Cable Edge Compute: Transforming Cable Hubs into Application-Centric Cloud

A Technical Paper prepared for SCTE•ISBE by

Rajiv Asati
Distinguished Engineer
Cisco Systems
rajiva@cisco.com

Alon Bernstein
Distinguished Engineer
Cisco Systems
alonb@cisco.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

Introduction

The mobile industry has popularized and has already started embracing the concept of Multi-Access Edge Computing (MEC). Clearly, the same concept can be applied to cable networks to benefit wide range of use-cases, especially the ones that are latency or bandwidth sensitive. This concept drives placing time sensitive applications e.g. IoT and/or bandwidth hungry applications e.g. CDN/cache at the network edge or Hub sites, closer to the customer. There is a tremendous opportunity for Cable MSOs in transforming their Hub sites into Next Gen Application-Centric Cloud sites.

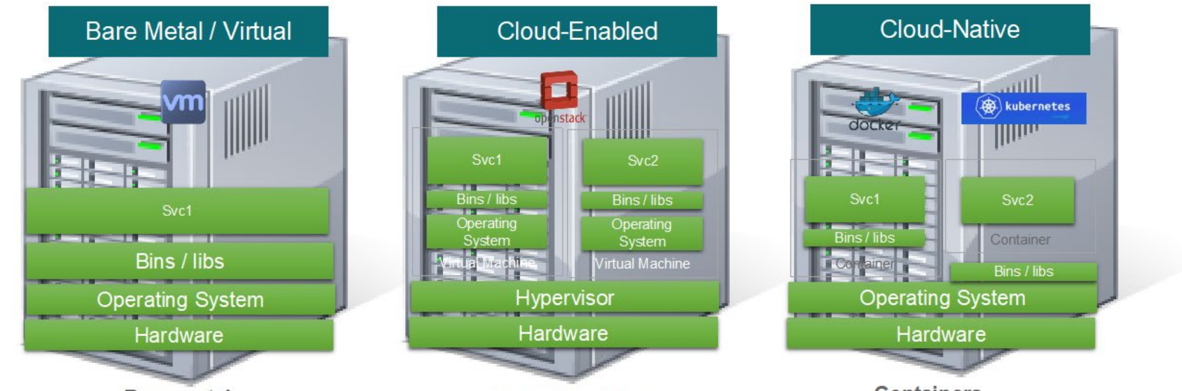
This paper defines the notion of Cable Edge Compute (CEC) along with 'Edge POD' and explores how it could be organized to serve a range of use-cases (e.g. what functions could be placed). The paper outlines the Architectural building blocks (suitable for CEC), their key attributes and captures unique opportunities, challenges and recommendations for Cable Operators.

1 Overview

Cable Multi System Operators/Service Providers are undergoing multi-pronged digital transformations.

A few are attempting to change the game, not just play it better, by morphing services and solutions offerings (connectivity-centric and/or consumption-centric) that ultimately yield the best customer experience, while recognizing the fact that they have to serve not only the ones with eyeballs/eardrums, but also the ones without them i.e. machines such as Internet of Things (IoT) sensors. It is somewhat obvious that wireless and/or wired endpoints, whether deployed in few tens or millions, whether mobile or not, require not only the optimal seamless connectivity constructs, but also the most optimal consumption experience, given that these endpoints consume one or more services (where each service in turn comprise of one or more applications).

In fact, Applications have become foundational to growth and experience, no matter where they run – private cloud, public cloud, hybrid cloud, as long as the customer SLAs are satisfied. (note that two of the phenomena fueling applications evolution are micro-services and cloud native constructs that have enabled more abstraction than ever before, as illustrated in the figure below)



Bare-metal

- Single Service mapped to dedicated hardware
- High performance is the key driver for this deployment.

Virtual Machine

- Hypervisor solve interdependencies between apps but…
- Heavyweight 100's of MB and slow to start 20 seconds or so
- Enabler for Optimial Resource & Agility

Containers

- Emulating OS environment for Services
- Are lightweight and fast to start (=subsecond)
- CI/CD: Life Cycle Management
- Enabler for Microservices and Cloud Native

Figure 1 Applications – Cloud Enabled vs Cloud Native

These trends have been fueling the usage of as-a-Service such as infrastructure as-a-service (IaaS), function as-a-service (FaaS), serverless aka backend as-a-service (BaaS) etc. for executing the application functions, as illustrated in the figure below. One or more application functions could execute in one or more cloud locations in a scale-out manner for whatever time-period, independent of the location (centralized, partially distributed, fully distributed).

