TIA

BRIEFING PAPER

WHERE IS THE EDGE?

WHERE IS THE EDGE?

Contributions by

Damon Boyd (Grainger)

Joseba Calvo (EPI)

Shizuko Carson (Fujitsu Network Communications)

John Eberhart (Crowntech Photonics)

Jacques Fluet (TIA)

Jayson Hamilton (Midpoint Technology)

Marc Cram (Legrand)

Alex Himmelberg (Edge Presence)

Kevin Monahan (Kevin P Monahan LLC)

Keith Rutledge (Compass Data Centers)

Mark Smith (Vertical Datacom)

Eric Swanson (Shared Services Canada)

Don Utz (Thermo Bond Buildings)

Tom Widawsky (HDR)

TABLE OF CONTENTS

INTRODUCTION

We have all heard about emerging business and consumer IoT/IIoT applications like artificial intelligence (AI), virtual reality (VR), autonomous vehicles, gaming, and content delivery. For these applications to fully and successfully work together, data processing needs to happen in real time at the Edge. The Edge is a computing model that uses a decentralized network consisting of various applications, technologies, platforms, and hardware to achieve overall seamless computation closer to end devices and users—either at home, a business, or even the local coffee shop. In other words, the Edge is everywhere we and our devices go.

Through proximity to IT resources, the Edge model increases network performance and reduces latency, which is simply defined as "the time it takes for a data packet to be transferred from its source to the destination." [1] Also known as lag and generally measured in milliseconds, latency is critical to supporting instantaneous data processing and analytics for emerging real-time applications. Just a few extra milliseconds (ms) of latency in a control session or hand-off to a segmented network could cause latency-sensitive technologies like an autonomous car to "buffer" and take too much time to make a decision resulting in higher potential for an accident.

Over the past decade, the ICT industry has defined the Edge using a diverse nomenclature and a variety of applications—encompassing everything from Far Edge, Near Edge and Enterprise Edge, to IoT Edge, Edge Taxonomy, the 4th Industrial Revolution, 5G Edge, and more. This has caused the true meaning of what really constitutes the Edge to get lost in translation.

For some, the Edge is perceived as a self-contained micro-modular data center, as shown in figure 1, located at the last mile; or, for some, it means a regional colo data center with the same elements of a traditional TIA-942 data center having robust infrastructure, power, cooling, security and protection. While an Edge data center can certainly take on these forms, they all typically have a common baseline function of hosting computation capabilities closer to end devices and users. This requires industry to look at proximity, function, and application versus just the physical location and form.

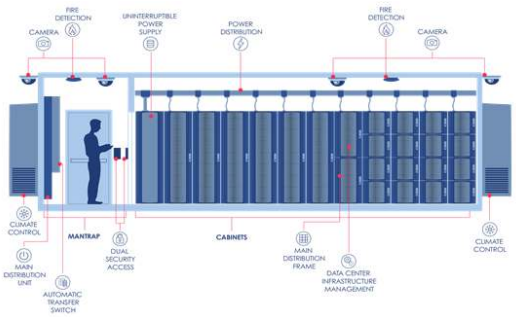


Figure 1: Common Edge Data Center at the Last Mile (Source: EdgePresence)

In recognizing the many different iterations and definitions of Edge in the ICT industry, the Telecommunications Industry Association (TIA) is working to help define a common language for the Edge compute model and the networks on which they operate. In this paper, we will explain how the Edge came to be and take a closer look at the different forms of the Edge, including the global distribution of data out to the Far Edge. We will explain the Far Edge model, its five defining characteristics, and the challenges that arise as we continue to decentralize and move closer to the connected users and devices.

HOW DID WE GET HERE?

It is important to reflect on how the need for the Edge came to be, as it provides the opportunity to understand scenarios that could have occurred along the way. In the past, there was long-term development in the way content was accessed, managed, and consumed. Looking forward, there is an expected continuation of exponential growth of data usage at the Edge. According to IT research firm Gartner, by 2025, 75 percent of enterprise-generated data will be created and processed outside of a traditional data center or cloud [2].

An evolutionary cycle of internet and technology advancements has occurred as networking infrastructure moved from the local end user out to the cloud and is now moving back closer to the end user to strengthen application performance.

Here is an overview for how internet technology has evolved to drive migration to the Edge:

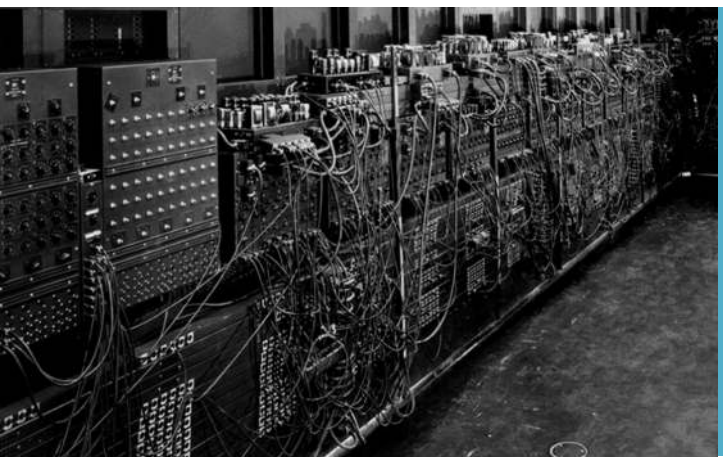
The Origination: While some may argue about who gets the credit for inventing the internet, in the early 1960s, the Department of Defense (DoD) funded the Advanced Research Projects Agency Network (ARPANET), which was essentially internet 1.0.

