# Application and Service Evolution

## A Pillar of Edge Transformation

### Reimagining the service edge starting with services and applications

With the emergence and rollout of 5G, service providers and enterprises are preparing for a new generation of dynamic, customized, and profitable business services. However, work remains before networks can actually deliver them. The key challenge is a need for greater intelligence and computational capabilities at the service edge. Although there's a broad agreement across the industry that the edge needs to transform, the details about what this transformation looks like or what the "service edge" actually becomes is more complicated.

In the 5G era, the service edge must enable new services and revenue models. In this paper, we briefly explain what this means.

## Contents



# Vertical market opportunities and network transformation

Most mobility networks have justified their rollouts based on consumer demand, but it's no longer the case for 5G. Service providers delivering consumer experiences are looking to the horizon for what comes next. The real growth potential lies in vertical market opportunities which would require a greater emphasis on business/enterprise service offerings (Figure 1).

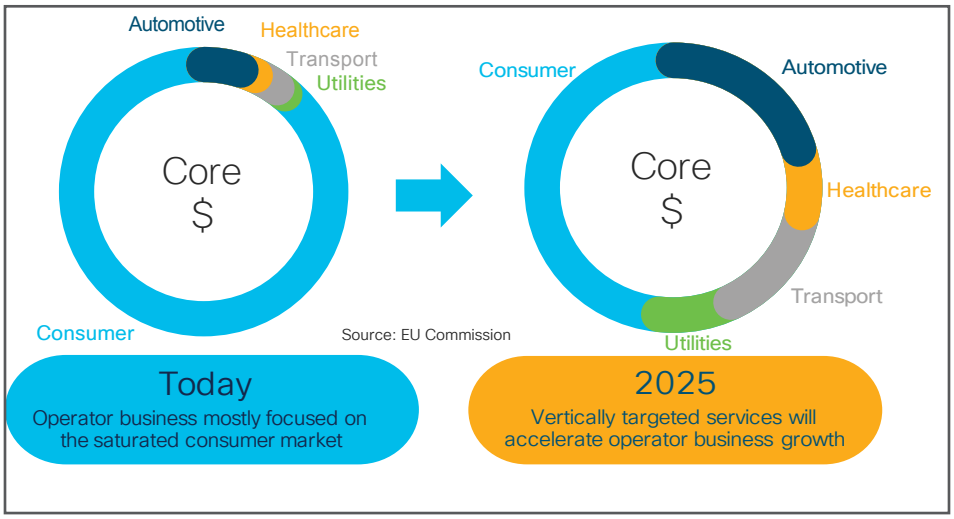



Figure 1. Business services

The advent of 5G brings huge improvements in bandwidth and latency and the ability to tailor service experiences for different vertical markets and individual users. Applications that harness these capabilities promise to transform industries such as manufacturing, automotive, healthcare, and transportation in addition to your balance sheet. But for any of these changes to occur, you must rethink the traditional model for building end-to-end service networks.

As your costs and complexity grow each year, current models for building and deploying basic network functions are becoming less viable. The massive wave of new devices and applications that accompany 5G will only exacerbate these problems. To deliver the high-quality experiences that consumers of 5G services expect and demand, you'll need to fundamentally reimagine the service edge.

You'll need to address:

- **User experience.** Applications such as those for connected vehicles, industrial IoT, gaming, and VR/AR will require very low latencies. Video needs pre-positioned content at the right location that avoids congestion, so a high-quality user experience can be achieved consistently. For some new services, you'll need to position data and application processing closer to subscribers, which could be human beings or machines.
- **Economics**. The growth in new connected devices will exponentially increase the amount of data generated. Backhauling all that data becomes expensive and impractical. The more you can process and offload data at the edge, the more you can reduce core transport costs. Other economic considerations are discussed in another paper, "Establishing the Edge."
- **Decomposition, disaggregation, and convergence.** Network functions such as mobile packet core and radio access network (RAN), and others are being decomposed into multiple entities to optimize resources and allow them to be placed more flexibly in different network locations. Software functions are disaggregating from dedicated hardware, and network functions can run on standard commercial servers. Network infrastructures (fixed and mobile, edge, and core) are converging.

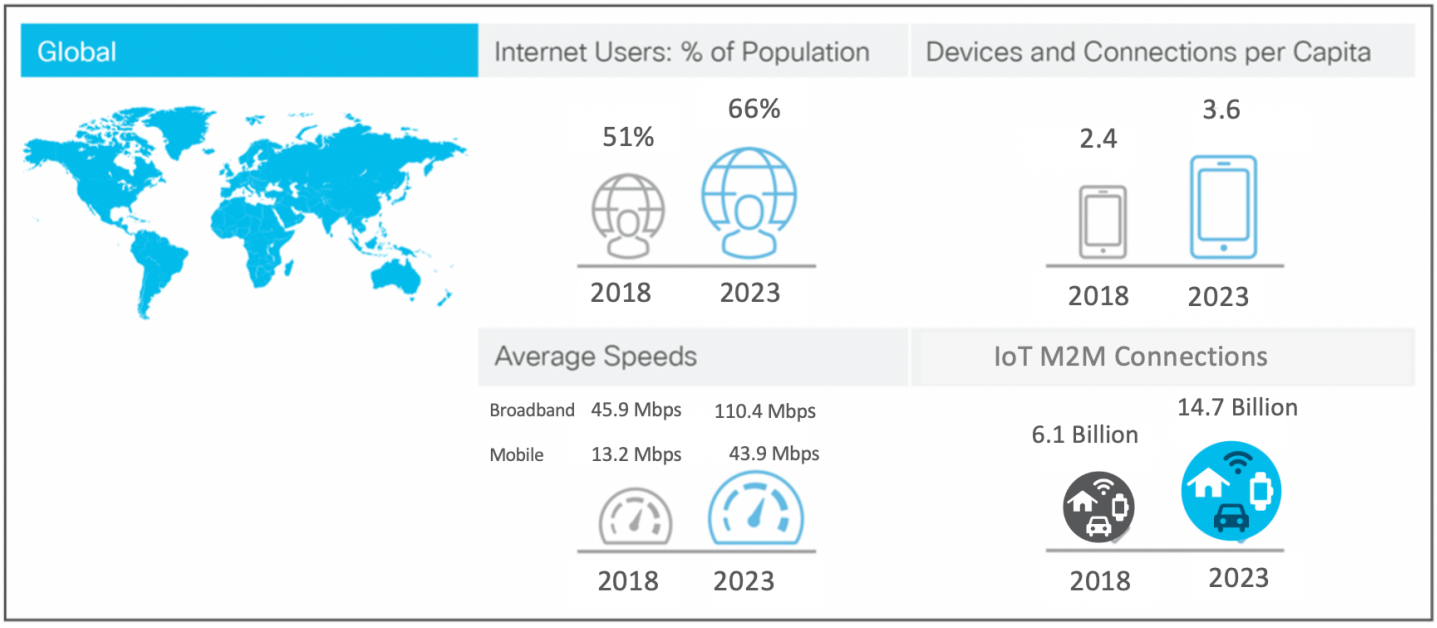
Together, these factors are spurring fundamental transformation of the network architecture and driving the need to position more compute power at the service edge, closer to subscribers. Just how close will vary across markets and the service requirements.

