Open source projects are enabling organizations to accelerate the adoption and deployment of edge applications. These communities also facilitate standardization across the industry, increasing the pace of innovation while mitigating the risk of vendor lock-in.

- The global IT power footprint for infrastructure edge deployments is conservatively forecasted to increase from 1 GW in 2019 to over 40 GW by 2028, with a CAGR (Compound Annual Growth Rate) of 40%. It is forecasted that by 2028, 37% of the global infrastructure edge footprint will be for use cases associated with mobile and residential consumers, with the remaining 63% supporting applications in vertical markets such as healthcare, manufacturing, energy, logistics, smart cities, retail and transportation.

Foreword

ENTERING THE HYPER-CONNECTED ERA

We are in the midst of a transition from the mobile internet to the hyper-connected era where nearly every object in our physical world can have computing and connectivity built in, whether it's a simple consumer doorbell or a complicated robotic manufacturing device. As the number of these smart devices grows, they will fuel automation and personalization, thereby transforming many industries.

This hyper-connectivity will also cause a transition from vertically-integrated, industry-specific solutions to horizontal platforms where all systems can process relevant data and exchange knowledge for Artificial Intelligence (AI)-based automation. Some speculate that this will be the impetus for the next Industrial Revolution, with AI, Machine Learning ML and distributed ledger technologies driving the decentralization of computing, communications and business processes.

While it's commonplace to refer to the Internet of Things (IoT), we will actually have an Internet of Systems, where devices serving different vertical applications within different systems need to communicate directly to exchange knowledge, autonomously and securely with no single point of failure, starting from end user "edge" devices (smartphones, appliances, TVs, cars, robots, sensors, etc.) across diverse networks and cloud platforms.

We are witnessing innovations in software, including much of it in the open source ecosystems led by LF Edge. It's not too hard to see the near-term future where hybrid edge clouds will enable API-first microservices with integrated services for authentication, authorization and identity management regardless of device type, operating system and network. Integration across software systems will create an open ecosystem that offers scalability, extensibility and interoperability along with the required flexibility to adapt to our changing demands in the future.

The Linux Foundation's State of the Edge project has become the industry's standard-bearer for open source research, and this kind of collaboration has helped fuel our ecosystem.

I hope that in 2021 we expand our vision beyond the classical view where client and server functions are delegated to different devices. We need to stretch the boundaries of edge to client devices such as smartphones, IoT devices and even smart sensors. These devices are more than capable of hosting microservices today and will be more capable tomorrow thanks to Moore's law along with the evolution of processor architectures.

We should also start thinking edge-out versus cloud-in when we build our applications. This approach will bring further efficiencies, data privacy, control and scalability required for the hyper-connected world where digital intersects with every aspect of our physical world.

Fay Arjomandi
Founder and CEO, MIMIK
2020 "Edge Woman of the Year"

From the Co-Chairs

The interconnected and cooperative nature of the edge has never been more apparent, whether we're talking about companies collaborating, networks interconnecting, or open source projects evolving.

In previous editions of this report, we have implicitly and explicitly argued that the complexity of the edge opportunity will require intense collaboration. In fact, the neutral and community-driven approach to the State of the Edge project has its genesis in the anticipation of this collaboration.

This past year has magnified both the need for collaboration and the advantages of participating. Across all parts of the edge ecosystem, collaboration has been helping to bring new ideas, technologies and commercial solutions to market. As a result, forecasts for the size of the edge computing market continue to grow.

To provide a framework for understanding the edge computing ecosystem, we structured this report around four key areas of innovation: **critical infrastructure**, **hardware**, **networks & networking**, and **software**. We asked four independent experts to contribute their insights, giving us a window into the most notable movements in each area as it relates to edge computing. In addition, over a dozen leaders contributed "postcards from the edge" (concise and bite-sized thought-leadership essays you will find sprinkled throughout) to help further our understanding of what the edge means.

Oftentimes, the best way to prepare for what is ahead includes understanding where momentum is building. Yet, as our authors reported, we find palpable change just about everywhere. Even in traditionally "slow and steady" areas like silicon or physical infrastructure, a series of dramatic shifts are forging new landscapes.

The edge, with all of its complexities has become a fast-moving, forceful and demanding industry in its own right.. By sharing our observations in this yearly report, we offer a glimpse of the future for our industry which is responsive, collaborative, open, and interconnected.

CALLS TO ACTION

If your organization is involved in edge computing as part of the ever-expanding ecosystem of vendors, service providers, integrators, data center operators and/or application developers, we want to help you promote your solutions. We encourage you to reach out to the **LF Edge Landscape** project to ensure your company is correctly represented, and join the flourishing edge computing community by participating in an **LF Edge project**.

Matt Trifiro
CMO, Vapor IO
Report Co-Chair

Jacob Smith
VP Strategy & Marketing, Equinix
Report Co-Chair

CHAPTER 1

A Lot Has Changed in a Year

Phil Marshall
Chief Research Officer, Tolaga Research

Last year, the 2020 State of the Edge report presented a use case driven market forecast of the growth and value of edge infrastructure from 2019 through 2028. The edge computing market is highly dynamic, fueled by a growing range of use cases with key service requirements, such as low latency performance, reduced bandwidth demands, and data security and sovereignty. To keep pace with the changing market, the 2020 forecast has been updated to reflect current market expectations.

Tremendous infrastructure investments are needed to support the growing device and infrastructure edge demand. We estimate that between 2019 and 2028, cumulative capital expenditures of up to $800 billion USD will be spent on new and replacement IT server equipment and edge computing facilities. These expenditures will be relatively evenly split between equipment for the device and infrastructure edges.

EDGE TAXONOMY

Although there are a variety of definitions for edge computing, The Linux Foundation's LF Edge has created a formal taxonomy that has received many accolades and is getting widespread adoption. The LF Edge taxonomy visualizes edge computing through the continuum of physical infrastructure that comprises the internet, from centralized data centers to devices.

By locating services at key points along this continuum, developers can better satisfy the latency requirements of their applications. Figure 1 summarizes the edge computing continuum, spanning from discrete distributed devices to centralized data centers, along with key trends that define the boundaries of each category. This includes the increasingly complex design tradeoffs that architects need to make the closer compute resources get to the physical world.

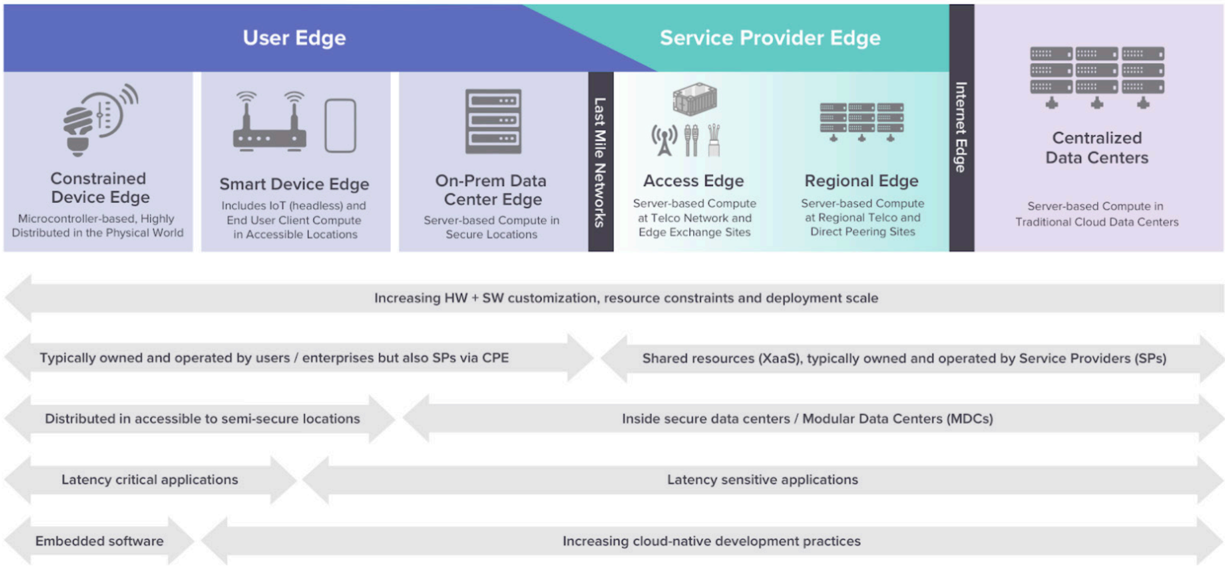

User Edge
Service Provider Edge
Constrained Device Edge
Microcontroller-based, Highly Distributed in the Physical World
Smart Device Edge
Includes IoT (headless) and End User Client Compute in Accessible Locations
On-Prem Data Center Edge
Server-based Compute in Secure Locations
Last Mile Networks
Access Edge
Server-based Compute at Telco Network and Edge Exchange Sites
Regional Edge
Server-based Compute at Regional Telco and Direct Peering Sites
Internet Edge
Centralized Data Centers
Server-based Compute in Traditional Cloud Data Centers
Increasing HW + SW customization, resource constraints and deployment scale
Typically owned and operated by users / enterprises but also SPs via CPE
Shared resources (XaaS), typically owned and operated by Service Providers (SPs)
Distributed in accessible to semi-secure locations
Inside secure data centers / Modular Data Centers (MDCs)
Latency critical applications
Latency sensitive applications
Embedded software
Increasing cloud-native development practices

FIGURE 1
Edge Continuum provided by LF Edge

Bringing Together the Global Cloud

Jonathan Seelig
Co-founder and CEO, Ridge

Applications that are going to really change the world will want more than just a small number of data centers in the same limited geography. The applications will care about being in a highly distributed cloud for latency, geographic, or data sovereignty reasons. An ad tech company with computationally intensive workloads will need to be closer to end users. A remote desktop company will require high levels of geographic distribution. Many of these applications are already being built and more are on the way.

It turns out the best space, power and connectivity in every geography everywhere in the world is owned and operated by a data center company in that place. The best data centers in Tokyo are owned by Japanese companies. The best data centers in Frankfurt are owned by German companies.

However, despite the incredible space, power, connectivity, and virtualization and compute offerings they have, very few of those data center companies are capable of offering the level of managed services that are the hallmark of what a modern cloud native application wants. The future is to bring together existing data centers in lots and lots of different markets, provide each with the means to deploy managed services, and create a truly global cloud.

*For more of Jonathan's thoughts on this topic, **catch his interview** with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

The LF Edge model focuses on the two main edge tiers that straddle the last mile networks, the "Service Provider Edge" and the "User Edge", with each being further broken down into subcategories. The **User Edge** consists of

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

And the **Service Provider Edge**, which consists of compute platforms, which are colocated with:

- Access sites which house network access equipment, such as cellular radio base stations, xDSL (Digital Subscriber Lines) and xPON (Passive Optical Network) access sites;
- Aggregation hubs, such as those which house DAS (Distributed Antenna Systems) and serve as an initial aggregation of transmission connections from the access sites;
- Regional data centers and central offices, where access controller, switching equipment and other service gateway functionality is commonly deployed.

ABOUT THE EDGE FORECAST

The forecast model focuses on 43 use cases spanning 11 verticals, including CSP, enterprise IT, residential and mobile consumer services, retail, healthcare, automotive, commercial UAV, smart grid, smart cities, and manufacturing. Other verticals, such as education and financial services and investments at the Device Edge will drive additional edge computing market opportunities that are not included in the forecast.

The forecast uses the power footprint of IT server equipment deployed at the Service Provider Edge as a primary measure to illustrate edge expansion. Server power footprint represents the rated power of all the equipment operating at the edge, though it does not represent actual electrical power consumed, which will be much less and depend on the power duty-cycle of the deployed equipment.

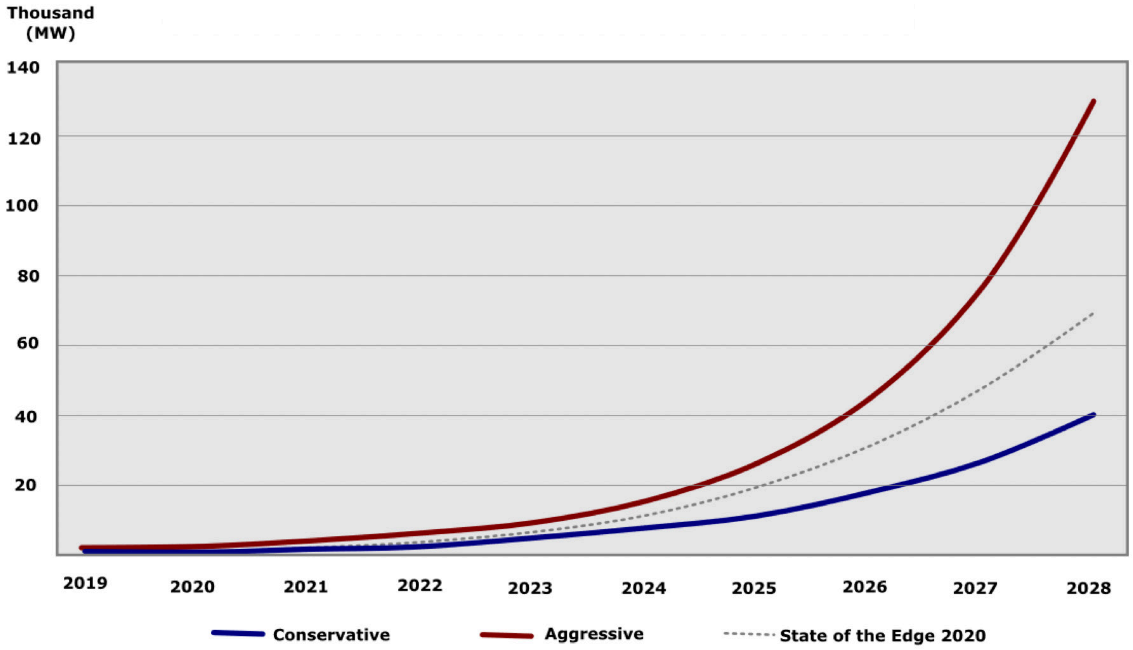


FIGURE 2
Total Global Infrastructure Edge IT Power Footprint

A CHANGED WORLD

The world has changed greatly since SOTE-2020 was published. The COVID-19 pandemic has disrupted the global status quo and is heralding a world of digital “haves” and “have-nots”. There are some areas where digital services are thriving, and other areas, such as the travel industry, which are languishing. In aggregate though, the pandemic is accelerating digital transformation and service adoption. Consumers and enterprises are using digital solutions to overcome pandemic challenges, such as lockdowns, social distancing and fragile supply chains. After we get through the pandemic, permanent change is inevitable for use cases where consumers and enterprises find continued value. However, exactly which digital use cases will prevail is challenging to predict.

To account for this uncertainty, both conservative and optimistic forecasts for the Infrastructure Edge have been developed. The conservative forecast has the global aggregate IT power footprint increasing from 1,078 MW in 2019 with a CAGR (Compound Annual Growth Rate) of 40% to reach 40,380 MW by 2028. The aggressive forecast results in a 70 percent CAGR to reach 120,840 MW by 2028. These forecasts are comparable to the forecasts in the SOTE-2020 report, which predicted a CAGR of 60.4% between 2019 and 2028.

REGIONAL REVIEW

The Infrastructure Edge is expanding in all global regions. The pace of this expansion depends on a variety of macro-economic, geographical, geopolitical and economic factors that were assessed in the forecast. By 2028, it is forecasted that 37.7% of the global Infrastructure Edge footprint will be in Asia Pacific. Asia Pacific has tremendous diversity, spanning some of the poorest and most wealthy countries in the world. Countries including China, Japan and South Korea are predicted to be significant contributors to edge computing adoption in the region.

Europe is forecasted to have 29% of the global Infrastructure Edge footprint by 2028, with over half being in Western Europe. Countries like Germany, France and the United Kingdom are predicted to be major contributors to Infrastructure Edge deployments in Europe

It is predicted that by 2028, 20.5% of the global Infrastructure Edge will be deployed in North America. North America is heralding the early adoption of the Infrastructure Edge, buoyed by its high technology industry as well as dominance in the internet and cloud computing. However, in the medium term, Asia Pacific and Europe will demand an increasing percentage of the global Infrastructure Edge, primarily because of their larger populations. Amongst the other regions, 7% of the global Infrastructure Edge will be deployed in Latin America by 2028, with significant investments in countries like Brazil and Mexico. The remaining 5.8% in 2028 is predicted to be deployed in the Middle East and African regions.

SHIFTING EDGE COMPUTING PRIORITIES

Digital service adoption accelerated in 2020 with specific use cases to address challenges created by the COVID-19 pandemic. This accelerated adoption is expected to have long-term implications for the edge computing market and the use cases that come to the fore. It is forecasted that by 2028, 36.5% of the global Infrastructure Edge footprint will be for use cases associated with mobile and residential consumers. This is down from the 45.1% share that was forecasted in the SOTE-2020 report and reflects increased edge services adoption in other verticals.

In 2028, it is forecast that 11.9% of the global Infrastructure Edge footprint will be associated with Enterprise IT use cases.

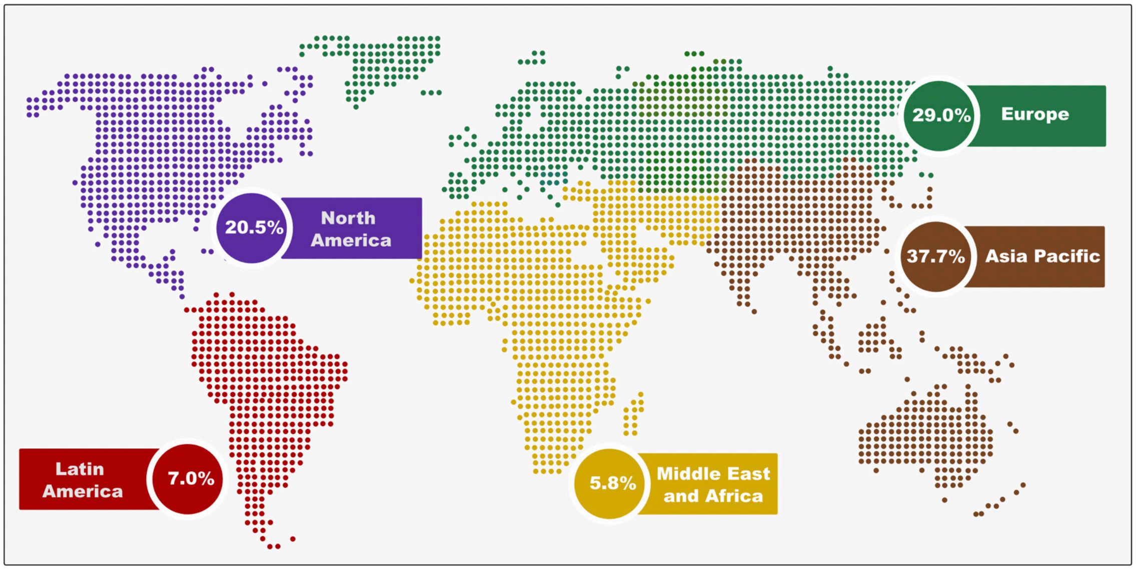


FIGURE 3
Regional Distributuion of Edge Deployments

In the short to medium term, infrastructure edge demand for Enterprise IT will be driven by cloud service use cases that are complemented and enhanced with edge computing capabilities.