Commercialization: The early 1980s saw the emergence of NSFNET, a network operated by the National Science Foundation (NSF), that connected a series of super computers together. Shortly after NSFNET's adoption, along with new protocols in the Domain Name System (DNS) and the adoption of Transmission Control Protocol/Internet Protocol (TCP/IP), the industry saw the decommissioning of the ARPANET and the removal of most restrictions to commercial internet service providers (ISPs). In other words, internet 2.0.

Dot-Com bubble: The early 1990s saw the birth of the World Wide Web (WWW) that led one of the biggest economic booms in U.S. history, fueling massive growth in business opportunities based on the internet. Almost in unison with the Dot-Com bubble was a shift in the way data was being delivered through the emergence of Content Delivery Networks (CDN). Many perceive CDN to be the first iteration of the Edge. It was the Dot-Com growth and CDN push that paved the way for internet service providers (ISPs) to deploy the dark fiber which would later be leveraged for migration to the cloud.

Migration to the Cloud: Around 2006, the enterprise migration to cloud architectures began as newer technologies with software layers allowed for the sharing of IT resources to maximize systems output. With resources being shared "as a service," IT purchases shifted from a capital expenditure (CAPEX) to an operational (OPEX) expenditure, driving the growth of Cloud Service Providers (CSPs) and applications moving out of the enterprise and into the cloud. Essentially, applications moved towards a centralized distribution module and further away from end users, creating the very problem that spawned the need for Edge computing.

The Edge Data Center (EDC): The explosive growth of mobile applications, along with the rise of the "as-a-service" business model and advancements in sensor technology, has led to the creation and consumption of data at exceptional speeds by more end devices and users. The subsequent development of A.I.-based applications which can leverage that data requires reduced latency and localized computation. The EDC infrastructure is now being considered a foundational technology for emerging IoT/IIoT applications and Industry 4.0.



Hopefully, the actual evolution of the data center and its long-endured journey to the Edge is clear. With current investment levels of roughly $4.8 billion a year (according to Grand View Research) into EDC infrastructure, product offerings, and service delivery, we can expect to see new monetization of network services in the future. However, as the Edge continues to evolve and its proximity closes the distance between the user and application, we are only seeing the beginning of the growth.

Figure 2: One of the First Data Centers - Photo courtesy of United States DoD

APPLICATIONS ARE MOVING TO THE FAR EDGE

Advancements in Edge computing are enabling the development of real-time applications that will require shorter proximity to IT and lower latency to properly function. As these technologies emerge, innovations for computation will continue to move assets closer to end devices and users and away from traditional data centers. We see evidence of this in the following forward-thinking Edge applications currently being developed and deployed:

- Precision-based agriculture practices with the ability to analyze soil PH, phosphorus, and moisture levels via on-field sensors to optimize crop output
- VR technology over Wi-Fi enabling an engineer to build a live 3D model of a new addition to a house
- Self-driving city taxis that analyze and adapt to traffic conditions to reduce traffic accidents and improve public services
- A surgeon on the east coast using robotics technology on the other side of the country to perform a life-saving procedure on a patient
- Mobile applications for real-time pandemic monitoring and public health support

These applications are just a few examples of new technologies being designed to enhance wellbeing, maximize device utilization, and even save lives. As IoT/IIoT technologies evolve with greater sensing technology, intelligence, and actionable capabilities, we will see an increase in the distribution of data out to the Far Edge on a global scale.

IT'S ABOUT PROXIMITY, NOT LOCATION

As more devices emerge or move to the Edge, they will require a more flexible network service delivery to maintain lower latency. Looking at it from the perspective of "IT everywhere," there are more end points, hubs, aggregations, and applications - all becoming more intertwined into a singular IT system.

Edge can no longer be just about a physical location as "the biggest benefit of Edge computing is the ability to process and store data faster, enabling more efficient real-time applications that are critical to companies." [3] With such a variety of meaning and locations within the Edge ecosystem, the true meaning of the Edge has gotten lost in translation and created significant complexity as shown in Figure 3 where we describe the 3 primary areas of Edge reach: Inner, Outer, and Far.

