WHITE PAPER

The Adaptive Network: A Framework for Understanding the Networking Implications of the Edge Cloud

Executive summary

Cloud services are pervasive. From individual users binge-watching Over-The-Top (OTT) video services to enterprises deploying Software as a Service (SaaS), cloud services are how people and organizations consume content and data. For years, large, centralized data center and cloud architectures have provided access to these services.

Now, a new generation of cloud-native applications is emerging in categories such as entertainment, retail, manufacturing, and automotive, which, in many cases, will be more compute-intensive and latency-sensitive. Traditional centralized cloud architectures will not meet the Quality of Experience (QoE) expectations for these applications and will require a more dynamic and distributed cloud model. As a result, compute and storage cloud resources need to move closer to the edge of the network, where content is generated or consumed, to meet the expected QoE. This new approach is referred to as Edge Cloud.

This shift to a distributed Edge Cloud model will result in an estimated three times as many data centers at the network edge as there are today, and will require the entire cloud ecosystem to think differently about the role of network connectivity[1].

This paper examines the drivers and implications of edge computing and explores how Ciena's Adaptive Network™ vision can provide an effective framework for the evolution to a distributed Edge Cloud architecture.

The world is changing, with data moving closer to the network edge

Imagine someone who is about to leave for a business trip making a simple request: *"Hey Siri, give me directions to the airport."* Within seconds, Apple Maps determines the shortest route and provides turn-by-turn directions. En route, a pop-up notification provides new directions to avoid traffic congestion ahead, adding just a few minutes to the trip. In each case, the Apple Maps request was likely served by a centralized data center that could be thousands of kilometers away, resulting in additional latency (delay) in processing the request. For a non-critical application like Apple Maps, such a response time is acceptable, and usually does not affect a user's ability to correctly navigate to their destination.

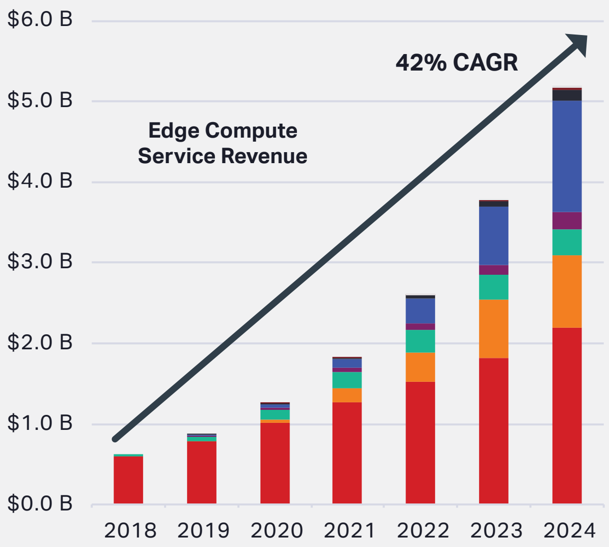
In another scenario, a shopper walks into their favorite grocery store, checking in via their smartphone app. Their movements—picking up and putting down items from the shelves—are captured by cameras embedded in the store ceiling. Computer vision AI analyzes these images to determine what the shopper has purchased and bills their credit card directly, eliminating the need for them to go through a cashier at the check-out counter. Significant compute resources will be needed either in the retail stores or at the edge to perform this near-real-time image processing to deliver a seamless customer experience.

An additional driver for computing at the edge is the value of processing the vast amounts of data generated locally by these applications and reducing backhaul traffic back to the central cloud. The goal is to reduce the latency and amount of backhaul traffic to the central cloud and more effectively serve the large-scale analytics required to push inferences and predictions to the devices at the edge, improving application performance.

Numerous emerging applications require lower latency than can be serviced via centralized data centers, as shown in Figure 1. Revenue for these applications is expected to see a 42 percent CAGR, from US$1.2B in 2020 to over US$5B in 2024, with the largest revenue drivers coming from video/content delivery networks, cloud gaming, and automotive applications[1].

[1] Mobile Experts: "Edge Computing for Enterprises 2019", July 2019.

	Application	Description
🟥	Video/CDN	Content needed to be cached at the local edge to improve performance
🟧	Cloud Gaming	Replace gaming consoles by streaming games from low latency cloud servers at the Cloud Edge
🟩	Industrial IoT/ Automation	Edge computing reduces latency for mission critical remote applications
🟪	Smart Venues	Processing needed at the venues to reduce response times
🟦	Automotive	Infotainment, Autonomous Driving
⬛	AR /VR	Low latency to reduce motion sickness



Source: Mobile Experts, "Edge Computing for Enterprises 2019", July 2019.

Figure 1. Top applications by revenue for edge computing

In addition to the need to support application latency and QoE requirements, further challenges come from the mobile side of the network. Mobile applications such as cloud gaming and augmented/virtual reality are increasingly compute-intensive, which negatively impacts mobile device battery performance. In this case, when compute and storage from the mobile device can be moved to a cloud data center closer to the user, users experience improved mobile battery performance.