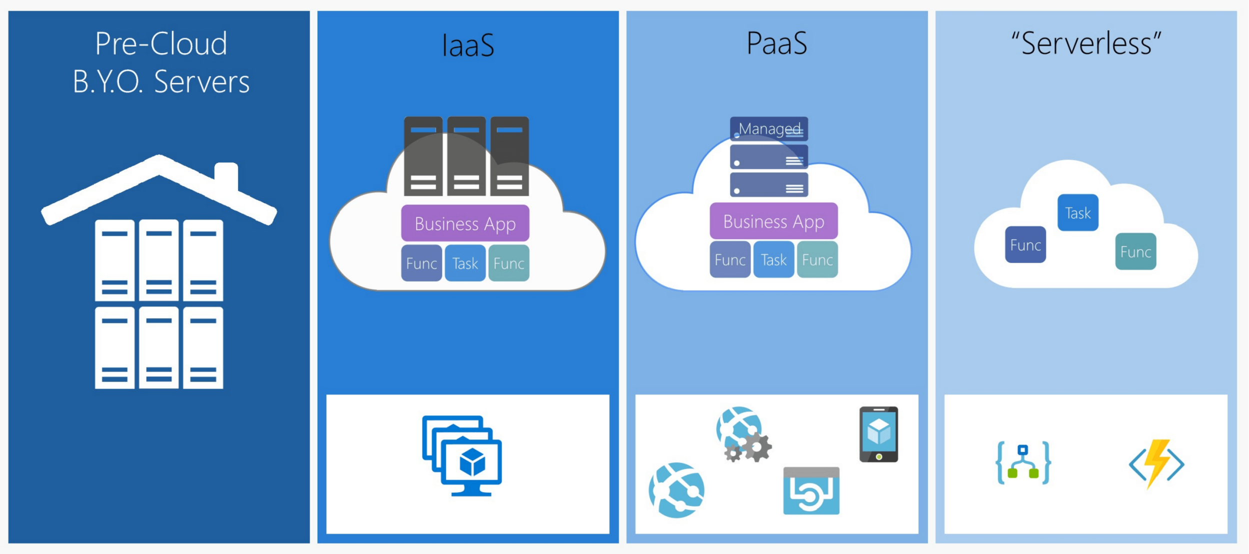


Figure 2 IaaS, PaaS, BaaS

Courtesy - https://stackify.com/function-as-a-service-serverless-architecture/

Application-Centric means that the network administrators manage a system for a set of applications rather than managing individual nodes like they did in the past.

The faster the Cable Operators leverage the **Application-Centric paradigm with flexible, distributed and intelligent network architecture**, the faster they get towards enabling superior customer experience. Applications could be related to internal usage (e.g. Infrastructure) or external usage (e.g. subscribers) or both.

The Operators could deploy Application-Centric Clouds in a centralized manner or distributed manner. It is important to point out that Cable Operators have had precious Hub Sites distributed across the footprints and they could be the ideal candidates for Application Centric Cloud. This is pertinent especially since the majority of revenue growth is now expected in B2B or B2B2x space, as illustrated in the figure below -