However, it is predicted that in the long-term, Infrastructure Edge demand will be driven by "edge native" use cases that can only function when edge computing capabilities are available. These edge native use cases depend on the maturation of key technologies, such as augmented and virtual reality, and autonomous systems, such as those for closed loop enterprise IT functions.

CSPs have driven early Infrastructure Edge demand as they virtualize and cloudify their networks. Initially core and transport networks are being virtualized with standards like NFV and SDN. This is generally a precursor for end-to-end network transformation that incorporates access network virtualization, such as Cloud Radio Access Networks (C-RAN). In addition, CSPs are uniquely positioned with geographically distributed network infrastructure, which is well suited for Infrastructure Edge implementations. Many CSPs are implementing Multi-access Edge Computing (MEC) technology to bring a variety of network-centric capabilities, often in partnership with third parties, including cloud service providers like Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP). In 2028, it is forecast that 10.9% of infrastructure edge deployments will support CSP use cases.

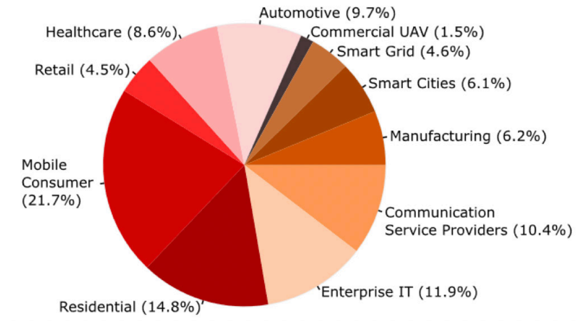


FIGURE 4
Global Infrastructure Edge Power

In 2020, digital healthcare was accelerated, with solutions that address the challenges attributable to the COVID-19 pandemic. Because of regulations and safety demands, digital healthcare innovation

Focus on the Bottom Line

My personal opinion — we all need to show the operators the money. When it comes to edge computing, they don't want to see a strategy about what hyperscalers are making in the cloud. They don't want to hear about the wonderful imaginary world where every car is self-driving and every person wears VR glasses. If they're going to invest in rolling out infrastructure at the telco edge, operators will need some low-hanging fruit that will allow them to recoup their investment quickly. Operators want to see the money in six months.

Finding those business cases today is hard work because it's always on a case-by-case basis. They are out there, but there's no standard formula that says every single telco should do this to put themselves in the right position and make money. You can certainly just make your sites available for Amazon, Google and AWS, and you're going to make money, and that's not necessarily a bad decision. However, I think in many cases, it's not the direction operators would prefer to go.

If you can show them an alternative that does not cede significant value to the hyperscalers and that allows edge computing to pay for itself, the operators will say, "Okay, I'm in." I think they see the promise of all the technology that's emerging now and with the right starting point they will see how they can start building towards a much bigger future.

For more of Alex's thoughts on this topic, ***catch his interview*** *with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

Alex Reznik
Chair, ETSI MEC ISG

is typically constrained. However, the pandemic has brought specific uses-cases into sharp focus, such as those relating to remote healthcare and assisted living. By 2028, it is forecast that 8.6% of the global Infrastructure Edge will support healthcare use cases. This is a significant increase on the 6.8% forecast in SOTE-2020.

Manufacturers have been pursuing multi-year strategies to implement digital services under the guise of frameworks such as Industry 4.0. The digital service priorities of manufacturers vary depending on their market conditions, operational objectives, product complexity, precision and customization demands, as well as the extent of brownfield and greenfield manufacturing facilities. For example, the automotive industry is confronted with tremendous disruption from companies like Tesla and has seen a surge in demand for electric vehicles. Electric vehicles require greenfield manufacturing facilities. Because these facilities are greenfield, they typically incorporate advanced digital operations, that capitalize on advancements in automation, wireless connectivity and autonomous systems, such as autonomous mobile robots. This requires extensive edge computing functionality, the lion's share of which is expected to reside at the Device Edge, particularly for Operational Technologies (OT). However, use cases that span geographical areas beyond the boundaries of individual factories, such as warehousing, supply chain and logistics, will also require Infrastructure Edge capabilities. By 2028, it is forecast that 6.2% of the global Infrastructure Edge will support manufacturing-related use cases.

Smart cities have been on the horizon for many years and are now using a range of digital services, such as those relating to security and surveillance, city operations (e.g., waste management) and traffic management. As more cities implement digital services, their value propositions become better understood and easier to justify elsewhere. While the Device Edge addresses many smart city use case demands, it is forecast that 6.1% of the global Infrastructure Edge in 2028 will support smart-city use cases. These use cases include smart buildings, lighting and traffic management and other digital services for public safety, venues and city operations.

Traditional retail companies are under tremendous pressure to innovate as ecommerce solutions grow in popularity. Permanent consumer behavior changes that favor ecommerce are likely to prevail after the COVID-19 pandemic and will continue to compromise traditional retail. Service continuity is important for retail, particularly for point-of-sale payment systems, which generally depend on Device Edge capabilities. Infrastructure Edge solutions are used for a range of use cases, including digital signage and Digital Out-of-Home (DOOH) experiences, immersive in-store solutions, proximity marketing and supply chain optimization. In 2028, it is forecast that 4.6% of the global Infrastructure Edge footprint will be associated with traditional retail use cases.

Advancements in energy utility technologies and renewables with smart grids, microgrids and distributed energy storage, are driving edge computing demands, both at the device and infrastructure edges. In 2028, it is forecast that 4.6% of the global Infrastructure Edge will support use cases associated with smart grids.

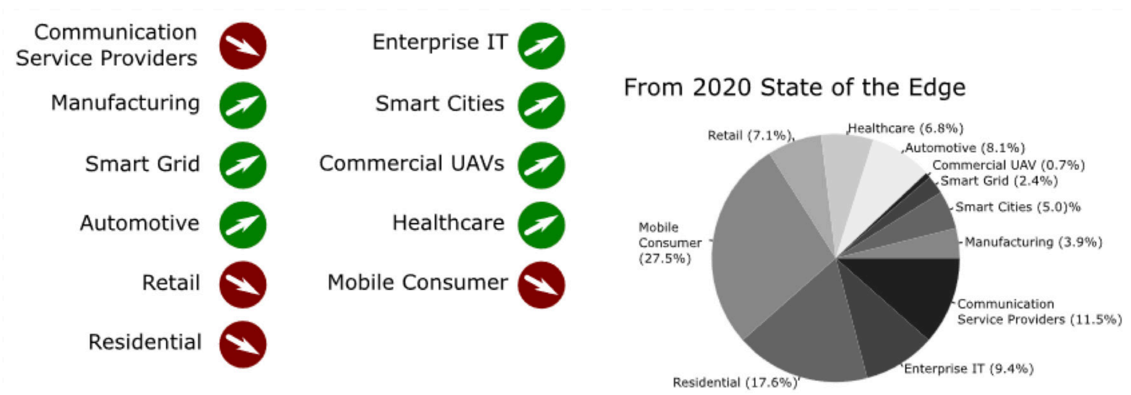


FIGURE 5

Comparison to State of the Edge 2020 Forecast

Commercial unmanned aerial vehicles (UAVs) are being used for a growing range of use cases, particularly as autonomous systems and Beyond Line-of-Sight (BLOS) operations become more widely available. Commercial UAVs are being used across a range of industries including agriculture, manufacturing, real estate, supply chain and logistics, and smart cities. By 2028 it is forecast that 1.5% of the global Infrastructure Edge will support commercial UAV use cases, which include mapping and surveying, photogrammetry and 3D and digital elevation modelling.

The demand expectations for the Infrastructure Edge have changed since 2019, when the forecasts for the SOTE-2020 report were developed. These changes reflect the lasting effects of the COVID-19 pandemic and shifting prospects as the edge computing market continues to mature. While the 2019 forecast falls within the range of the current global Infrastructure Edge forecast, the distribution of edge computing amongst various use cases and verticals has changed. In particular, the percentages of the global Infrastructure Edge footprint for CSPs, retail, residential and mobile consumer use cases have decreased in the latest forecast, while the percentages for use cases in all the other verticals have increased. Notable amongst these increases are:

- Manufacturing, which increased from 3.9 to 6.2 percent, as companies bolster their supply chain and inventory management capabilities and capitalize on automation technologies and autonomous systems.

- Healthcare, which increased from 6.8 to 8.6 percent, buoyed by increased expectations for remote healthcare, digital data management and assisted living, and;

- Smart cities, which increased from 5.0 to 6.1 percent, in anticipation of increased expenditures in digital infrastructure in the areas such as surveillance, public safety, city services and autonomous systems.

A PROMISING FUTURE

Edge computing is forecast to grow tremendously in the coming years as digital services mature and require distributed computing resources. A vast variety of use cases depends on edge computing and this will continue to increase as the market develops. Both the device and infrastructure edges will play important roles. Early market momentum for the Infrastructure Edge has come from CSPs as they virtualize their networks. However, in the longer term, we expect that mobile and residential consumer use cases will contribute to 35-40% of global infrastructure edge demand.

The current global pandemic illustrates how unforeseen events can dramatically impact nascent technology markets such as edge computing. The forecasts developed for this report reflect long-range market expectations in the context of what is known today. As we monitor market progress, we will continue to pay particular attention to trigger events and regulations, ecosystem stakeholder priorities and engagement, and the operational and commercial impact of edge computing initiatives along with related technology developments. Edge computing is an exciting market and while it has its challenges, it has a strong foundation that is underpinned by robust use cases with compelling value propositions.

CHAPTER 2

Critical Infrastructure at the Edge

Philbert Shih
Managing Director, Structure Research Ltd.

Edge computing is only possible if the underlying critical infrastructure is performant, redundant and seamlessly integrated. As the old saying goes, a chain is only as strong as its weakest link and this could not be a more fitting commentary for the world of internet-based computing infrastructure. Each and every piece of the puzzle has a role to play and if one part breaks down, things can go wrong in a hurry. At the edge, critical infrastructure is diverse, vendor-neutral and comes in various form factors. Each layer is built on the other and forms a synergistic relationship. From the underlying wholesale data center to the cloud infrastructure that it houses; to the multiple sources of connectivity that connects end users and moves data from the core to the edge; to the real estate that is able to support all these complex requirements. At the edge, each critical infrastructure building block is important in and of itself, but they work together as part of a single integrated ecosystem.

DATA CENTERS

The explosive global growth of public cloud infrastructure and the emergence of edge computing have shifted the third-party data center market. Wholesale data centers that were built to handle the needs of traditional enterprises are simply not equipped to handle the next-level requirements of edge computing and public clouds. As a result, new types of critical infrastructure have emerged to service these needs.

Rack Density

Almost everything important to data center design comes down to three elements: space, power and cooling. The industry's journey has been to pack increasing amounts of power into ever smaller spaces without overwhelming the cooling systems, which themselves have become increasingly more efficient. The end result of all these optimizations has been to increase the density of IT equipment a data center is capable of handling.

The "rack" in "rack density" refers to a standard 7-foot equipment rack, the kind found in virtually every data center around the world, while "density" refers to how much equipment can be packed into that rack. Rack density, then, becomes a way of expressing the power density of a data center and this concept is being extended to new form factors.

For comparison purposes, a typical enterprise data center might average 6-12 kW per rack, whereas some hyperscale data centers can handle densities upwards of 50 kW per rack. A data center with 50 kW rack density can deliver four times the power of a typical enterprise data center on a rack-by-rack basis.

At the edge, rack density becomes important because space is scarce and expensive. The higher the density, the more we can do at the edge.

Hyperscale data centers are often built in a campus-style format and serve a handful of customers. A typical campus environment consists of several multi-MW data halls that are tailored to the growth needs of hyperscale customers. These data centers tend to be newer and purpose-built to handle not just the massive capacities required, but also the power densities required to perform increasingly complex cloud computing functions. Examples of hyperscale data center providers include Compass Data Centers, CyrusOne, Digital Realty, Equinix and Vantage Data Centers. Hyperscale data centers continue to grow rapidly as the adoption

The Many Flavors of Edge

My main perspective on the edge is a bit controversial: latency is perhaps not the be-all and end-all we thought it was initially. It's actually factors like data sovereignty, security, control and interconnection that are more important than people thought they were a few years ago.

Something I'm seeing is the slow convergence of views of different edge worlds — from the cloud and data center world, the telecom world, and up to a point in the IoT and the device world — although I still find people talk across each other quite a lot.

When I start a conversation about the edge, I always calibrate where people are on the scale of things. Some think of the edge as a megawatt data center in a Tier 3 city. Other people think the edge is a milliwatt processor on a sensor. And there's another bifurcation. For some use cases, microseconds matter. For many others, as long as this year's latency is better than last year's latency, that's good. I think the edge has maybe nine orders of magnitude in both latency time and power, about all of which people say, "That's the edge." Different magnitudes of edge apply in different conversations.

If there is one thing you can count on, it's that the edge is going to be messy. Everything is going to be deeply inelegant based on a pragmatism that's messed up by acquisitions and awkwardness around the physical world: issues such as planning constraints from local authorities, overlapping jurisdictions, property rights and existing incumbency.

*For more of Dean's thoughts on this topic, **catch his interview** with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

POST OFFICE Postcards from the Edge experts from around the world 2021 AIR MAIL ★★★

Dean Bubley
Founder and Director,
Disruptive Analysis
@disruptivedean

of cloud computing shows no signs of slowing down. This development is important from an edge perspective because computing at the edge will need to integrate with existing hyperscale cloud infrastructure in what will be a symbiotic relationship. Application functions and architectures are becoming more decentralized and will split between the core and edge.

Regional data centers have existed for some time, but are gaining prominence as cloud infrastructure moves out to the edge. These facilities are frequently found in second-tier markets that are generally underserved from an infrastructure perspective. Regional data centers rarely attain the size and density of hyperscale data centers, but their location in second tier metros makes them convenient to deliver infrastructure in smaller increments. Examples of providers in this category include DataBank, Flexential and TierPoint.

Interconnection-oriented data centers offer data center capacity that is also equipped with extensive interconnection capabilities. Providers like Cologix, Equinix and QTS house dense ecosystems of networks and cloud on-ramps that customers want to connect through. The greater the density of networks and clouds available through an interconnection fabric, the more options and flexibility become available to end users to create hybrid environments or to build between the core and edge. Interconnection-oriented data centers are found in markets of all sizes, but are most common in first- and second-tier markets. Interconnection will enable the core and edge to work in tandem and will underpin distributed architectures.

Micro-modular edge data centers are smaller facilities that range in size from street-side cabinets to cargo container-like structures that house limited quantities of server infrastructure. By being built in a factory and having a smaller form factor than typical data centers, micro-modular data centers are relatively easy to move and deploy, which makes them a good fit for housing IT infrastructure at the edge. These special-purpose edge data centers are small enough to fit in atypical data center locations, such as on the rooftops of commercial buildings, in parking lots, in business parks, on university campuses, on the real estate surrounding wireless towers or at the cross-sections of fiber routes. Their size also makes it possible to deploy them quickly across multiple locations to create, for example, multiple availability and failover zones in a major metro or to support the latency-sensitive services, such as Open RAN (O-RAN). In the last year, micro data centers have moved from concept to reality. Operators like EdgeMicro, EdgePresence and VaporIO have deployed with partners at the base of wireless towers or near high-traffic locations in major population clusters like sports stadiums. The common thread is proximity to wireless and terrestrial connectivity, resulting in maximum performance and the ability to backhaul traffic to the core.

While the data center industry continues to evolve and mature, one of the most profound shifts is from single-site redundancy to multi-site high-availability. While there has always been meaningful separation (across all these different types of data centers) between premium facilities and those that come up short, there is a new recognition that the 2n+1 redundancy that is common in large-scale facilities is both cost- and space-prohibitive in edge environments. Instead, edge data centers are beginning to rely on real-time monitoring and software-based failover to maintain reliability. Software systems like Kubernetes have been designed to manage and scale container-based applications across disparate environments, including the edge. Software-based resilience combined with critical infrastructure ensures that if there is an issue, operations will transition seamlessly. It is about building and designing so that downtime is not just minimized, but mostly factored out of the equation altogether.

Streetside Cabinets represent the smallest data center form factor at the edge. These cabinets are designed to hold anywhere from a quarter-rack to two full racks of modern data center equipment in highly-remote locations. Streetside cabinets are easier to deploy than micro-modular or full-sized data centers because they are less capital-intensive, have fewer requirements for permitting, and can often use available power feeds already available at the site.

WIRELESS TOWERS AND CABLE HEADENDS

Wireless towers and cable headends have quickly become key infrastructure components of the edge computing ecosystem. These facilities provide a crucial point of connectivity, as they are the "jumping off" point for the last mile networks. Because of that, the real estate that surrounds these facilities is an ideal place for edge compute infrastructure to reside. Placing compute infrastructure in close proximity to a wireless tower or cable headend can minimize the distance travelled between the last mile network and the primary processing functions of an application that is housed in an edge data center on the same property. The end result of computing at the edge in this fashion is significantly reduced latency and improved performance.

Wireless towers, cable headends and similar facilities are basically the front line of next generation application architectures. End-to-end applications will perform real-time processes at the edge and connect back to the core for less performance-sensitive functions such as storage, archiving and analysis, which themselves are executed in the core.

The growing importance of real estate in close proximity to wireless towers, cable headends, and key fiber aggregation points, brings a new type of player into the ecosystem: landowners. Similar to how landlords built wholesale data centers on their land to house early generation hosting and cloud infrastructure providers within their walls, landowners are providing strategic real estate for edge compute infrastructure, just in smaller increments and in a vastly greater number of places (when the edge ecosystem is more mature). Landowners provide the real estate, micro data center operators provide the colocation facilities, and then cloud providers, system operators and end users provide the IT equipment and applications that operate in those facilities.

Edge facilities will also create new types of interconnection and peering. Similar to how data centers became destinations and meeting points for networks, the micro data centers at wireless towers and cable headends that will power edge computing often sit at the crossroads of terrestrial connectivity paths. These locations will become centers of gravity for local interconnection and edge exchange, creating new and newly efficient paths for data.

TERRESTRIAL (FIBER) CONNECTIVITY

Terrestrial connectivity is also a pillar of the critical infrastructure value chain, along with data centers and edge facilities. Edge sites are being set up in locations where access to fiber is convenient and can route traffic from the edge to the core and back. This is a critical piece to the puzzle that is bringing yet more service providers and vendors into the ecosystem. Similar to how the carrier-neutral data center provides interconnection, edge data centers are doing the same. The main difference is the complexity in identifying and procuring the most optimal fiber paths. In a highly centralized public cloud world, there are fewer pathways to deal with. In the edge world, multiple locations with smaller increments of compute and storage will require many more routes as edge moves out not just to underserved markets, but other new and emerging markets around the world and in regions that are relatively underdeveloped from a connectivity perspective.