Additionally, in response to privacy concerns, several governments are mandating cloud service providers store their customers' data in country. For a market like Europe, many cloud companies have traditionally hosted their cloud resources in countries like the UK and serviced customers centrally all over Europe. With the advent of governments erecting digital boundaries, cloud providers are being mandated to host those cloud resources in edge data centers within each country, closer to their users.

The industry approach to address these challenges is to create a more distributed and dynamic cloud model, which involves moving cloud resources from centralized data centers closer to the user, in data centers at the edge.

This paper addresses the network implications of a distributed cloud approach, referred to as the Edge Cloud. Ciena defines the Edge Cloud as a interchangeable cloud ecosystem that encompasses edge computing components (storage and computation) from multiple vendors as well as a scalable, application-aware network interconnecting edge data centers that can sense and adapt to applications needs securely, and in real time.

Where exactly is the edge?

While many in the industry are attempting to establish static definitions of where the edge exists, the reality is that the edge will reside at any number of locations, depending on the QoE expectations and resource requirements/availability for a given application. The location of the Edge Cloud will vary depending on the perspective of a user, network operator, or application provider.

This paper refers to the following groups of locations where an application could physically reside, as shown in Figure 2:

1. **Metro Edge:** A mix of large, multi-tenant data centers (Global Content Networks [GCNs] and Data Center Operators [DCOs]) and Communications Service Provider (CSP) hub central offices retrofitted as data centers, situated in regional/metros to service that market

2. **Far Edge:** A mix of CSP central offices, Cable Multi-Service Operator (MSO) head-end or Mobile 5G Distributed Unit (DU) locations, situated closer to the user

3. **User/On-premises Edge:** A mix of large and small enterprise locations, including enterprise data centers and branch offices; could extend to transportation hubs, mining sites, and manufacturing plants

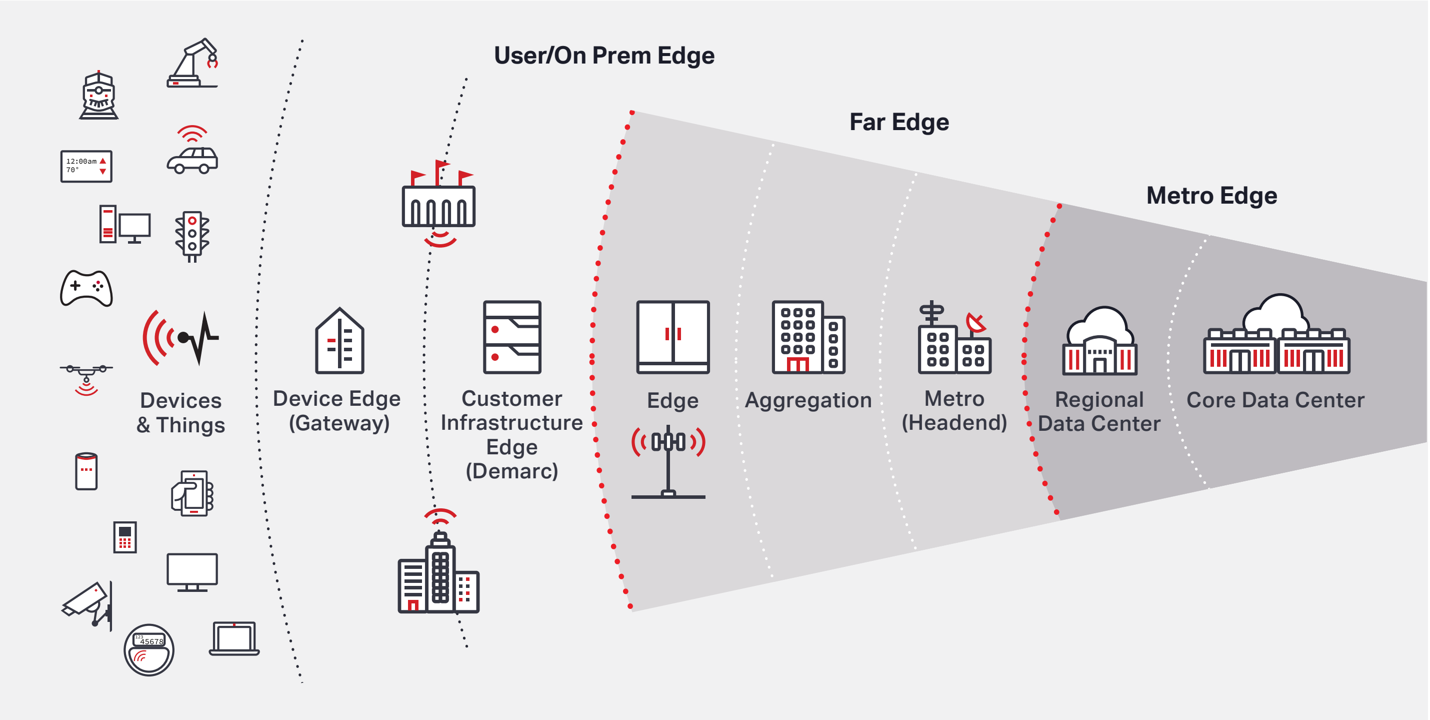
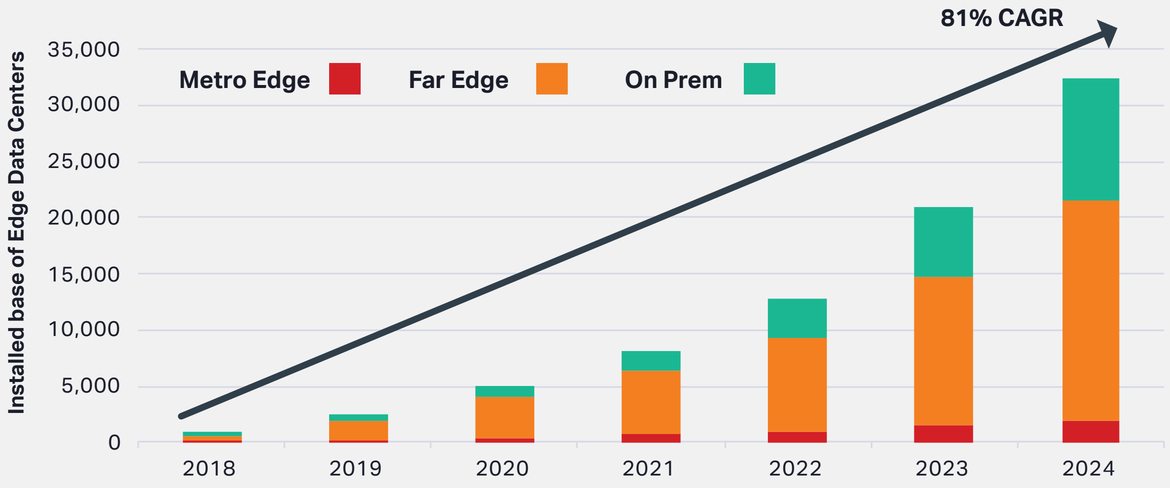