Ultimately, you'll need to find the sweet spot for the service edge(s) in your environment. You need to determine the right combination and placement of edge resources to deliver the best economics.

## Building a more intelligent edge

Of all the technology trends affecting the design of your network, none are more significant than the rise of an evolved edge architecture. Standards organizations such as ETSI refer to this evolution as multi-access edge computing, or MEC. Your MEC strategy will provide much of the core functionality to deliver the next generation of services and user experiences.

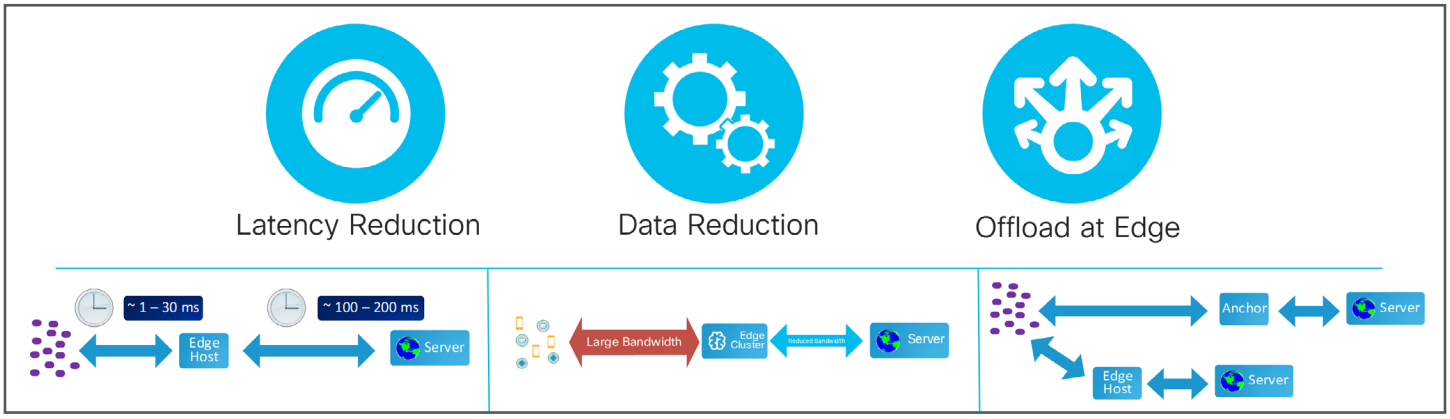
Figure 2. Global internet business traffic forecast (Source: Cisco VNI Forecast 2018-2023)



There are multiple benefits in positioning compute, storage, and networking capabilities at the edge of the network including:

- Latency reduction achieved via moving packet processing closer to the end user
- Data reduction and reduced bandwidth usage via prepositioned content and distributed data processing
- Offloading of traffic (or specific traffic) at the edge of the SP network

Figure 3. Drivers for MEC




## 5G evolution and innovations

5G is the primary driver for many of the current MEC discussions and of network evolution in general. The use cases have profound implications for where you deploy compute power and storage in the network. Innovations are also occurring in several areas, including:

- Cloud RAN (C-RAN) or virtual RAN (vRAN), realized through initiatives such as Open vRAN
- Mobile packet core control and user plane separation (CUPS), which is not strictly a 5G concept but being realized via the 5G evolution
- Network slicing

## Service edge requirements

When you're contemplating technology evolution at the edge, it's important to stay focused on the services. What new services will you be delivering, and what kinds of edge capabilities will you require?

Service provider services can be classified into three categories including:

- **Infrastructure**: These services are mostly related to existing network functions that will be decomposed, disaggregated and virtualized. Examples include C-RAN/vRAN, CUPS-based BNG/EPC, and cloud-based cable modem termination system (CMTS).
- **Operator branded services**: These services relate to offerings the service provider gives users to differentiate their brand. Examples include content streaming using a content delivery network (CDN), live TV, and IoT services.
- **Business services**: These services relate to addressing specific vertical markets and associated with the business-to-business and business-to-consumer markets. Examples include online gaming, augmented reality and virtual reality services (AR/VR), and third-party application hosting.

Most activity today is related to evolving the infrastructure services that could provide the baseline platform upon which operator branded services and business services will be deployed in the future. Addressing evolving infrastructure use-cases such as C-RAN/vRAN with latency requirements around 100

microseconds is mandating a more distributed service architecture (100 microseconds is the target budget allowed for the transport between the decomposed RU (Remote Unit) and DU (Distributed Unit) components). In the near term, even services like mobile video and AR/VR and gaming will continue to push the limits of traditional edge designs with latency requirements in the order of 10s of milliseconds (ms).

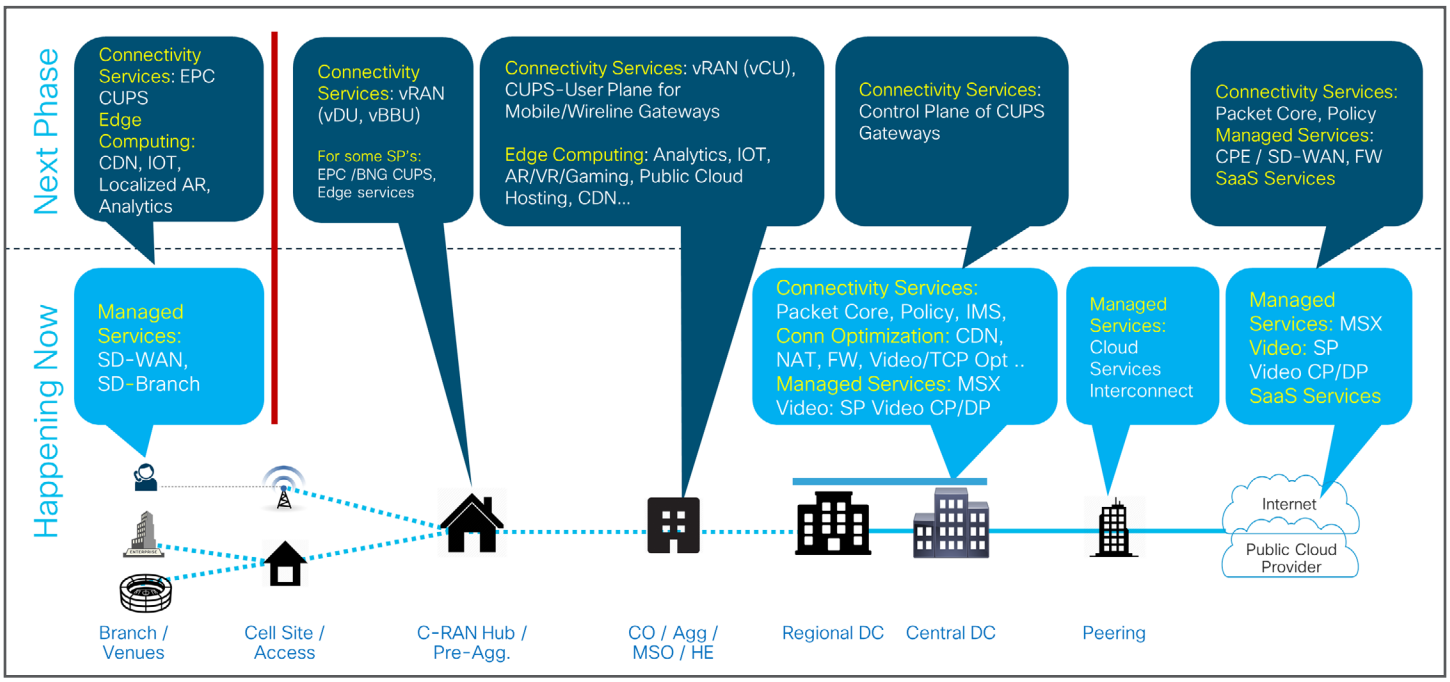
| Use Case | Minimum One-Way Delay |
|---|---|
| Mobile video | ~75 ms |
| Mobile AR | 10 ms |
| Mobile VR | 20 ms |
| Interactive gaming | 50 ms |
| Voice-over-IP | 200 ms |