NON-TERRESTRIAL (SATELLITE) CONNECTIVITY

Non-terrestrial connectivity is another piece of this emerging value chain. Satellites orbiting the earth capture and store data such as weather patterns and natural disaster information that can be retrieved through ground station antennas. This data then traverses networks on the ground and ends up being processed and analyzed by application functions that run on hyperscale and edge computing infrastructure. Satellite-based connectivity integrated with cloud infrastructure is in its infancy, but speaks to what the future is going to look like. Computing functions will be split between the edge and core and across multiple locations, with data and end users connecting from multiple sources of connectivity: wireless, terrestrial and non-terrestrial.

NETWORKING AND INTERCONNECTION

With copious amounts of data being generated at the edge, we need new networking capabilities and infrastructure to support the capacity and latency needs of edge applications. Traditional networking service providers, such as Lumen and Verizon, are expanding network capacity at the edge, while startups such as Vapor IO, are implementing next generation networks, edge exchange points, and creating new, more direct network routes.

Interconnections, the physical and virtual linking of networks, are like the connective tissue of the internet infrastructure ecosystem. Interconnection services are available through third party data center operators and enable customers to integrate infrastructure environments — whether on-premises or in public, private and bare metal clouds — both within and across data centers. The rise of public cloud infrastructure has increased the criticality of interconnection services because growing numbers of organizations want to take advantage of the scale and comprehensive tool sets that cloud is able to deliver. They are not necessarily moving all-in on cloud, but they see the advantage in connecting and scaling with cloud for specific workloads. Interconnection fabrics that reside in data centers now have on-ramps to multiple cloud infrastructure platforms, on the same campus or to private network connections that get there very efficiently.

The rise of edge is going to require interconnection to move from its traditional centralized Internet Exchange (IX) model, typically in primary locations within major metros, to an edge exchange model. End users and devices out at the edge are far away from primary IX points and the distance it takes for traffic to travel to these locations degrades performance and also increases transport costs significantly. To solve this problem, interconnection of networks will need to happen in edge data centers near the last mile network in much closer proximity to the end user. We will see edge exchanges emerge to allow peering and data sharing at the edge without necessarily involving the core. Edge interconnection does not operate in a silo. Edge interconnection will allow more traffic to remain local, but it will also become an interdependent extension of the traditional IX. Critical infrastructure at the edge will require not just data centers, but the exchange of traffic seen in centralized cloud computing deployments. And applications will be built accordingly. Applications at the edge will see functions split between the edge (real-time processing) and the core (primary application functions, analysis, archiving) and the data will move back and forth through interconnection services. Having an exchange point at the core that is also tied to the edge will enhance performance and drive cost efficiencies.

SERVICE PROVIDERS

The final component of the critical infrastructure ecosystem supporting edge computing is the set of infrastructure service providers. There are three primary groups: data center operators, last mile operators, and cloud and edge computing service providers.

This third group procures, deploys and leases the server infrastructure in edge locations, whether in a small

Three Parts Coverage

Applications need a robust data infrastructure and dependable data layer. At the edge, the problem is how to make this data layer reliable in a fully distributed way, across potentially hundreds of locations. I see three parts of the industry that are rapidly maturing in parallel, converging towards a singularity that’s going to create an explosion of value.

The first part is capital. Capital is getting much smarter about this problem and why it matters. Capital is being deployed across the infrastructure layer as well as the smart software stacks that we need. Capital is picking platforms that are comprehensively addressing both the data as well as the compute side of the problem.

The second part is customers. They’re also getting smarter. When we started talking about the promise of the edge, customers scratched their heads because they thought all they had was a latency problem. There has been a shift, and customers now recognize there are other issues costing them real dollars and cents, and they need to meaningfully improve performance.

The third part is the ugly secret about the cloud: it’s easy to get in, but it’s really hard to get out. If you haven’t built an application with scalability in mind, something like sloppy coding costs a lot of money in the cloud. The edge is a place where you can absorb and process certain things at a lower cost point than in the cloud.

*For more of Chetan’s thoughts on this topic, **catch his interview** with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

Chetan Venkatesh
CEO and Co-founder,
Macrometa Corp

increment like a micro data center or a more traditional facility, but in a strategic edge location. Many of these computing service providers have poured resources into the development of advanced software management platforms. This enables edge computing services to be delivered on-demand and remotely just like with cloud computing. Computing service providers are the most common tenants in early generation edge deployments, which is not a coincidence: they are the ones working directly with end user customers and ultimately hosting and managing the workloads and applications that are being placed at the edge. Edge compute service providers come in different shapes and sizes. There are startups moving into the edge compute game (like Vapor IO, EdgePresence, Packet, acquired by Equinix, or Ridge). There are

Content Delivery Networks (CDNs) that already have a data center footprint and are looking to convert some of their capacity into edge computing infrastructure (e.g., Fastly, Limelight Networks). Established cloud and hosting providers are also moving into edge computing (such as DediPath and Hivelocity).

PARTNERSHIPS

At the end of the day, it is partnerships that will bring all these disparate pieces together. Wireless tower operators have considered owning data centers (and in a few cases acquired them), but the most efficient way for them to get to market is to partner. Micro data center vendors are partnering with the service providers that host and manage cloud infrastructure and bringing them on as tenants. Meanwhile, micro data center operators are working with wireless tower companies to access the real estate at tower sites. Real live deployments were brought online in 2020 even amid a pandemic.

Carrier neutrality within all these critical infrastructure pieces is paramount and the only way edge is going to work. It ensures the diversity and choice that maximizes the number of pathways, locations, performance levels and service provider options that different types of application requirements will demand.

Partnerships within the edge ecosystem will not traverse these three segments exclusively. There will be other combinations. The public cloud is not going to lead build out of edge infrastructure, but they will want to ensure they are part of the value chain, being the core that is tightly integrated with the edge, as well as shipping software to run on edge infrastructure. This thinking brought AWS to partner with micro data center provider Vapor IO along with Crown Castle, which has access to an extensive wireless and terrestrial fiber footprint. They are also working with telco providers to access and place edge infrastructure at 5G network aggregation points. Expect more partnerships and ecosystem development in the coming years. The pieces are varied and require high degrees of coordination.

RETURN ON INVESTMENT

The build out of edge computing infrastructure will require significant amounts of supporting capital. Currently, capital is moving quite conservatively, but is making bets on micro data center operators that are building the new form factors and regional edge data center providers that represent the sector's more immediate upside as edge momentum builds. It makes sense for capital to bet on these foundational elements. They require significant CAPEX to build out, but will stabilize and generate meaningful returns once the sector starts to emerge. Capital is also supporting service providers that are building the management platforms that will make edge computing possible. This is a significant opportunity, but with a wider range of possible outcomes. Given the upside, money will continue to move in this direction as the sector flourishes and the strategic value of edge computing platforms become clear to a large universe of strategic entities that will want to invest and ultimately acquire assets.

The state of critical infrastructure in the edge ecosystem can be best described as fluid. Some of the categories are relatively new and there is experimentation with new technologies in the areas of liquid cooling, floating data centers and satellite communications delivery. Because of the links with cloud computing, there is already a foundation to build off. And this makes progress all the more feasible. In many respects, it is not about re-inventing the wheel, but adjusting familiar models for new requirements and scenarios. The challenge will be how to take all these disparate pieces and piece them together. The edge computing ecosystem is still developing and coordination and partnership will determine how fast this will push forward.

CHAPTER 3

Edge Hardware

Mary Branscombe
Freelance Technology Writer

Edge computing comprises combinations of systems that span a wide range of locations and conditions, and support a diverse set of use cases. While one use case might demand high-powered GPUs for AI (Artificial Intelligence), another use case might demand low power consumption to lengthen battery life. The location of equipment, such as a micro edge data center or a wall-mounted industrial cabinet, places different constraints on the hardware. All of these factors result in a wide range of edge hardware that will continue to diversify in 2021.

Sharpening the Edge: The LF Edge Taxonomy and Framework.

Intended for readers interested in both the technical and business aspects of edge computing, a recent white paper introduces a set of open source software projects hosted by the Linux Foundation (LF) and its subsidiary organization LF Edge. It outlines the LF Edge taxonomy and framework and describes opportunities for companies to participate in and benefit from these projects, accelerating the development, deployment and monetization of edge compute applications.

KEY TRENDS IN SERVERS AND PROCESSOR PLATFORMS

Hardware deployed at the edge has historically been purpose-built for specific workloads, frequently CDNs (Content Delivery Networks) or IoT. As edge computing grows in popularity and new use cases emerge, general purpose infrastructure is also being deployed to run cloud-like workloads. IDC predicts that by 2023 edge networks will represent over 60% of all deployed cloud infrastructure[1]. Adding to the trends that are already driving edge growth, the impact of the pandemic on workforce and operations practices will continue to accelerate the delivery of infrastructure, application and data resources in edge locations through 2021 and into the following years.

Familiar data center companies and cloud providers will add edge offerings, but form factors will be increasingly diverse. Many of the first at-scale edge deployments systems have been built using micro-modular edge data centers, for example, and those are quickly being augmented by new form factors, such as street-side cabinets and light pole attachments. To address the entire continuum of requirements, particularly at different points within networks where compute is required, edge hardware will vary from full-scale racks in a telco central office to a smart camera on a factory production line or in a warehouse connected over private 5G, or ruggedized for outdoor locations from race tracks to oil rigs, with locations such as factories, offices and even planes and ships turning into micro data centers.

Increasingly, IT and OT (Operational Technology) are converging and this trend is especially visible at the edge.

2021 will bring more variety inside the box as well: Arm server processors, AI processing chips, GPUs, SmartNICs (Smart Network Interface Controllers) and FPGA (Field-Programmable Gate Array) boards will be increasingly common. The wider range of workloads performed at the edge will increase the heterogeneity of hardware, resulting in a wide range of CPUs, as well as new kinds of hardware and network accelerators.

Arm-based silicon has long been common at the edge in IoT devices but Arm's new Neoverse platform includes offerings targeted as server and storage processors, as well as for network hardware. AWS has invested heavily in its Arm-based Graviton instances, which has helped to validate Arm as a general-purpose server processor architecture. AWS has also announced a 1U version of its Outposts hyper converged systems that include Graviton2 processors. Microsoft is working on its own Arm-based hardware for use in its CDN and other edge scenarios, but has yet to bring it outside its own data centers. Apple's recent announcement of its Arm-based M1 Apple Silicon processors is likely to add focus to this area, with the power and performance gains visible to consumers (and providing easy local access to the architecture for millions of software developers), resulting in a better overall understanding of Arm's platform capabilities.

Meanwhile, Intel is pushing Atom, Pentium and Xeon D SoCs created for IoT as competitors to both Arm and AMD in user edge devices, such as for in-camera analytics and real-time inspection for industrial applications. FPGAs, Xeon and Arm cores are all showing up in SmartNICs, which Nvidia is keen to rebrand as DPUs: Data Processing Units that offer CPU offload for IO, storage, security and even virtualization as well as network acceleration, and can be combined with GPU capabilities as a multi-purpose hardware accelerator. There is also the possibility, driven mainly by Chinese vendors, that the open source RISC-V silicon architecture may also have a role to play here.

In recent years, hyperscale cloud providers have turned to FPGAs both for network offload to free up CPU resources that can be sold to customers and, increasingly, for AI acceleration, because the hardware can be reprogrammed to adapt to improvements in Machine Learning (ML) algorithms. Not many organizations have the technical capacity to build and run similar FPGA systems themselves, but as they're packaged into SmartNICs, DPUs or other accelerators they will become more widely accessible. And with power efficiencies that are more like an ASIC (Application-Specific Integrated Circuit) than a GPU, they're well suited to edge computing.

Intel, Nvidia and AMD are also making moves to offer full hardware and software stacks for data centers, including at the edge. Intel is launching its first discrete GPU, alongside IoT-specific hardware, AI acceleration hardware, FPGAs and SmartNICs, Xeon instructions designed to boost inferencing speed on CPUs, and composable network switches based on silicon photonics. All these options are tied together by Intel's oneAPI programming model.

To compete with that vertical stack, AMD and Nvidia are embarking on significant acquisitions. Having already bought Mellanox for its SmartNIC and networking expertise, Nvidia announced the intention to acquire Arm. This would allow Nvidia to offer GPU and tensor core acceleration as IP that can be licensed in addition to offering its own integrated hardware options like the EGX edge AI platform, complete with a SaaS (Software-as-a-Service) Fleet Command control plane.

AMD's purchase of Xilinx will bring it FPGAs, including the Alveo SmartNICs and accelerators that the company has been positioning as competition for Nvidia GPUs. And the other specialist hardware acceleration that's widely incorporated into SoCs with Arm cores isn't going away.

This rise in hardware acceleration reflects the importance of AI and other computationally-intensive workloads at the edge, where data is gathered and needs to be acted on, for any industry that needs to bridge the physical and digital world in real time rather than after the fact. This will only increase as the

Unique Hardware Requirements for the Edge

A lot of edge data may never even reach the cloud. Cargo ships have gone from hundreds of sensors on board to tens of thousands. We will never push all that data to the cloud over a satellite connection. We need cloud-like compute power on the ship that can run machine learning models.

But wait, it gets more fun….

Cargo ships are only in port a day or two at most, at the whim of weather and port traffic. How will you schedule a technical crew to install the hardware? The answer is, you can't. Data center hardware will not work in this edge use case.

Jeffrey Ricker
CEO, Hivecell

amount of data generated grows. In 2018, only 10% of enterprise-generated data was created and processed outside traditional centralized data centers or cloud services. By 2025, 75% of that data is expected to be created at the edge[2], in factories, hospitals, retail stores and cities, with much of it processed, stored, analyzed and acted on at the edge[3].

Over the next few years, hardware at the edge will be endemic rather than unusual. The Precambrian explosion of form factors and silicon architectures will continue for some time yet, but there's a tension between the heterogeneity of hardware appropriate for the diverse workloads that will run at the edge and the desire for enough uniformity to allow developers to build applications that will run in the largest number of locations. Moreover, standardization in form factor will make deployments easier and potentially less costly. Having a robot that could install a new Open19[4] server into a rack when more capacity is required is still some way off, but pre-wired racks that allow power, cooling and networking to be connected at the same time already speed up build out and repair times.

COVID Accelerates Remote Operation.

COVID-19 has only underlined the importance of minimizing the need for onsite expertise in data centers of any type, from core to edge, and this has hastened the development of tools for remote monitoring, provisioning, repair and management, which could greatly reduce the cost of edge computing.

THE IMPACT OF OPEN HARDWARE

Standardization in form factor is an important strategy for making deployments easier and potentially less costly, including finding scalable points of shared infrastructure investment. Similar to how the hardware in hyperscale data centers has been designed specifically around that use case with the Open Compute Project, efforts are underway to design for the physical and operational realities of the infrastructure edge. One approach — pioneered by LinkedIn, Flex, HPe and Vapor and now a project at the Linux Foundation with members such as Equinix and Cisco — is Open19.

Open19 provides open source standards for a form factor that works with standard 19" racks, which are ubiquitous at existing regional data centers and telecom central offices. Open19 features cable-free installation for both power and data by leveraging a "blind mate" cable system. This enables compute infrastructure to be installed and maintained separately from the physical rack, network and power infrastructure. This can dramatically lower operational and maintenance costs in remote field locations, as well as allow for heterogeneous deployments of specialized hardware within a rack or sub-unit of a rack.

Having a robot that could install a new Open19 server when more capacity is required is still some ways off, but pre-wired racks that allow power, cooling and networking to be connected before any compute is installed is already speeding up build out and repair times at early adopters like Equinix.

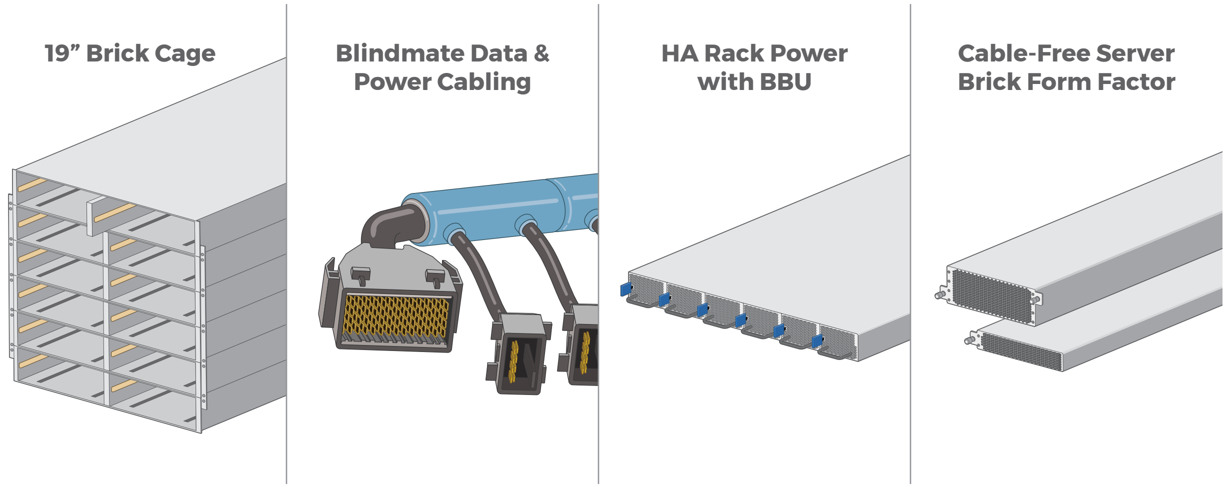

19" Brick Cage
Blindmate Data & Power Cabling
HA Rack Power with BBU
Cable-Free Server Brick Form Factor

FIGURE 6
Open Hardware

INFRASTRUCTURE LIFESPAN AND RELIABILITY

The majority of edge locations will have hardware deployed for a long service life: the 5-7 years of cloud hardware or longer, but with the expectation of even less frequent maintenance and physical servicing. The extreme example is Microsoft's experimental underwater data center, Project Natick, which in 2020 produced encouraging results not just for the expected cooling efficiency but also for the stability and uptime of servers running in a nitrogen atmosphere with no disturbances for maintenance: one-eighth the failure rate of the same servers in a standard Azure data center on land, making a virtue out of a necessity.[5]

Ephemeral Data Flows on the Edge

Simon Crosby
CTO, SWIM.AI
@simoncrosby

A big change I see coming is that we're moving to a world of big data where data flows are boundless and need to be processed on the fly. There are huge problems with this that remain to be solved.

You used to be able to store all of the data and then analyze it later, and now you can't. There's just too much of it. We're in an era where we have data flows that never stop. Businesses must constantly process and analyze streaming data to get continuous intelligence, to make your organization more responsive, or whatever your need happens to be. The cloud has been tremendously successful, but stateless computing is a million times slower than the CPU, which means you're getting results in hours versus milliseconds.

In addition, all this boundless data has mainly ephemeral value. You can't store it then get to it later, because later the data is useless. You don't care about the past. What you care about is using data to predict what's going to happen next. So, we have infinite data with ephemeral value that you need to continually compute on by building a model on the fly from the data. Algorithmically and mathematically, it's a whole new approach.

*For more of Simon's thoughts on this topic, **catch his interview** with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

It's no contradiction that bare metal (using unconfigured hardware and avoiding virtualization) is so popular for edge compute though: it requires a high level of sophistication but with firmware that allows remote autoconfiguration, bare metal offers a combination of performance and ease of delivery. Just as cloud IoT services have simplified development, deployment and management for IoT devices, so unified control planes will be key to leveraging edge compute hardware.