Figure 2. Locations of the edge

The emergence of the Edge Cloud is blurring these lines, as partnerships form among a wide variety of Edge Cloud players with the ability to deliver in-cloud resources that span multiple clouds, service providers, or solutions developed by the GCN or end-customer themselves.

Today, there are approximately 10,000 data centers globally. With the expanding shift toward Edge Cloud, forecasts (as shown in Figure 3) show that there will be up to three times as many new data center locations at the Metro/Far Edge and User/On-premises Edge groupings within the next four years[1].



Source: Mobile Experts, "Edge Computing for Enterprises 2019", July 2019.

Figure 3. Growth of edge computing data centers

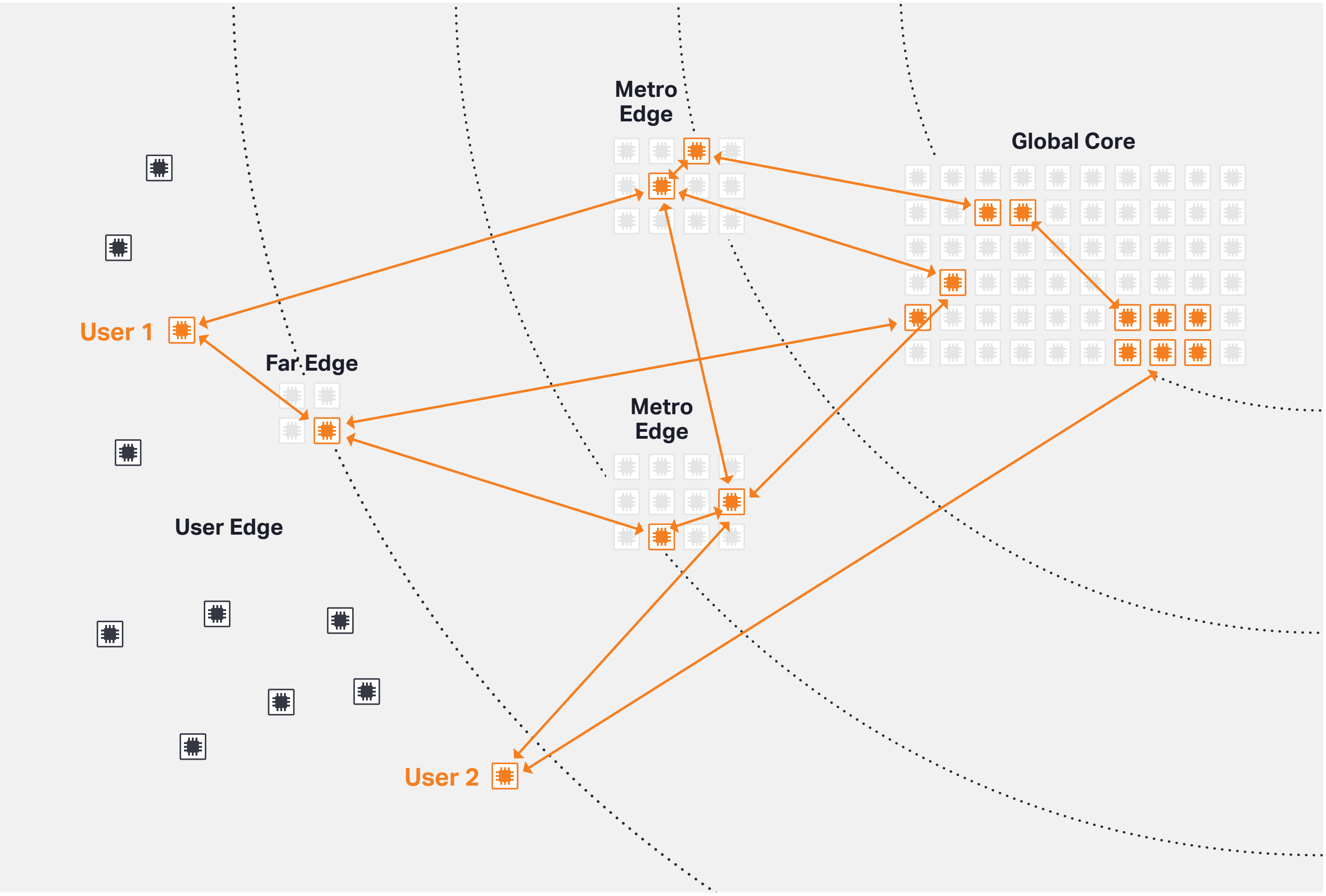


Figure 4. Dynamic nature of applications enabled by Edge Cloud

It is not just the definition of the edge's location that is fluid; it is the nature of applications at the edge of the network that could be static or dynamic. Each square in Figure 4 represents a computing resource where an application, microservice, or network function could reside to support an end-user (enterprise or consumer). A user application could tap into compute resources in multiple Edge Cloud locations during the lifecycle of using that application. Different users could be tapping into cloud resources at any of the Far Edge, Metro Edge, and/or Global Core locations, depending on the nature of the application and availability of cloud resources needed to meet the QoE requirements for the duration of the application session. It is the dynamic nature of how applications will move across different Edge Cloud locations that necessitates new networking requirements to support the Edge Cloud.

Who are the Edge Cloud providers?

In addition to the application developer, several providers that will leverage the Edge Cloud—Hyperscalers, GCN operators, DCOs, and CSPs—will all need to form new business relationships to enable it. As the race to build out the Edge Cloud expands, these ecosystem providers will need to work together. Partnerships are already emerging and will continue to play out in the coming years.

The following section examines how the different Edge Cloud providers are expected to evolve their data center strategies to enable the Edge Cloud (Figure 5).

