Establishing the Edge

A new infrastructure model for service providers

Discover edge computing considerations for an efficacious services architecture

As the demands on your networks grow, simplicity, scale, and agility are essential. However, it's the end user experiences with services, content, and applications that drive the perceived value of your services.

This paper is the first in a series of papers that examine the aspects of building a new services infrastructure, which uses computing resources closer to human or machine subscribers. This paper provides a viewpoint from the network, describes some use cases and considerations towards meeting the latency requirements required for a new services infrastructure.

Contents

Why an edge computing model is needed

Today, more applications are moving to the cloud, and multiple clouds are being deployed. With the explosion of end-points, mobility, and nomadic computing, the volumes of data used for analytics, machine learning and automation can result in high costs to transport to central locations for processing. Traditional service provider architectures can no longer meet modern needs.

At the same time, connecting network has become critical in delivering high-quality experiences, application performance, and security across data, services, and applications. To solve these issues, a new services edge architecture is emerging that is based on distributing computing capacity to the edge of the network. This architecture results in lower latency with respect to subscribers. Throughout this paper, we use the term edge computing to refer to the general architectural shift, which most standards developing organizations (SDOs) such as ETSI refer to as multi-access edge computing or MEC.

Defining edge computing

Edge computing is the architectural principle of moving services to locations where they can:

- Yield lower latency to the end device to benefit application performance and improve the quality of experience (QoE).
- Implement edge offloading for greater network efficiency.
- Perform computations that augment the capabilities of devices and reduce transport costs.

For example, lower latency can improve the QoE of certain consumer applications. These applications range from the delivery of HTTP(S) content and video to augmented reality and virtual reality. Offloading may reduce the cost of networking and improve QoE as metro area peering points to the Internet are becoming more ubiquitous. The benefits in terms of customer retention and churn reduction are clear, but there also is an aspect of fundamental service enablement. In addition, some use cases associated with security and the Internet of Things (IoT) may see benefits associated with edge processing and edge analytics. These benefits are apparent when considered against the alternative of transporting vast amounts of data to a centralized data center.

Two types of workloads are being deployed at the edge. Service product workloads are directly associated with service products that generate service revenue. The second type of workload is associated with infrastructure and are workloads that directly enable a better network. Two examples of infrastructure workloads are cloud radio access network (RAN) and user plane offload using a decomposed mobile core, such as in the 5GC or in the LTE CUPS architecture. Other examples of edge infrastructure workloads exist in cable broadband and Gigabit-capable Passive Optical Networks (GPON) access.

Advantages of edge computing

The edge computing services architecture is intimately associated with an ecosystem and can't exist without a value chain. Ensuring greater openness allows a new ecosystem to emerge and new more efficient business models to develop. The benefits span the entire ecosystem of applications and services providers, network operators, enterprises, and consumer customers. In the ecosystem, you might have a business-to-business model (B2B) where the operator develops service products that are consumable by other businesses. For example, the operator could develop a service product for public cloud providers based on a low latency edge tenancy supported within the operator network. The public cloud provider, which has an established channel into the enterprise, could offer its own X-as-a-service (XaaS) capabilities to those enterprises.

A similar business model can be established with content delivery network (CDN) providers. In this case, as an alternative to the points of delivery (PODs) supported by CDNs, the operator could offer a tenancy within their network so the CDN provider could build their POD at a low-latency location. This approach would offer significant value to the CDN provider and result in a revenue opportunity for the operator. A final example is the connected vehicle business. Automobile vendors seek to establish more value with connectivity to the vehicles they sell. To establish this connectivity, the vendors could establish a branded presence by consuming a tenancy in a mobile edge cloud.

B2B opportunities don't exclude the possibility of operators supporting business-to-consumer (B2C) services. In this case, the operator becomes the branded delivery mechanism of consumer services. Consumer video is the most salient example of this type of product opportunity since it's well known that video streaming benefits from low latency to the subscriber. These examples show the advantages of an open edge computing model.

- A mobile network operator could open their RAN edge to partners to create and rapidly deploy innovative new applications and services to business and consumer mobile subscribers. This approach reduces the load on the operator's core network and could be a much less costly way to host applications and services.
- Applications could extract real-time information about local access conditions. By deploying services and caching content at the network edge, congestion in the core can be mitigated and local demands more efficiently met.
- Application developers and content providers can benefit from closer proximity to subscribers and real-time access network data to provide a superior user experience.

In a B2B model, the consumers of the services of a network operator can include the following:

Content publishers	Connected car vendors	Industrial and automation
Public cloud providers	Gaming, Augmented and virtual reality vendors	Utilities
Mobile virtual network operators (MVNOs) and MVNO enablement platforms	Municipalities	CDN providers
Government organizations including public safety	IoT-connected device platforms	Enterprises

Supporting your partner ecosystem

As a service provider, supporting the value chain and your selected partners requires developing an operational model that makes it easy for partners to develop and deploy their own solutions. Such an operational model also needs to support partners' access to data and analytics that are appropriate for ensuring the health, quality of experience, financial reconciliation, and trust of the services.