Resilience and disaster recovery will also evolve at the edge. Traditional ways of maximizing data center uptime tend to rely on fully redundant (e.g., 2n+1) mechanical systems, which are often too costly and consume too much space in edge locations. In edge environments, system uptime will be delivered as much by high-availability software and AI-assisted automation as by today's standard of relying mostly on physical redundancy. Software orchestration allows developers to spawn workloads in multiple locations, and high-availability software systems can use real-time and predictive telemetry feeds to route traffic to locations optimally available, as well as to restart services in nearby locations when a failure is detected.

Although each edge resource will be acting semi-autonomously, processing and acting on local data, the facilities and hardware to support a robust edge application will likely be distributed across many edge locations, making it essential to manage collections of edge components as distributed systems rather than individually, and developers will want layers of abstraction to reduce the effort of targeting individual edge locations or specific hardware. New technologies have emerged to assist developers in managing these distributed systems, including purpose-built application stacks, such as those offered by **LF Edge's Akraino**, designed to accommodate conditions at the edge, as well as sources of telemetry for making real time decisions, such as the open source **Synse** API.

This intertwining of hardware and software to deliver the abstractions that make edge resources easier to consume will be tightest as hybrid cloud infrastructure arrives at the edge. Offerings like Google Anthos, AWS Outposts and Snowball Edge, Azure Stack Edge and Azure Private Edge Zone wrap commodity hardware into appliances that promise the same cloud services and control plane as in the public cloud, with a consistent application and operations experience. For many organizations, the choice of edge hardware will be dictated by the cloud services and the cloud native applications they need to bring to the edge, or in some cases by how well that hardware supports private 5G networks.

VIRTUAL NETWORKING

Networking hardware is becoming increasingly software-based. The Software-Defined Networking (SDN) and Network Functions Virtualization (NFV) trends, familiar in data centers, are being driven and adopted by telcos looking to standardize around commodity "white box" server equipment, allowing workloads to be placed in more locations on lower-cost hardware. That includes on customer premises where consolidating multiple proprietary hardware appliances onto a single, general purpose white box device is particularly important. Universal Customer Premise Equipment (uCPE) isn't new but it's becoming less of a prediction and more of a reality. Many existing uCPE devices use Intel processors, but this is a space Arm is targeting heavily.

Trends in I/O for Virtual Networks

Modern cloud networking is one of the points where software and hardware are blending in new ways. Adding compute to NICs allows applications to offload functionality to hardware; for example, embedding security rules at an adapter-level rather than in separate firewalls or other security appliances. Similarly, new technologies like the Project Zipline hardware-based compression tools make SmartNICs key components in at-scale distributed systems, facilitating the transfer of large amounts of data from edge to core. Technologies like DPDK (Data Plane Development Kit), support delivering code to these devices, enabling tools like Service Mesh to interact with hardware, supporting policy-based networking. At the same time, support for SR-IOV (Single Root I/O Virtualization) in recent hardware makes PCIe hardware, including SmartNICs, accessible from virtual machines and containers, sharing hardware resources securely.

Telcos have the networking expertise to take advantage of SmartNICs, which are only now becoming accessible to enterprises as more standardized software support arrives, but they should be well suited to the edge.

Nvidia claims that the power budget for a Data Processing Unit, or DPU, is only incremental over that of the NIC that would have been in the server anyway, so these devices may prove an efficient way of adding acceleration and freeing up CPU cores on edge servers in locations with limited power availability. SmartNICs that can run a network OS like SoNIC can also remove the need for a separate switch in some edge locations, meaning non-technical staff could install a server by plugging in power and Ethernet cables.

STORAGE

Just as SmartNICs offload network-specific processing from the CPU, computational storage is emerging as a way to perform data processing and simple storage-related applications like compression, encryption, backup or search at the exact location where the data is stored. That will have power and performance advantages for data-intensive applications running at the edge (which is, again, where the data is often generated) and as the embedded compute in the storage array becomes more sophisticated, it can support more workloads.

ML training usually reads data from storage and writes the model back into storage. Moving the training algorithm into on-device compute resources that can perform search and aggregation on the storage device can free up the CPU for other work or reduce power requirements by allowing the workload to run on a lower spec device. The lower latency may even improve training performance.

Running inferencing in computational storage could analyze and label video as it's saved, which can be a useful performance booster on devices with no space or power budget for dedicated video processing chips or a GPU.

Computational storage can even run edge versions of cloud services on IoT devices, because it comes in SSD and NVMe (Non-Volatile Memory Express) packaging and fits into devices with no other options for acceleration. With Samsung joining the handful of lesser-known vendors already offering computational storage devices, this technology is poised to become mainstream in the longer term.

While the current trend in edge compute often involves tighter integration of acceleration, hyperscale cloud providers are starting to investigate disaggregated architectures. To reduce the inevitable fragmentation of the familiar multi-tenant approach where compute, storage, networking and memory become a set of composable fabrics, Rack-Scale Architecture (RSA) deploys CPU, GPU, hardware acceleration, RAM, storage and network capacity separately. Resources are then composed dynamically to fit workloads at arbitrary scale; even the components of the motherboard might be modularized and separated. Non-volatile memory models, like NVMe and Intel's Optane, provide an interesting set of components that blend memory-like performance with persistent storage, allowing high performance operations where power reliability may be an issue. With the addition of NVMe over TCP standards to the Linux kernel, disaggregation of compute, RAM and storage is becoming more compelling and reliable. Space and other constraints make these technologies especially interesting for edge environments.

The silicon photonics required to connect components at low enough latency for more extreme disaggregated architectures will reach the edge as part of 5G infrastructure. So, connectivity permitting, in the long run some degree of disaggregation in future generations of edge hardware could allow truly distributed workloads to take full advantage of the diverse nature of edge compute and storage.

PLACEMENT TRADE-OFFS

Hardware considerations at the edge are often about getting the best power-vs.-performance trade-offs for the compute required for the workload. As more edge resources are deployed and workloads become more distributed across the edge or between edge and cloud, there will be a variety of trade-offs to consider, like placing a particular workload close to users or devices to minimize latency, or which operations should run at the edge to preserve the battery life of a consumer device used for gaming or AR (Augmented Reality). Again, this kind of dynamic scheduling will rely on software orchestration that's still being developed, and also on accurate metrics about latency, network conditions and power efficiency for determining where it's most efficient to run a particular computation. This will also take into account which operations are the most latency sensitive and which can safely run several milliseconds away without degrading the performance of what may well be critical infrastructure.

Rebooting Businesses at the Edge

Edge computing promises to drive new experiences through ultra-low latency services, however the 2020 pandemic forced many organizations to focus on critical business needs.

Still, we saw a number of customers adopting edge technologies to both get their businesses back online and improve resilience through remote orchestration of autonomous field operations. Deployments included solutions to remotely monitor oil well heads and securely extract and pre-process factory data to ramp production of PPE equipment.

Key for customers was investing in flexible, open architecture that enabled them to rapidly spin up remote orchestration of existing applications while also setting the stage for deploying new cloud-native innovations at the edge over time.

For more of Jason's thoughts on this topic, [catch his interview]() with Matt Trifiro on the Over the Edge podcast, from which this was adapted.

2021 AIR MAIL

Jason Shepherd

Governing Board Chair,

LF Edge and VP Ecosystem,

ZEDEDA

CHAPTER 4

Networks and Networking

Phil Marshall
Chief Research Officer, Tolaga Research

A Few Key Terms

The edge computing industry uses many terms of art when discussing edge networks. Some of these are misinterpreted or misunderstood. Here are a few of the most common terms that practitioners need to understand:

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. Typically measured in milliseconds, latency represents the network delay that data will incur at single or repeated points in time between two or more endpoints. Latency is a key metric for optimizing edge applications and for understanding the responsiveness of a network. Latency is distinct from jitter, which refers to the variation of latency over time. Network data cannot travel faster than the speed of light which increases latency over long distances, hence the edge networking proclivity for bringing the compute closer to the data.

NETWORK HOP

As data packets move from one network segment to another, they typically pass through a device, such as a router or a switch. The device might also translate one physical media to another, as when a wireless radio receiver converts radio waves to fiber optics. Because these network devices often have buffers and involve processing, they tend to introduce latency and jitter into the data transmission. As a general rule: The more hops, the higher the latency and the greater the jitter.

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.

TAIL LATENCY

The small percentage of network data transmission latency that is much larger than the average. Tail latency is measured in terms of percentile, with a larger percentile representing the smaller proportion of more extreme latency conditions. 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.

Some of the most interesting use cases, especially those that have life-safety implications, have more stringent performance requirements than today’s widely deployed networks can accommodate. Notable examples include autonomous driving solutions, real-time command-and-control functions for healthcare (e.g., remote robotic surgery) and enhanced emergency response and public safety solutions. The critical network infrastructure for these use cases requires new levels of reliability, such those provided by hardware and software "hardening," redundant connectivity, high-availability programming techniques and improved utility backup. These added infrastructure requirements can be costly, particularly when implemented at scale.

Low-latency networks are an essential part of the edge computing infrastructure, and must be implemented with sufficient density and local proximity so that network connections are terminated close to the end-point devices. As general-purpose edge networks emerge, they will reshape how the internet routes data. For example, as more data is created locally and kept locally, interconnection density will explode at the edge, the public and private internet backbones will extend to the edge, peering and data exchange will happen within a hop or two of the access network, which is also where the new generation of edge CDNs and edge cloud systems will operate.

HOW FAST IS FAST?

The analysis of performance starts with the laws of physics which dictate that signals will travel at the speed of light as radio waves through the air and at about 70% of the speed of light as optical signals through fiber optic cables. Moreover, every time the signal passes through a router, switch or any other type of networking equipment (typically referred to as a "network hop"), there is an increased likelihood of introducing unpredictable delays, often called "jitter". The amount of jitter, latency and tail latency on any given network route significantly impact whether or not a particular application can be delivered.

But challenges in edge networking go beyond latency and jitter. For example, in order for a wireless access network to take advantage of edge computing, peering and exchange points, the wireless network must be capable of terminating data connections locally. This is known as local breakout and can vary greatly in complexity amongst geographical regions and between different technologies. For example, in some countries there are no existing domestic peering points between service providers for competitive reasons. Mobile service provider engagement and advanced control/user plane separation with technologies like CUPS[6] (Control and User Plane Separation) are needed to achieve local breakout for mobile services. Furthermore, existing technologies like 4G-LTE can be used with some edge computing applications but require a 100ms preamble for connection synchronization, which determines the minimum latency that can be achieved. A key innovation for 5G is to enable a flexible preamble to address the need for lower-latency connections. Moreover, as edge computing demands grow, the existing backhaul networks may be incapable of handling the speed and capacity demands they generated, which will require middle-mile network providers to upgrade their networks so they can add new, faster and higher-capacity routes with more discrete predictability.

THE NETWORK ECOSYSTEM

Edge networks are constructed using both fixed and wireless connections, as well as both public and private solutions. In some cases, particularly for on-premises implementations, new and upgraded networking equipment, such as universal Customer Premise Equipment (uCPE), will be needed to capitalize on current and emerging edge devices. Moreover, new equipment may be required to support edge services, including content and application distribution networks (CDN/ADN) in addition to cloud-enhanced and edge-native services.

As edge computing continues to flourish, network bandwidth and performance demands will grow significantly both in terms of the sheer capacity needed and the diversity of service requirements that must be supported. Network capacity and performance challenges are not new and have driven continuous technology innovation, which in the last decade has seen a migration towards virtualized and containerized network architectures.

Incentives for Edge-Native Applications

Mahadev Satyanarayanan (Satya)
Carnegie Group Professor of Computer Science

How much end-to-end latency is acceptable is very much a function of the application. But it's a two-way street: the applications that get written depend on what today's technology can offer. If the application's demand gets too far ahead of the technology, then the application will die because it's not viable. It's crucial that technology and applications do this dance. One gets a little ahead of the other briefly, the other catches up then slips further ahead, and the process repeats. It's a virtuous cycle where each influences the other.

Edge-native applications are those that are so dependent on the edge that they simply do not work without edge computing. One type of edge-native application being developed is called wearable cognitive assistance, essentially combining augmented reality and AI. I'm confident that in five years there will be use cases of these applications as a way to scale out expertise in industrial troubleshooting.

To accelerate this timeframe, let's push the venture capital domino. The return on investment that incentivizes the creation of edge-native applications will have a long-term payoff. The ultimate beneficiaries of edge computing should make strategic investments and apply different success criteria from what is traditionally applied by venture capitalists. The investment will be valuable even without a hockey stick growth curve, because you are building long-term demand for the core product you're creating, which is edge computing itself.

For more of Satya's thoughts on this topic, ***catch his interview*** *with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

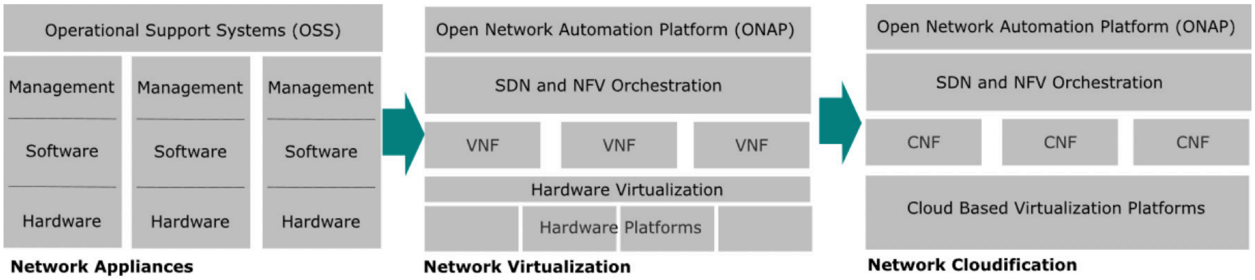


FIGURE 7

Evolving Network Architectures for Performance and Agility

Historically, networks have been built using dedicated network appliances with vertically-integrated hardware, software and management functions. For the most part, network devices have consisted of tightly coupled hardware and software which required vertical integration to deliver the necessary performance. However, advancements in network virtualization techniques and cloud technologies are enabling enterprises and service providers to migrate their network functions off of dedicated appliances and onto modern "white box" servers to deliver increased network performance and agility.

Key Terms in Network Virtualization

As cloud computing took over data centers, including telecom infrastructure, Virtual Network Functions (VNFs) became the accepted way to run network workloads in cloud environments. While VNFs are equipped to run in virtualized and cloud environments, they are not necessarily fine-tuned for the cloud. With better understanding of how to optimize network workloads in cloud environments, the focus has shifted towards design principles that are represented by Cloud Native Network Functions, regardless of what technology is chosen (containers, virtual machines or a hybrid of these) to package and deliver them..

CLOUD NATIVE NETWORK FUNCTIONS (CNFs)

A cloud native network function (CNF) is a cloud native application that implements network functionality. A CNF consists of one or more microservices and has been developed using Cloud Native Principles including immutable infrastructure, declarative APIs and a "repeatable deployment process".

NETWORK FUNCTIONS VIRTUALIZATION (NFV)

According to ETSI, NFV is the principle of separating network functions from the hardware they run on by using virtual hardware abstraction.

VIRTUALIZED NETWORK FUNCTIONS (VNF)

According to ETSI, VNF is an implementation of a Network Function (NF) that can be deployed on a Network Function Virtualization Infrastructure (NFVI).

Network virtualization includes several key technologies and principles, such as Software Defined Networking (SDN) and Network Function Virtualization (NFV). SDN and NFV provide separate network control and packet forwarding capabilities for VNFs, which can be implemented via VMs or containers and orchestrated with cloud native platforms, public and private. Network management and orchestration capabilities, such as the Open Network Automation Platform (ONAP), are needed to contend with the unique operational requirements for virtualized and cloudified networks.

Virtualized and cloudified network technologies have matured in recent years, buoyed by their ability to deliver tremendous financial gains. For example, in September 2019, AT&T's CEO, Randall Stephenson, commented that his company had at that time seen 17 quarters of year-over-year cost savings between 7% and 8%, with 75% of its core network having been virtualized. Cost savings through virtualization are enabling enterprises and service providers to improve the economics and flexibility of their networks as they scale to support a wider range of edge use cases.

NEXT-GENERATION SD-WAN: A CRITICAL BUILDING-BLOCK FOR EDGE

As edge computing solutions proliferate, they need agile connectivity to efficiently adapt to different operating environments and service demands. This is particularly true where edge solutions depend on Wide Area Network (WAN) connectivity across heterogeneous network environments. For example, mobile services, such as those associated with autonomous vehicles and mobile gaming, must perform reliably over WAN environments. Software-Defined Wide Area Networking (SD-WAN) solutions enable the agile management of network resources so that they can be provisioned when and where they are needed.

The control and data plane separation enabled with SDN has spawned SD-WANs, to enable virtualized network overlays with the capability of managing and orchestrating end-to-end WAN connectivity that leverages multiple available network technologies.

As a technology, SD-WAN was developed initially with a focus towards enabling enterprises to offload MPLS network traffic onto lower-cost broadband connections, ensuring sufficiently reliable end-to-end IP packet performance. Legacy SD-WAN solutions have served enterprises well over the years, but have several limitations including complicated OA&M (Operations, Administration and Management) and security management requirements for large scale implementations.

Next-generation SD-WAN solutions overcome many of the shortcomings of their predecessors. These solutions are particularly important for cloud and edge computing because of the magnitude and variety of services that must be supported at scale. The next generation SD-WAN solutions incorporate OA&M automation capabilities to enable SD-WAN implementations to scale without incurring massive operational costs. These automation capabilities leverage a variety of technologies including AI and ML to optimize network performance and reduce the need for human intervention. Next-generation SD-WANs are also shifting up the stack. No longer content to focus on end-to-end performance at the networking layer (OSI Layer 3), newer SD-WANS also offer functionality at the application layer (OSI Layer 7). Managing and orchestrating SD-WAN connectivity at the application layer will provide a new level of agility and efficiency to support the vast range of edge services and operating environments.

Since SD-WAN solutions have traditionally focused on network connectivity, security policies are typically implemented with vendor-specific point solutions. These solutions create scalability challenges and potential security vulnerabilities, particularly for large implementations. This has seen the development of holistic and cloud native security regimes for next generation SD-WANs, and the concept of the Secure Access Service Edge (SASE) framework, which was first described in a **Gartner research report** in 2019. In contrast to

Beyond the Cloud to the Edge

POST OFFICE Postcards from the Edge experts from around the world
2021
AIR MAIL

Victor Bahl, PhD
Technical Fellow & CTO
Azure for Operators,
Microsoft
@SuperBahl

As someone who incessantly thinks about how to evolve technology and where we should be heading, one question that I have always thought about is “what happens after the cloud?” We answered this question by conceiving edge computing over a dozen years ago, now I call that — classic edge computing. We cut down the latency by placing computers close to where the data is being generated. The newer direction of edge computing is about making computing part of the networking infrastructure fabric. Edge computers that make things like autonomous driving, AR and VR experiences, fast action cloud gaming, IoT analytics, live video analytics and many more such applications possible will now also provide core networking services. For example, the signal processing functions that are part of the cellular radio access network, routing and forwarding mechanisms that makes machine-to-machine communications possible and network functions that implement the cellular mobile core network will all co-exist on the same servers as the applications I just talked about. Overall, edge and cloud will form a computing continuum with the edge doing much more than previously envisioned. I believe this is where we are headed.