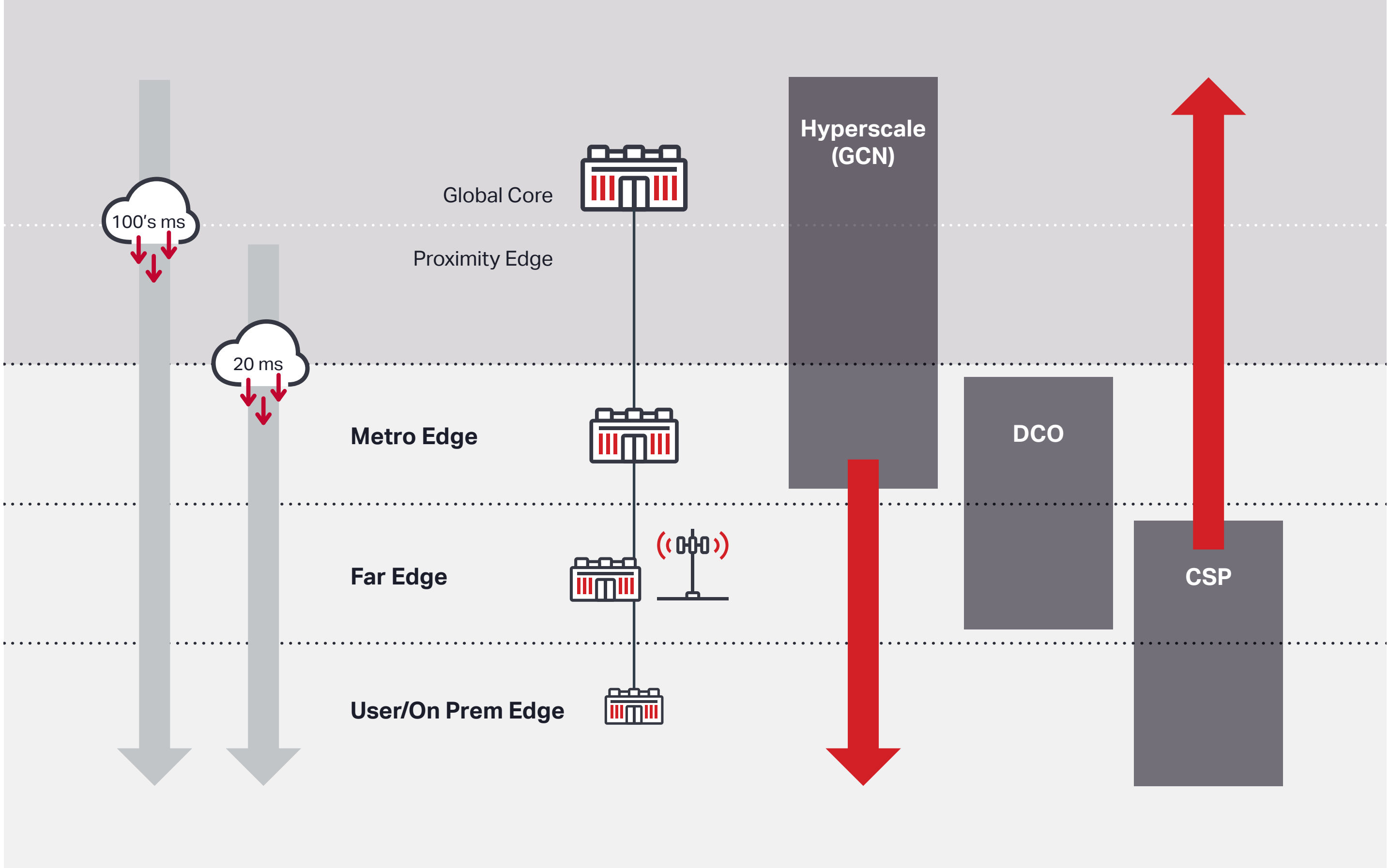


Figure 5. Edge Cloud providers versus edge data center site

GCNs: GCNs have built out an extensive suite of hyperscale data centers (global core) as part of a decentralized cloud architecture. They have also partnered with DCOs to expand their metro footprint into the edge of the network. However, to scale into more distributed and lower-latency cloud services, they will need to further expand their edge footprint into the Far Edge and User/On-premises Edge locations, either by building their own or via partnerships with CSPs who dominate this space.

Strategic partnerships between GCNs and CSPs are already emerging, such as in the case of AT&T with Google Cloud and Azure, or Verizon with AWS. Due to the dynamic nature of how Edge Cloud services will be consumed, GCNs will expect DCOs' and CSPs' networks to provide greater awareness of the Edge Cloud in the context of the GCN-powered applications running on top of them. For example, GCNs streaming OTT video services in their clouds must be aware that there are sufficient network resources from the CSPs to deliver a consistent video streaming QoE to consumers.

DCOs: DCOs have built out vast numbers of data centers in metro locations to enable their enterprise customers to outsource their data center infrastructure, offer space and power to cloud providers, and provide an exchange to content and cloud providers. Their primary revenue sources are derived from real estate, power, and connectivity, but they recognize the need to move up higher in the cloud stack to bring additional value and margins to their businesses. They will continue to play a key role in expanding the Edge Cloud.

CSPs: CSPs currently dominate in providing connectivity and infrastructure to the end-user—whether enterprise or consumer—and have built out thousands of central office/head-ends over the years to do this. They are also virtualizing their internal networks in an edge architecture with local proximity to the end-user. As mentioned above, CSPs are pursuing new partnerships with the GCNs, creating awareness of the underlying network resources connecting their Edge Cloud locations to the applications provided by the GCNs, and vice versa.

Requirements to deliver Edge Cloud services

The dynamic nature of Edge Cloud requires the various ecosystem players to think differently about their networks. To be successful with an Edge Cloud deployment, GCNs, DCOs, and CSPs need to understand the requirements their networks will face, and how they must respond. Key Edge Cloud network requirements are outlined as follows:

1. **Application awareness**: Application networks will define the next-generation networking approach for cloud services and applications. Currently, applications operated or hosted primarily by GCNs run over a virtualized infrastructure that is abstracted from the physical infrastructure. To enable efficient operation of networked applications across geographically distributed edge computing resources, it is necessary for application and physical infrastructure networks (overlay and underlay, see Figure 6) to be aware of each other's characteristics and requirements.

2. **Network and application workload visibility and placement:** Meeting the dynamic demands at the network edge requires both application and network providers to have greater visibility at the infrastructure and application layers, to observe where congestion is forming and where issues are anticipated. This level of visibility must be supported across all layers and within a multi-vendor environment.

3. **Security:** As applications become more distributed and dynamic, maintaining a consistent security posture has become increasingly more complex. When cloud resources and applications are centralized in a data center, enterprises can standardize technical and physical security more easily. An Edge Cloud approach introduces greater complexity by forcing the ecosystem to grapple with both emerging zero-trust security models and physical security parameters, but for three times more edge data center locations in a widely distributed footprint.