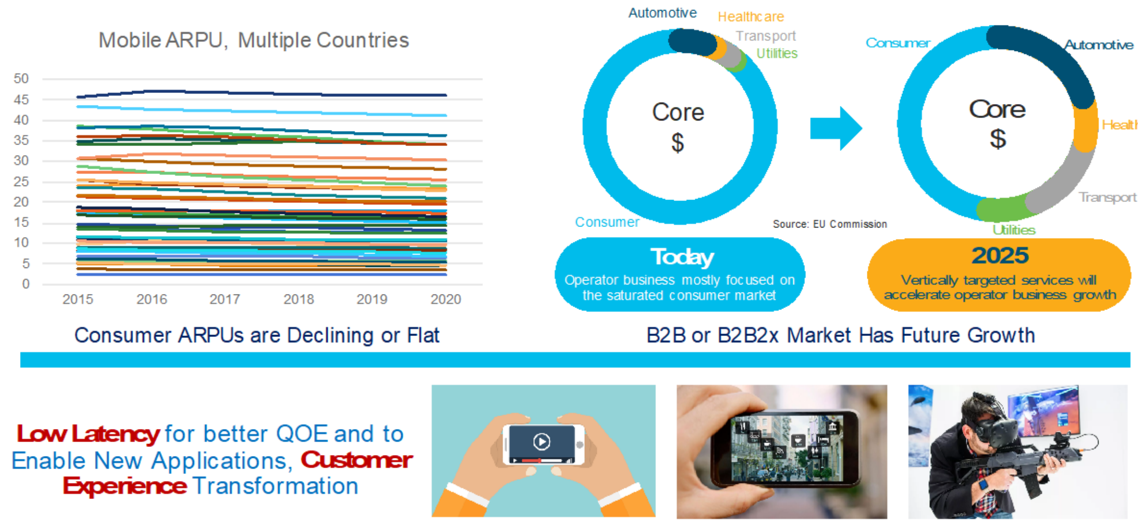


Figure 3 Customer Experience Driving Applications Enablement Closer and Closer

Minimizing the latency is now more important (than increasing bandwidth) for improving QoE, and by hosting applications closer to the customers can greatly help. Increasing/throwing more bandwidth data rate doesn't help much after a while. For example, browser application could load a webpage in around 3500msec with 200ms E2E latency, but in around 2000msec with 100ms E2E latency – 45% improvement. However, increasing the bandwidth bandwidth from 5Mbps to 10Mbps yields around 5% improvement in page loading experience. See Reference [9].

Cloud Services Providers related to Residential and Enterprises e.g. CDN operators, Gaming etc. are pushing hard to get closer to the eyeballs and last-mile networks.

1.1 What – Definition, Background and MEC Relation !

"Cable Edge Compute" is intended to represent "IT & Telco centric" cloud-computing capabilities at the edge of the MSO network and in close proximity to Cable subscribers.

Cable Edge Compute (CEC) is a form of Multi-Access Edge Computing (MEC) that is applied to Cable MSO environment.

Cable Edge Compute aims to improve subscribers' experience by cutting out the often long and imperfect network path between the subscriber's device and the location where the application they are accessing is hosted, as much as possible, in order to lower latency, increase reliability and improve overall network efficiency.

The concept of placing computing power near the customer's devices with the primary goal of improving customer experience while reducing latency, backbone capacity, etc. and getting better scale and availailbity. Edge Computing nodes are on the outer region of the core network or its backbone. Almost any device with computational power that is near or at the customers' devices location can act as an edge computing device, as long as it's practical.

1.1.1 MEC Background!

Multi-access Edge Computing (MEC), has picked up a variety of names:

- edge computing,
- edge cloud,
- fog computing,
- mobile edge computing,
- Etc.

So, what is MEC? It is about hosting one or more applications on compute, network and storage resources that are placed closer to the subscribers (residential or enterprise). Per ETSI, MEC is "an evolution of cloud computing [that]brings application hosting from centralized data centers down to the network edge, closer to consumers and the data generated by applications." In other words, MEC is a cloud-based IT service environment at the edge of the network.

> MEC was originally coined to benefit applications and subscribers with mobile access, however, it has since evolved to cover multiple types of access, including wireline access.

MEC essentially brings cloud capability (not only compute, but also networking and storage) to the network's edge and helps unlock superior experience to the "things" including eyeballs, sensors etc. It enables real-time, high-bandwidth, low-latency access to applications and subscribers, allowing operators to open their networks to a new ecosystem and value chain.

MEC in the context of Cable access can be referred to as Cable Edge Compute.

1.1.2 Cable Cloud vs. IT Cloud

Clouds are about hosting application functions. However, depending on the applications, the Cloud could be designated as Cable/Telco Cloud or IT Cloud, independent of whether deployed on-premise or not.

Cable/Telco Cloud is about hosting Telco related infrastructure services such as Subscriber Edge Functions (e.g. CCAP, BNG, SGW, PGW), Access Network Functions (e.g. RPHY, OLT, eNB) as well as end-user services such as IMS voice, SBCs, video, media content, etc. It is worth pointing out that Cable Applications such as CCAP could be deployed as a single network function (e.g. single container) or multiple functions (e.g. multiple containers).

> Cable/Telco Cloud applications tend to require high throughput (10Gbps+) and low latency/jitter centric infrastructure.

IT Cloud is intended to host IT related services such as Operations Support System (OSS) applications, Billing Support System (BSS) applications, end-user portal, media storage, etc. IT Application such as Portal could be deployed as a single function (e.g. single container) or multiple functions (e.g. multiple containers).

> IT Cloud applications tend to require high compute and storage centric infrastructure.

Of course, the data centers implementing Cable/Telco Cloud or IT Cloud exhibit high degree of resiliency, faster convergence, etc. It is likely that the line between Cable/Telco Cloud and IT Cloud would continue to diminish and soon, there won't be any meaningful difference between Cable/Telco Cloud and IT Cloud, though Security posture may mandate them to stay separated.

Reference [1] and [3]

1.2 Why – Benefits?

Many Operators have 1000s of Hub Sites already deployed/operational across the country. These Hub sites already have Network Edge functions for certain services such as Internet Data etc. and are best suited for transformation into Application Centric Cloud that can host additional B2B and B2C services such as RAN, IoT, Gaming, AR/VR etc. and offer superior customer experience.