Looking ahead, a variety of emerging low-latency and uRLLC use cases will offer significant growth and profit potential if your evolved services edge can meet increasingly demanding requirements.

| Future Use Case | Required Latency |
|---|---|
| Factory automation (real-time control of production line machines and systems) | .25–10 ms |
| Intelligent transportation (autonomous driving) | 0–100 ms |
| Robotics and telepresence (remote control with synchronous visual/haptic feedback) | 10–100 ms |
| Healthcare (biotelemetry, telediagnosis, telesurgery) | 1–10 ms |
| Smart grid | 100 ms |

The choices you make in the use cases you pursue, and the latency and other requirements of those applications will largely dictate the type of new edge capabilities you build and where you distribute those capabilities in the network. Some service elements like CUPS control plane, policy, and SD-WAN managed services will continue to run out of centralized and regional data centers. Others such as vRAN, edge computing for analytics, and IoT will increasingly be located farther out in the network, such as in central offices/exchanges, preaggregation sites, and beyond.

Figure 4. The span of service edge and MEC locations

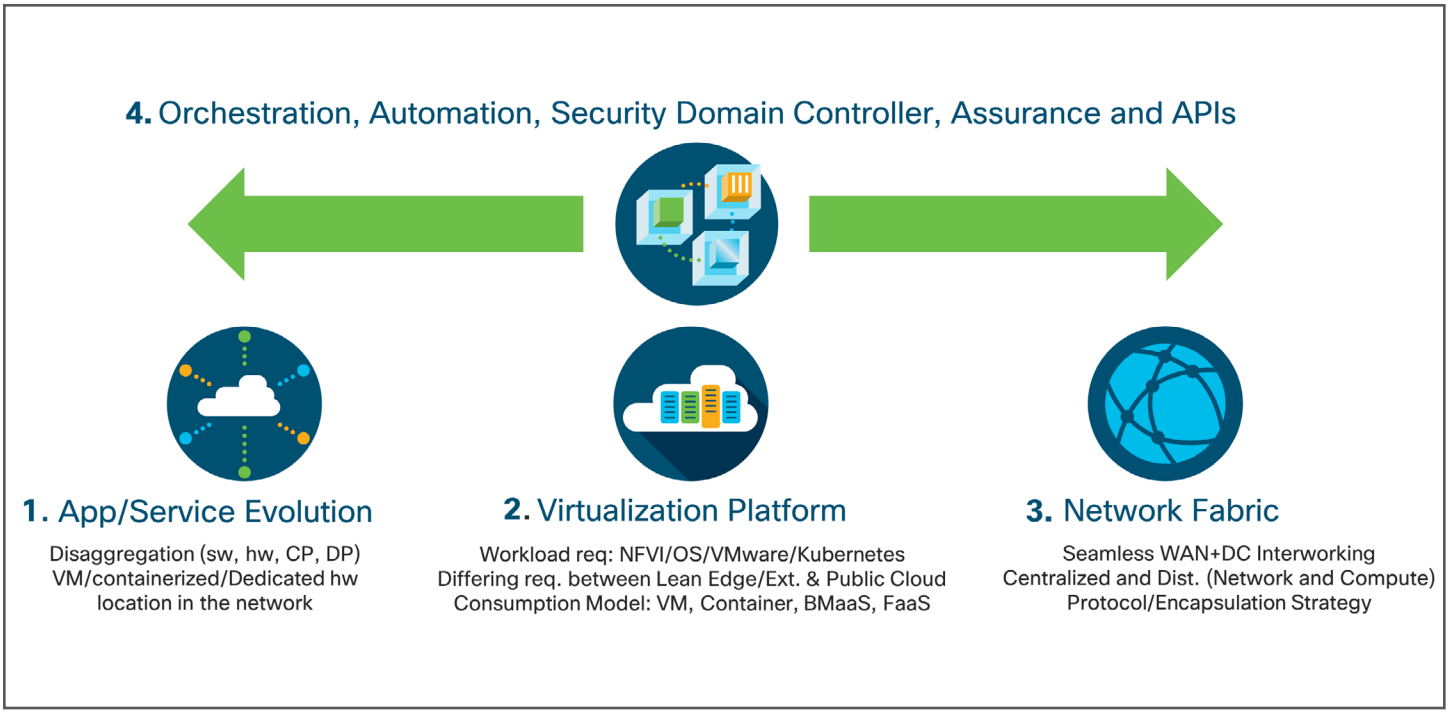


# Application and service evolution and the next-generation service edge

Every decision about edge transformation should start with the service and what it requires. What will the network functions delivering this service look like? Will they be disaggregated, and how? Will the network components delivering the service be cloud-native? Hybrid? Most important, where will they be placed in the network?

In reimagining the edge, consider the applications you'll be running. Although use cases such as autonomous vehicles, ubiquitous AR/VR services, and online gaming are interesting to discuss, the industry will initially focus on foundational infrastructure use cases. Applications such as C-RAN/vRAN, decomposed mobile packet core, CUPS-based broadband network gateway (BNG), virtualized CMTS, Remote PHY, and Gi-LAN will create the baseline on which future applications will be built.

