White paper
Cisco public

The Four Pillars of Service Edge Transformation

Make smart choices to evolve your service networks

Reimagining the service edge

With the emergence of 5G, analysts and enterprises are preparing for a new generation of dynamic, customized, and profitable business-to-business (B2B) services. However, work remains to be done before service provider networks can actually deliver them. The key challenge is a need for greater intelligence and computational capabilities at the service edge. If you poll the industry, you'll find broad agreement among service providers, enterprises, and vendors that the edge needs to transform. But ask for details about what that transformation looks like or what the "service edge" actually is and things become more complicated.

In the 5G era, the service edge must evolve to enable new services and revenue models. Although work in this space is ongoing and the industry has yet to reach definitive answers on several important questions, four fundamental pillars of service edge transformation need to occur.

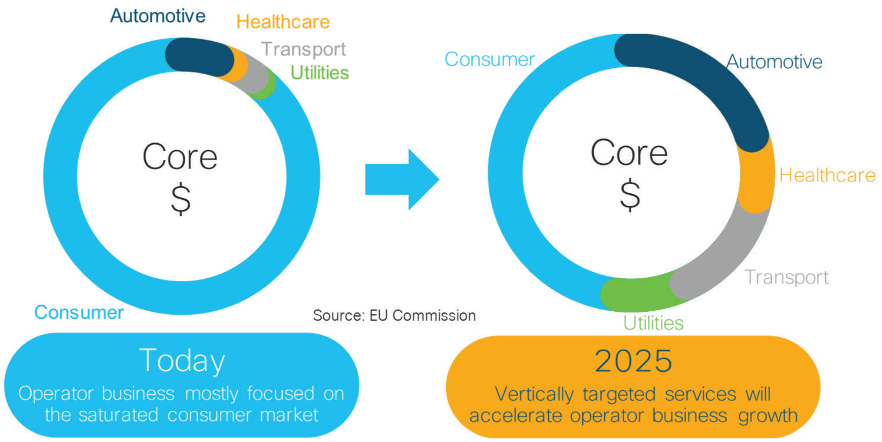
Contents

Industry trends drive 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. Most likely, your margins and average revenue per user (ARPU) for consumer services have been flat or declining for several years. The real growth potential lies in 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. End-to-end orchestration that is possible with 5G will be most important in delivering high quality experiences economically.



Source: EU Commission

You will need to address:

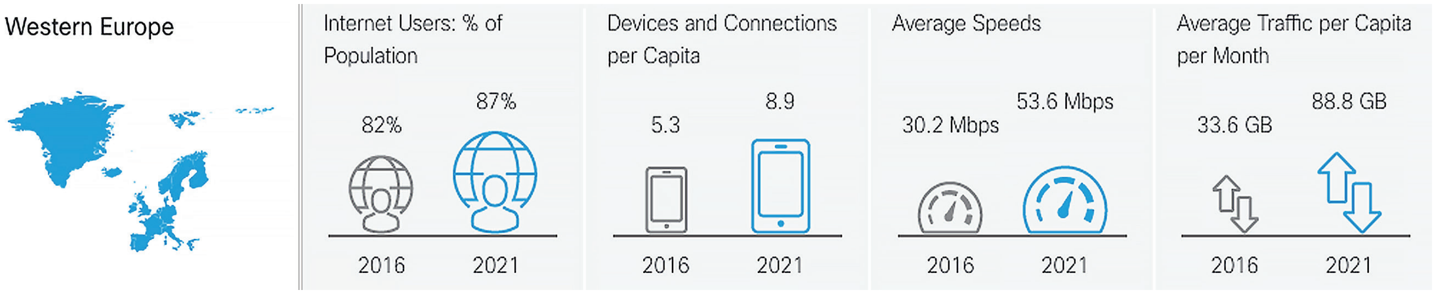
- **User experience.** Many 5G applications such as those for connected vehicles, industrial IoT, gaming, and VR/AR will require very low latencies. Video, on the other hand, is less about latency and more about bandwidth with pre-positioned content at the right location that avoids congestion, so a high-quality user experience can be achieved consistently. For some new services, you'll need to position data and application processing closer to subscribers, whether the subscriber is a human being or a machine.
- **Economics.** An explosion of new connected devices because of the growth of the Internet of Things (IoT) will exponentially increase the amount of data generated. Backhauling all that data to a centralized national or regional data center becomes expensive and impractical. The more you can process and offload data at the edge, the more you can reduce core transport costs.
- **Decomposition, disaggregation, and convergence.** Network functions such as packet core, remote radio unit, and others are being decomposed into multiple entities to optimize resources and allow them to be placed more flexibly in different network locations. With virtualization, 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 as well.

Together, these factors are spurring fundamental transformation of the network architecture and driving the need to position more compute power at the service edge, so it's closer to subscribers. Just how close will vary across markets and the service requirements. If you keep edge resources too centralized, you can't get the performance and intelligence you need to deliver next-generation user experiences. But if you distribute edge intelligence too far out, you introduce other problems with increased operational complexity and problems associated with remote network sites such as lack of power and space and a need for environmentally hardened solutions.

Scale is another issue. When implementing cloud services and network functions in central offices (COs), you deploy and manage these capabilities at a few hundred sites. When you push them out to C-RAN hubs/pre-aggregation nodes, you're now managing thousands. If you go even farther out to cell sites, it's now tens of thousands. Truck rolls to cell sites already make up the vast majority of operating expenses. Positioning even more sophisticated intelligence at these sites isn't economically viable. Ultimately, you'll need to find the sweet spot for the service edge 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, which standards organizations like ETSI refer to 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.



Source: Cisco Visual Networking Index (VNI) Forecast.