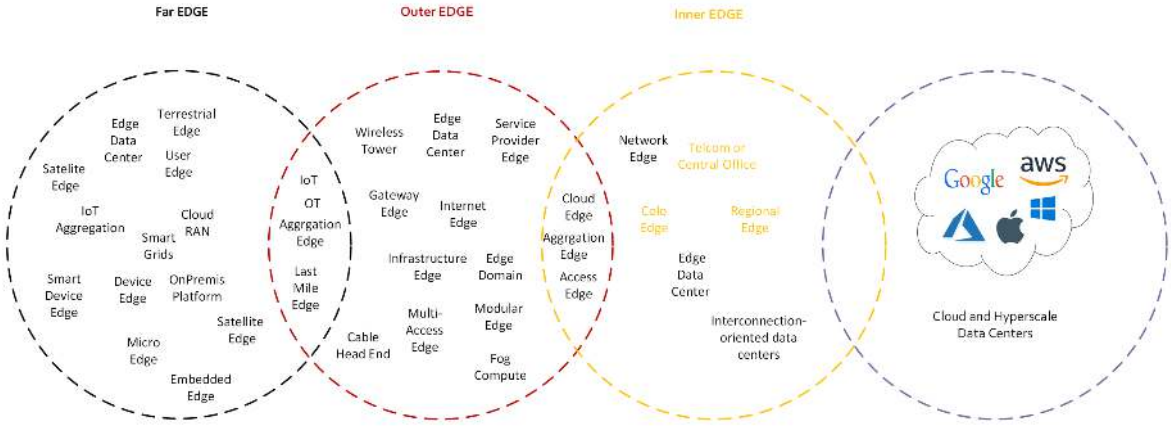


Figure 3: The Complexity of the Edge (by Jayson Hamilton)

IT proximity rather than physical location is what truly enables reduced latency and increased network performance. Being too focused on actual physical location makes it difficult to accurately characterize an Edge data center placement as there is a constant shifting and migration of applications that reside on the following three Edge layers:

Inner Edge

An Inner Edge data center has a primary function of aggregation for Far Edge compute processing or as a rural aggregation site away from the core data center. These data centers typically house primary IT systems and could contain less than 50 racks, residing in mid-sized or modular datacenters. Inner Edge data centers can operate under the TIA-942-B data center standard to reflect traditional compute models in rural or mid-sized markets. With primary IT systems housed at the Inner Edge, there is a need for high compute processing that has these data centers often resembling the functions of a core data center. Common applications for Inner Edge would include Edge colocation, AI, or VR.

Outer Edge

An Outer Edge data center offers access, radio, or cellular services as primary functions further away from the traditional datacenter. These data centers are typically small, self-contained, or single modular data centers with some level of hardened systems in compliance with codes and standards (e.g., NFPA, NEC, UBC, IBC, AHJ). Free standing exterior cabinets and indoor cabinets are subject to ingress protection (IP) or NEMA ratings due to the harsh environments. They typically serve as initial processing points for low latency or aggregation for Far Edge devices, housing small IT systems and containing up to 10 racks with low- to mid-level compute processing capacity to create IT proximity to the end device or application. Common applications for Outer Edge would be 4G/5G, Wi-Fi 6, and Edge colocation at a cell tower or access site.

Far Edge

The Far Edge is commonly defined as the actual Edge device - utilized for local or cache computation for an Edge application. Edge data centers are used for device aggregation as a primary function on the Far Edge - achieved by utilizing smaller devices and newer technologies to bring local computation into rural, isolated, or desolate areas. This could also apply to aggregation of Far Edge devices between the central office (CO) and Distributed Unit (DU) on a front haul network. Edge devices can support applications that require ultra-low latency in the network and therefore require close proximity to eliminate the distance between application and end user. Common applications for the Far Edge include smart factory operations and quality control, enhanced stadium experiences using augmented reality, remote location monitoring, and data filtering (i.e., mining operations, space station).

A SIMPLER FAR EDGE MODEL

Recent efforts within TIA's Data Center Work Group and TR-42 Engineering Committee to add an addendum to the current ANSI/TIA-942-B data center standards for establishing relevant Edge standards led to the defining characteristics and the development of a Far Edge model through proximity, function, and application use rather than by a physical location.

As shown in Figure 4, the Far Edge model is an all-encompassing, holistic layered model that identifies network performance, acknowledges IT proximity, and standardizes the language and locations in the Inner, Outer, and Far Edge layers of the Edge ecosystem. While existing Edge names or locations are defined by their functionality or technology, this Far Edge model is designed to reduce confusion and help develop greater commonalty across the IT proximate locations of where local application data is processed.