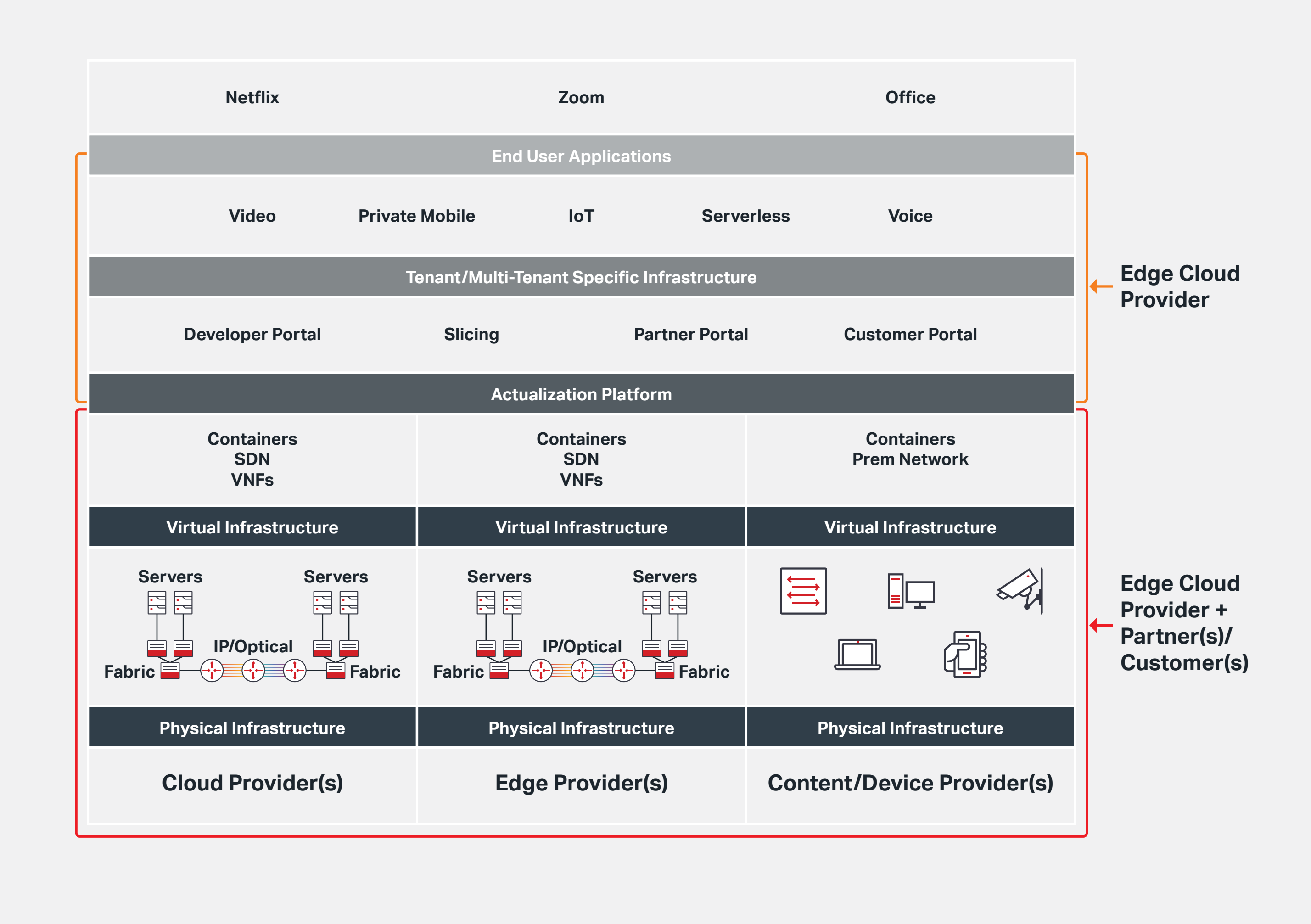


Figure 6. Components of the Edge Cloud ecosystem

4. **Analytics:** Real-time data is a key business asset, and the data pool will continue to grow as more data is collected from end-points, especially with the growth of IoT devices. Data analytics needs be considered from two perspectives:

    - Data-in-motion analysis – real-time/near-real-time ingestion and analysis of data; key meta-data will be sent northbound for storage, which will occur close to the source, such as at the edge
    - Data-at-rest – data will be aggregated/fused with other data sets and analyzed for various business intelligence in a centralized data center

    However, the data is not useful unless it can be analyzed to provide actionable insights, and the results operationalized. AI, Machine Learning (ML), and analytics tools are critical for understanding changing edge application demands and improving network and application performance.