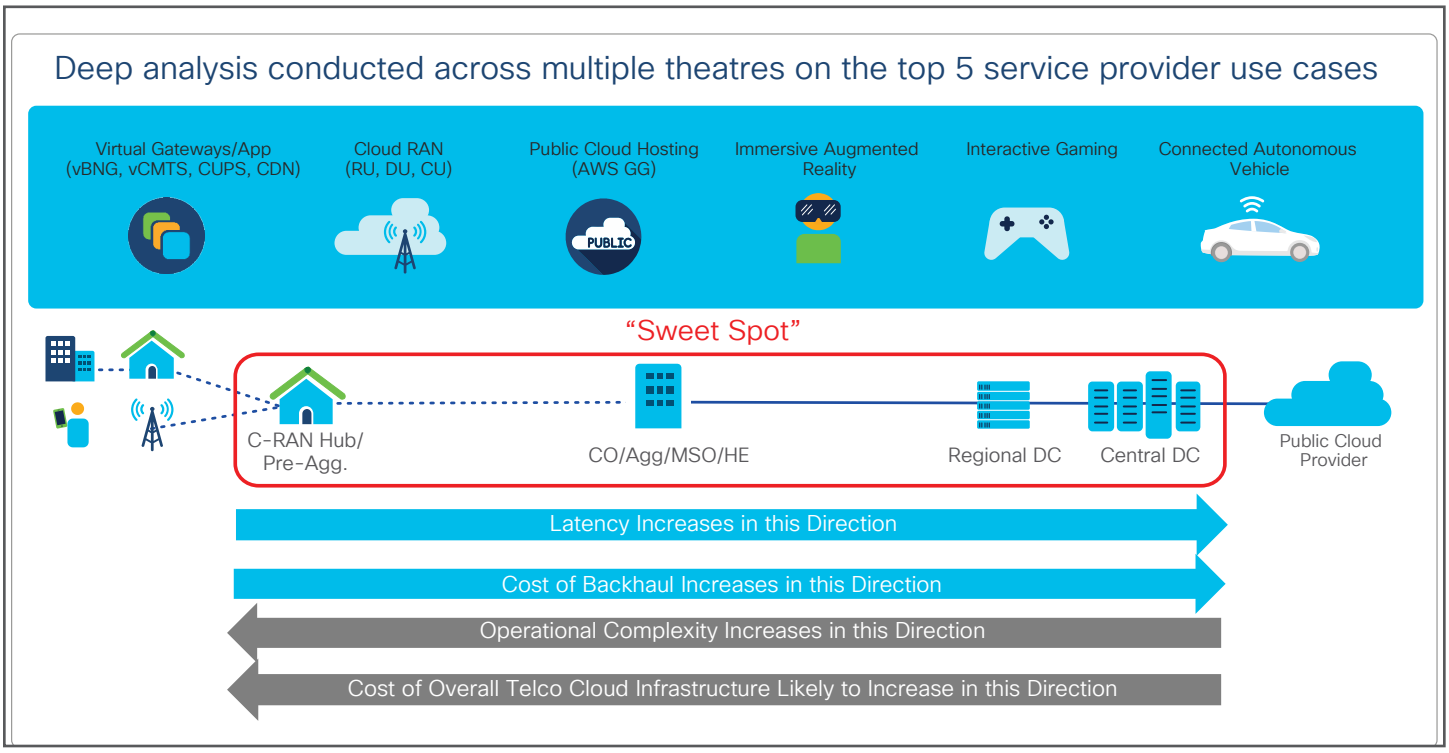
Figure 5. Edge architecture



## Application requirements

In this section, we discuss a sample of five prevalent applications and examine their requirements. The following diagram depicts the major network locations and potential “sweet spot” opportunities for the component functions of the five use cases. This analysis is based on a survey of operators in multiple theaters around the world. Deep analysis was conducted across multiple theaters on the top five service provider use cases.

Figure 6. Major network locations and opportunities

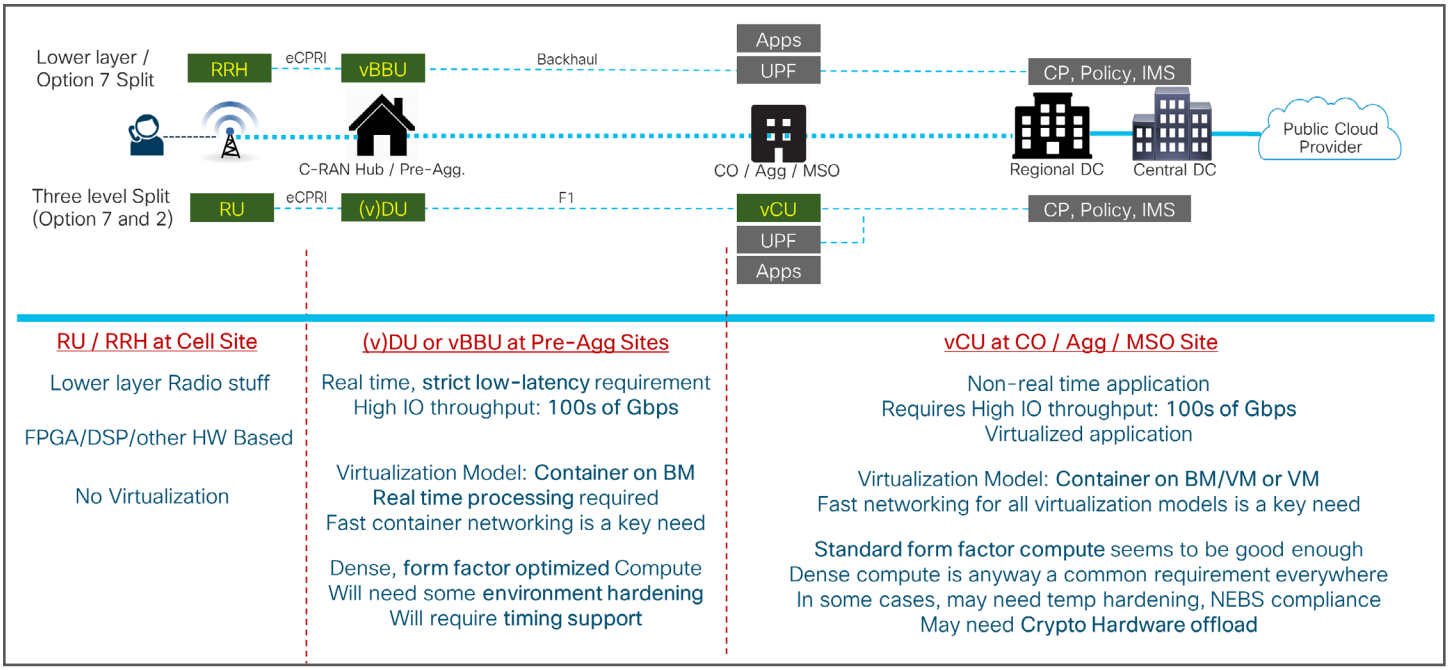


## Cloud RAN/vRAN

The first use case is C-RAN/vRAN (Figure 7). Clearly, the functions such as remote radio head or remote radio unit are associated with the antenna system at the cell site. There is no virtualization need and it is very much hardware based.

For the virtual distributed unit (vDU) or the virtualized baseband unit (vBBU), latency becomes important. In most designs, the transport alone contributes ~100 microseconds. Where you locate these functions then needs to be no more than about 12-14 kilometers from the cell site at preaggregation sites. Throughput is high, in the order of 100s of Gbps. A lot of the initial deployments use containers for the virtualization model; real-time processing and fast networking are critical. vDU is intensive, which leads to a dense form factor. In many cases, environmental issues need to be addressed

Figure 7. Cloud RAN/vRAN



including hardening which has economic implications. Timing still needs to be done here.

The virtual central unit (vCU) has quite high throughput (100s of Gbps) but not the strict latency needs of vDU. Optimal location for the vCU can be at the central office (CO) or equivalent aggregation site. Typical deployments use containers or virtual machines for the virtualization model and standard but dense compute achieves better economics. In some scenarios, some level of NEBS compliance and temperature hardening may be required.