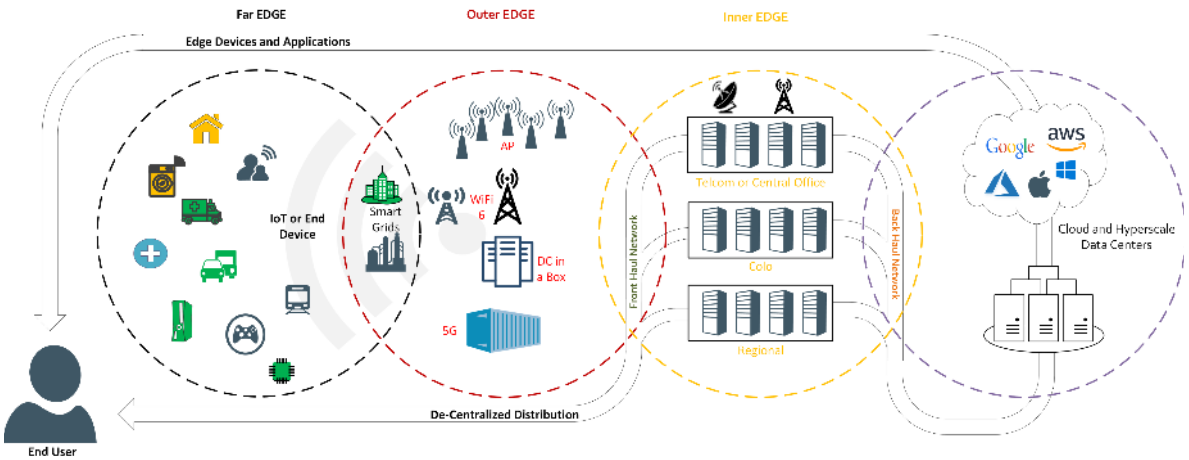


Figure 4: The Simplicity of the Layered Edge Drawing (by Jayson Hamilton)

Our TIA working group identified the following five characteristics that define the Far Edge model and establish the network's position to distribute data across all three layers of the Edge.

Decentralized Network

The decentralized network allows for traditional compute models to be moved closer to the applications' I/O. The Edge compute model utilizes decentralized compute systems across decoupled networks to increase network performance and to reduce application latency. In this model, the core data center is not used for the primary function of local computation. The core data center could still be used for secondary functions such as data aggregation, backup, or reporting.

Network Capacity

Capacity is commonly thought as being coupled with data rates of the network services. While the need for higher throughput can be a driving factor of the Edge network, requirements do not always revolve around data rates. There can be a need for replicated network continuity that needs the same bandwidth throughput to enable seamless transition to and from multiple network operators. This can also be seen as data is aggregated across the back, mid, or fronthaul.

IT Proximity

Proximity has the greatest impact on the Edge network's primary purpose, to reduce latency and increase performance levels. IT proximity means reduction of distance between the end user and the local computation asset or assembly of data. This is key to support an application's ability to process local data in real time without distribution across a traditional or cloud network. Shifting the focus away from a singular physical location to IT proximity provides a flexible and more adaptable model to follow.

Application Use

In today's world, the application and user are one. We live in a mobile-driven society that demands reliable application delivery everywhere we go. Depending on how an application is used will determine its importance and priority level, and in turn, will drive demand accordingly. The local computation of data in an Edge compute model helps meet mobile demands.

A Secured Edge

Moving further out to the Far Edge creates separation from an organization's skilled labor and impacts their ability to make timely changes or respond to an IT incident. To ensure the survival of the Edge data center or device, it must have a minimal level of remote access with both physical and logical protection. The functions of remote access and monitoring are therefore vital to the survivability of applications at the Far Edge as they allow organizations to make the required IT changes needed to ensure optimal application performance. Physical protection is critical since the common components of an IT system are active electronics that will fail to perform if exposed to harsh environments or are physically damaged. Logical protection is twofold—by providing an adequate level of encryption for cyber security, but also data protection or data replication are necessary to prevent data loss.

Understanding the five defining characteristics—decentralization, capacity, IT proximity, application use, and security—help develop best practices for the local data sets based on their needs and the application's requirements. The transition to the Far Edge model using the Inner, Outer, and Far Edge layers as the three primary IT proximate locations creates a reference model while also promoting industry collaboration to build a universal edge model.

CHALLENGES OF THE FAR EDGE

The Edge computing model can be considered disruptive because it requires a different network deployment approach. As the Edge computing continues to expand to the Far Edge, keeping a singular IT system operational within the requirements of its primary function becomes increasingly challenging.

Moving to the Far Edge uncovers new challenges, including bandwidth capacity, backhaul avoidance, session control, and localized power strategies, each of which are described below.

Bandwidth Capacity

Today's networks do not have adequate capacity levels to accommodate the expected growth in the number of connected devices at the Far Edge. Thankfully, recent advancements in fiber optic technology are enabling higher data transmission rates across longer distances. Not only has cabling technology advanced to provide reduced attenuation and improved chromatic dispersion, but advancements in encoding technology have enabled a shift from Non-Return to Zero (NRZ) data modulation to newer 4-level pulse amplitude modulation (PAM4).

NRZ modulation traditionally transmits 2 logical bits over a single channel using a Hi/Low signal. In contrast, PAM4 transmits over four voltage levels, representing two logical bits of data per symbol over a single channel. As shown in Figure 5, this has enabled a 1 and 0 to be transmitted through more logical bit combinations like 00, 10, 01, and 11 instead of utilizing high and low signals, resulting in increased throughput and performance of the link.[4] For example, NRZ allows only a maximum of 25 Gbps (Gigabits per second) per electrical lane, while PAM4 allows for 50 Gbps and is expected to soon support 100 Gbps lanes, which has paved the way to 400 and 800 Gbps transmission speeds.