To achieve the scale required and manage the growing complexity, the operational model needs to include automation across the ecosystem and be able to meet end customers' changing needs. Open application program interfaces (APIs) must be a necessary part of the overall operational model. Standardizing on an operational environment can help you achieving your objectives. An entire edge computing platform can be designed based on a Network-as-a-Service (NaaS) business model. In this model, APIs are essential along with the necessary software that controls and manages the platform and supports your ecosystem.

Edge computing infrastructure workloads

Three major architectural shifts underpin the emergence of the edge computing network infrastructure:

- Decomposition. Network functions are separated (control/signaling and user/data) for optimization of resources.
- Disaggregation into software and hardware. Software-centric solutions that use off-the-shelf or white-box hardware, which can be procured separately.
- Infrastructure convergence. Fixed and mobile networks share a common 5G core-based infrastructure for efficient operational practices.

The 5G system promotes the emergence of an edge infrastructure that combines decomposed subscriber management from a converged core with the data plane of a wireline access node, for example: DSLAM/OLT, as well as upper layers of the 3GPP radio stack.

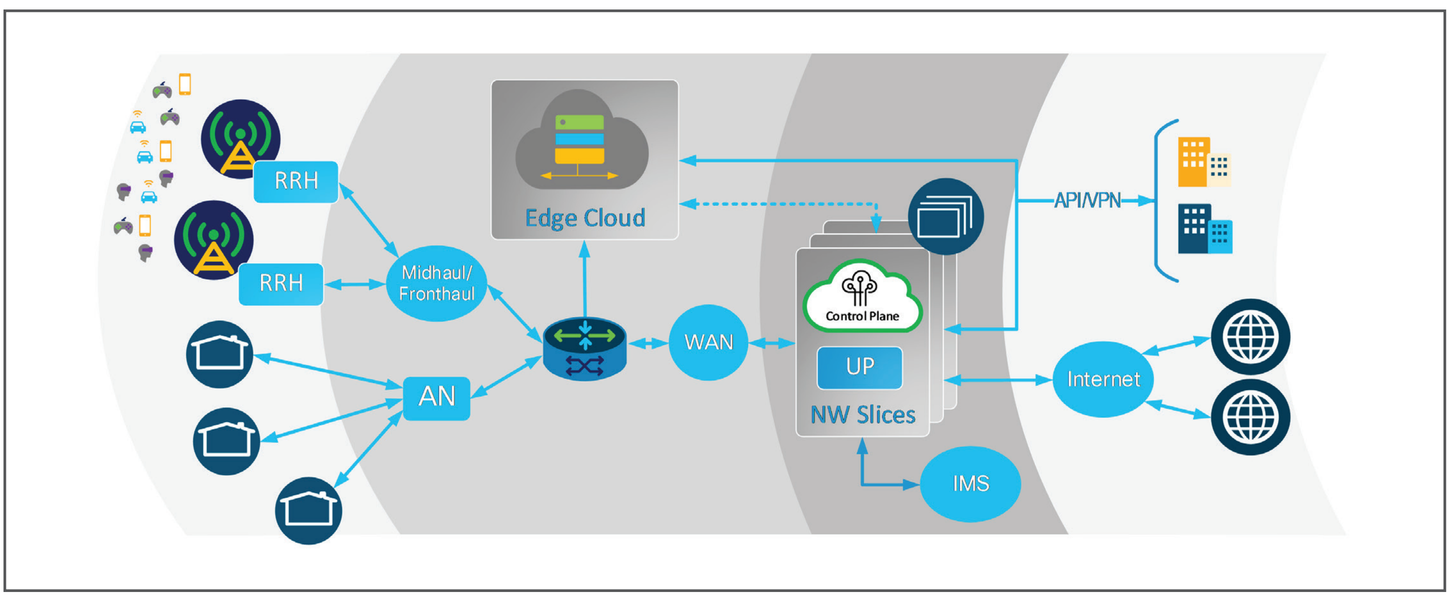

Figure 1. Emergence of the infrastructure edge

Edge computing use cases are driven by the need to optimize infrastructure through offloading, better radio, and more bandwidth to fixed and mobile subscribers.

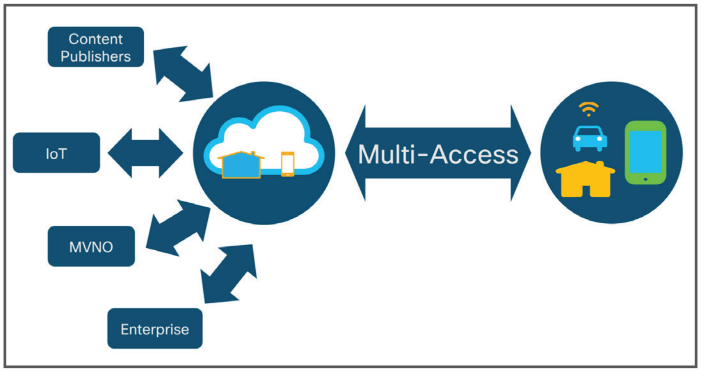

Figure 2. The edge

Some organizations are testing edge computing at the cell-site itself. At first glance, this approach might appear reasonable because it puts the computing as close as possible to the mobile subscribers. However, several issues result:

- It is operationally complex because of the typically large number of cell sites.
- It's expensive due to enclosures, power, and HVAC needs. Specialized servers may be needed versus tapping mass scale production servers.
- New trends in radio are for leaner cell-site architectures comprised primarily of lean elements such as remote radio heads. Note: Cloud radio access networks (C-RANs) don't have packet-awareness at the cell site.