The Hub sites are usually already quite fiber rich and are employing innovative technologies such as distributed CCAP, Remote PHY, Full Duplex DOCSIS etc. These are quite complementary to the notion of Application Centric Cloud and Edge Computing that essentially brings cloud capability (compute, storage, network) to the network's edge and helps unlock superior experience to the "things" including eyeballs, sensors etc.

Cable operators can now be enabled to be the cloud providers, taking a page from the success of companies such as Amazon, Google, etc., and leveraging the networks & Hub Sites assets in a new way, and really strive to yield win-win:

1. Better Utilization – If the hub sites are transformed into application centric cloud sites, then they could allow hosting not only walled-garden services, but also 3rd party services on-demand and take advantage of geographic closeness to the subscribers.

   In particular, Edge Computing is seen as key to massive IoT deployments and as crucial for analyzing large amounts of data coming from increasingly connected things.

2. Increased Security – If data is processed closer to the customer site instead of far away (public cloud, for ex), then the risk of data theft or illegal access is significantly reduced. One can localize mission-critical data processing to help meet security requirements.

3. Decreased Latency – If data is processed closer to the customer edge of the network in near-real time, then propagation delay could be significantly reduced.

4. Increased Control – If one is able to dictate which data stays local and what goes external for processing with utmost granularity to the device level, then it increases the overall optimality.

5. Better Operations – If key constructs of the network are virtualized on a common x86 hardware platform, then it could improve the operations in terms of seamless service creation environment and efficiency.

1.3 Where is the Edge? Centralized vs. Distributed ?

Since 1990s, the “Edge” has referred to the point where a "customer connects to the provider."

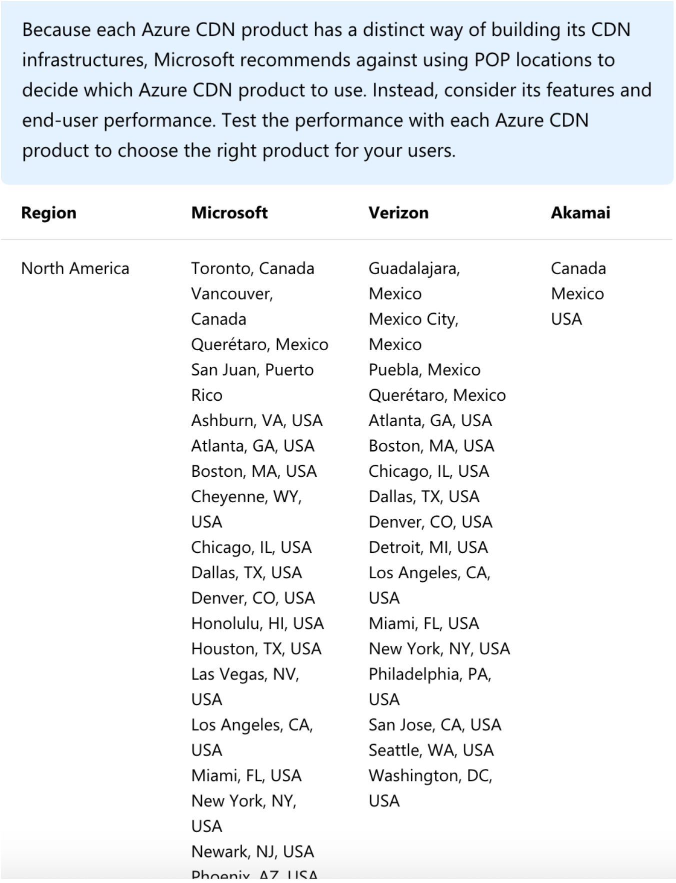
> The provider being the organization providing a service such as broadband, telephony, video, mobility etc. to the customer belonging to enterprise, residential, retail etc.

However, since the last decade or so, the “Edge” has increasingly referred to the point “where the service is located’, largely because of the emergence of cloud service providers (CSP), which are more concerned about where the cloud services can easily run at scale.

Two things have changed – (1) more focus on workload centric services (less focus on network centric connectivity), and (2) more focus on proximity of the workloads wrt its users

So, while it is debatable where the Edge exactly is (the answer may vary quite a lot depending on who we ask – Content Providers/Aggregators e.g. Netflix, vs. Public Cloud Providers e.g. AWS, vs. Online Gaming Provider e.g. TakeTwo etc., vs. Subscribers e.g. eyeballs/ears/sensors etc.), it is important to describe the Edge in the context of Cable Operators paradigm. Arguably, many regard the Edge where the Subscriber Session Control function (e.g. CMTS) is instantiated – usually the Hub Site.