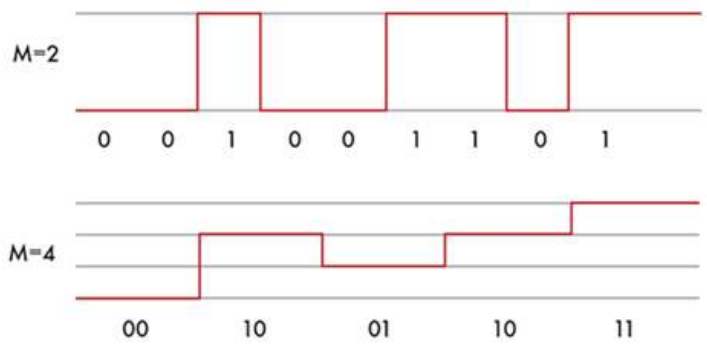


Figure 5: NRZ represented by M=2 and PAM4 represented by M=4

Increased bandwidth capacity in front and backhaul networks enables the level of scalable growth needed at the Edge. Additional information about NRZ vs. PAM4 modulation schemes can be found in the references section at the end of this paper.[10]

While the shift from NRZ to PAM4 helps to explain some of the advancements in bandwidth capacity, the industry is also seeing a shift at the software layer through a practice known as "network slicing."

In essence, network slicing allows the creation of multiple virtual networks on a shared physical infrastructure. Utilizing Software Defined Networking (SDN), Data Compression and Network Function Virtualization (NFV) to act as the underlying physical network provides more flexibility and greater potential for operational automation.

Backhaul Avoidance

The backhaul network is the back end of a traditional network connecting the Central Office (CO) to the Centralized Units (CU) for data aggregation of mobile data usage. The fronthaul is the continued distribution from the Distributed Units (DU) to Cell Towers, Nodes, or Radio Units (RU) as an access layer to the end device or user.

Traditionally, these network components have depended upon each other, but as we move out further to the Edge, these networks can start to decouple from each other as shown in Figure 6.

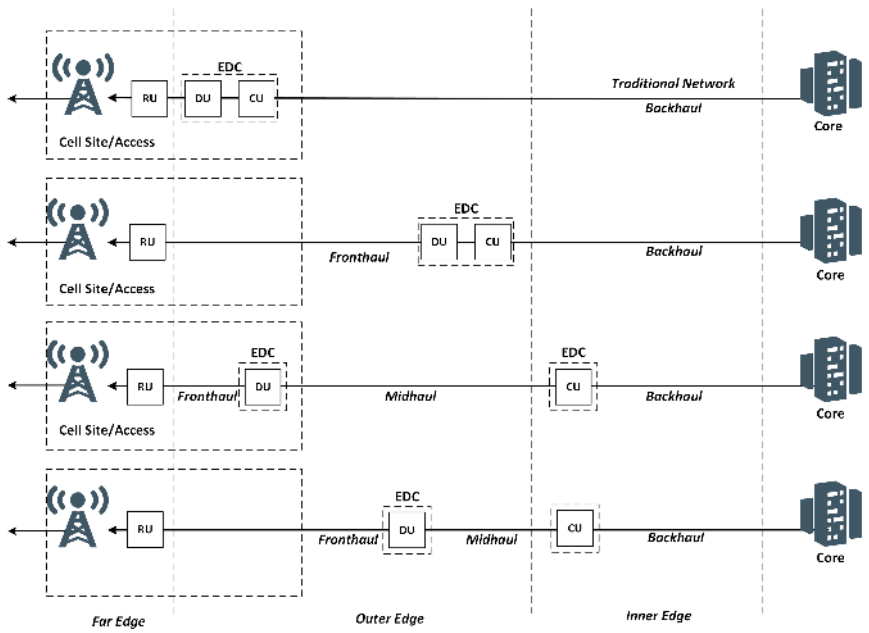


Figure 6: Examples of Fronthaul Networks and EDCs on the Edge (by Jayson Hamilton[5])

At the Edge, the fronthaul network can become the primary content distribution medium, as the baseline costs of distribution across a backhaul network can, at times, exceed the costs of the application. According to a report by ACG Research on hybrid cloud economics, there is a more than 1600% difference between transport costs at the core versus the Edge.[6] This is obviously driven by the application or the critical nature of the data set. For example, the current generation of Virtual Reality (VR) headsets run best with a 25 Mbps (Megabit per second) connection. Next-generation VR headsets are projected to require between 50 Mbps and 400 Mbps connections with some advanced VR applications requiring up to 1 Gbps of bandwidth.

For some applications, the costs associated with sufficient backhaul bandwidth to support the application correlate with the revenues and can be a strain on the IT budget. The question then shifts to whether the application can be deployed without the expenses of the backhaul network to understand costs and determine if the IT budget or revenue streams can support the application. As the users, devices, and common components are added up to calculate the raw bandwidth requirements, optimization solutions like compression or multiplexing can be applied to identify the number of transport circuits needed. All these factors establish the baseline cost for the backhaul network needed to determine its financial impact on application delivery.