By positioning new compute, storage, and networking capabilities at the edge, you can:

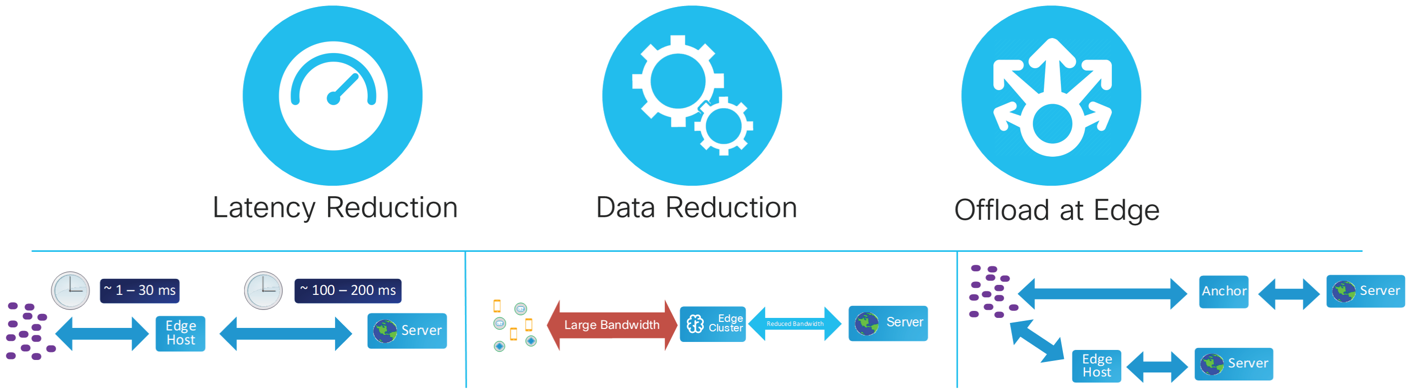
- **Reduce latency.** MEC allows for latencies of 1-30 milliseconds between edge services and consumers. When traffic is routed to a centralized data center, it takes 100-200 milliseconds, so the lower latency improves the quality of experience (QoE) and enables new business services.
- **Reduce bandwidth.** With MEC, edge nodes can perform data analytics such as machine learning inference to compensate for less-capable user devices by reducing bandwidth and/or offloading compute.
- **Offload data at the edge.** MEC lets you route data from edge hosts along less-expensive and lower-latency paths towards services such as cloud-hosted business applications.

5G evolution and innovations

The emergence of 5G is the primary driver for current MEC discussions and of network evolution in general. At the highest level, 5G networks must support three overarching business services use cases Enhanced Mobile Broadband (eMBB) includes fixed access use cases. Massive Machine-Type Communications (mMTC) will be used for IoT deployments. And Ultra-Reliable Low Latency Communications (URLLC), will support delay-sensitive applications like self-driving vehicles and automated public safety solutions.

These use cases have profound implications for where you deploy compute power and storage in the network. Innovations also are occurring in several areas, including:

- **Evolution of xHaul/CRAN:** A new RAN model is emerging that supports different functional splits. This model has resulted in the emergence of entities such as radio units (RUs), centralized units (CUs), and distributed units (DUs). These components no longer need to be collocated; they can be flexibly placed in the most optimal location in the network. This evolution ultimately includes the ability to also virtualize these entities as part of a cloud RAN.
- **Open vRAN:** Vendors and industry groups are advancing open virtualized RAN (vRAN) solutions, which allow different RAN components to work together through the use of open interfaces that can be extended into a broader software-defined network architecture.
- **Control and user plane separation (CUPS):** The mobile packet core is being decomposed. It combines centralized control plane functions with multiple user planes that can be deployed anywhere in the network and augmented with inline services.
- **Network slicing:** Central to the revenue potential of 5G is the ability to run multiple logical networks as virtually independent business operations with their own SLAs over the same infrastructure.



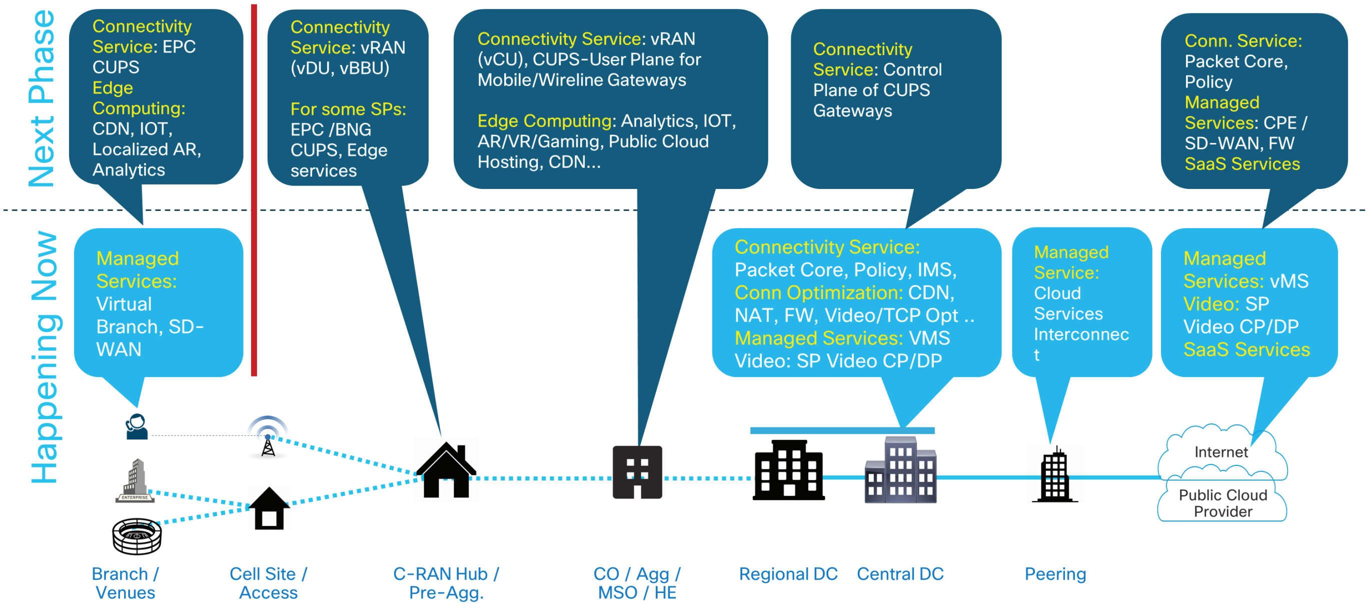
Service edge requirements

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

Service provider services can be classified into three categories including:

- **Infrastructure:** These services are mostly related to existing network functions that will be virtualized and decomposed. Examples include vRAN/Cloud RAN, CUPS-based BNG/EPC, and cloud-based CMTS.
- **Operator branded services:** These services relate to offerings the service provider offers to 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 are related to 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 will provide the baseline platform upon which the operator branded services and business services will be deployed in the near future. The service deployment designs employed in service provider networks today tend to be centered around regional and central locations. Addressing evolving infrastructure use-cases such as vRAN with latency requirements around 100 microseconds is mandating a more distributed service architecture. In the near term, even services like mobile video and AR/VR, gaming will continue to push the limits of legacy edge designs with