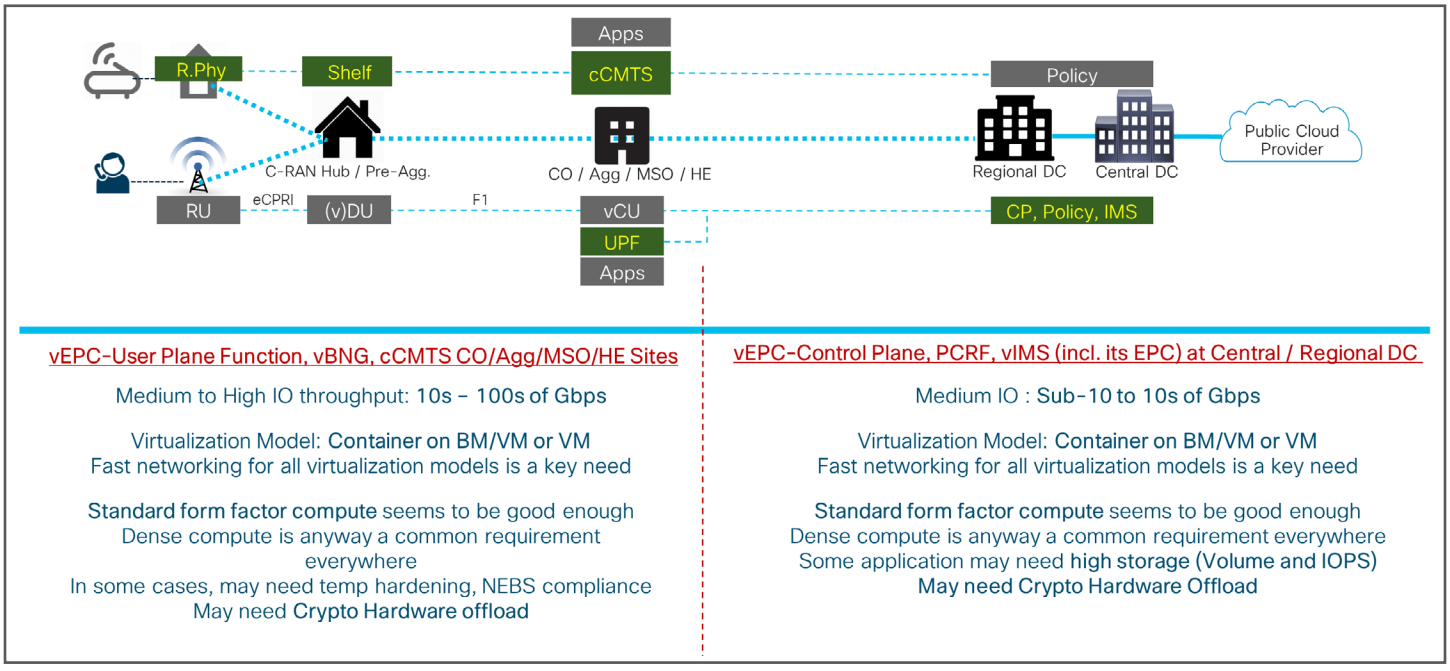
## Virtualized service gateways

The next use case is virtualized service gateway (see Figure 8). In this case, throughput is medium to high, 10s to 100s Gbps. Deployments are starting to be based on control and user plane separation. Both containers and virtual machines are being used for the virtualization model. Fast networking is a necessary need no matter the virtualization used. The one function that perhaps is slightly different from the others is BNG which, for some operators, continues to be based on network processing units (NPU) rather than be virtualized (vBNG).

Optimal locations for these functions include the aggregation site (CO for example) for vEPC user plane, vBNG, and cloud CMTS. The vEPC control plane, PCRF, and vIMS are best placed at the regional or central data center or even as far as a public cloud.

Standard but dense compute form factor is sufficient although in some case NEBS compliance and temperature hardening may be required. It is also important to note that crypto hardware offload might be required.

Figure 8. Virtualized service gateway



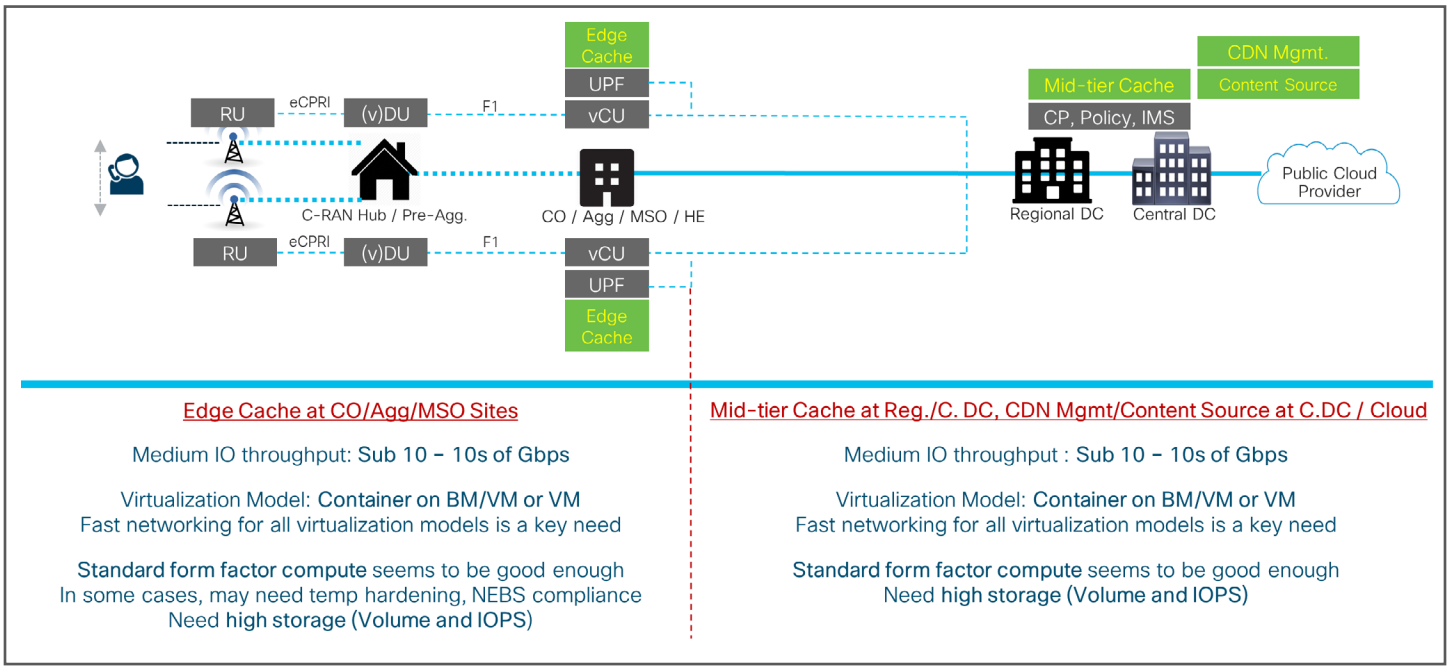
## Managed mobile video with CUPS and edge CDN

The third use case we examine is managed mobile video with CUPS and edge CDN as shown in Figure 9. The important element of this use case is caching and where optimally to locate it. Throughput is medium, sub 10s to 10s of GBPS. Virtualization models being deployed include both containers and virtual machines. In all cases, fast networking is an essential need.

The caching function needs to follow the user plane function (UPF) because the UPF needs to be invoked before the inline services. The big impact of caching is of course storage capabilities. In a centralized environment, it can be easier and more economical to locate storage. The location for caching is mostly dependent on where you have the tightest area for bandwidth and congestion.

Pushing caching further out toward the subscriber at aggregation sites can potentially avoid or reduce congestion.

Figure 9. Managed mobile video with control user plane separation



## IoT and public cloud hosting

An interesting use case given its topical nature is IoT and, relevant to this discussion, the IoT gateway. An example of another solution component is AWS Greengrass which we have used along with our IoT solutions (Figure 10).