In many cases, the cost of backhaul may exceed the cost of the Edge infrastructure. While the conceptual analysis is simple, the details can be fairly complex and should be addressed in another more specific paper.

Application Mobility and Session Control

Application mobility requires maintaining network connections as a device moves from one location to another, such as cellular phones where voice and data connections are handed off from one cell tower to another. Today, this transfer works seamlessly—it has become rare for a connection, a call, or a video conference to “drop” during a session transfer. While the transfer feels seamless to the end user, it’s important to understand that there is a lot going on behind the scenes that involves multiple connection points and equipment to make transfers happen in milliseconds, so they are imperceptible to the user.

The cellular network was established globally by a limited number of telecommunications providers and now functions essentially as a single converged network. However, incorporating the Edge computing model demonstrates a need for all-encompassing and expansive multi-tenant networks with numerous network operators. This will have a significant impact on the Edge in the form of a rigorous set of requirements needed to enable application mobility. While ubiquitous control functions (e.g., access, permission, billing, throttling, and others) have not yet been developed for the Edge, they will be required eventually. These functions will be analogous to those utilized in cellular networks today and they are foundational elements for the financial incentives that will drive innovation as it did for cellular network development.

Power and Cooling

Power and cooling are critical for data centers, which also holds true at the Edge. Having automated environmental controls with remote management and alerting is also essential, especially at the Outer and Far Edge where specialist resources are often limited. Similarly, reliability, availability, and survivability (RAS) of Edge services depends on both power and cooling being resilient and well maintained; most likely requiring service contracts with appropriate support and services levels in place. Environmental impacts also need to be addressed during maintenance activities as this practice includes preventing rapid rates of change (temperature and humidity), dust intrusion and condensation.[7]

Power and cooling requirements at the Far Edge need to be carefully addressed early in the design phase. Depending on the resiliency needs of the Edge data center, there may also be a need for emergency backup power and cooling equipment, ideally reliable form(s) of sustainable electricity generation (e.g., solar, wind, wave) with adequate battery storage. The latter solutions, in conjunction with innovations in liquid cooling, should be encouraged as they help reduce the overall operational carbon footprint. Responses from the last Data Center World Annual Survey demonstrated that, "trends are indicating a broader focus on performance, density, and efficiency, with 65% seeing an increase in renewable energy utilization, and 87% having implemented or plan to implement Edge locations within the next three years." [8]

Thermal modeling is an essential tool that helps ensure Edge data centers operate efficiently and within expected tolerance ranges. Thermal management must consider multiple facets of cooling, from the individual chips to the blades to the servers to racks and rows of racks. Space becomes an even more critical design element within the Edge Data Center due to tradeoffs between power density and the physical space element. A former scientist at the Lawrence Berkeley National Laboratory stated, “By 2028, … the total global power footprint of Edge IT and Edge data centers would be 102,000 megawatts (102 gigawatts).” [9]

There are concerns that powering the Edge at scale could potentially create an enormous carbon footprint. Hyperscale data center operators are already addressing sustainability through their commitment to be carbon neutral by 2030 via more sustainable practices such as committing to using only renewable energy sources. Most telecommunications firms are joining with the data center community in committing to becoming carbon neutral in the same timeframe. TIA also offers businesses an affordable and useful sustainability tool called the Sustainability Assessor, which helps businesses create and measure performance in meeting sustainability goals.

Figure 7: 80 megawatts in capacity, including Blue Cypress in Indian River County. (Source: FPL)

SUMMARY

We have covered several subjects in this whitepaper. While still in early days of the edge computing model, global data centers will continue to evolve at a significant pace. There will be an exponential number of deployments and continued growth with the global Edge Data Center market, which is expected to reach more than $40B by 2028 according to Zion Market Research. Additionally, IDC predicts that more than 55 billion IoT devices will create 73 trillion gigabytes of data by 2025. There are no signs of slowing in organizations from every vertical industry and sector which are deploying IoT devices to monitor a plethora of statistics and events to meet the current demand.

To keep up with this growth, the industry needs to look at IT proximity rather than a physical location as it reduces latency and increases network performance. By taking a holistic and simplified approach to define the Edge based on the proximity, function, and application, we can help reduce conceptual complexity and move the industry towards a more consistent set of standards for ensuring high-performance, resilient, and secure Edge data centers.

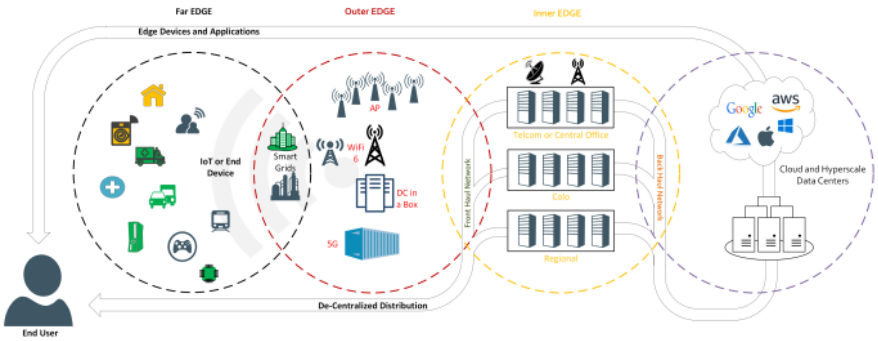


Figure 8: The Simplicity of the Layered Edge Drawing (by Jayson Hamilton)

About Telecommunications Industry Association (TIA)

The Telecommunications Industry Association (TIA) represents more than 400 global companies that enable high-speed communications networks and accelerate next-generation ICT innovation. Through leadership in U.S. and international advocacy, technology programs, standards development, and business performance solutions, TIA and its members are accelerating global connectivity across every industry and market. TIA is accredited by the American National Standards Institute (ANSI).