Instead of focusing on proximity, instead you can focus on addressing latency requirements. A good IP design can cure latency issues from a centralized metro location to the cell site. The economics are more important for the location of the edge in edge computing. You need to consider capital expenditures (CapEx) and operating expenses (OpEx) to ensure a good IP network design.

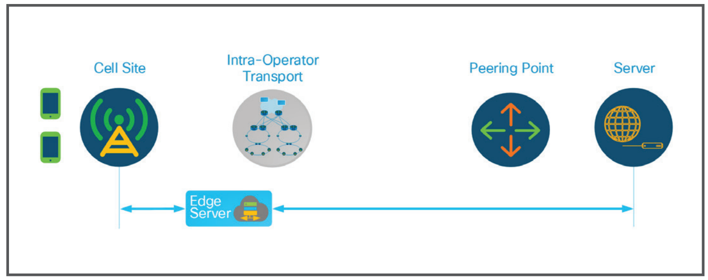

Figure 3. Edge closer to cell-site

An edge server closer to the cell site means less IP network growth but more cost (OpEx and CapEx) for the edge servers because of the larger number of locations that need to be supported.

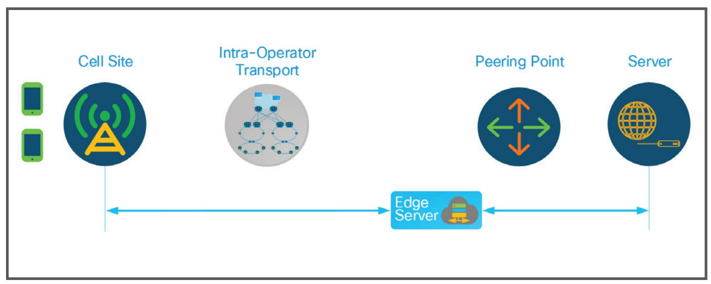

Figure 4. Edge server further from the cell site

An edge server located farther from the cell site means that the operator must deploy more IP network capacity. But it can lower edge server costs due to the economies of scale of the centralized metro location. The higher network capacity is easily manageable through priority queues on latency-sensitive traffic.

To determine the location of the edge computing node, consider these questions:

- Can the more efficient location be the customer premises equipment (CPE)?
- Can the more efficient location be located on the customer premises?
- If the edge location is optimally in the network, is there enough low-latency queuing to the endpoint device?
- For mobile access use cases, can the locations be mapped to a present or future C-RAN central unit (CU) location?

Use cases for edge computing

The use cases for edge computing fall into one of four categories:

- Infrastructure use cases: Resources the operator can use to create edge services.
- Operator branded services: Consumer-oriented services provided with the operator brand name.
- Services to businesses: Services offered to enterprises such as public cloud providers or IOT.
- Private radio for enterprise: Special use cases designed to support enterprises with private radio (LTE, NR, Wi-Fi) for low-latency and security.

To determine the uses cases to support, consider these questions:

- Is it a low latency or a data reduction use case?
- What are the tangible benefits?
- Is the use case fixed or mobile?
- Can the use case be mapped to a consumer or business targeted product?
- Who are the right business partners?
- What is the realization of the product?
- What does the ecosystem look like?

At this point in the emergence of edge computing, wireline and wireless infrastructure use cases dominate, and the challenges are mostly in the operator monetization model for service workloads. These challenges will be overcome in time as more robust ecosystems form, more use cases are discovered (or discarded) through lessons learned, and the economics better understood.

A new approach

The current way networks are built is no longer sustainable. A new approach is necessary, one that is more open and easily places the computing capacity needed for a set of services, to where best located. The end user experience drives the perceived value of your services and is directly related to how the network performs and how the required latency is achieved. Building the foundation for Edge Computing requires you to consider many factors which influence your services architecture.

Low latency benefits many types of services towards a good user experience. Low latency isn’t equivalent to close proximity. A properly designed IP network supports low latency while supporting optimal economics. Interesting edge locations for edge computing include metro data centers and repurposed central offices or public exchanges, not cell sites. In the enterprise environment, an enterprise’s locations might be the optimal locations.

At this stage in the evolution of edge computing, infrastructure use cases dominate the requirements for edge computing nodes and associated activities. Meanwhile, you also need to develop your business model and partner ecosystems. Standardizing on an operational environment will help you achieve your objectives. An entire edge computing platform can be designed based on a Network-as-a-Service business model in which APIs are essential plus the necessary software that controls and manages the platform and supports your ecosystem.

Learn more

For more information, visit the [5G Service Delivery] page.