Few of public cloud providers have 20+ Edge locations in the USA and building more. They continue to expand their Edge presence, either by placing the applications and/or compute capacity in peering points and into the MSO/SP networks. However, most of their Edge locations are far from the access & aggregation/regional networks, and not as close to the subscribers (yet) as Cable Hub sites are. Nonetheles, one of the public cloud providers has partnered with an incumbent Telco SP to expand their number of Edge locations, as illustrated in the figure below.



Because each Azure CDN product has a distinct way of building its CDN infrastructures, Microsoft recommends against using POP locations to decide which Azure CDN product to use. Instead, consider its features and end-user performance. Test the performance with each Azure CDN product to choose the right product for your users.

Region	Microsoft	Verizon	Akamai
North America	Toronto, Canada Vancouver, Canada Querétaro, Mexico San Juan, Puerto Rico Ashburn, VA, USA Atlanta, GA, USA Boston, MA, USA Cheyenne, WY, USA Chicago, IL, USA Dallas, TX, USA Denver, CO, USA Honolulu, HI, USA Houston, TX, USA Las Vegas, NV, USA Los Angeles, CA, USA Miami, FL, USA New York, NY, USA Newark, NJ, USA Phoenix, AZ, USA	Guadalajara, Mexico Mexico City, Mexico Puebla, Mexico Querétaro, Mexico Atlanta, GA, USA Boston, MA, USA Chicago, IL, USA Dallas, TX, USA Denver, CO, USA Detroit, MI, USA Los Angeles, CA, USA Miami, FL, USA New York, NY, USA Philadelphia, PA, USA San Jose, CA, USA Seattle, WA, USA Washington, DC, USA	Canada Mexico USA

Figure 4 Public Cloud Edge Locations - Example

Source - https://docs.microsoft.com/en-us/azure/cdn/cdn-pop-locations

Subscriber closeness is a key advantage that Cable MSOs can utilize.

Centralized or Distributed or both

It is important to highlight the pros & cons of centralized vs distributed in the context of what dictates the customer experience – latency, bandwidth etc. and what dictates the cost & complexity in MSO environment, as illustrated in the figure below -

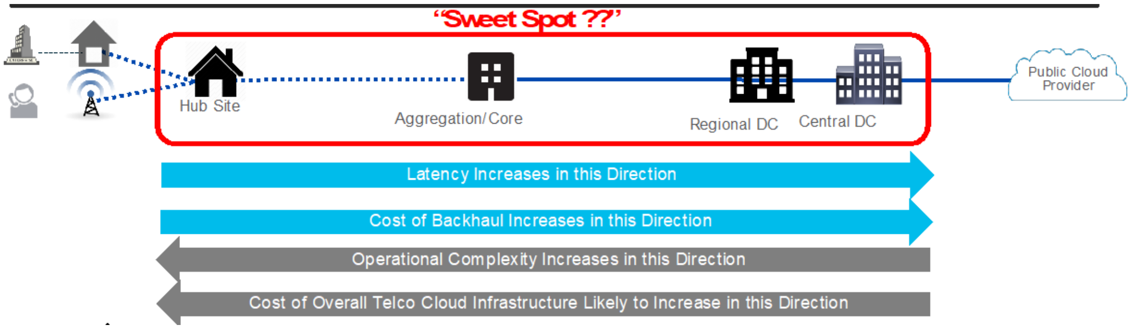

"Sweet Spot ??"
Hub Site
Aggregation/Core
Regional DC
Central DC
Public Cloud Provider
Latency Increases in this Direction
Cost of Backhaul Increases in this Direction
Operational Complexity Increases in this Direction
Cost of Overall Telco Cloud Infrastructure Likely to Increase in this Direction

Figure 5 Centralize or Distribute the Edge Clouds - Pros & Cons

It is somewhat clear that the applications that require SLAs comprising lower latency and higher bandwidth would demand Edge Clouds distributed in the network. For example, on-demand/online video consumption requires tons of downstream bandwidth, whereas physical security/monitoring (e.g. video surveillance) requires tons of upstream bandwidth. For example, the below figure illustrates a home network usage with growing upstream WAN traffic share.