To learn more about the TIA's Data Center Working Group or developing standards for Edge Data Centers (EDCs), please reach out to datacenterinfo@tiaonline.org.

*** Disclaimer:** The information and views contained in this paper are solely those of its authors and do not reflect the consensus of opinion for TIA, TIA members or of TIA Engineering Committee TR-42. This paper is for information purposes only and it is intended to generate opinion and feedback so that the authors and TIA members may continue to learn, refine, and update this content over time. The Telecommunications Industry Association does not endorse or promote any product, service, company, or service provider. Photos and products used as examples in this paper are solely for information purposes and do not constitute an endorsement by TIA, our members or committees.

APPENDIX A: GLOSSARY

Note: TIA collaborates with the Linux Foundation in maintaining the Open Glossary of Edge Computing. Terms included in this paper are listed below. For the full list of Open Edge Compute Glossary terms refer to the official repository: https://github.com/State-of-the-Edge/glossary ome may argue about who gets the credit for inventing the internet, in the early 1960s, the Department of Defense (DoD) funded the Advanced Research Projects Agency Network (ARPANET), which was essentially internet 1.0.

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

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

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

Availability
The ability of a system to maintain above average levels of uptime through design and resiliency characteristics. At the Edge, availability is even more important due to the unique requirements of the use cases and the compute support necessary. Due to the distributed nature of Edge data systems, they will be able to provide the high availability required.

Autonomous Vehicle
A vehicle capable of navigating roadways and interpreting traffic-control devices without a driver actively operating any of the vehicle's control systems. In the case of the autonomous vehicle, the contextual Edge becomes an integrated component of in-vehicle computing and applications that provide autonomous and extreme low-latency processing.

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.

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.

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.

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

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

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.

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

Cooked Data
This data set is raw data that has been analyzed and processed, also known as cooked.

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.

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 EDC 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.

Decentralized, Distributed Architecture OR Distributed Edge Cloud Architecture
Distribution of computing power and equipment at the Edge of the network spread across different physical locations to provide closer proximity to the use cases of said compute. Components are presented on different platforms and several components can cooperate with one another over a communication network in order to achieve a specific objective or goal. A distributed system is a system whose components are located on different networked computers, which then communicate and coordinate their actions by passing messages to one other."

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

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.

Distributed Computing
A Compute Model that has shared software resources across multiple computing systems to allocate processing needs. Using multiple distributed computing systems helps you improve efficiency and performance because you can pool resources, so they don't max out processing units or graphics cards.

Edge Appliances
Edge appliances refer to gateway devices that do not sensors themselves to be able to generate data for example to measure temperature or pressure, listen audio or record video. Edge appliances require sensors to be connected to them. The benefit of using these devices is that they are often capable for taking input from several sensors simultaneously and running several AI models in parallel.

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 but not exclusively, in close proximity to the last mile network, on both the infrastructure and device sides.

Edge Data Center (EDC)
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.

Edge Devices
Edge device is a device that has one or multiple sensors and can run a local compute model, process data, and take actions without a roundtrip to cloud.

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.

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.

Far Edge
A potential site location where the compute model moves even closer to the end user but utilizing smaller devices and newer technologies to bring local computation further away from the core site and into rural, isolated, or desolate areas. These Edge sites are relative to and commonly defined by Utility, Radio or Cell Sites.

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

Internet of Things (IoT)
A large increase in machine-to-machine communication will produce large increases in bandwidth requirements. Machine response times are many times faster than human interactions, creating even more opportunities to benefit from ultra-low latency communication. Data traffic patterns will change, with local peer-to-peer relationships and localized IoT data/control systems generating data that does not need to traverse backhaul links that will be hard pressed to carry the flood of traffic to come. Distributed EDCs will transform large volumes of raw data into valuable information that can then be transferred to more centralized tiers where required.

IoT Devices
IoT Devices refers to sensors that can gather and transferring data, but not powerful enough to process data in a device. These apps send telemetry to your IoT hub, and optionally receive messages, job, method, or twin updates from your IoT hub. You can connect virtually any device to IoT Hub by leveraging different device SDKs.

Infrastructure Edge
Edge computing capability, typically in the form of one or more Edge data centers, which is deployed on the operator side of the last mile network. Compute, data storage and network resources positioned at the infrastructure Edge allow for cloud-like capabilities like 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.