The pace of innovation is incredibly fast, but it takes time for paradigm shifting ideas to seep in especially when they’re possibly disruptive. The cloud was a big idea. Edge is an equally big and perhaps an even bigger idea. It’s about enabling ubiquitous low-latency computing. Thankfully, I no longer have to convince people on the need for edge computing. These days, what I see is that most business-minded people are thinking hard about how best to monetize the edge — how to light up these awesome new services.

As I look around, I notice many edge computing companies emerge and they are developing impressive products. I think over the next few years, the pace of product innovation will increase. Edge is not only here, it has momentum and it’s on its way to becoming pervasive.

For more of Victor’s thoughts on this topic, ***catch his interview*** *with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

SD-WAN, SASE has a cloud native design and integrates networking and security capabilities into a unified architecture by managing the connections between individual endpoints and service edge nodes. The service edge nodes are provisioned in edge and cloud data centers with SASE software stacks. With this approach, SASE simplifies end-to-end security for edge enabled services.

EDGE NETWORKS, EDGE EXCHANGES, INTERCONNECTION

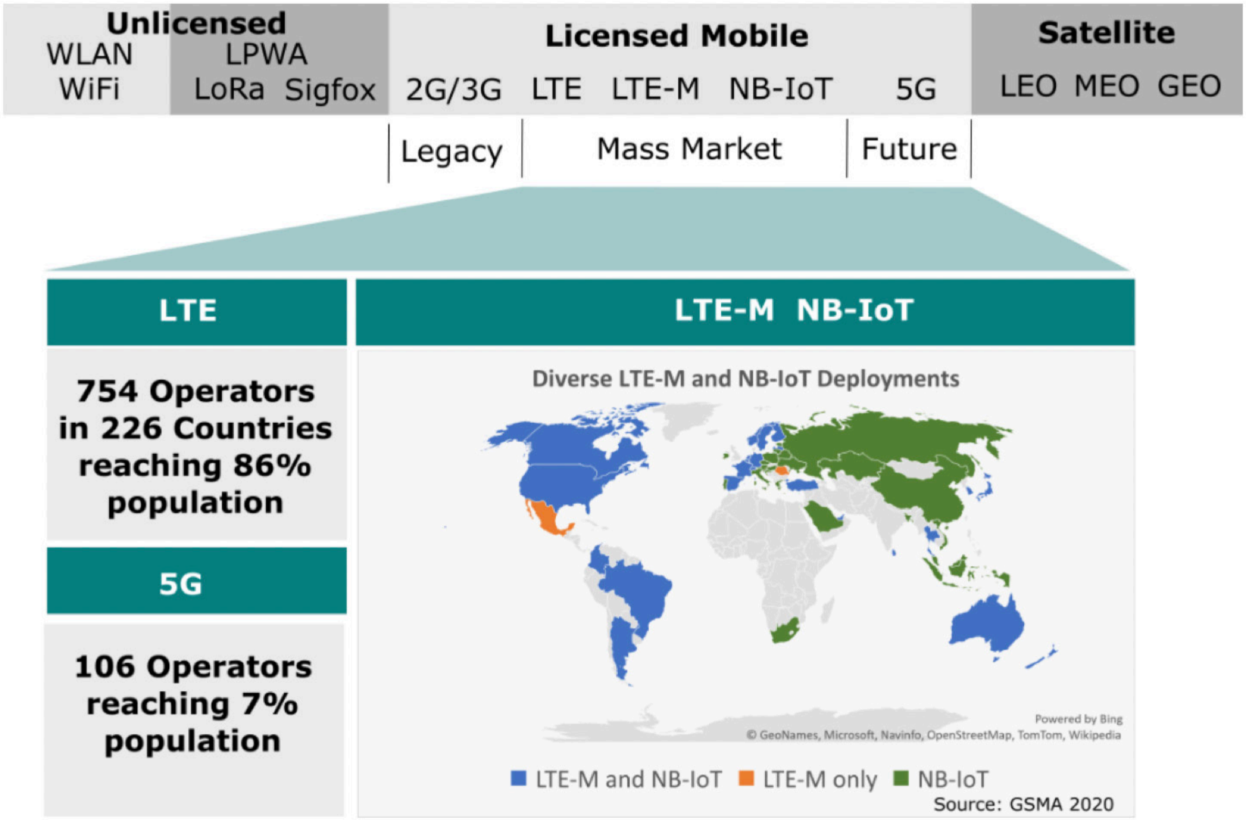
Edge computing is making new demands on networking infrastructure and dedicated edge networks are emerging to service those needs. Legacy backhaul networks are already overloaded and newer backend services may need to migrate from one edge node to another to remain sufficiently proximate to their data and devices, particularly if those devices are in motion. At the edge, network routes are monitored for latency and congestion and can be dynamically reconfigured to support QoS goals. Edge networks are becoming increasingly agile at delivering services that adapt to actual application and network conditions in real time, often by incorporating the data center and cloud technologies, such as VXLANs (Virtual Extensible LANs), which allow for the creation of dynamic Layer 2 networks. A dynamic Layer 2 network can ease the live migration of backend services from one edge server to another — from one Layer 2 switch to another — while maintaining the same IP address, allowing for a continuity of service that would be otherwise difficult or impossible.

The ability for two networks to meet and exchange data has always been an essential part of the internet and the delivery of cloud services. The global growth of internet exchange points — those places on the planet where networks converge and exchange data — has been one of the largest contributors to the increase in the internet's efficiency, both in terms of cost and speed, as well as its ability to scale. As edge computing grows in demand, there will be a corresponding growth in demand for new network routes as well as need for more locations to exchange data in close proximity to the edge. Today, a city might have only a few locations where networks converge to exchange data, such as at regional IX points and carrier hotels. As edge computing infrastructure gets built out, many of the facilities will be used as edge exchanges to cross-connect networks. Networks will converge on these locations to support edge services that benefit from low-latency exchange, shorter fiber runs and fewer network hops. As computing gets more distributed at the edge, so will network cross-connections.

THE WIRELESS EDGE

Edge computing networks depend on a wide variety of wireless technologies to provide connectivity to untethered devices that are being deployed at the edge. Best effort edge services are supported by unlicensed wireless technologies like WiFi and Low Power Wireless Access (LPWA). WiFi delivers broadband capabilities over local area environments and is being continually upgraded with recent versions that include WiFi6 and 802.11be. LPWA provides wide area coverage in unlicensed spectrum, albeit with only narrowband capabilities. Meanwhile a growing number of edge services, particularly those that are edge-native, such as collaborative autonomous vehicles and mobile immersive gaming, depend on superior network availability, reliability, bandwidth and latency performance that might exceed the capabilities of unlicensed spectrum technologies.

Satellite connectivity is commonly used as a backbone network for edge computing services that require global geographical coverage, such as for maritime or oil platform applications. Satellite performance has increased significantly in recent years with the advent of High-Throughput Satellite (HTS) technology. Satellite constellations are placed in Low, Medium and Geostationary Earth Orbits (LEO, MEO, GEO). Higher orbiting GEO satellites have greater geographical footprints and require a reduced number of satellites for global coverage, but experience larger connection latencies.


Unlicensed
WLAN
WiFi
LPWA
LoRa Sigfox
Licensed Mobile
2G/3G LTE LTE-M NB-IoT 5G
Satellite
LEO MEO GEO
Legacy
Mass Market
Future
LTE
754 Operators in 226 Countries reaching 86% population
5G
106 Operators reaching 7% population
LTE-M NB-IoT
Diverse LTE-M and NB-IoT Deployments
Powered by Bing
© GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom, Wikipedia
🟦 LTE-M and NB-IoT 🟧 LTE-M only 🟩 NB-IoT

Source: GSMA 2020

FIGURE 8

Typical Spectrum Usage for Wireless Connectivity

Licensed spectrum technologies such as 4G-LTE, LTE-M (Long Term Evolution category M1), NB-IoT (Narrowband IoT) and 5G are candidates for wireless edge connectivity in cases such as fleet management and for medical devices where licensed spectrum technology tends to be more desirable. The required resources might be obtained via commercial relationships with mobile network operators, acquired in markets where private and industrial licenses are available, or deployed in shared spectrum licenses such as the CBRS (Citizens Broadband Radio Service) band in the United States.

Each approach has its benefits and drawbacks. Telecom operator relationships might prove the easiest option, but require operator engagement and support for the enterprise edge projects being pursued. Private networks are likely to be the most expensive option, but provide enterprises with a great deal of control over their networks. CBRS and shared spectrum solutions are appealing but potentially prone to performance challenges in areas where spectrum resources are highly contested.

Legacy 2G and 3G technologies have been used in the past, primarily for connecting edge IoT devices. Today 4G-LTE is widely available and its LTE-M and NB-IoT variants have been deployed across the Americas, Europe and in parts of Asia Pacific. Today 4G-LTE and its variants are generally the most suitable technologies for wireless edge connectivity in licensed spectrum bands. The LTE-based solutions that are being deployed today will remain operational for the next decade or more.

It is still early days for 5G, but it is gaining momentum in many markets across the globe. One of the most interesting aspects of 5G, as it relates to edge computing, is that 5G itself is an edge computing use case.

Edge is about the Distribution of Services

Don't get caught up in defining the location of "the edge." Edge is where your consumer or opportunity needs it to be.

Edge solutions facilitate engagement and protections for the customer base of globally delivered SaaS applications. Edge can enable sales and safe service delivery in any or all countries where there are regulatory, privacy or security concerns associated with locally-generated content. Cloud isn't the edge and managed hosting isn't the edge, at least not the edge that's needed for many solutions. Edge is more about facilitating the distribution of services to exactly where they're needed than it is about any given technology.

Mark Thiele
CEO and Founder, Edgevana

Virtualized network functions that support 5G require a highly-distributed data center infrastructure that allows for white box servers to be placed within a few milliseconds of the radio heads to support virtual BBU (Baseband Unit) functionality. These same data centers (typically micro-modular data centers) can house sufficient server capacity not only to operate a 5G network but also to run the cloud services that can deliver applications at the edge over that same 5G network. This model of deployment has the potential to change the economics of 5G network infrastructure, as the facilities can do double-duty, running the 5G network as well as hosting the applications and services which ride on top of it.

5G essentially comes with a four-prong value proposition that includes high bandwidth connectivity, ultra-reliable low latency connectivity (uRLLC), network slicing (to allocate resources based on performance demands) and massive machine-type connectivity (mMTC). On the face of it, these are all compelling attributes for edge networking, but they come with some costs and caveats. In particular:

- The bandwidth performance of 5G depends on wideband spectrum resources being available. Commonly this wideband spectrum is at high operating frequencies with more challenging radio coverage.
- Most current 5G deployments are based on non-standalone (NSA) implementations with 4G core network functionality that does not support uRLLC, network slicing or MTC capabilities. In 2021, many mobile operators plan to accelerate their 5G core network deployments to offer end-to-end 5G standalone (SA) capabilities, albeit within their 5G wireless coverage footprints.
- Many of the high-performance features in 5G, such as uRLLC and network slicing, are not solely dependent on technical feature availability, but also depend on 5G network designs, including site densities and infrastructure hardening, backup and redundancy. For example, conventional telecommunication network designs that are based on 'five-nines' availability might not address the

performance demands for business and mission critical infrastructure, such as that required for public safety, collaborative autonomous vehicles or command and control systems.

THE PATH FORWARD

Much of the network technology used today for edge computing has been designed for different purposes, such as communications and internet connectivity, rather than machine-type connectivity. However, technologies like 5G, next-generation SD-WAN and SASE have been standardized and are well suited to address the multitude of edge computing use cases that are being adopted and are contemplated for the future.

As digital services proliferate and drive demand for edge computing, the diversity of network performance requirements will continue to increase. These requirements must be carefully evaluated relative to the capabilities of the available network technologies, the engagement of key stakeholders and other factors including the total cost of ownership and geographical coverage requirements. This is particularly the case for edge-native use cases that can only exist in edge computing environments and will dominate the market in the long term.

CHAPTER 5

Software at the Edge

Simon Bisson
Freelance Technology Journalist

Software is at the heart of edge computing. It’s how applications are delivered, how edge hardware is managed and how workloads move around networks. The ecosystem is building a new stack to run at the edge of our networks, taking lessons from the hyperscale cloud, from IoT, from metro data centers, and from content delivery networks and the web, mashing them all together and building something new to suit new hardware, new networks, and a new generation of applications.

2020 was the year the hyperscale cloud providers doubled down on the edge, throwing their weight behind data center and metro-scale deployments of their cloud native platforms to take advantage of growing interest driven both by long-term trends and the more immediate needs of the pandemic. It was also when they noticed the work being done on WebAssembly and began to focus on it as a universal run time for edge binaries. Consistency became the watchword, putting workloads in the spotlight, where they rightfully belong.

As the COVID-19 pandemic continues to roll around the world, 2021 looks much the same as 2020 for edge compute, although with an element of consolidation of development done in more haste than usual. Open source technologies like Kubernetes will continue to build new foundations for edge applications, taking them from data center to microcontroller, and even into the air where they’re providing a new structure for software on military aircraft. If there’s one message for developers looking at moving to the edge, it’s to spend the year looking at how to make your code portable and scalable, so that it’s ready to support workloads as they migrate from edge node to edge node and back to the hyperscale cloud that’s become the hub of modern software.

If there’s one theme driving the current state of software on the edge of the network, it’s a focus on running workloads as close to where they’re needed as possible. Latency in the public cloud can be high, and when it comes to use cases like cloud gaming, real-time IoT analysis, autonomous driving and the like, putting compute close to where it’s needed ensures sufficiently agile responsiveness and reduces the risk of degrading the user experience.

Putting workloads on the edge means adapting how we build and run applications, finding ways to build and deploy code that can run on the increasingly diverse types of hardware in locations ranging from servers in metropolitan data centers to microcontrollers in devices on customer premises. While much of what we do on the edge is similar to how we handle compute across the rest of the network, the edge also presents new challenges that we need to confront, such as managing highly-distributed applications and data, as well as orchestrating edge operations at significant scale.

THE NEW EDGE STACK

Cloud native technologies and the DevSecOps movement are having a significant impact on the way users manage systems and software, allowing common abstractions to deliver software at the edge and automate their operations. Broadly speaking, one can think of the edge stack as three distinct layers that require different engineering skills.

At the lowest level is the systems layer, populated by firmware, operating systems and hypervisors, providing the technologies needed to work directly with edge hardware, whether IoT-class devices on the user edge or an increasing collection of specialized infrastructure at the service provider edge.

The next layer encompasses implementation and management. This is where tools like VMware's vSphere, Microsoft's AKS (Azure Kubernetes Service), open source Kubernetes and OpenStack, or Red Hat's OpenShift and RDO (RPM Distribution of OpenStack) operate, providing the services needed to support modern applications. This space is expanding to include specific edge solutions, like Azure's IoT Hub or Node-Red, with platform-level support for event-based operations.

At the top sits tooling designed to deploy and operate applications. Built into Continuous Integration/ Continuous Deployment (CI/CD) pipelines and using methodologies like GitOps, they provide a layer that allows effective management of distributed applications at the scale needed for edge networks.

Across all three layers is the need for a common observability layer, providing information tailored to the needs of different stakeholders. Traditional logging and monitoring services are still a key component of modern application operations, with tools like the ELK stack (Elasticsearch, Logstash and Kibana) offering log consolidation, querying and dashboards and Prometheus providing monitoring. Using ML alongside log and metric analysis allows prediction of failures and spotting of security breaches.

Edge technologies are particularly amenable to using cloud native tooling to provide an alternative approach to management and control, as much management tooling isn't designed for distributed architectures or for the global scale that can be necessary for edge architectures.

CODE AT THE EDGE

Circumstances at the edge of the network put strict requirements on packaging, delivering and deploying code on devices that are often in costly and hard to reach remote locations. That means that code pushed to the edge probably needs to be self-contained: each build must be complete in itself, and any changes require a complete package of not just code, but also any configurations, required libraries and software-defined environments, ensuring the container or VM can be run anywhere without dependencies.

With latency being a key issue for edge applications, it's important to keep as much functionality at the edge as possible. This can have a significant effect on application design, as well as changing the way code is managed. Dynamic applications that require direct connections to APIs can be risky, so many applications will combine code and content into single packages to ensure the best possible performance.

For many applications, using containers with a thin host OS, like Flatcar Linux, keeps resource demands to a minimum, while user interfaces can take advantage of approaches like JAMstack (JavaScript, APIs and Markup) to pre-render content and cache data locally, using in-page JavaScript to handle interactions and display content as required.

These self-contained approaches to deliver edge applications simplify management: code can be built, tested and deployed as one piece. Any changes will require a complete refresh. Errors can be batched up and reported back to developers for fixes.

WORKLOAD ORCHESTRATION

Running real-time workloads across the highly-distributed infrastructure presented by edge computing introduces many complicated challenges to developers and operators. How do you decide which workloads should run where? How do you handle failovers and geo-redundancy? How do you move services in response to devices in motion, such as to maintain continuity and service-guarantees with a mobile device moving across a geography, as in a drone or autonomous vehicle?

Challenges in Management at Scale

If edge is just an endpoint of a spider-web-like architecture, it opens up a set of new challenges while also putting some well-known ones into the spotlight again.

Cloud computing grew beyond the traditional data centers and the computing power needed to be available as close to the end users, whether humans or machines, as possible. By extending cloud infrastructure, it becomes crucial to be able to manage the large, geographically distributed architectures in an efficient and highly automated way.

Ildikó Váncsa
Ecosystem Technical Lead,
Open Infrastructure
Foundation
@IldikoVancsa

Many orchestration technologies, open source and otherwise, have emerged to tackle these types of complex scheduling problems. These orchestration systems take into account increasingly sophisticated levels of edge criteria for workload placement, automating decisions in real-time, abstracting away the complexity from developers and operators, who would prefer to simply specify the SLAs they require. A custom scheduler for edge computing might contemplate many sophisticated attributes requested by workloads. In addition to the typical scheduling attributes such as requirements around processor, memory, operating system, and occasionally some simple affinity/anti-affinity rules, edge workloads might also specify some or all of the following:

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

New networking and data center systems have begun delivery telemetry via real-time data feeds. Modern scheduling algorithms can ingest this telemetry and use it to make real-time and predictive workload placement decisions that spare the developer from needing to understand the underlying complexity.

The most popular orchestration solutions today tend to be based on the open source Kubernetes. There are edge orchestration algorithms emerging in open source, such as within the CNCF (Cloud Native Computing Foundation), as well as private solutions from companies like Rafay Systems and VMware.