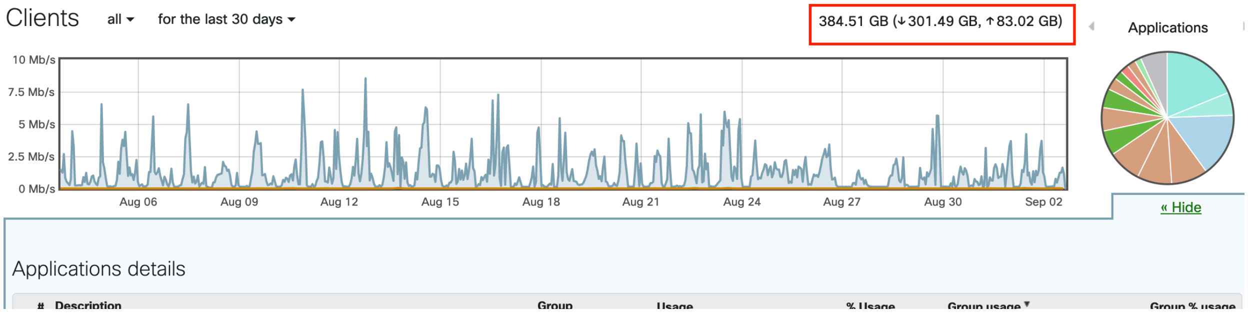


Figure 6 Upstream Bandwidth could be ~30% of downstream Bandwidth

It is worth noting that almost all of ISP networks are built to optimize downstream traffic consumption, not upstream traffic consumption.

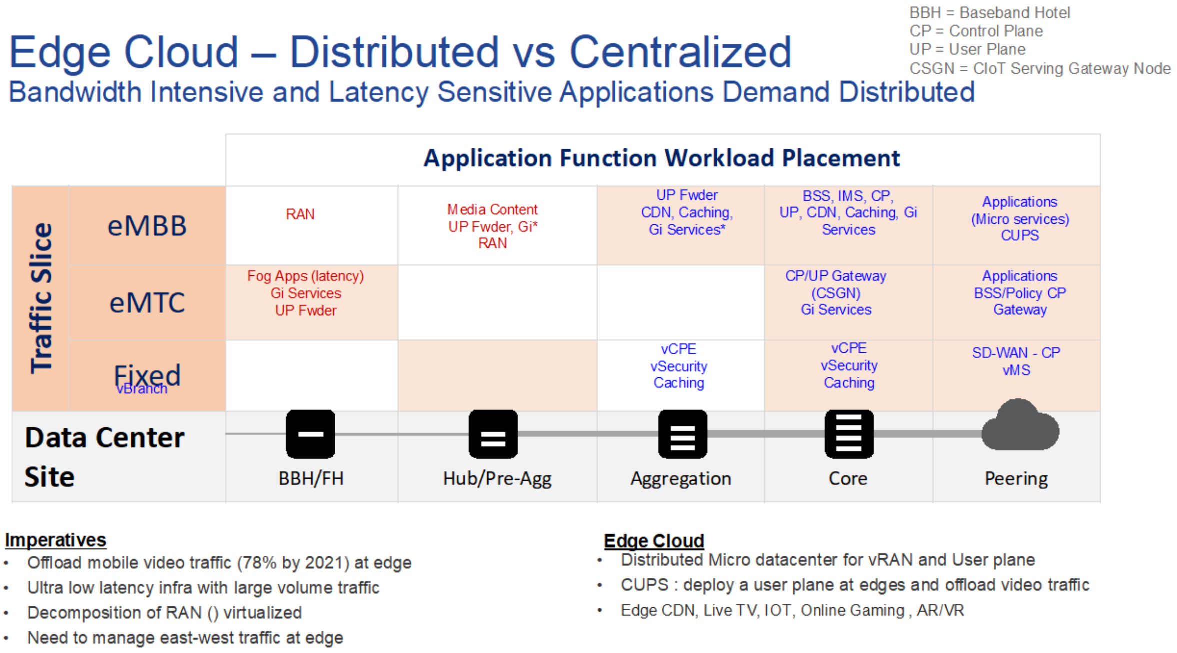
There are number of factors to consider in order to decide whether to distribute the Edge Clouds or not, as tabulated below –

Table 1 Factors that can Influence Distributed Edge Clouds

	+ve Factors	-ve Factors
1	Reduction of Latency	Operational Complexity
2	Reduction of xHaul Bandwidth	Higher Infra Costs
3	Location Awareness	Location Availability
4	Regulatory / Compliance	Security Concerns
5	Localized Impact of Fault	Technology Maturity

All in all, it is imperative to consider different applications – network centric, IT centric and customer centric and where all should the Edge Clouds be placed in the network in order those applications, considering the +ve/-ve factors from the above.