Throughput requirement is relatively low at sub-10 Gbps levels. Many IoT gateways have been deployed on enterprise premises. However, what we've observed with our IoT solutions is that there's no performance degradation in having the IoT gateway at an aggregation site (ex: CO) versus on customer premises.

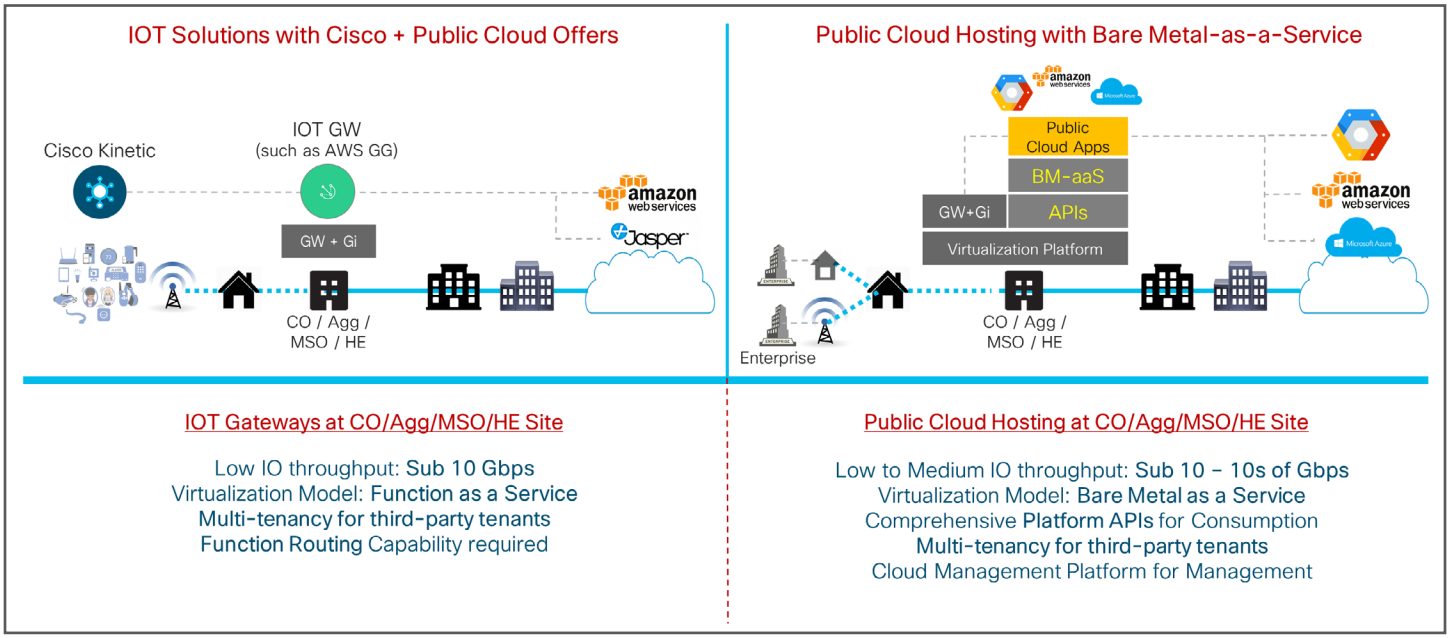
Offering this type of service is somewhat different from previous use cases in that it would be offered in a "function-as-a-service" model. Multitenancy is an important feature to support for this use case. To get the right traffic to the gateway, you would need a function routing capability.

Another scenario example using AWS, is public cloud hosting with bare-metal-as-a-service (BMaaS). The issue to consider is whether there's a benefit to putting it into a service provider network?

What we observed in locating the service capability at the aggregation point was that throughput is low to medium, sub10s to 10s of Gbps. However, the performance of the applications greatly improved because they were closer to the user.

Requirements include a virtualization model supporting BMaaS, a comprehensive API platform for user consumption, multitenancy, and a cloud management platform that supports your own operator needs as well as those of the enterprise clients. It is feasible for network operators to offer these services. Business-to-business partnerships and a viable ecosystem are essential.

Figure 10. IoT and public cloud hosting



## AR and interactive gaming services

The final use case is AR and interactive gaming services (Figure 11). Once again, the virtualization model is primarily containers with virtual machines also being deployed in some cases for interactive gaming services. Multitenancy is important to support such services economically. AR services have low to medium (sub 10 to 10s Gbps) throughput requirements while interactive gaming needs medium throughput (10s of Gbps). A lot of the machine learning back end needs specific hardware.

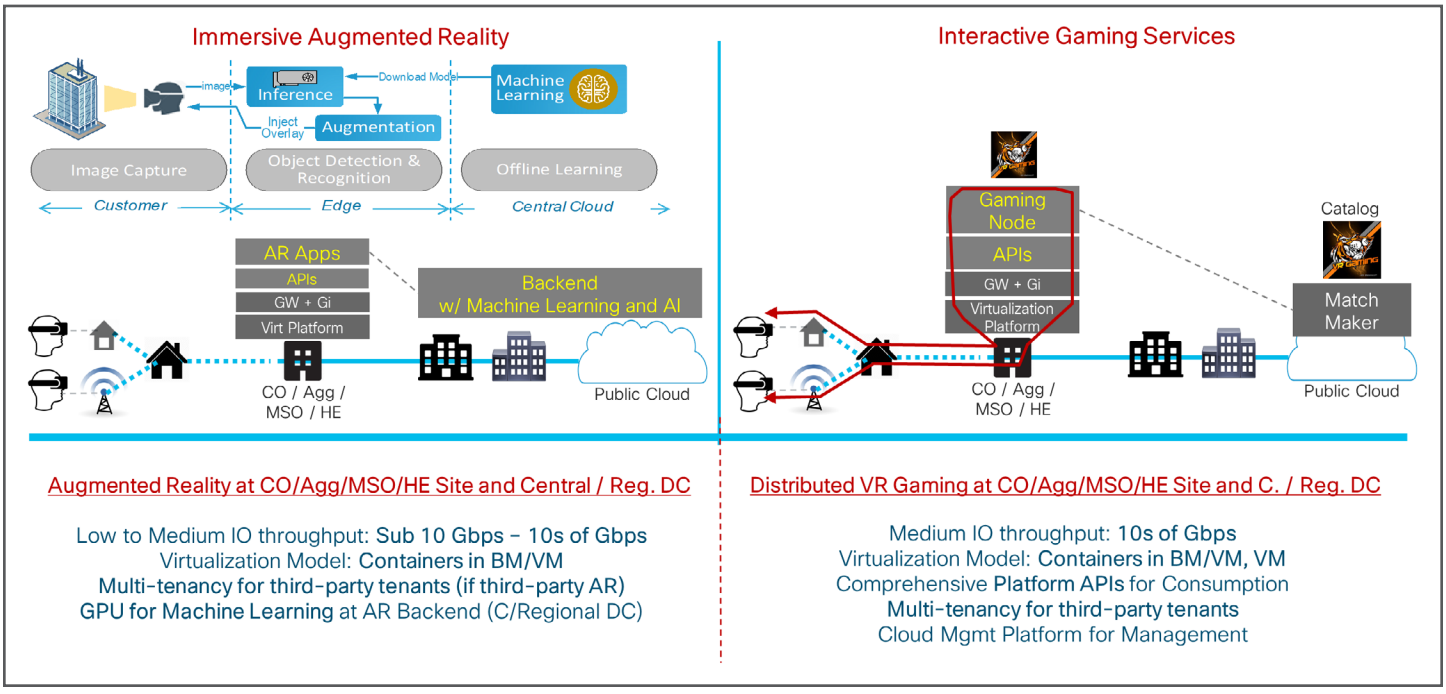
We've observed that the performance within a service provider network at the aggregation sites (ex: CO) meets the performance expectations of interactive gaming companies. This finding dispels a "belief" propagated earlier in the industry that the service is too far into the network to provide the required latency and performance.