5. **Scalable capacity for inter/intra-edge data center networks:** The adoption of edge computing creates new expectations for always-on, high-performance applications within and between the edge data centers and the central cloud. It is expected that the ratio of traffic in the metro part of the network will grow at a faster rate than the core, and the adoption of computing for high-bandwidth computation at the edge will only accelerate this trend. This core-to-metro capacity inversion will demand a scalable packet optical infrastructure for within and between these edge data centers, with continued connectivity to the central cloud.

6. **Intelligent Edge Cloud orchestration and automation:** To optimize the utilization of Edge Cloud resources while satisfying the demand for and requirements of dynamic applications, intelligent automation—with a bird's-eye view of the network and Edge Cloud—is required. Separate instances of orchestration will work across cloud/edge virtualization, platform, infrastructure, and application to place and interconnect the application components in suitable Edge Cloud hosts based on end-user location; application resource, QoS, and service specifications; host capability, capacity, cost, and availability; network capacity and performance; and operator, regulatory, tenant, and other constraints. Unlike the centralized cloud that delivers to a large number of customers, each edge application is specific to a much smaller subset of customers and must dynamically and automatically respond instantaneously to every local customer's requirements in an environment with limited resources.

7. **Edge Cloud slicing for multi-tenancy:** One of the opportunities for network providers from the Edge Cloud is to dynamically allocate different cloud and network resources for each tenant within and across their edge data centers. One can refer to this as Edge Cloud slicing, delivering end-to-end compute, storage, and network resources to the edge based on the tenant's application and SLA requirements.