Next Phase
Happening Now
Connectivity Service: EPC CUPS
Edge Computing: CDN, IOT, Localized AR, Analytics
Connectivity Service: vRAN (vDU, vBBU)
For some SPs: EPC /BNG CUPS, Edge services
Connectivity Service: vRAN (vCU), CUPS-User Plane for Mobile/Wireline Gateways
Edge Computing: Analytics, IOT, AR/VR/Gaming, Public Cloud Hosting, CDN…
Connectivity Service: Control Plane of CUPS Gateways
Conn. Service: Packet Core, Policy
Managed Services: CPE / SD-WAN, FW
SaaS Services
Managed Services: Virtual Branch, SD-WAN
Connectivity Service: Packet Core, Policy, IMS,
Conn Optimization: CDN, NAT, FW, Video/TCP Opt ..
Managed Services: VMS
Video: SP Video CP/DP
Managed Service: Cloud Services Interconnec t
Managed Services: vMS
Video: SP Video CP/DP
SaaS Services
ENTERPRISE
Internet
Public Cloud Provider
Branch / Venues
Cell Site / Access
C-RAN Hub / Pre-Agg.
CO / Agg / MSO / HE
Regional DC
Central DC
Peering

latency requirements in the order of 10s of milliseconds (ms). The current guideline used for designing such network is to achieve a user to application latency of between 10-10ms.

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 (bio-telemetry, tele-diagnosis, tele-surgery)	1-10 ms
Smart grid	100 ms

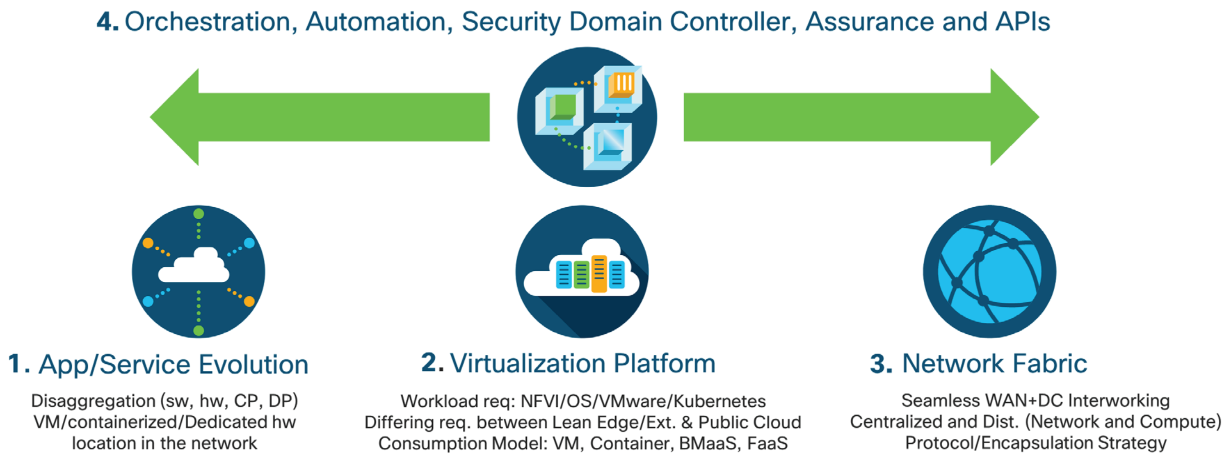
The choices you make in the use cases you pursue, and the latency and other requirements of those applications will largely dictate the type of new edge capabilities you build and your approach to distributing those capabilities in the network. Some service elements like CUPS control plane, policy, and 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 live farther out in the network in COs, pre-aggregation sites, and beyond.

Envisioning the next-generation service edge platform

To meet these requirements, a new edge platform is needed. This platform is not a specific commercial product because the service edge will likely be defined differently by different operators and across different services. However, a platform needs to incorporate four basic pillars of service edge transformation: application and service evolution, the virtualization platform, the network fabric, and orchestration and automation. To enable the next generation of business services with their associated user experiences a reimagined edge must take each of these four pillars into account.

Application evolution

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



4. Orchestration, Automation, Security Domain Controller, Assurance and APIs

1. App/Service Evolution

Disaggregation (sw, hw, CP, DP)
VM/containerized/Dedicated hw
location in the network

2. Virtualization Platform

Workload req: NFVI/OS/VMware/Kubernetes
Differing req. between Lean Edge/Ext. & Public Cloud
Consumption Model: VM, Container, BMaaS, FaaS

3. Network Fabric

Seamless WAN+DC Interworking
Centralized and Dist. (Network and Compute)
Protocol/Encapsulation Strategy

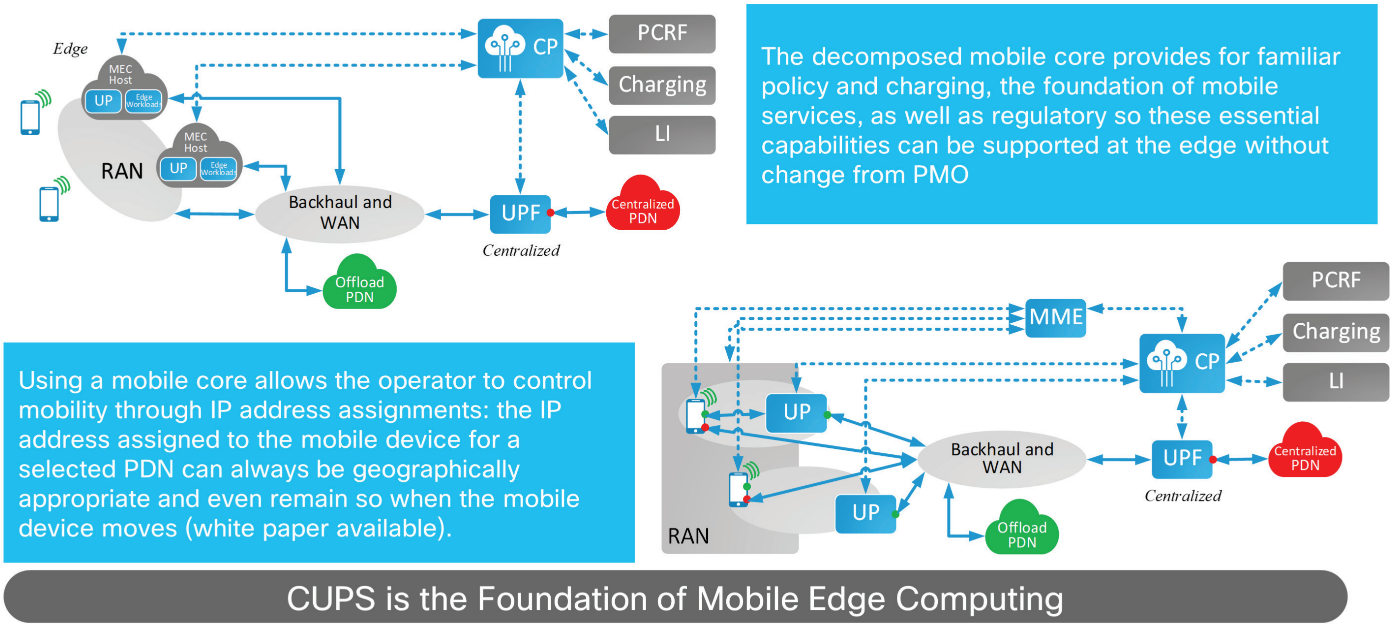
components delivering the service be cloud-native? Hybrid? Most important, where will they be placed in the network?