Another key requirement is developing easily consumable APIs, business to business partnerships, and a robust ecosystem.

# Application evolution

What is important in supporting these new service models is the decomposed CUPS. To support the new infrastructure services and the new business services envisioned and more, CUPS has become very important. We need to be able to have a centralized control plane and a highly distributed data plane at least to those sites applicable to meeting the application and service needs economically.

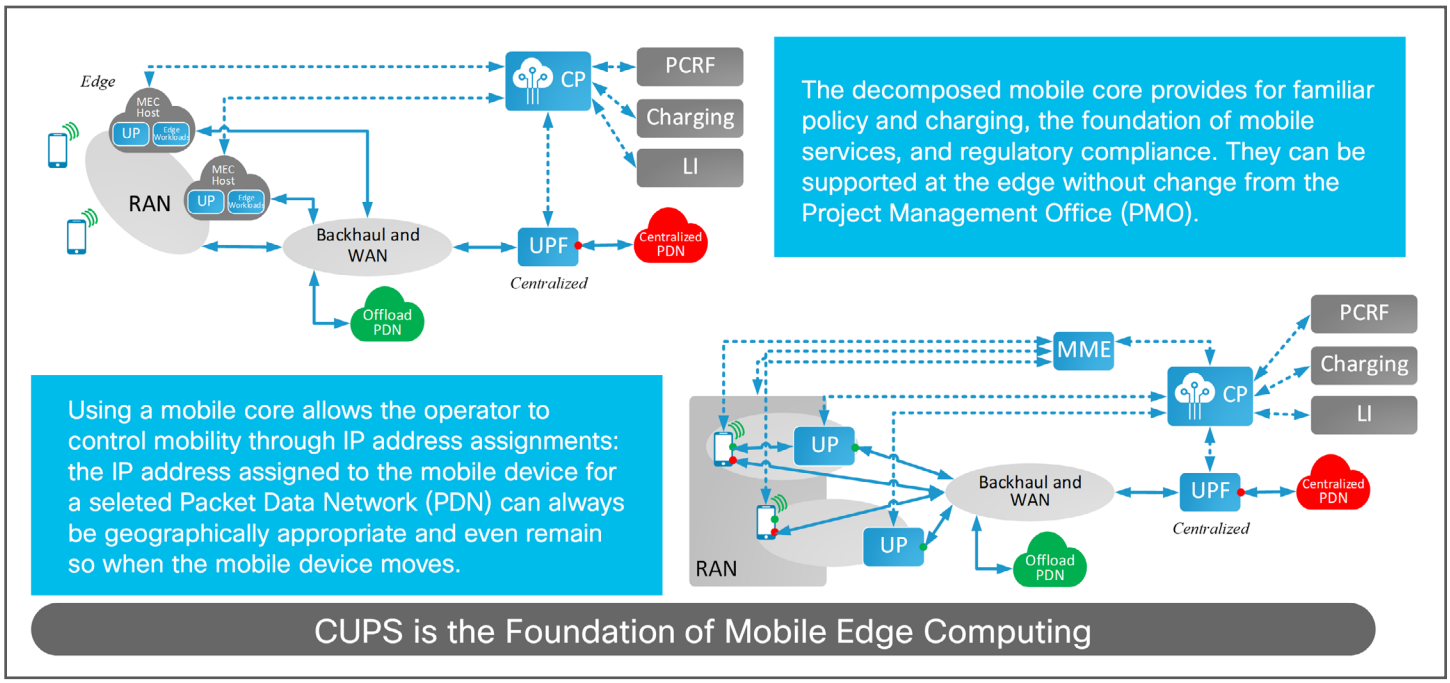
Figure 11. AR and interactive gaming services



## Decomposed CUPS for the edge

Recently, the industry has made significant strides in user plane functions, terminating sessions in distributed architectures, and generally making CUPS a concrete reality. The positioning of the UPF will depend on the service requirements but could be positioned out as far as the customer premise. The control plane function (CPF) can be located in the centralized sites regardless of the location of the UPF. CUPS as depicted in Figure 12 is a stepping stone toward rearchitecting your software and moving to a containerized model.