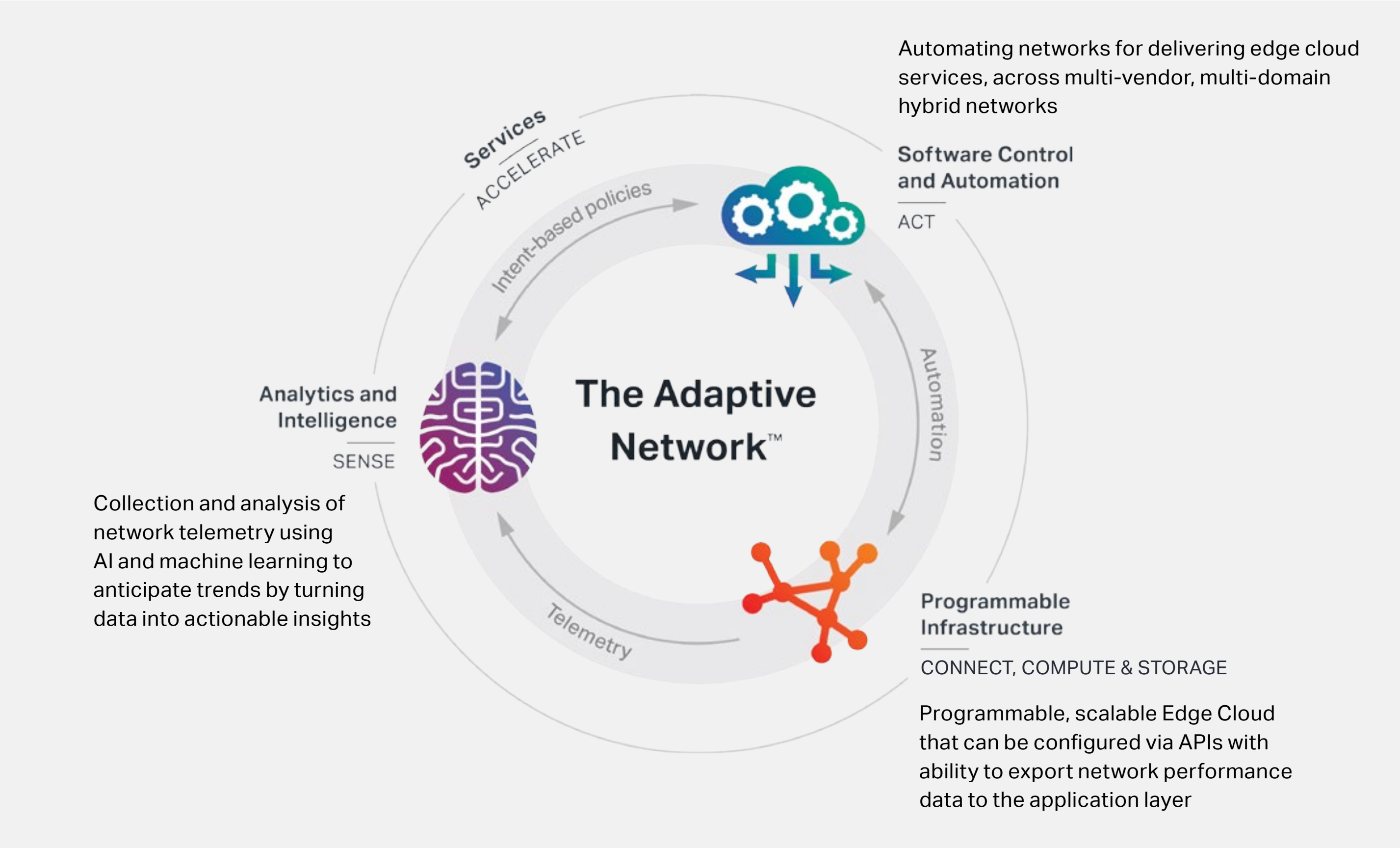
What is Edge Cloud? →

The Adaptive Network™ approach to enabling the Edge Cloud

The key challenge for edge providers is to efficiently and intelligently manage the network and application resources for Edge Cloud data centers during peak periods of usage. Ciena's Adaptive Network vision provides a framework for Edge Cloud that allows providers to collectively achieve an end-to-end network that grows smarter and more agile every day, with the scale required to respond dynamically to the pressures being placed upon it.

The Adaptive Network enables Edge Cloud providers to optimize their existing infrastructures while incorporating new technologies and ways of working to meet new requirements of Edge Cloud. The Adaptive Network is built on four key foundational elements—Programmable Infrastructure, Analytics and Intelligence, Software Control and Automation, and Services—that enhance network and business outcomes independently, but are a force multiplier when working together.

Programmable Infrastructure: A programmable edge packet and optical infrastructure is one that can be accessed and configured via common open interfaces, is highly scalable and instrumented with the ability to export real-time network performance data to the application layer of the Edge Cloud, and can adjust its resources, as needed, to meet the demands of the application layer. This will be key for enabling an application-aware network and providing scalability for interconnecting Edge Cloud fabrics between and within Edge Cloud data centers. Also, network slicing at the infrastructure layer will be essential for providers to enable Edge Cloud multi-tenancy services to different cloud providers and application overlays.


Services
ACCELERATE
Intent-based policies
Software Control and Automation
ACT
Automation
Programmable Infrastructure
CONNECT, COMPUTE & STORAGE
Telemetry
Analytics and Intelligence
SENSE
The Adaptive Network™

Automating networks for delivering edge cloud services, across multi-vendor, multi-domain hybrid networks

Collection and analysis of network telemetry using AI and machine learning to anticipate trends by turning data into actionable insights

Programmable, scalable Edge Cloud that can be configured via APIs with ability to export network performance data to the application layer

Figure 7. The Adaptive Network: a framework to enable the Edge Cloud

Analytics and Intelligence: Edge Cloud is an extension of cloud computing and operation practices, which rely heavily on automation informed through interpretation of massive telemetry/Key Performance Indicators (KPI) streams from underlying resources. Application awareness and automation will depend significantly on the collection and analysis using AI of network, server, and virtualized (container) resources, and the ability to anticipate trends by turning mountains of data into actionable insights. Leveraging these insights will deliver an application-aware network that can sense and adapt to edge applications' needs securely, and in real time.