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 vRAN/cloud RAN, decomposed mobile packet core, CUPS-based BNG, vCMTS, Remote PHY, and Gi-LAN will create the baseline on which future applications will be built.

The following technology innovations are examples of how decomposition and disaggregation will affect how services and applications will be deployed in the future.

Control and user plane separation (CUPS)

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 user plane function (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.



The decomposed mobile core provides for familiar policy and charging, the foundation of mobile services, as well as regulatory so these essential capabilities can be supported at the edge without change from PMO

Using a mobile core allows the operator to control mobility through IP address assignments: the IP address assigned to the mobile device for a selected PDN can always be geographically appropriate and even remain so when the mobile device moves (white paper available).

CUPS is the Foundation of Mobile Edge Computing

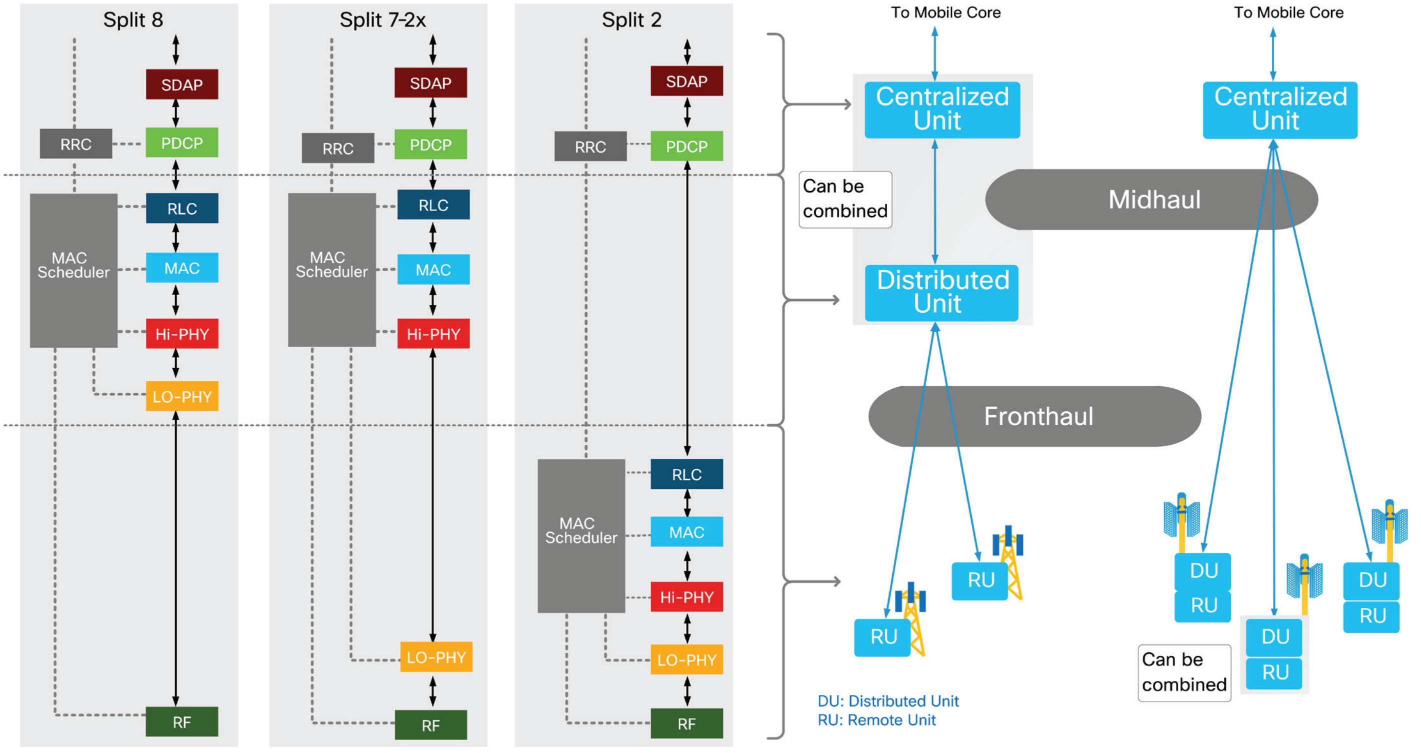
vRAN and Cloud RAN Options

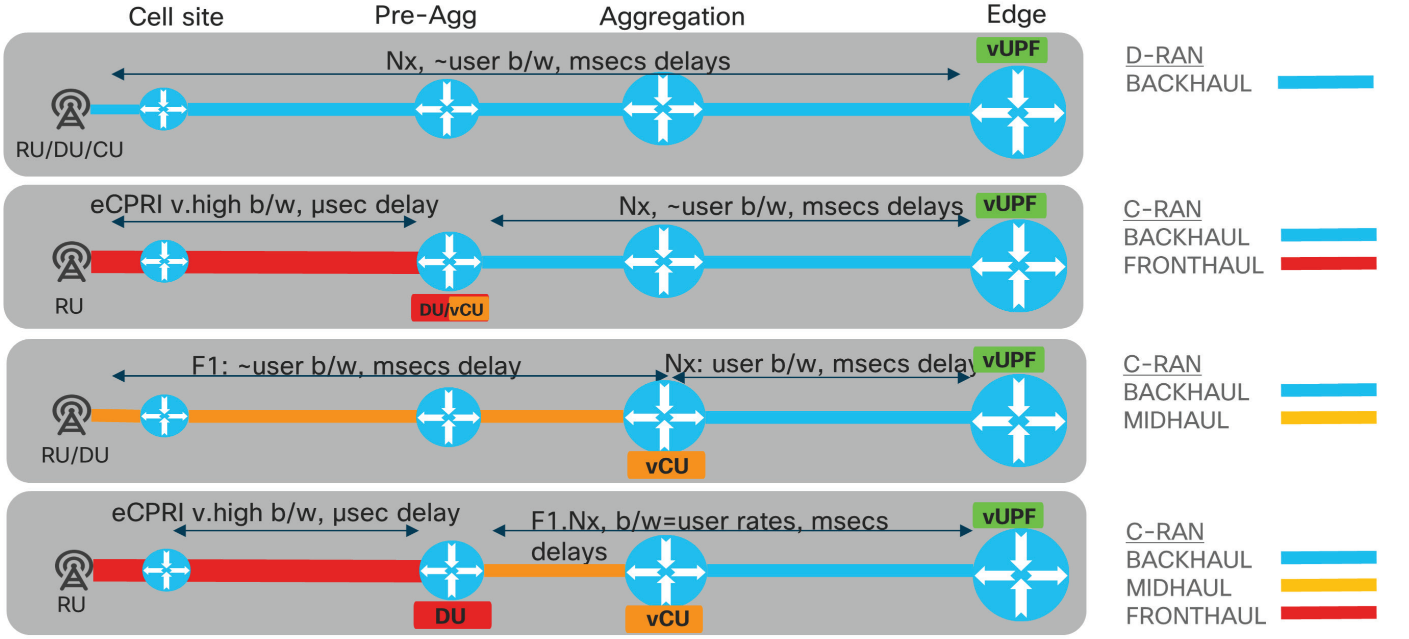
RAN design choices will dictate the placement and virtualization of network functions and, ultimately, the fundamental IP transport architecture. As you explore new RAN splits, you will need to consider where to place and whether to virtualize DU (distributed unit) and CU (central unit) components.

Multiple RAN deployment options and multiple phases of deployments are likely to occur. Increasingly, some functions will be moved from the cell site to another part of the network, such as pre-aggregation or aggregation nodes. In some cases, virtualized CU in aggregation nodes are being deployed, while leaving RU and DU components at the cell site. Stringent bandwidth and latency requirements (in the order of 100 microseconds) will mandate that the positioning of the DU/vDU in a very distributed part of the network (typically within 10-12 km of the cell site locations).

While there are technical benefits to centralized RAN/cloud RAN deployments, many of these decisions will be driven by OpEx considerations. By removing components with active intelligence and complexity from distributed cell sites and getting down to a single, basic connection, you can radically reduce your ongoing overall OpEx costs.