SERVERLESS

Modern serverless and function as a service (FaaS) platforms are especially well-adapted for running ephemeral edge computing workloads. They instantiate quickly and hide a lot of the complexity required to manage the underlying server and orchestration layers. Since these functions themselves do not persist data, they are employed to process data that is already being stored at the edge — which can make them ideal for ingesting or extracting value from edge data efficiently. Serverless architectures are best used for elements that can be processed in short bursts and which don't require any state to be preserved. Also, because edge resources are expected to be more expensive than resources operating at the core (due to higher operational costs and higher price elasticity), the serverless functions may provide ways to lower the TCO of edge services because they can be deployed as needed, then disbanded quickly, such that the developer is only paying for the brief moments the functions run.

DATA AT THE EDGE

It's not just code we need to consider when delivering software on the edge; it's also data. While many edge operations, especially those focused on the Industrial Internet of Things (IIoT), process and route streaming data to data center and public cloud-hosted applications, we still need to think about how we work with data. With edge-to-core bandwidth a significant constraint, it's important to understand the costs of moving data and building application architectures that are designed to handle distributed data. Moreover, edge applications are likely to be running on devices with limited storage, so need to have an appropriate cache of data for current workloads in order to avoid latency issues.

Developers can solve these challenges by writing client code that can work with core distributed databases like Google's Spanner or Microsoft's Cosmos DB with their alternative consistency models. As the first option, using Cosmos DB's session consistency to handle queries from edge clients can limit updates to current operations, while background consistency actions ensure that all global instances are up to date.

NoSQL databases like Couchbase are introducing their own consistency models for distributed operation, and can run on edge hardware as well as in core systems. Being able to work directly on a local store, with the database engine handling synchronization, offers significant speed improvements, especially when used in conjunction with predictive caching.

Startups like Macrometa are building database technologies using edge-first principles, which, if successful, will introduce new methodologies for storing and retrieving data and state in dynamic edge environments and it is expected that one or more companies will offer edge storage-as-a-service at global scale.

MESSAGING-BASED ARCHITECTURES

Message-based architectures like Actor/Message work well for edge environments, where event-driven operations are common. Research work at Microsoft Research (MSR) and other establishments has led to methods of handling full ACID (Atomicity, Consistency, Isolation and Durability) transactions in distributed operations, an approach which could lead to improved methods of working with distributed data and managing edge cached content. Messaging architectures can also help with unreliable connections, using queues and event grids to manage events.

SUPPORTING SMALLER DEVICES

A new generation of capable low-power devices run full OSs and are able to support modern application environments. For example, the latest generation of Raspberry Pi's Compute Module has an available PCIe (PCI Express) channel with support for SSD storage, and now has OS support for Ubuntu's server Linux releases, along with the MicroK8s Kubernetes environment.

A cluster of ARM-based devices can make an effective low-power server environment using this or K3s. Devices powered using PoE (Power over Ethernet) connections can be deployed quickly requiring minimal skills. Prepared images can quickly phone home for configuration, downloading Helm charts and containers as required.

Investments at Cloudflare and other metro edge vendors are extending the Web Assembly tooling that allows compiled code to run on JavaScript engines, offering a set of standard interfaces that allow Web Assembly code to run outside the browser. WASI, the Web Assembly Systems Interface, is likely to become an important edge technology as it allows you to use the same runtime on a wide range of devices, from ARM microcontrollers to Intel servers. You can build code in familiar languages and environments, including Rust, Go and C#, compiling using standard build chains into WASI (WebAssembly System Interface) pseudo-assembly language before deploying and running.

In Search of the Killer App

After almost four years since it first burst on the scene, most IT folks have heard of edge computing. Nearly everyone agrees that this technology is here to stay and will deliver amazing solutions, be it in Industry 4.0, healthcare, telco or retail. But we're still on a quest for the killer app.

But edge hasn't yet delivered widely on its two most promising facets - high bandwidth and low latency. While there are many edge applications, people are still looking for that killer app. Why is that? Is it a case of unrealistic expectation or did we over promise? I believe we just have to find the right recipe of combining AI, 5G, analytics and hitherto unheralded technology.

2021 AIR MAIL

Ashok Iyengar
Garage Solution
Engineering: Network Edge
Computing, IBM Cloud

Microcontrollers are an important class of edge device, and as such need to be considered in any application development strategy. However, software deployment and management remain a significant issue, especially when deploying firmware at scale. Twilio's acquisition of Electric Imp adds its software distribution platform to the company's portfolio of IoT solutions, and offers a way of using wireless networking and a central deployment platform to manage software updates and configuration.

Microsoft's Azure Sphere takes a similar approach, building on a secure application implementation and a custom embedded Linux with hardware-enforced trusted boot and digital signatures to ensure application integrity. Using an end-to-end software deployment model, applications are distributed from Azure to devices, with tooling to manage pools of devices using a cloud service as an artefact repository.

New technologies are adding improved software support to platforms like these using familiar cloud native management layers. Microsoft's Krustlet approach uses Kubernetes to deploy and manage Rust-based applications on microcontrollers, running them using WASI as a common runtime. Applications can be automatically added to new edge points automatically as they're discovered and added to a device pool. As WASI currently lacks a standard deployment model, Krustlet allows you to use Kubernetes' familiar node management features to deliver and update code to devices.

Another Microsoft Kubernetes project, Akri, works with devices that are only exposed via APIs or drivers. These leaf devices can be identified using standard protocols, and added to a resource pool where they can be consumed by Kubernetes applications. For example, video camera streams can be discovered and then distributed across application nodes.

THE HYPERCONVERGED EDGE

The move to using virtual and containerized applications on relatively small form-factor hardware makes HCI systems ideal for edge compute, and hyperscale cloud providers are offering tools to integrate them into hybrid cloud systems with deployment to branch offices and other edge locations.

Microsoft's Fall 2020 update of its Azure Stack HCI software suite makes this trend clear, with support for Microsoft's Azure Arc application deployment and management tool along with a version of Azure's managed Kubernetes all tied into a GitOps workflow. Similarly, Amazon Web Services expanded its Outposts lineup to include a 1RU variant, leveraging their ARM-based Graviton processors (RU or "rack unit" is a measurement that refers to the space between shelves on a standard server rack; 1RU is equal to 1.75"). Google has recently ported its Anthos suite to bare metal, and is investing to bring partner applications to the edge with Anthos for Telecom.

Hivecell, based outside of New York City, offers a condensed hyperconverged solution for far edge locations. Their self-networking and fully managed devices promise an expandable "edge as a service" for nearly any environment. In an example of how cloud pricing models are influencing the edge market, Hivecell provides hardware, software, application deployment and support as part of a unified service.

All of these offerings share the goal of providing an easy to install, easy to manage cloud native platform at the edge of the network. Bundling all the software elements needed for an edge deployment into a single platform is a logical step. It reduces risk for the end user and at the same time provides a known target for software deployments, allowing developers to build and deliver packaged software using familiar tools and methodologies, from installers like MSIX to Kubernetes technologies like Helm and CNAB (Cloud Native Application Bundle).

These turnkey solutions solve many deployment issues for end users and service providers alike, as they no longer have to deploy and manage all aspects of the hardware. Instead, it allows them to work directly with customers to support the customers' own edge compute needs.

EDGE CLOUD MARKETPLACES

Industry pundits have long warned of a “chicken-and-egg” scenario that could potentially constrain the commercial success of edge technologies. On the one hand, a lack of widespread edge infrastructure deployments would limit the business incentive for third-party developers of edge applications. On the other hand, a shortage of ready-to-run applications would mean a lack of demand for edge infrastructure.

Throughout 2020, large cloud vendors announced various offerings to bring their ecosystems to hybrid, on-premises and edge environments. In May IBM introduced its Edge Application Manager, looking to orchestrate and deliver its Cloud Pak ecosystem in edge environments. In December, Google expanded its Anthos hybrid cloud platform with an ecosystem of edge applications available from multiple Independent Software Vendors (ISVs), aiming to unlock use cases in multiple vertical segments. Meanwhile, Amazon Web Services kicked off their annual re:Invent conference by announcing smaller 1u and 2u form factor versions of Outposts, as well as “anywhere” versions of ECS (Elastic Container Service) and EKS (Elastic Kubernetes Service) for running managed workloads on-prem.

These announcements, as well as an ever-expanding open source ecosystem, are making off-the-shelf, edge cloud-hosted applications available to a wide range of companies and use cases beyond early adopters like telecom.

LF EDGE Projects

The Linux Foundation (LF) is a non-profit technology consortium founded in 2000 to standardize Linux, support its growth and promote its commercial adoption. LF and its projects have more than 1,500 corporate members from over 40 countries. LF also benefits from over 30,000 individual contributors supporting more than 200 open source projects.

The Linux Foundation's LF Edge was founded in 2019 as an umbrella organization to establish an open, interoperable framework for edge computing independent of hardware, silicon, cloud or operating system. The project offers structured, vendor neutral governance and has the following mission:

- Foster cross-industry collaboration across IoT, Telecom, Enterprise and Cloud ecosystems;
- Enable organizations to accelerate adoption and the pace of innovation for edge computing;
- Deliver value to end users by providing a neutral platform to capture and distribute requirements across the umbrella;
- Seek to facilitate harmonization across edge projects.

PROJECT TAXONOMY

Each edge tier represents unique tradeoffs between scalability, reliability, latency, cost, security and autonomy. In general, compute at the user edge reflects dedicated, operated resources on a wired or wireless local area network (LAN) relative to the users and processes they serve. Meanwhile, the Service Provider Edge and Public Cloud generally represent shared resources (XaaS) on a wide area network relative to users and processes.

In many applications, User Edge workloads will run in concert with Service Provider Edge workloads. Workloads on the User Edge will be optimized for latency criticality, bandwidth savings, autonomy, safety, security and privacy, whereas workloads on the Service Provider Edge will be optimized for scale. For example, an AI/ML model might be trained in a centralized cloud data center or on the Service Provider Edge but pushed down to the User Edge for execution.

The boundaries between edge tiers are not rigid. As mentioned previously, the Service Provider Edge can blend into the User Edge when CPE resources are deployed on-premises in order to provide a user with connectivity and compute as a managed service. Meanwhile, the User Edge can also extend to the other side of the last mile network, as in the case of enterprise-owned private cloud data centers. While the edge boundaries are fluid, they are instructive: certain technical and logistical limitations will always dictate where workloads are best run across the continuum based on any given context.

Regardless of the definitions of various edge tiers, the ultimate goal is to provide developers with maximum flexibility, enabling them to extend cloud native development practices as far down the cloud-to-edge continuum as possible, while recognizing the practical limitations.

The following sections dive deeper into LF Edge and how each project within the umbrella is working to realize this goal.

Democratizing the Edge

I firmly believe the best approach is to start with a problem you're trying to solve. Identify something you can't do but that you need to do. Find those gaps, and that will get you to the problems to address.

A solution we're working on is democratization: that is, to make IoT and edge accessible to more people through open source. As we try to reach a larger landscape of applications and solutions that have different demands, the needs change. There's a lot of engineering effort going into building protocols that can be standardized. The approach is a stone soup: let's come together, build it together and then leverage it together. By making these open source projects and the infrastructure available, you're opening the floodgates to many more adopters, to many more applications and solutions coming to market sooner.

Look at the possibilities that adoption has already opened up. Edge combines so much we know, whether it's security, cryptography, collecting sensor data from different protocols, blockchain, or machine learning. There's going to be a moment where the edge is just part of the fabric of the internet, when we stop thinking of the edge as something separate. That's the beautiful thing. The edge is not narrow. The edge is everything.

*For more of Malini's thoughts on this topic, **catch her interview** with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

POST OFFICE Postcards from the Edge experts from around the world 2021 AIR MAIL ★★★

Malini Bhandaru
Open Source Lead for IoT
and Edge, VMware

PROJECT UPDATES

As with other LF umbrella projects, LF Edge is a technical meritocracy and has a Technical Advisory Committee (TAC) that helps align project efforts and encourages structured growth and advancement by following the **Project Lifecycle Document (PLD)** process. All new projects enter as "**At Large Stage**" projects, which are projects that the TAC believes are, or have the potential to be, important to the ecosystem of Top-Level Projects, or the edge ecosystem as a whole. The second "**Growth Stage**" is for projects that are incubating on their way to reaching the Impact Stage, and have identified a growth plan to reach that level. Finally, the third "**Impact Stage**" is for projects that have reached their growth goals and are now on a self-sustaining cycle of development, maintenance, and long-term support.

Simplifying the Lives of Developers

In order to accelerate edge computing adoption, edge platforms need to remove the burdens for developers and operations engineers when it comes to managing the complexities associated with infrastructure provisioning, workload orchestration, traffic routing, scaling and monitoring, all while minimizing impact on application design.

This is the critical role that edge platforms play in the future of the internet.

Stewart McGrath
CEO and Co-Founder, Section
@stewmcgrath

STAGE 3 - IMPACT PROJECTS

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 and IoT, for both provider and enterprise edge domains. These Blueprints have been created by the Akraino community and focus exclusively on the edge in all of its different forms. They are ready for adoption as-is or can be used as a starting

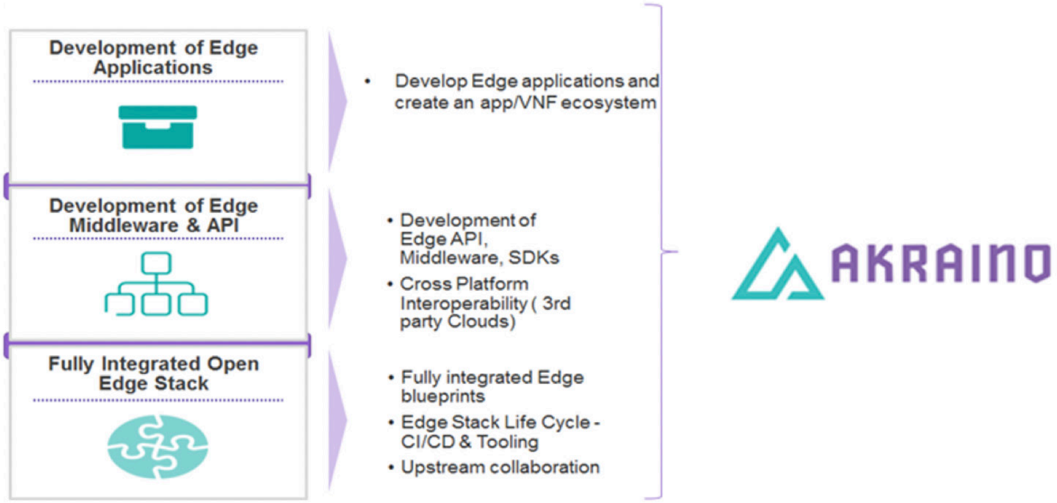

Development of Edge Applications
Development of Edge Middleware & API
Fully Integrated Open Edge Stack
• Develop Edge applications and create an app/VNF ecosystem
• Development of Edge API, Middleware, SDKs
• Cross Platform Interoperability (3rd party Clouds)
• Fully integrated Edge blueprints
• Edge Stack Life Cycle - CI/CD & Tooling
• Upstream collaboration
AKRAINO

FIGURE 9
LF Edge Akraino Project Scope

Distributed Data Handling and Provenance

With all of our edge data beginning to fly through vast networks, questions arise….

How can we verify that the data hasn’t been spoofed or tampered with at some point in transit? How can we prove after the fact that data was handled in a manner compliant with regulations? How can we visualize the path of a piece of data as it traversed from box to box, application to application? Is it possible to establish an objective measurement of confidence in the security and veracity of the data once it has reached its final destination? And finally, how in the world are we going to store it all?

No one vendor has the ability to survey the entire landscape of technology innovation and confidently answer all of these questions. It will be the collaborative effort of partners that will make the edge real.

Trevor Conn
Director of Software Engineering, Dell Office of the CTO

point for customizing a new edge blueprint. Currently, there are 20 Akraino Blueprints that are tested and validated in real hardware labs supported by users and community members.

The Akraino community has worked together to provide shared resources to developers and open source participants, that ease development across the different hardware platforms and architectures. The project collaborates with multiple upstream open source communities/SDOs such as Airship, OpenStack, ONAP, ETSI MEC, GSMA, TIP, CNCF and ORAN. Akraino supplies a fully integrated solution that supports zero-touch provisioning and zero-touch lifecycle management of the integrated stack.

In August 2020, **Akraino Release 3 (R3)** delivered a fully functional open source edge stack that enables a diversity of edge platforms across the globe. With R3, Akraino brings deployments and PoCs from a swath of global organizations and enables innovative support for new levels of flexibility that scale 5G, IIoT, telco and enterprise edge cloud services quickly, by delivering community-vetted and tested edge cloud blueprints to deploy edge services. New use cases as well as new and existing blueprints provide an edge stack for Connected Vehicle, AR/VR, AI at the Edge, Android Cloud native, SmartNICs, Telco Core and Open-RAN, NFV, IOT, SD-WAN, SDN, MEC and more.

Adding new blueprints, as part of Akraino Release 4 (R4) and releasing an Akraino API portal are some of the main activities scheduled for 2021.

Some of the selected new Blueprints for Akraino Release 4 (R4) are related to Connected Vehicle, AR/VR oriented Edge Stack for Integrated Edge Cloud (IEC), The AI Edge,5G MEC/Slice System to Support Cloud Gaming, Public Cloud Edge Interface (PCEI) and Enterprise Applications on Lightweight 5G Telco Edge.

For more information, please visit the **Akraino** website.

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 7+ 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, Home Edge and Open Horizon. EdgeX is part of the Akraino Edge Lightweight IoT (ELIOT) Blueprint and tested under the Akraino Community Lab. Open Horizon is building an integration project that will demonstrate the delivery and management of EdgeX Foundry as a containerized solution in stages. In collaboration with Home Edge, a centralized device can be designated as a primary device to store the data from different devices.

In 2021, the project will continue to improve its virtual presence with a new website that will allow new and existing users ways to become part of the ecosystem, to participate in the community and to contribute to the project. The EdgeX China efforts will also be critical to support the growing community in the region. China-focused initiatives will include documentation, website and developer channels in the local language.

For more information, please visit the **EdgeX Foundry** website.

STAGE 2 - GROWTH PROJECTS

Project EVE aims to do for the edge what Android did for mobile by building an open, curated and universal operating system for solution deployments at the distributed edge, outside of conventional data centers. Developed within Project EVE, EVE-OS enables lifecycle management and remote orchestration of any application and hardware, scales down to constrained compute nodes and incorporates a zero-trust security model to meet the unique physical and cyber security requirements of edge computing deployments in the field.

The project scope includes:

- Delivering the flexible and modular EVE-OS with built-in security
- Providing a reference controller implementation
- Specifications and definition of open orchestration APIs

Supporting Docker containers, Kubernetes clusters and virtual machines, EVE-OS enables organizations to extend their cloud-like experience to distributed edge deployments in IoT, AI, networking and security use cases, while also supporting legacy software investments. It provides an abstraction layer that decouples software from the diverse landscape of edge hardware to make application development and deployment easier, secure and interoperable. The hosting of Project EVE under LF Edge ensures vendor-neutral governance and community-driven development.

In 2020, the EVE community doubled to over sixty unique contributors. EVE-OS has been deployed in various pilot and production environments spanning the manufacturing, oil and gas and renewable energy verticals and is emerging as the anchor point for a growing ecosystem of hardware, software and services providers that are delivering solutions at the distributed edge.