Software Control and Automation: Automating the placement of Edge Cloud workloads to meet the demands of applications in near-real time will be critical to meeting the key constraints and goals of Edge Cloud. Through the implementation of SDN, NFV, and open APIs, providers can simplify the act of managing, securing, and automating their networks end to end for delivering Edge Cloud services across multi-vendor, multi-domain hybrid networks.

Services: Technical and professional services are required to help providers determine their best strategy and architecture for Edge Cloud, and to build, operate, and continually improve their networks, accelerating their journey to the Adaptive Network.

The Adaptive Network vision
Learn more

Summary

These are still the early phases of the deployment of Edge Cloud and the evolution to a distributed cloud architecture. The edge should not be thought of as a specific data center location, but will reside at any number of locations, depending on the QoE expectations and resource requirements/availability of a given application. The location of an application could shift to different edge data centers during its lifecycle, driving the need to intelligently scale infrastructure both within and between edge data centers and to the central cloud, while automating workloads between locations at the edge of the network.

While GCNs have successfully demonstrated their ability to scale within a centralized public and hybrid cloud model, moving to a distributed Edge Cloud model will require partnerships with DCOs and CSPs to take advantage of their extensive infrastructure and footprints closer to end-users. For a distributed Edge Cloud model to reach its full potential, new networking requirements will need to be addressed. Most notably, the application layers of the cloud stack must be dynamically aware of the resources across network layers, while the network layers must maintain awareness of the changing dynamics at the application layer.

Ciena's Adaptive Network solutions play a critical role in some of the largest inter-data center and cloud architectures in the world today. With the number one market share in both Global DCI and Metro DCI, Ciena is well-positioned to bring its depth of experience and leadership in the cloud and DCI markets to the edge. The Adaptive Network provides a further framework for all edge ecosystem providers to follow by addressing the challenges of the dynamic Edge Cloud model and leveraging a highly programmable and scalable infrastructure, analytics, and automation to dynamically scale both network and application cloud resources as required to meet end-user expectations. Following the Adaptive Network framework can help ensure that the performance of an Edge Cloud model can scale and adapt to meet the ever-changing demands of the network edge.

? Was this content useful? Yes No