Cisco and partners worked with Rakuten to deliver the industry’s first cloud 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 pre-aggregation 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 could be effectively creating the foundation for a distributed edge platform on which you can build and run future services and applications.



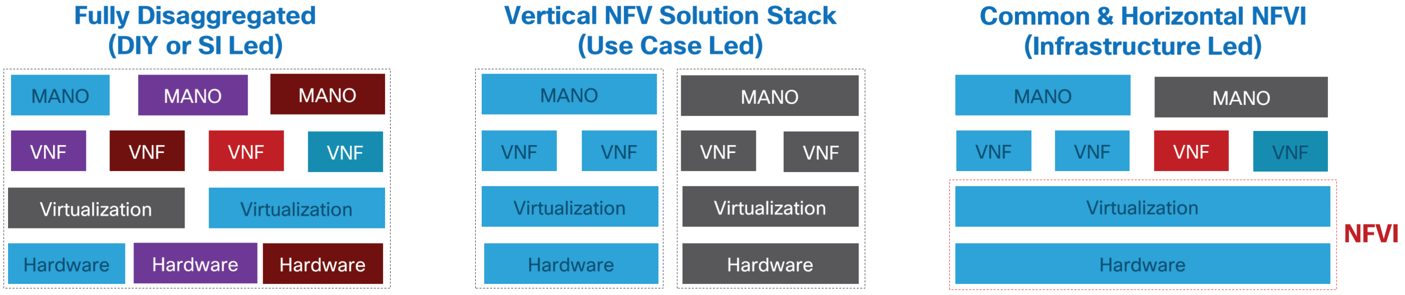

Cell site
Pre-Agg
Aggregation
Edge
Nx, ~user b/w, msecs delays
vUPF
RU/DU/CU
D-RAN
BACKHAUL
eCPRI v.high b/w, μsec delay
Nx, ~user b/w, msecs delays
vUPF
RU
DU/vCU
C-RAN
BACKHAUL
FRONTHAUL
F1: ~user b/w, msecs delay
Nx: user b/w, msecs delay
vUPF
RU/DU
vCU
C-RAN
BACKHAUL
MIDHAUL
eCPRI v.high b/w, μsec delay
F1.Nx, b/w=user rates, msecs delays
vUPF
RU
DU
vCU
C-RAN
BACKHAUL
MIDHAUL
FRONTHAUL

Virtualization platform

Once you know what the service will be, you need to decide which virtualization platform will support it. What are the workload requirements? What is the consumption model? Virtual machines? Containers? Function-as-a-service? Will it be cloud-native? Are you implementing a model you'll use for the foreseeable future? How will it evolve? Where does virtualization go in the network, and will the platform differ depending on where you are?

Network functions virtualization (NFV) supporting OpenStack has transitioned through multiple incarnations. Initial projects favored a fully disaggregated approach that required system integration to get the components working as a solution. The next approach was to support a vertically integrated stack where, in most cases, a single vendor proposed a full stack to support a specific function or group of functions. While this approach helped time-to-market (TTM) launches, it led to multiple silos within single organizations, inability to scale to multiple VNF types, and questionable savings from traditional approaches.

Recent successes have come from providing a common NFV Infrastructure (NFVI) layer upon which multiple VNFs from different vendors can be deployed with commonality at management/MANO layer. This current



approach to virtualization tends to be highly centralized. It's typically limited to a handful of central or regional data centers. Although NFV has been shown to deliver significant advantages in these environments, you can't simply replicate the same approach in remote parts of the network and assume you'll see the same results. Deciding how the virtualization platform will be distributed to support specific use cases is among the biggest questions the industry is working to answer.