Figure 12. User plane separation



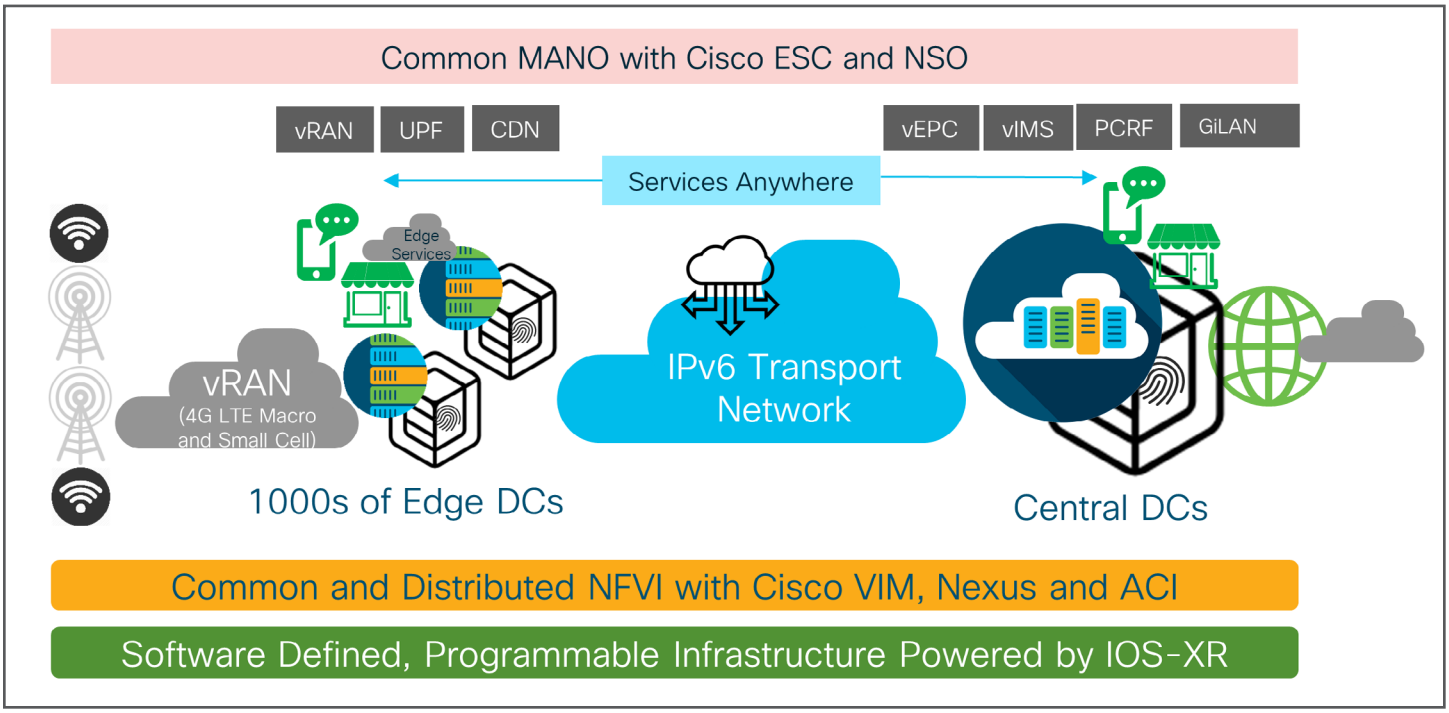
## vRAN ecosystem overview

C-RAN is happening now. Figure 13 depicts a fully end-to-end C-RAN enabled network that has been implemented in 2019. The scale of the distribution of virtual functions such as vDU and UPF is in the thousands of edge sites. As discussed previously these sites are in the preaggregation location inside the network from the cell sites. The average usage for C-RAN is as low as 1–2 servers.

Important elements include a common distributed NFVI, a programmable IP infrastructure, and end to end common management and orchestration with easily consumable APIs. Such a highly virtualized, distributed, and software defined platform is well prepared for all types of new applications and services.

Cisco and partners worked with Rakuten of Japan to deliver the industry's first C-RAN deployment with a fully virtualized environment from RAN to core, which included edge computing and software-defined operation. The project also included deploying vDU capabilities in ~3,000 preaggregation sites. Deployments like this one show the advantages of the model, how services can scale, and how to orchestrate services end to end. These lessons will have a significant influence on future service architectures. In addition, you are effectively creating the foundation for a distributed edge platform on which you can build and run future services and applications.

Figure 13. A fully end-to-end C-RAN enabled network



## CMTS evolution – cloud-native

Cloud native is becoming more commonplace. Functions, especially network functions, cannot be simply ported to a virtualized infrastructure and expected to deliver optimal performance; they need to be reengineered as cloud-native. CMTS is no exception and we have implemented our CMTS solutions as cloud native. Figure 14 depicts our cloud CMTS solution as a reference architecture. You will note that it also includes open source components.

In such a practical and deployable architecture, it is important not to have bottle-necks in the underlay, such as networking and data plane.

This example represents a fundamental rethinking about how a real-time carrier-class product is built:

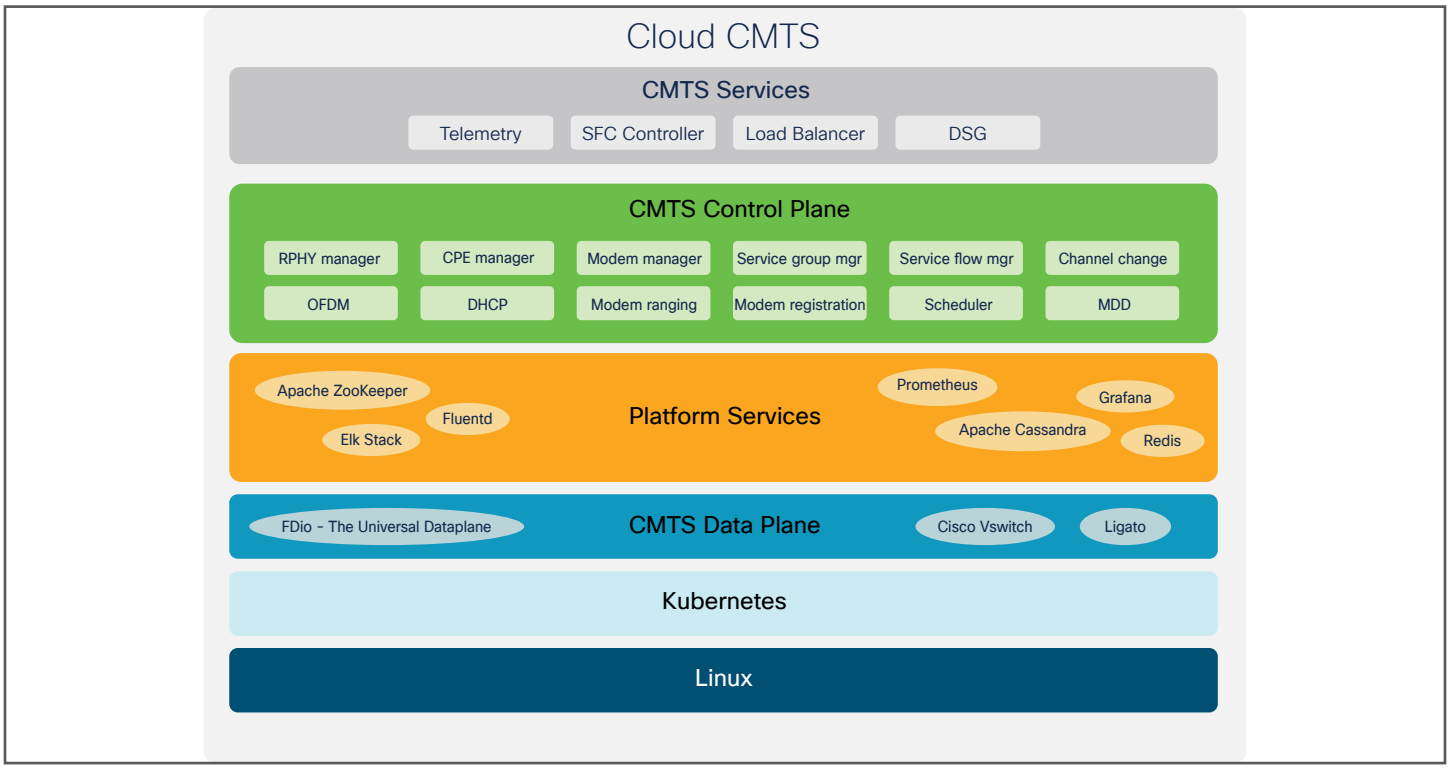
- Cloud-native microservice-based
- Best of web development innovations applied to real-time embedded DOCSIS

It incorporates our deep expertise across important technology areas including:

- DOCSIS leadership
- Expertise in full cloud-native stack, for complete solution

This architecture enables cable operators to transform by embracing cloud operations. It also allows rapid service development with highly streamlined DevOps.

Figure 14. Cisco cloud CMTS solution



## Learn more

Take a closer look at our Edge Computing solutions to determine if they're the right fit for your organization.

## To transform your services edge, start with the services and applications

Service providers and enterprises are preparing for a new generation of dynamic, customized, and profitable business services. The emergence and deployment of 5G is a prominent driver, but a key challenge is the need for greater intelligence and computational capabilities at the service edge.

In the 5G era, the service edge must enable new services and revenue models. Whatever the state of the industry is, in providing answers to all the questions that arise, an important pillar for edge transformation in the evolution of applications and services and represents the starting point of your journey. It's directly related to your ability to provide a robust and compelling platform on which you can deploy and scale profitable new services.

In many of the use cases that we examined, it's apparent that you need to build business partnerships and develop a broad ecosystem around each of the use cases and the services within.