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.

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.

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

Inner Edge
A potential site location for an Edge data center that is being utilized is in a traditional environment such as an existing 942-B data center, building or any facility that uses traditional compute models. This Edge model can be a proximity site inside a traditional 942-B data center environment in a rural area or even at the bottom of a basement where traditional applications reside. These Edge sites are relative to and additionally commonly by Core and Aggregation Sites.

IP Aggregation
The use of compute, data storage and network resources at a layer beyond the infrastructure Edge to separate and route network data received from the cellular network RAN. Although it does not provide the improved user experience of local breakout, IP aggregation can improve performance and network utilization when compared to traditional cellular network architectures.

Intelligent device
This is any equipment, machine or instrument that has its own computing capability. Aside from personal computers, examples include cars, home appliances and medical equipment.

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.

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).

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.

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.

Lights-out management (LOM)
The ability remotely monitors and manage servers. Hardware for this setup comes in a module and logging software, which tracks microprocessor utilization and temperature within the Edge data center.

Microcontrollers
A microcontroller is a small computer on a single metal-oxide-semiconductor integrated circuit chip. A microcontroller contains one or more CPUs along with memory and programmable input/output peripherals.

Network functions virtualization (NFV)
A network architecture concept that uses the technologies of IT virtualization to virtualize entire classes of network node functions into building blocks that may connect, or chain together, to create communication services.

Outer Edge
A potential site location where the compute model is moving away from traditional data center space and into a containerized or similar type harden solutions outside on the Edge near rural areas. The outer Edge brings local computation closer to the end user. These Edge sites are relative to and commonly defined by Aggregation and Access Sites.

Raw data
This data is collected but not applied to any specific use case; sometimes, it is automatically filed into a database before any type of processing. This raw data can come from a variety of devices and infrastructure components that are a part of the company's IT network.

Redundancy
Redundancy is typically defined as the inclusion of extra components which are not required for function, but to provide continued availability in the event of a failure. Part of the RAS Model.

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.

Survivability
Survivability is defined in how it relates to the percentage of time a facility is actually available versus the total time it should be available to provide operational continuity and resiliency.

Unmanned Edge Data Centers
Due to their size and distributed nature, most Edge data centers are expected to be unmanned, creating new security (surveillance, anchoring, alarms, etc.), maintenance (additional monitoring systems and sensors), operational (increased automation), and resiliency considerations that do not exist with manned data centers.

APPENDIX B: REFERENCES

I. **Tools, T. (2020, June 19).** Network Latency: Definition, Causes, Measure and Resolve Issues. Retrieved January 12, 2021, from https://www.tek-tools.com/network/resolve-network-latency-issues

II. **Haranas, M. (2020, January 16).** 8 Biggest Data Center Trends and Technologies to Watch For In 2020.from https://www.crn.com/slide-shows/data-center/8-biggest-data-center-trends-and-technologies-to-watch-for-in-2020/3

III. **Shaw, K. (2019, November 13).** What is Edge computing and why it matters. Retrieved January 26, 2021, from https://www.networkworld.com/article/3224893/what-is-Edge-computing-and-how-it-s-changing-the-network.html

IV. **J. Van Kerrebrouck et al.**, "NRZ, Duobinary, or PAM4?: Choosing Among High-Speed Electrical Interconnects," in IEEE Microwave Magazine, vol. 20, no. 7, pp. 24-35, July 2019, doi: 10.1109/MMM.2019.2909517. https://ieeexplore.ieee.org/document/8726447

V. **Lavoie, L. (2020, April 07).** 5G fronthaul networks add flexibility. Retrieved April 14, 2021, from https://www.5gtechnologyworld.com/5g-fronthaul-networks-add-flexibility/

VI. **Marangella, P. (n.d.).** The Next Evolution of the Cloud is at the Edge. Edge Industry Review. https://www.Edgeir.com/the-next-evolution-of-the-cloud-is-at-the-Edge-20210519/amp.

VII. **ASHRAE. (2020, Sept).** Edge Computing: Considerations for Reliable Operation. Retrieved from ASHRAE TC 9.9 Technical Bulletin: https://www.ashrae.org/file%20library/technical%20resources/bookstore/tb_Edgecomputing_sep2020.pdf

VIII. **Kleyman, B. (2021, March).** EVP of Digital Solutions at Switch, AFCOM Contributing Editor. The 2021 State of the Data Center Report. Retrieved May 2021

IX. **Koomey, J. (n.d.).** Just how much energy will Edge data centers consume? Retrieved May 05, 2021, from https://www.iceotope.com/lab/just-how-much-energy-will-Edge-data-centers-consume/

X. **J. Van Kerrebrouck et al.**, "NRZ, Duobinary, or PAM4?: Choosing Among High-Speed Electrical Interconnects," in IEEE Microwave Magazine, vol. 20, no. 7, pp. 24-35, July 2019, doi: 10.1109/MMM.2019.2909517. https://ieeexplore.ieee.org/document/8726447

TIA

BRIEFING PAPER

THANK YOU!

WHERE IS THE EDGE?