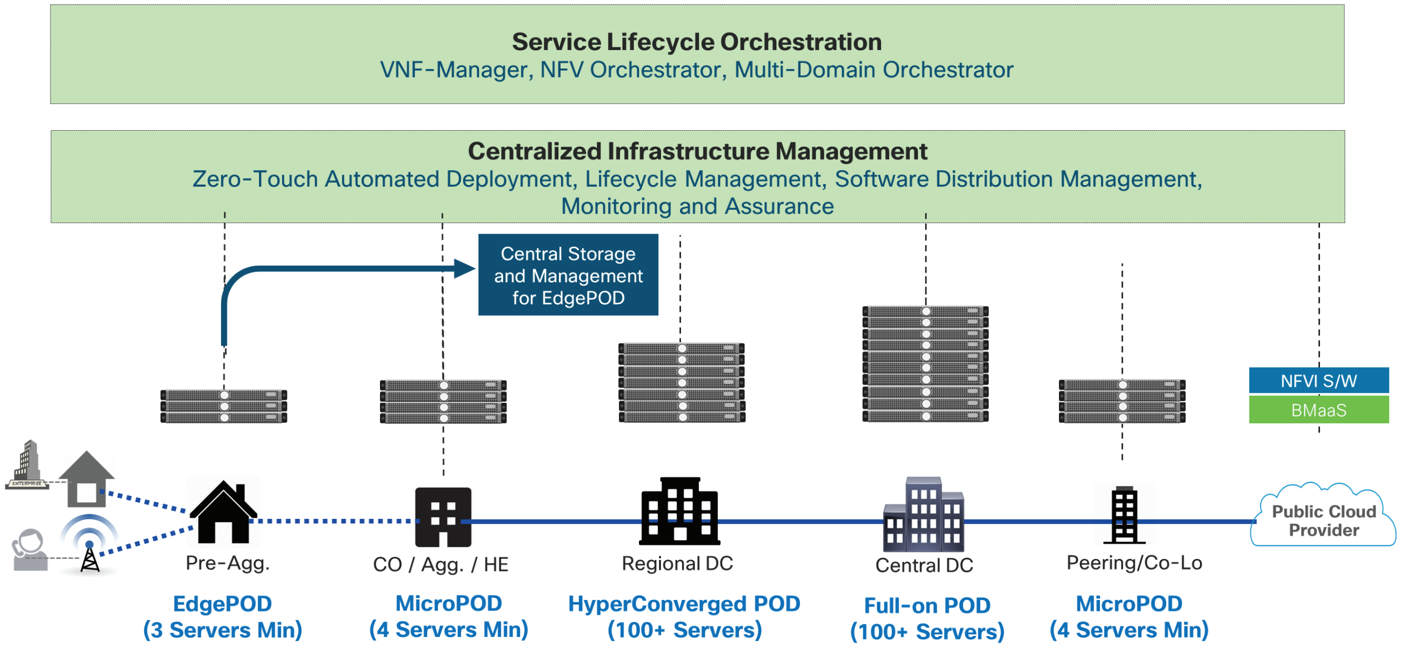
Distributed deployment architectures

What is the best way to move from a centralized data center to distributed virtualization? At many edge locations where you're looking to deploy virtualized functions, it's not viable to deploy the same kind of full-service points of delivery (PODs) you'd find in a central or regional data center. Work is being done to scale NFV PODs to support new environments like COs and pre-aggregation nodes. The advent of MicroPOD and EdgePOD solutions will allow you to support OpenStack but with a much smaller number of servers at CO/pre-aggregation sites that reflect the number of VNFs and services required to be supported at these locations. For example, three servers, versus 64 or even 128 in a centralized data center.

Distributing NFV out to remote sites involves more than reducing the number of servers. You should consider these questions. How will these solutions deliver real-time performance? How can we optimize the stack to support scaled-down solutions? How do you manage and orchestrate virtualized functions end to end?

It's important to deploy a virtualization platform that can adapt easily to the full range of service and operational requirements without sacrificing performance, tooling consistency, or deployment and lifecycle automation.

Ultimately, you should be able to activate the capabilities you need at any given time, in any given location, based on service requirements. And, no matter which virtualization platform you're using in a given location, you should still expect fast networking, common policy, consistent security, and orchestration to automate the process end to end.


Service Lifecycle Orchestration
VNF-Manager, NFV Orchestrator, Multi-Domain Orchestrator
Centralized Infrastructure Management
Zero-Touch Automated Deployment, Lifecycle Management, Software Distribution Management, Monitoring and Assurance
Central Storage and Management for EdgePOD
NFVI S/W
BMaaS
Public Cloud Provider
Pre-Agg.
CO / Agg. / HE
Regional DC
Central DC
Peering/Co-Lo
EdgePOD (3 Servers Min)
MicroPOD (4 Servers Min)
HyperConverged POD (100+ Servers)
Full-on POD (100+ Servers)
MicroPOD (4 Servers Min)