Key project goals for 2021 include ongoing collaboration with the Kubernetes/K3S community to scale up to clustered edge deployments, significant improvements to the OSS reference implementation controller, expanding the supported hardware and "runs on EVE" application ecosystem and integration with other LF Edge and industry efforts. The community is also targeting overall optimizations of the Virtual Machine Manager (VMM), storage layer, and registry artifacts for VM and container workflows, as well as improved support for alternative hypervisors like ACRN.

For more information and links to documentation, visit the **EVE website**.

Fledge is an IIoT open source platform that is easy, economical, scalable and secure for building and operating industrial applications for condition monitoring, predictive maintenance, increased efficiency (Overall Equipment Effectiveness, or OEE), higher quality, situation awareness and safety.

Fledge is a mature, field tested codebase that has been deployed in industrial use cases since 2018. It offers more than 50 industrial protocols, data mappings and sensor plugins, as well as more than 20 integration solutions connecting to Enterprise Resource Planning (ERP), logistics, Manufacturing Execution Systems (MES), historians, databases and cloud providers.

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

The Devil is in the Details

Do not make the mistake of thinking of edge computing as just a mini cloud. It requires an entirely different way of looking at infrastructure, applications, networks and most importantly operational tools for deploying and supporting edge services.

Verizon has had network edge products on the market for close to four years now, yet we are still uncovering delivery and management gotchas. The devil is in the details, so while it might be tempting to cut corners, it is essential to complete extensive testing to verify that your edge applications will perform as expected under production conditions.

2021 AIR MAIL

Beth Cohen
Cloud Product Technologist,
Verizon

operators can build, manage, secure and support both Supervisory Control and Data Acquisition (SCADA) and 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 2021, Fledge will focus on two sets of features:

- Life-cycle management of deployments at scale: enabling provisioning, management and discovery at scale of IIoT applications in a secure manner.
- AI integration: incorporating edge-based ML life-cycles. Users will start with labeling and harmonizing data to build and train models, then execute, distribute, improve and upgrade such models for industrial use cases at the edge.

For more information, please visit the **Fledge website**

Home Edge is a robust, reliable and intelligent home edge computing open source framework, platform and ecosystem. It provides an interoperable, flexible and scalable edge orchestration and computing services platform, with APIs that can also be used with libraries and runtimes to enable various user scenarios for the home.

This year, Home Edge successfully released its third major release called Coconut, constituting two key features. Data Storage provides a dedicated entity to store data from connected devices in a distributed manner. Multi-NAT Edge Device Communication enables broader subnet communications inside and outside of a home network, extending its use case to multiple remote control and discovery scenarios.

Since its launch in 2019, Home Edge has been consistently collaborating with other LF Edge projects, and specifically closely co-worked with EdgeX Foundry in 2020 to deliver the Data Storage feature in the Coconut release. Home Edge will extend its collaboration with additional projects to develop its full feature lineup from the Home Edge architecture.

In 2021, Home Edge will firstly focus on providing more protocols in communicating with various cloud entities such as MQTT and CHIP (Connected Home over IP) through the connectivity abstraction layer. In addition, as a stretch goal, Home Edge will attempt to deliver further advanced features such as cloud synchronization with the edge, and the integration with other existing ML frameworks to enable the intelligent management and control use cases in the home.

For more information, please visit the **Home Edge** website.

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:

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

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.

For more information, please visit the **State of the Edge** website.

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. It provides the ability to enable the autonomous management of more than 30,000 edge devices simultaneously. It also includes the ability to handle multi-tenancy for up to 1,000 organizations (clients) in a single hub.

Open Horizon collaborates with other LF Edge projects such as EdgeX Foundry and Secure Device Onboard (SDO), an automated "Zero-Touch" onboarding service 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.

In 2021, Open Horizon will focus on

- Engagement with new volunteer contributors;
- Development of a mentorship program, with 12 new mentorships planned for the year;
- Addition of 3 project partners;
- Collaboration with additional LF Edge projects such as Akraino, Fledge, Home Edge and Project EVE;
- Stage 2 maturity.

For more information, please visit the [Open Horizon website](#).

STAGE 1 - AT LARGE PROJECTS

Baetyl (pronounced "Beetle") seamlessly extends cloud computing, data and services to edge devices, enabling developers to build light, secure and scalable edge applications.

Baetyl offers a general-purpose platform for edge computing that manipulates different types of hardware facilities and device capabilities into a standardized cloud native runtime environment and API, enabling the efficient management of application, service and data flow through a remote console both on-cloud and on-premises. Baetyl also equips the edge operating system with the appropriate toolchain support, reducing the difficulty of developing edge applications with a set of built-in services and APIs.

Early in 2020, the project reached a technical milestone with Baetyl 2.0, which featured the long-awaited remote management system and support for the Kubernetes ecosystem.

In 2021, Baetyl will focus on enhancing technical benefits such as ease of use, remote management and application compatibility. The project will reduce the number of installation steps and facilitate the use of Kubernetes on edge devices. Baetyl-Cloud will further integrate with the Kubernetes control plane and introduce Custom Resource Definition (CRD) objects to enable the use of the same set of APIs to simultaneously manage the cloud and the edge.

Additionally, the project will increase collaboration with other LF Edge open source projects, including EdgeX Foundry and Fledge, which will both run on Baetyl.

For more information, please visit the **Baetyl website.**

Secure Device Onboard is an Automated "Zero-Touch" Onboarding Service that securely and automatically onboards and provisions an edge device. The device only needs to be drop-shipped to the point of installation, connected to the network and powered up, then SDO does the rest. This zero-touch model simplifies the installer's role, reduces costs and eliminates poor security practices such as shipping with default passwords.

Secure Device Onboard joined LF Edge in mid-2020 as a Stage 1 "At-Large" project. It provides easier, faster, less expensive, secure onboarding of devices. It expands the Total Available Market (TAM) for IoT devices and in turn accelerates the resulting ecosystem of data processing infrastructure. Most "Zero Touch" automated onboarding solutions require the target platform to be decided at manufacture, whereas SDO provides increased flexibility.

Secure Device Onboard collaborates with other LF Edge projects such as EdgeX Foundry and Open Horizon, a platform for managing the service software lifecycle of containerized workloads and related ML assets. It enables the autonomous management of applications deployed to distributed web-scale fleets of edge computing nodes and devices without requiring on-premises administrators.

In 2021, Secure Device Onboard will focus on:

- Graduating to Stage 2 maturity;
- Completing the transition from an Intel-proprietary SDO architecture to the FIDO specification;
- Adding 10 new active volunteer contributors;
- Adding three project partners who contribute significant value;
- Begin integration projects with SDO supporting the FIDO specification with existing LF Edge projects, including Open Horizon and two others;
- Bring two new members to LF Edge.

For more information, please visit the **Secure Device Onboard** website.

Partner Updates

OPEN INFRASTRUCTURE FOUNDATION

The OpenStack Foundation transitioned into becoming the **Open Infrastructure Foundation** in October 2020. The Foundation is a not-for-profit organization with the mission to support open source communities who write code that runs in production. Over the past decade of building and supporting the OpenStack project, we learned a lot about how individuals and organizations are utilizing cloud computing and the open source platform that the community is developing and saw how OpenStack is just one piece of the open infrastructure puzzle. The name change reflects the extended scope to help further communities to create building blocks to power infrastructure.

While open source software is becoming more and more popular, it doesn't only refer to the availability of the source code, but the collaborative way to create it. Infrastructure is one of the layers that benefits the most from open source as everyone relies on infrastructure, though not the component that enables the easiest differentiation. The concept of open infrastructure relies on the availability of open source components to each building block of the system, which can help with challenges like interoperability.

Interoperability has always been a challenge, achieving limited success to address through standardization or merely building a solution from products delivered by one vendor. Cloud computing broke down single-vendor environments while edge computing puts all the unsolved issues back under the spotlight due to the scale and complexity of edge deployments. Edge computing environments consist of hardware and software components from various sources often forming a massively distributed system which is growing organically over time. Being able to fit the pieces together is critical for success in the edge computing space, and this is where open source projects play a key role.

The Open Infrastructure Foundation supports multiple projects that are focused or relevant for edge computing:

OPENSTACK

OpenStack is an open source cloud platform that is widely adopted around the globe in various industry segments. While the software was initially used in the data center thanks to its modular architecture, it can be extended to run workloads on the edge as well. As the platform is capable of running bare metal, containerized and virtualized applications as well it fits very well with the varying needs of the different use cases in this space.

STARLINGX

StarlingX is a fully-integrated platform that is optimized for edge and IoT use cases. The project creates a fusion between OpenStack and Kubernetes, utilizing container technology to run the infrastructure services as well to make the platform more flexible and robust. This architecture decision also provides the possibility to run only containerized workloads at the edge, utilizing the relevant Kubernetes components with a minimal footprint. While the project is integrating a lot of well-known open source projects together, it is also a development project to fill the gaps in the infrastructure layer in areas such as hardware and software management, orchestration and fault management. One of the key features of the project is the distributed cloud architecture that is focusing on keeping the central and edge sites in sync while also providing autonomy on the edge. It is critical for handling error scenarios where the connection between a central location and the edge is lost, in order to maintain full functionality at the edge for the time period of this condition. The architecture choice is in line with the Edge Computing Group's Distributed Control Plane model.

EDGE COMPUTING GROUP

The **Edge Computing Group** (ECG) is a top-level working group supported by the Foundation. The goal of this group is to provide a better understanding of edge computing and its demands on the infrastructure layers. The scope of the working group is on the infrastructure layer while not being limited to any industry segments. ECG is collecting use cases to identify and analyze requirements in order to be able to define reference architecture models based on common characteristics. The group's activities include the testing and evaluation of architecture models using relevant open source projects as building blocks, as well as the documentation of learnings and outcomes.

OPEN-IX

The **Open-IX Association** (Open-IX) is a non-profit industry association that promotes the expansion and proliferation of interconnection and internet exchange points (IXPs), primarily through advocacy, standards and tools.

2020 was a big year for tools and standards at Open-IX.In June, they launched the beta version of the much anticipated **Interconnection Navigator** which allows industry practitioners, analysts and researchers to explore the explosive growth of Internet Exchange Points (IXPs). This interactive tool leverages publicly available data from **PeeringDB** to build graphical representations of how interconnection has evolved over selectable time periods and geographies. Offering granularity down to individual data centers and IXPs, the Interconnection Navigator is designed to be used by research analysts, industry executives and infrastructure procurement teams.

As an Accredited Standards Developer under The **American National Standards Institute** (ANSI), Open-IX officially submitted its two infrastructure standards for consideration in 2020. Both of these standards were accepted by ANSI:

OIX-1 for Internet Exchanges: Accepted by ANSI in June 2020, this specification which governs technical, physical and operational standards is the first ever global certification for Internet Exchange Points. OIX-1 Certification sets a minimum level of service and engineering and ensures fair, reasonable and non-discriminatory access to interconnection services. OIX-1 was developed by a broad consensus of world-class IXP managers, engineers and their customers.

OIX-2 for Data Centers: Accepted by ANSI in September 2020, this standard which represents a high engineering standard for concurrent maintainability and open access is the first global certification for data centers desiring to serve as points of network interconnection. OIX-2 Certification sets a minimum level of service and engineering for data centers, ensuring fair, reasonable and non-discriminatory access to interconnection services. OIX-2 was developed by the broad consensus of world-class data center managers, engineers and their customers.

Open-IX recognizes that edge computing will have an important impact on interconnection. To address this evolving dynamic, Open-IX has continued and expanded the work of its Edge Committee which is in the process of creating Edge Data Center standards. These edge standards will empower and inform stakeholders as they manufacture, procure and deploy data processing infrastructure for applications residing in non-traditional environments. This effort has received broad support from a constituency including edge data center operators, manufacturers and consumers, and we plan to release the work of the committee in the coming months and eventually submit it to become an ANSI standard as well.

As a volunteer-led and powered organization, Open-IX is always seeking broader engagement. To get involved, either through volunteering or sponsorship or through certification, please see our **website** or contact us at **info@open-ix.org**.

CNCF (CLOUD NATIVE COMPUTING FOUNDATION)

The mission of the CNCF is to make cloud native computing ubiquitous. As we have seen in the last few years, Kubernetes and cloud native computing adoption has skyrocketed in the enterprise. The simple paradigms around observability, loosely coupled systems, declarative APIs and robust automation that have made cloud native technologies so successful in the cloud are even more important for the edge.

There are multiple initiatives within the CNCF that seek to address the needs of edge computing. Both KubeEdge and OpenYurt seek to bring Kubernetes to the edge while K3s is a Kubernetes distribution focused on resource constrained devices. Tinkerbell helps bridge the divide between the physical bare metal and software worlds. Kubernetes has an IoT Edge Working Group dedicated to discussing, designing and documenting the use of Kubernetes for developing and deploying IoT- and edge-specific applications. The CNF Working Group is defining how cloud native networking applications are helping telcos and enterprise IT organizations understand how to make their networks more cloud native. We expect the collaboration to continue as the edge market grows and matures, taking in the lessons learned from the cloud native world.

Kubernetes is often called the "Linux of the Cloud" and we believe we will see a similar evolution pattern, as Linux originally started out as a hobbyist operating system but was stretched to meet new use cases in other environments like mobile and embedded as end users contributed their new unique requirements and features.

In the end, the CNCF strongly believes that the cloud native movement and Kubernetes will evolve and impact the edge in a positive fashion. We look forward to collaborating across many organizations and projects to realize that vision to truly make cloud native computing ubiquitous in not only the enterprise but the edge.

ETSI MEC

In 2014, a small team representing a number of forward-looking companies in the mobile space published a white paper which put forward the idea that the cloud would move to the edge. The paper postulated that when it came to the practical aspects of hosting such an edge cloud for public access, the telecom industry was in pole position. Telcos have the unique combination of global reach and customer proximity, both geographical and network-topological. For such an "edge-cloud at the telco access", the white paper coined the term "Mobile Edge Compute" (MEC), which was subsequently tweaked to "Multi-access Edge Compute" while retaining the same acronym.

This group of experts and companies also recognized that a number of significant challenges would need to be overcome for MEC to realize its potential. Not the least of these was the need for a set of standards to ensure interoperability at the key points of the yet-to-emerge MEC ecosystem. To address this issue head on, and proactively, they formed a group to define such standards under the ETSI umbrella, known as "ETSI MEC".

From the onset, the mission was clear, albeit never stated explicitly: to ensure that the MEC market was not delayed due to a lack of well-defined industry standards. However, as the team worked on defining these standards, another aspect of MEC became evident: uniquely among the major telecom industry Standards

Development Organizations (SDOs) they had to address a key set of customers who do not interact with standards in a typical fashion. These customers are the application developers and so the mission expanded to ensure that the MEC market was not delayed due to a lack of well-defined industry standards or tools that enable application developers to develop to these standards.

In six years ETSI MEC has made good progress. A core set of standards has long been available and continues to evolve. Importantly, the team has gone where few other SDOs have gone before, open sourcing the APIs for simple inclusion in software (see **ETSI Forge**), providing open sourced test scripts (also at **ETSI Forge**) and lately enabling a **Sandbox**, where application developers can experience the interaction of a cloud service with ETSI MEC-defined APIs.

As the first commercial MEC deployments are becoming a reality, the ETSI MEC team is seeing the fruits of their labor. Pending public announcements of commercial adoption in vendor products and operator systems, the use of ETSI MEC APIs by Akraino, 5G Automotive Association (5GAA) and others is indicative of traction. In September 2020, ETSI extended the term of ETSI MEC for a further two years. As the project anticipates these two years, the team plans to remain focused on the core mission, with the "use" part becoming increasingly important.

Appendices

Landscape

The **LF Edge Interactive Landscape** is 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. The project accepts community inputs and is overseen by the LF Edge State of the Edge Landscape Working Group.

Glossary

The Open Glossary of Edge Computing is a freely-licensed, open source lexicon of terms related to edge computing. It has been built using a collaborative process and is designed for easy adoption by the entire edge computing ecosystem, including by open source projects, vendors, standards groups, analysts, journalists, and practitioners. The Open Glossary is maintained by a working group under the LF Edge State of the Edge project.

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; 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: Multi-Access Edge Computing (MEC).

Access Edge

The sub-layer of the Service Provider Edge closest to the physical last mile networks, zero or one hops from the RAN or cable headend. For example, an edge data center deployed at a cellular network site. The Access Edge Layer functions as the front line of the Service Provider Edge and typically connects to the Regional Edge layer upstream in the hierarchy. Edge compute at the Access Edge consists of highly-distributed server-class infrastructure located at front- and mid-haul sites, such as cell towers, cable distribution plants, aggregation and pre-aggregation hubs, central offices and other facilities which house network access equipment such as cellular radio base stations, as well as xDSL and xPON equipment. Access Edge data centers are often of the micro-modular variety because of their ease of deployment and self-contained operation. Because of the need to support ultra-low-latency workloads, including those that require a predictable connection to the last mile network, Access Edge facilities are typically sited within

15km of the radio heads or cable head ends and are best used for workloads that require latencies in the sub-1ms - 30ms range.

See also: Aggregation Edge.

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.

Aggregation Edge

The layer of Service Provider edge one hop away from the Access Edge. 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 between the Regional Edge and the Access Edge to allow for greater collaboration, workload failover and scalability than can be provided by a single data center location.

See also: Access Edge.

Base Station

A network element in the RAN (Radio Access Network) 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 cloud native network function (CNF) is a cloud native application that implements network functionality. A CNF consists of one or more microservices and has been developed using Cloud Native Principles including immutable infrastructure, declarative APIs, and a "repeatable deployment process". An example of a simple CNF is a packet filter that implements a single piece of network functionality as a microservice. A firewall is an example of a CNF which may be composed of more than one microservice (i.e. encryption, decryption, access lists, packet inspection, etc.).

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. 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: Distributed Antenna System (DAS) Hub, Infrastructure Edge.

Cloud Service Provider (CoSP)

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.

Colocation

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

See also: Shared Infrastructure.

Computational Offloading

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

See also: Traffic Offloading.

Constrained Device Edge

A subcategory of the User Edge consisting of microcontroller-based devices that are highly resource-constrained and distributed in the physical world. These devices range from simple, fixed-function sensors and actuators that perform very little to no localized compute, to more capable devices such as Programmable Logic Controllers (PLCs), Remote Terminal Units (RTUs) and Engine Control Units (ECUs) addressing time- and safety-critical applications. Devices at this tier leverage embedded software and have the most unique form factors.

See also: IoT Edge, On-Premises Data Center Edge, Smart Device Edge, User Edge.

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

Edge Security is No Game

Security at the edge is no game. The gaming industry is not only one of the most lucrative and fastest growing sectors in the world, it's also one whose popularity and scope make it ripe for cyber attacks.

A prominent game developer had to find a way to reduce the growing threat of Layer 7 DDoS attacks for an application we helped them deploy globally. They were able to find the solution that could be delivered by deploying web application firewalls directly at the edge. They used an off-the-shelf product from a leading security partner who offered a kill chain-based approach that stopped application layer attacks at the source when deployed at the edge.