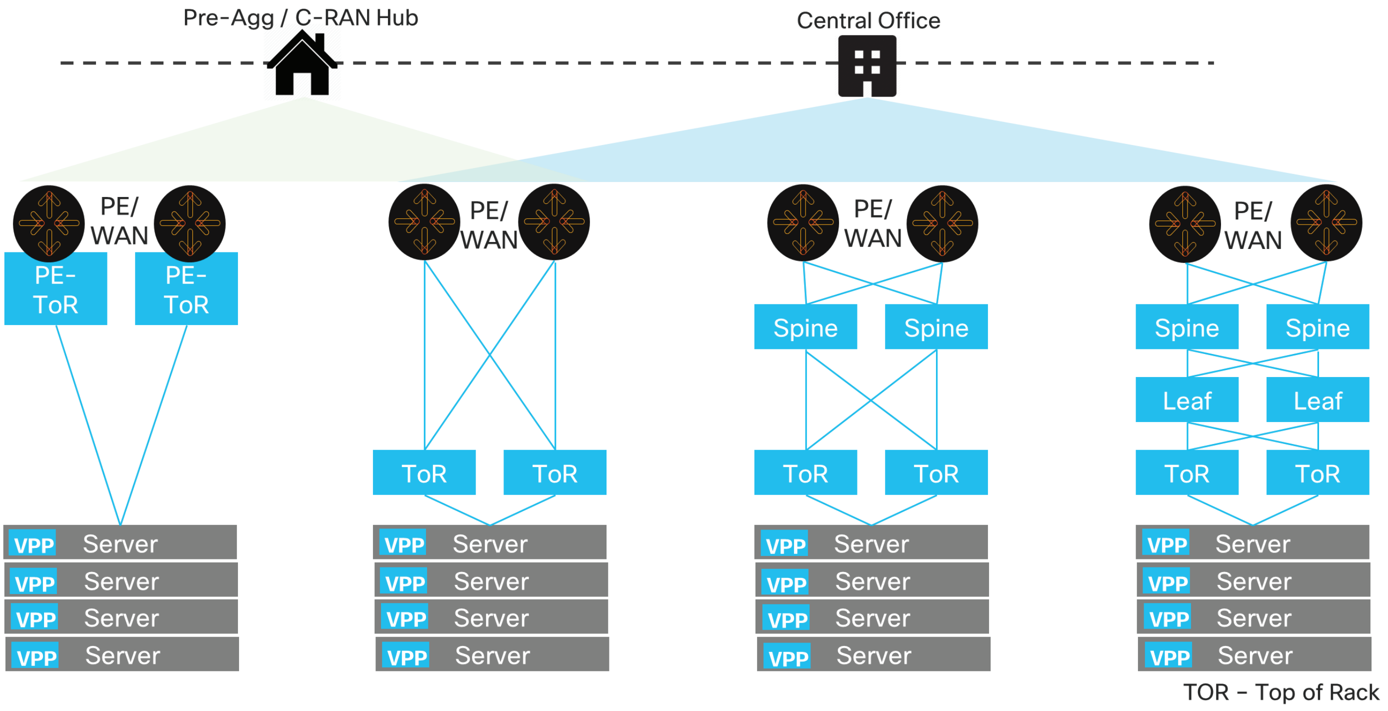
Network fabric

The transformed service edge requires a new vision of a network fabric that connects myriad edge computing nodes such as COs, pre-aggregation, and regional data centers with centralized data centers. Indeed, as you shift towards a distributed edge computing model, the network fabric connecting these points becomes more critical to ensure the necessary latency and economics.

To support new edge capabilities, compute capabilities that previously lived in centralized data centers will need to be distributed. The actual design of those network data centers will vary depending on their location in the network and the scale of VNF, services, and servers required.

In more centralized data centers and COs, you'll see more traditional spine-and-leaf and top-of-rack (ToR) designs. As you distribute these capabilities farther out in the network to pre-aggregation/C-RAN hub sites, the designs will change, becoming simpler and more converged.

These network data centers that will be distributed into the network will require different characteristics from traditional data centers. Often, they will be seen as an extension of the network design and will be needed to support new NFV use-cases, as opposed to legacy or IT applications. The traffic tends to be north-south and only needs to scale to 10s and maybe 100s of tenants. You may use network protocols such as IP/MPLS or segment routing (SR) and also collapsing multiple functions such as provider edge (PE)/TOR into a single platform. You also may need to support full timing/synchronization capabilities with the ability to support low latency services.



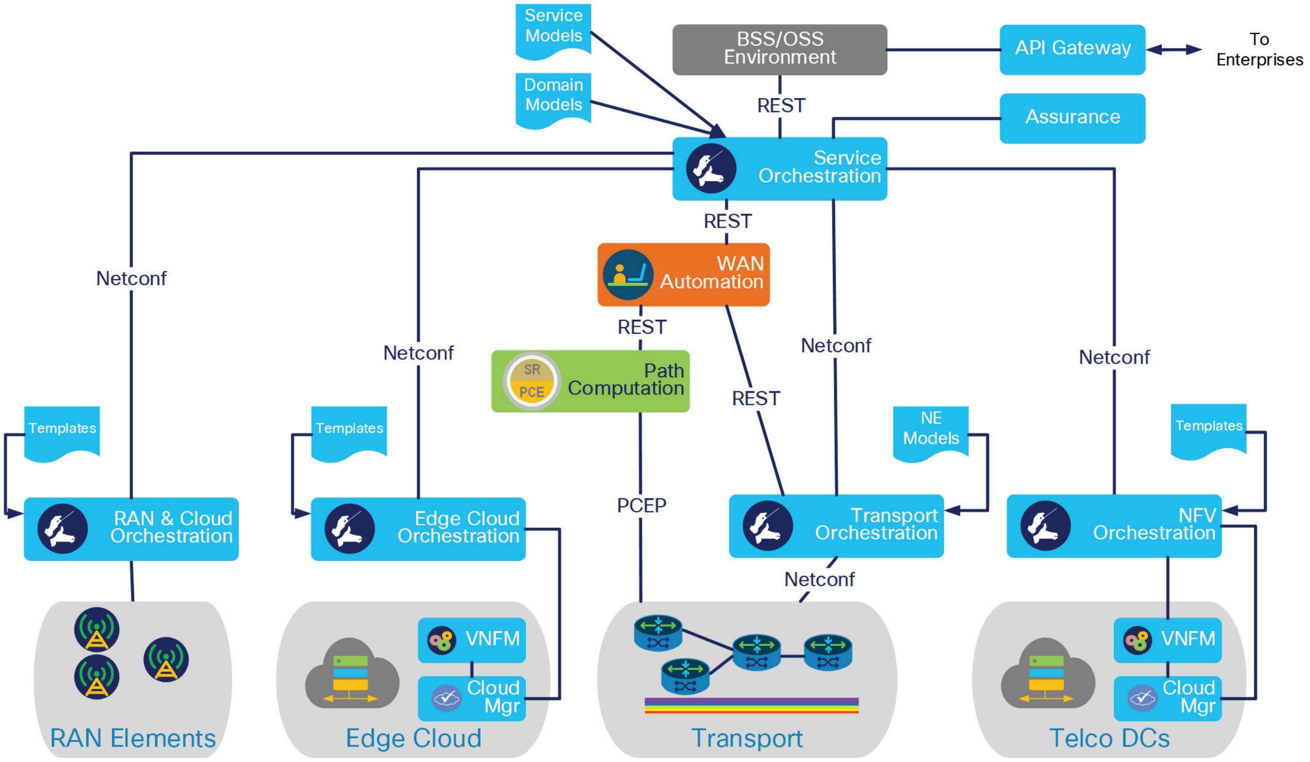
Automation and orchestration

As you distribute more network functions and capabilities out towards the edge, it's important to consider how you'll automate this increasingly complex environment. How will you secure and assure services? And, increasingly, how will customers consume it?

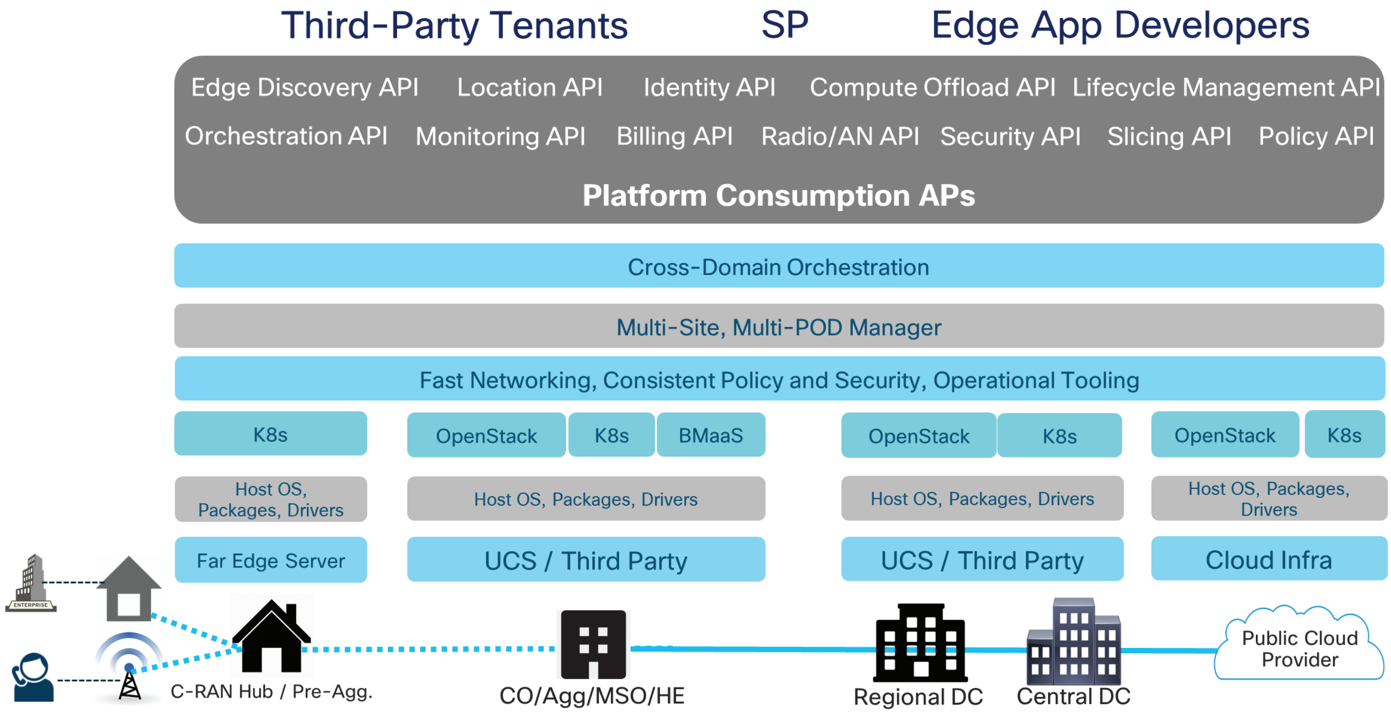
The transformed service edge will need to rein in the inherent complexity that comes with distributing capabilities like virtualization and multi-cloud out to more locations in the network. The only way it can exist in real-world service provider environments and be economical, is if you have as close as possible to a fully autonomous infrastructure.

Service edge transformation therefore requires cross-domain orchestration. However, cross-domain orchestration doesn't necessarily imply a single all-encompassing orchestration layer. You can still rely on separate controllers to orchestrate each individual domain. But in a reimagined edge, you will need the capability to seamlessly stitch these domain-specific controllers into a single solution that you can use to automate operations end to end. This process needs to happen in an open and standardized fashion that supports multi-vendor environments with physical and virtual network functions.

Cisco is one of the few vendors that touches every element of the service provider network, including transport, data center, and edge, so we're well positioned to enable end-to-end orchestration. In fact, we currently have deployments running in real-world customer networks where we deliver end-to-end cross-domain automation from the interface unit under the antenna system (at the cell site), all the way to the centralized data centers.



To enable end-to-end automation when more intelligence is distributed to the edge, the industry needs to address how these capabilities will be consumed. Platform consumption APIs will act as the connective tissue linking all stakeholders in the ecosystem. Service providers, third-party tenants, and edge application developers are all linked with new network architecture capabilities. This type of versatile API model is still being developed across the industry, but it will be central to your ability to monetize your distributed cloud platforms.


Third-Party Tenants
SP
Edge App Developers
Edge Discovery API
Location API
Identity API
Compute Offload API
Lifecycle Management API
Orchestration API
Monitoring API
Billing API
Radio/AN API
Security API
Slicing API
Policy API
Platform Consumption APs
Cross-Domain Orchestration
Multi-Site, Multi-POD Manager
Fast Networking, Consistent Policy and Security, Operational Tooling

Learn more

For more information, visit

- www.cisco.com/go/edge
- www.cisco.com/go/vni

Start preparing now for tomorrow's edge services

As the demands of mobile applications grow and the impact of 5G networks continues to emerge, it's clear tomorrow's edge services will look different from today's. Providing edge networks that can facilitate those services will require new approaches and careful consideration of the placement of distributed and virtualized resources. The choices you make now to evolve your edge networks will dictate the applications you can support and the ways you can build and differentiate your services.

The industry doesn't yet have all the answers regarding the best way to implement technology like MEC or the optimal location to distribute new edge capabilities. In fact, the best location will vary depending on your customers and the use cases you prioritize. Regardless of the unique attributes of your market and strategy, however, the core pillars of successful edge transformation remain the same.

Your applications and services should always be top of mind, and their requirements should dictate every edge decision you make. You will need to settle on a virtualization platform early on and decide what that looks like at different locations in the network. You'll need a next-generation network fabric to run services seamlessly across data centers, WAN, edge, and access. And, as you distribute more functionality and capabilities out toward the edge, you will need to be able to automate this increasingly complex environment.

As the industry continues working to transform edge networks and services, we will continue to raise new questions. However, by keeping these four pillars in mind, you can make sure you're answering the most important ones first.