Rob Roskin
Managing Principal, Lumen
Solutions Architecture

gateway or similar device in the field to collect and process data from devices. May also use limited spare compute and data storage capability from user devices such as smartphones, laptops and sensors to process edge computing workloads. Distinct from infrastructure edge as it uses device resources.

See also: Infrastructure Edge.

Device Edge Cloud

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

See also: Edge Cloud.

Distributed Antenna System (DAS) Hub

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

See also: Edge Data Center.

Edge Cloud

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

See also: Cloud Computing.

Edge Computing

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

See also: Infrastructure Edge.

Edge Data Center

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

See also: Virtual Data Center.

Edge Exchange

Pre-internet traffic exchange occurring at an edge data center, often at or near the Access Edge. This function will typically be performed in the edge meet me room of an 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 regional or centralized internet exchange.

See also: Internet Exchange Point.

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 built natively to leverage edge computing capabilities, which would be impractical or undesirable to operate in a centralized data center. Edge-native applications leverage cloud native principles while taking into account the unique characteristics of the edge in areas such as resource constraints, security, latency and autonomy. Edge native applications are developed in ways that leverage the cloud and work in concert with upstream resources. Edge applications that don't comprehend centralized cloud compute resources, remote management and orchestration or leverage CI/CD aren't truly "edge native", rather they more closely resemble traditional on-premises applications. A traditional SCADA application within a nuclear power plant that has no connection to the cloud would not be considered an Edge-Native Application. May use edge capabilities to provide data ingest, data reduction, real-time decision support, or to solve data sovereignty issues.

See also: Edge-Enhanced Application

Fog Computing

An early edge computing concept that stipulates compute and data storage resources, as well as applications and their data, be positioned in the most optimal place between the user and Cloud with the goal of improving performance and redundancy. The term fog computing was originally coined by Cisco as an alternative to edge computing, but has more recently fallen into disuse in favor of more precise terms.

See also: Workload Orchestration.

Gateway Device

A device on the User Edge which operate as conduit for other local devices, with the goal of aggregating and facilitating data transference from devices in the field, many of which are battery-operated and may operate for extended periods in a low-power state. Gateways connect to these devices and collect data for forwarding to On-Premises Data Centers or for transit across the last mile network.

See also: Resource-Constrained Device.

Hard Real-Time

Related to a use case or application that requires deterministic responses, where messages must arrive on time and in a predictable fashion and a failure to do so could result in critical or life-threatening malfunction. Resources like PLCs, RTUs and ECUs have been used in industrial process control, machinery, aircraft, vehicles and drones for many years, requiring a Real-Time Operating System (RTOS) and specialized, fixed-function logic. Examples of hard real-time functions include controlling an industrial lathe, applying a vehicle's brakes or deploying a vehicle's airbag; these functions are universally performed at the User Edge because they can't rely on control over a last mile network regardless of the speed and reliability of that connection.

See also: Real-Time, Soft Real-Time.

Infrastructure Edge

Recently replaced by the term Service Provider Edge in the LF Edge taxonomy, the Infrastructure Edge historically referred to 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, Service Provider Edge.

Interconnection

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

See also: 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 (IX point or 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.

IoT Edge

A subset of the Smart Device Edge composed of headless (i.e., has no user interface in regular operation) compute resources targeted at IoT use cases.

See also: Smart Device Edge, User 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.

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.

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.

Management and Orchestration (MANO)

In the context of edge computing, this is the management and orchestration of edge devices and edge applications over their entire lifecycle, including provisioning, monitoring, updating, operating and securing apps and data. Different edge tiers require similar principles but often depend on different tool sets due to inherent technical tradeoffs like available compute footprint, autonomy in periods of lost last mile connectivity, uptime needs, time criticality and so forth.

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.

Mixed-Criticality Workload Consolidation

The practice of consolidating hard real time or latency- and safety-critical workloads alongside soft real time and latency-sensitive workloads such as AI/ML models on common edge infrastructure.

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).

Trustworthy in All Things

I don’t know if we’re ever going to get to the cloud version of the edge. Cloud thinking is “How can we form a monolith? How can we build an application that can run from cloud all the way to the IoT?”. But with all the individualized use cases in IoT, we’re not going to have monolithic applications or even architectures. Eventually, we’re going to realize that what we really need are a bunch of highly personalized, customized applications: the very opposite of a monolith.

More and more things in the IoT will be eaten up by the cloud. But there will always be some devices and applications that won’t be part of the cloud, that in fact can’t be part of it for safety, security, privacy, reliability and other reasons. This includes things that can never go down, that have safety protocols you can’t do at scale, or that lives depend on. People are underestimating how many of those things there are.

This is why we have to be thinking about developing toward trust by building more safeguards, boosting reliability, establishing reasonable data retention policies and putting legal frameworks in place for both companies and individuals when things go awry.

For more of Stacey’s thoughts on this topic, ***catch her interview*** *with Matt Trifiro on the Over the Edge podcast, from which this was adapted.*

Stacey Higginbotham
Founder and Editor,
Stacey on IoT
@gigastacey

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.

Near Real-Time

Applications or use cases that benefit from discrete, low-latency timing, but which have some tolerance for timing that is low-latency but not Hard Real-Time.

See also: Soft Real-Time.

Network Functions 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.

North, South, East and West Data Flow

Refers to the directionality of data flow across the edge to or from the centralized cloud data center continuum. Northbound refers to data flowing “upstream,” such as from resources deployed at the User Edge to resources deployed at the Service Provider Edge and centralized cloud; whereas, southbound refers to data flowing in the opposite direction. Eastbound and westbound data flow refers to intercommunication between resource peers at the same/similar locations along the overall continuum.

On-Premises Data Center Edge

A subcategory of the User Edge consisting of server-class compute infrastructure located within, or in close proximity to, buildings operated by end users, such as offices and factories. IT equipment in these locations is housed within traditional, privately-owned data centers and Modular Data Centers (MDCs). These resources are moderately scalable within the confines of available real estate, power and cooling. Tools for security and MANO are similar to those used in cloud data centers.

See also: Constrained Device Edge, IoT Edge, Smart Device Edge, User Edge.

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).

Perishable Data

Data that is most valuable if acted on in the moment, and which can potentially be discarded once processed in order to reduce the cost of connectivity through the last mile network. Applications and connectivity can be optimized by processing data from sensors locally and then sending only relevant information to the Service Provider Edge or cloud, instead of raw streams of data.

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).

Real-Time

An application or use case that benefits from or requires discrete, low-latency timing.

See Also: Hard Real-Time, Soft Real-Time.

Regional Data Center

A data center positioned in scale between a centralized data center and a micro-modular data center, which has been built to sufficient size and is conveniently located to serve an entire region. Physically further away from end users and devices than the Access Edge, 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, Regional Edge.

Regional Edge

A subcategory of the Service Provider Edge consisting of server-class infrastructure located in regional data centers which also tend to serve as major peering sites. Regional edge sites are commonly in the form of Multi-Tenant Colocation (MTCO) facilities owned by companies like Equinix and Digital Realty, but can also take the form of a backhaul facility owned by a telco network that has been upgraded to house server-class IT equipment. Regional Edge sites usually provide, also, a regional Internet Exchange (IX) point, where tenants can connect to other networks and reach nationwide internet backbones. Large cloud providers, web-scale companies, CDNs and other enterprises place servers and storage in these facilities to reduce the latency and network hops that would otherwise be required to reach a centralized data center, but which do not require the ultra-low latencies available at the Access Edge or at the User Edge.

A rich confluence of data passes through these locations. Edge computing resources in a regional data center can work in conjunction with resources at the Access Edge and the User Edge to deliver different tiers of latency. As a general rule, Regional Edge data centers are capable of supporting edge workloads that can tolerate latencies in the 30ms - 100ms range.

See also: Access Edge, Regional Data Center, Service Provider Edge.

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.

In a Pandemic, We are All on the Edge

Edge computing is often viewed from the perspective of a central cloud. That's probably how we coined the name. But the COVID-19 pandemic turned that worldview on its head. In our individual lock-down environment, each of us is an edge node of the Internet and all our computing is, mostly, edge computing. The edge is the center of everything.

I call this human-centric computing and passionately believe that this should be how we view, design and deploy our systems in the first place. Think about all the speakers (or listeners), the cameras (or snoopers), the phones, watches (or trackers). That's my edge, your edge, and in the next decade we shall make them centered around people. Will you join me to make edge a human-centric edge?

POST OFFICE Postcards from the Edge experts from around the world
2021
AIR MAIL
★★★

Wenjing Chu
Senior Director of Open
Source and Research,
Futurewei Technologies

Service Provider

An organization which provides customers with access to its network, typically with the goal of providing that customer access to the internet via a last mile network. A customer will usually connect to the access network of the service provider from the User Edge side of the last mile via a fiber optic cable or a wireless cellular modem.

See also: Access Network.

Service Provider Edge

One of the two main edge tiers in the LF Edge taxonomy, used to specify edge computing capability deployed on the service provider side of the last mile network. The Service Provider Edge consists of IT equipment placed adjacent to or in support of service provider networks in a metropolitan region and encompasses the physical geography between the access networks and the nearest internet exchange (IX) points. The Service Provider Edge further subdivides into the Access Edge and Regional Edge and is typically capable of delivering edge computing with sub-100ms latencies. Formerly referred to as the Infrastructure Edge.

See also: Access Edge, Infrastructure Edge, Regional Edge.

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: Colocation.

Smart Device Edge

A subcategory of the User Edge consisting of compute hardware located outside physically-secure data centers but still capable of supporting virtualization and/or containerization technologies for cloud native software development. These resources span consumer-grade mobile devices and PCs to hardened, headless gateways and servers that are deployed for IoT use cases, such as in challenging environments that include factory floors, building equipment rooms, farms and weatherproof enclosures distributed within a city (see IoT Edge). While capable of general-purpose compute, these devices are performance-constrained for various reasons including cost, battery life, form factor and ruggedization (both thermal and physical) and therefore have a practical limit on processing expandability when compared to resources in an upstream data center.

There is an increasing trend for these systems to feature co-processing in the form of Graphics Processing Units (GPUs) or Field-Programmable Gate Arrays (FPGAs) to accelerate analytics, with the added benefit of distributing thermal dissipation which is beneficial in extreme environments. Resources at the Smart Device Edge can be deployed and used as separate devices (e.g., a smartphone or IoT gateway on a factory floor) or they can be embedded into distributed, self-contained systems such as connected/autonomous vehicles, kiosks, oil wells and wind turbines.

See also: Constrained Device Edge, IoT Edge, On-Premises Data Center Edge, User Edge.

Soft PLC

A virtualized Programmable Logic Controller (PLC) that can be consolidated onto common infrastructure alongside additional virtualized and/or containerized applications for data management, security and analytics applications running in parallel and interacting with higher edge tiers. This consolidation of mixed-criticality workloads requires specific considerations in the abstraction layer to ensure separation of concerns between functions.

Soft Real-Time

Associated with latency-sensitive applications, such as video streaming, where the application relies on low-latency networking to provide a good user experience but where a networking failure or delay will not cause a critical and potentially life-threatening malfunction. Applications with soft real time requirements are often delivered from the Service Provider Edge for convenience and economies of scale.

See also: Hard Real-Time.

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).

Thick Compute

In the context of edge computing, refers to higher-end gateways and server-class compute located usually at the Smart Device Edge and the On-Premises Data Center Edge. Can be deployed inside or outside of secure data centers.

See also: Thin Compute.

Thin Compute

In the context of edge computing, refers to more constrained edge compute resources in the form of gateways, hubs and routers that have only minimal processing power and which are usually used in conjunction with other, more powerful (Thick Compute) devices to perform computations. Part of Smart Device Edge, typically deployed outside of physically secure data centers.

See also: Thick Compute.

Tiny ML

Deploying limited-function Machine Learning (ML) inferencing models in microcontroller-based devices, typically at the Constrained Device Edge. Requires highly specialized toolsets to accommodate the available processing resources. An example is an ML model that enables a smart speaker to recognize a wake word (e.g., "Hey Google/Alexa/Siri") locally before subsequent voice interactions are processed by servers further up the compute continuum.

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.

User Edge

Edge computing capability which is deployed on the user side of the last mile network, also referred to as the Device Edge. One of the two main edge tiers in the LF Edge taxonomy, consisting of server, storage and networking, as well as devices, deployed on the downstream side of the last mile networks. User Edge resources are adjacent to end users and processes in the physical world and encompass a wide range of equipment types, including gateways, servers, and end user devices.

Workloads on the User Edge often work in conjunction with resources on the Service Provider Edge but are able to achieve lower latencies and conserve broadband network bandwidth by processing data without requiring it to pass across the last mile networks. Compared to the Service Provider Edge, the User Edge represents a highly diverse mix of resources. The User Edge contains the subcategories On-Premises Data Center Edge, Smart Device Edge and Constrained Device Edge. See also: On-Premises Data Center Edge, Smart Device Edge and Constrained Device Edge.

See also: Constrained Device Edge, On-Premises Data Center Edge, Smart Device Edge.

Vehicle 2 Everything (V2X)

A superset of V2I which refers to V2I-like technologies that allow a connected or autonomous vehicle to connect to more than its infrastructure, including to other vehicles, street side cabinets, and traffic devices.

See also: Vehicle 2 Infrastructure.

Vehicle 2 Infrastructure (V2I)

The collection of technologies used to allow a connected or autonomous vehicle to connect to its supporting infrastructure such as a 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: Vehicle 2 Everything.

Virtual Data Center

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

See also: Edge Data Center.

Virtualized Network Function (VNF)

A software-based network function operating on general-purpose compute resources which is used by NFV in place of dedicated physical equipment. In many cases, several VNFs will operate on an edge data center at the infrastructure edge.

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.

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.

Endnotes

1 DC webinar: Market Insights Opportunities, Jan 16th 2020.

2 Gartner: What Edge Computing Means for Infrastructure and Operations Leaders.

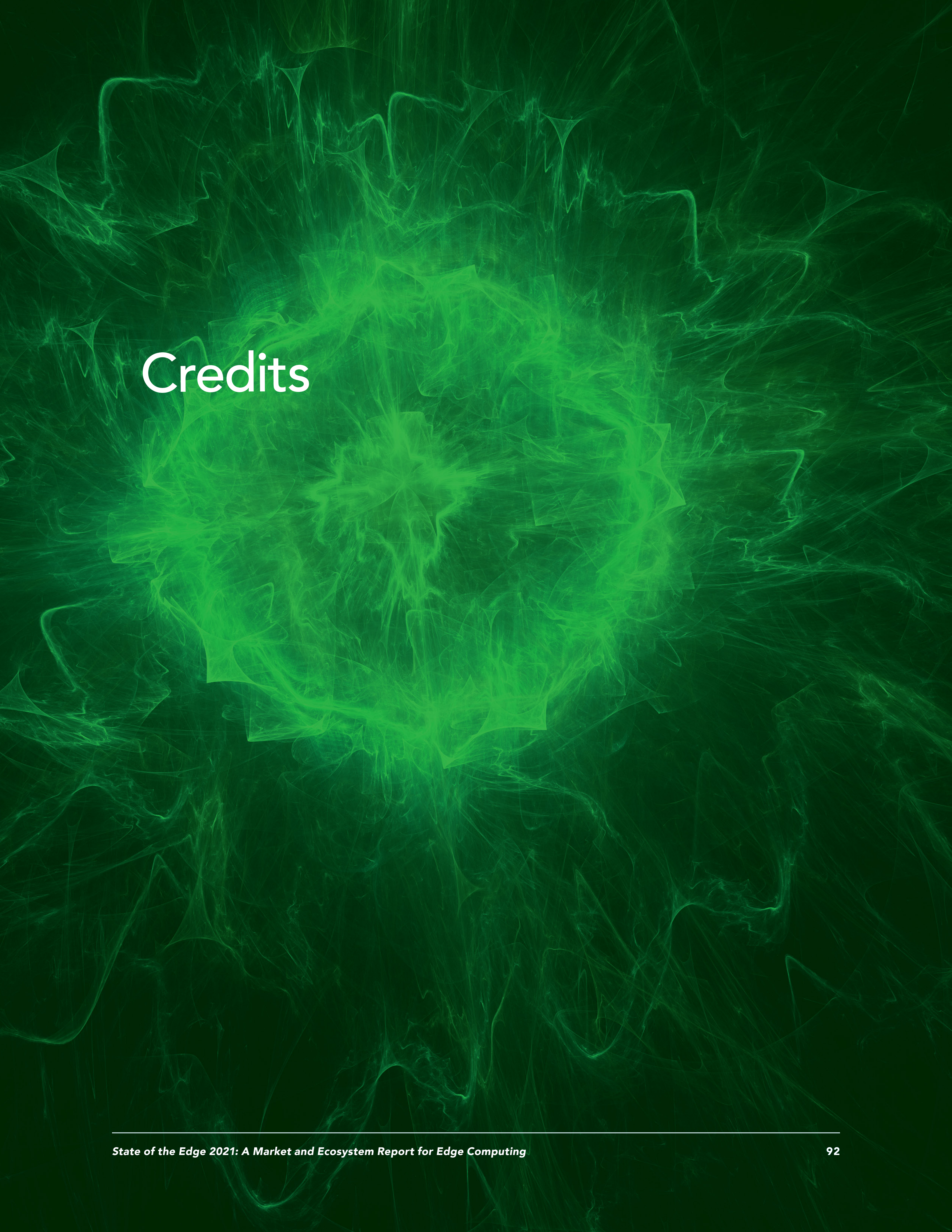
3 DC FutureScape: Worldwide IoT 2020 Predictions 4: By 2023, 70% of enterprises will run some level of data processing at the IoT edge.

4 The Open19 Project is an industry specification that defines a cross-industry common server form factor, creating a flexible and economic data center and edge solution for operators of all sizes.

5 Details of Project Natick.

6 Control/User Plane Separation (CUPS) in mobile networks.

Credits



A special note of thanks to the Linux Foundation staff and the LF Edge community for supporting the State of the Edge project and this 2021 report. In particular, Maemalynn Meanor and Brett Preston have been very generous, helpful, and resourceful.

Additionally, a hearty bow of respect to the State of the Edge report working group for all the late nights and early mornings — for their ideas, creativity, introduction, writing, proofing, and various other contributions to help this report be up-to-date, future-facing, and relevant to the community. Ildiko Vancsa (Open Infrastructure Foundation) and Molly Wojcik (Section): you are the best! As well, Bill Mulligan for his valuable edits and comments, and Scott Anderson for helping wrangle.

To our writers, Mary Branscombe, Simon Bisson, Phil Marshall, and Philbert Shih, thank you for putting up with short deadlines, and for your willingness to engage in healthy discussions about the how and why of the research areas.

Charlie Ashton, now with Napatech and formerly with 21K Consulting, this year's Managing Editor has been our rock — steadily and graciously guiding the report to completion. It's no wonder that Charlie's reputation for detail, professionalism and industry knowledge preceded him by such a wide margin.

And last but not least Jacob Smith (Equinix) and Matt Trifiro (VaporIO) for their continued leadership of the State of the Edge project. As project co-chairs, their leadership continues to shape the vendor-neutral stance and vision for this report.

We couldn't have done it without any of you!

STATE OF THE

EDGE

2021

THE LINUX FOUNDATION

stateoftheedge.com