The below figure captures some of the applications (note that it doesn't cover all of applications) -



BBH = Baseband Hotel
CP = Control Plane
UP = User Plane
CSGN = CIoT Serving Gateway Node

Edge Cloud – Distributed vs Centralized

Bandwidth Intensive and Latency Sensitive Applications Demand Distributed

		Application Function Workload Placement				
Traffic Slice	eMBB	RAN	Media Content UP Fwder, Gi* RAN	UP Fwder CDN, Caching, Gi Services*	BSS, IMS, CP, UP, CDN, Caching, Gi Services	Applications (Micro services) CUPS
	eMTC	Fog Apps (latency) Gi Services UP Fwder			CP/UP Gateway (CSGN) Gi Services	Applications BSS/Policy CP Gateway
	Fixed vBranch			vCPE vSecurity Caching	vCPE vSecurity Caching	SD-WAN - CP vMS
Data Center Site		BBH/FH	Hub/Pre-Agg	Aggregation	Core	Peering

Imperatives

- Offload mobile video traffic (78% by 2021) at edge
- Ultra low latency infra with large volume traffic
- Decomposition of RAN () virtualized
- Need to manage east-west traffic at edge

Edge Cloud

- Distributed Micro datacenter for vRAN and User plane
- CUPS : deploy a user plane at edges and offload video traffic
- Edge CDN, Live TV, IOT, Online Gaming , AR/VR

Figure 7 Example Application Functions for Edge Cloud Locations

It is important to keep E2E network architecture and the relevance of Hub sites perspective. Please refer to the architecture section 3.

2 Cable Edge Compute – Application Functions

Edge Computing is one of the disruptive paradigms in the network architecture that enables a myriad of industry-specific use cases. By becoming edge cloud providers, Cable operators can leverage their nation-wide wired access networks (in additional to any wireless access) to shift their relationships with application developers as well as the customers who consume those services, and ideally, to become more like the agile, cloud and application-centric operators focused on innovation.

Of course, application functions would vary depending on the use-cases that are targeted. A few may be more relevant than the others. The applications fall within the below three categories

1. Infrastructure Use-Cases – RAN, BNG, CMTS, PGW/SGW etc.
2. Services B2C Use-Cases – CDN, LiveTV, IoT, Gaming, AR/VR, AI/ML etc.

3. Services B2B Use-cases – CDN Hosting, Online Gaming, Surveillance etc.

2.1 Infrastructure Use-Cases

2.1.1 (Virtual) RAN Functions

The way mobile networks have been built for decades are evolving and improving. Radio Access Network (RAN) with traditional cell sites, where the traditional monolithic eNodeB(s) have resided, are getting fast modernized with eNodeB getting decomposed into RU, DU and CU [4]. Operators have been developing RAN strategies around cost-effectiveness rooted in virtualization, cloud-native etc. In fact, they are the key tenets of the 5G RAN and made largely possible by “edge computing” with end-to-end automation for both infrastructure and services.

Modern RAN is centered around disaggregation and decomposition at multiple fronts e.g. two-layer split with cell site having only Remote Radio Units (RUs) and Antennas, virtualized Distributed Unit (vDU) functions processing lower layers of the radio stack and virtualized Central Unit (vCU) functions processing upper layers of the radio stack, will reside. RU – vDU connectivity being fronthaul and vDU – vCU connectivity being midhaul.

In essence, vRAN infrastructure service could comprise of 2 workload applications – Distribution Unit (DU) and Control Unit (CU). Each could represent a singular function or plural functions that could potentially be executed at the same site or different sites depending on the mobile operator’s design, as illustrated in the figure below:

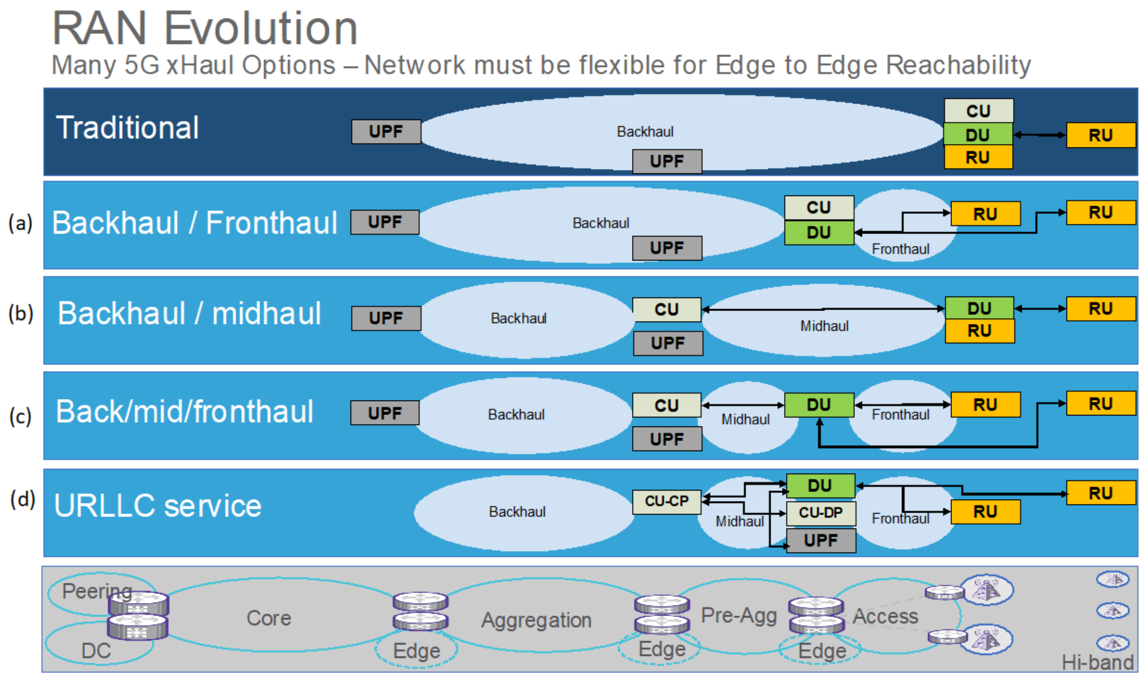
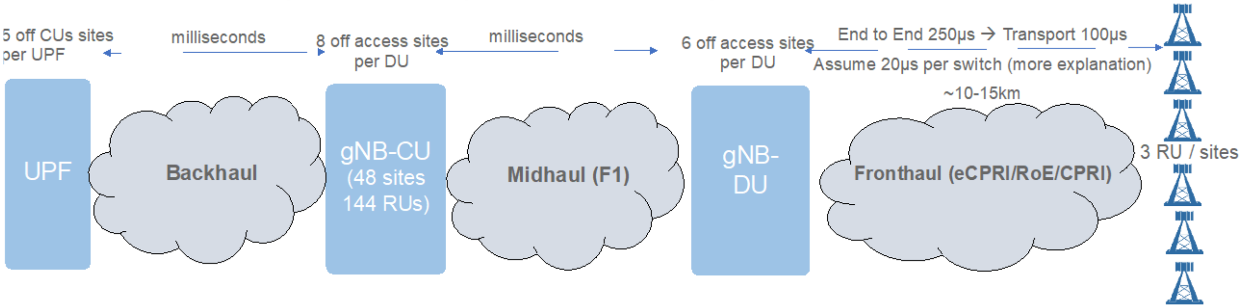


Figure 8 Mobile RAN Evolution

Note that the RAN fronthaul interface (RU-DU) deals with digitized RF signals, hence, it has a very high throughput (3Gbps+) and extremely low latency requirement (<250us) depending on the chosen split design, as illustrated in the figure below –

RAN Evolution

5G bandwidth implications - (based on NGMN)



Per UPF site

F1 split = 2880Gbps

Per CU site

In:F1 split = 576Gbps

Out: Nx split=576Gbps

Per DU site

F1 split = 72Gbps

Per DU site

3GPP best case = 396Gbps

3GPP worst case = 1548Gbps

eCPRI = 450Gbps

Per Access site

3GPP best case = 66Gbps

3GPP worst case = 258Gbps

eCPRI = 75Gbps

NOTE: No statistical gain assumed

DL	4 Gb/s*
UL	3 Gb/s*
Latency	1.5 – 10ms

(a*)

DL	10 – 22 Gb/s
UL	16 - 22 Gb/s
Latency	250us

(b**)

DL	38 – 86 Gb/s
UL	54 – 86 Gb/s
Latency	250us

(c)

DL	10 – 22 Gb/s
UL	54 – 86 Gb/s
Latency	250us

Split numbers are 3GPP numbers based 5G 100Mhz b/w (3GPP TR 28.801 v14): eCPRI figure approx 25 Gbps (UP: ~20Gbps CP: ~5Gbp) (e-CPRI specification v1)
RRU: 100 MHz, 256 QAM, 8x8 MIMOIQ BW (7-16)bit, 32 Antenna Port) 3GPP TR 38.801 V14.0.0 (2017-03)

Figure 9 5G Bandwidth in Access, Aggregation

The point to take away here is that the usage of Cable Edge Hub sites could be suitable with fiber connectivity between the cell sites and edge sites.

Also worth noting that RAN focusing on the usage of higher frequency bands such as mmWave (24-86 Ghz) (whereas the majority of current RAN deployments are around 2 GHz or below) want to benefit from increased radio efficiency/capacity etc, however, they also have to put up with corresponding limited coverage and strict line-of-sight consideration, which ultimately mean that RAN will likely have a lot more cell sites in a given area i.e. a lot more investment – in fact, according to a research report [http://www.delloro.com/products-and-services/mobile-radio-access-network#5-year-forecast-report], the 5G New Radio (NR) would propel the RAN market to around $160B over the next 5 years.

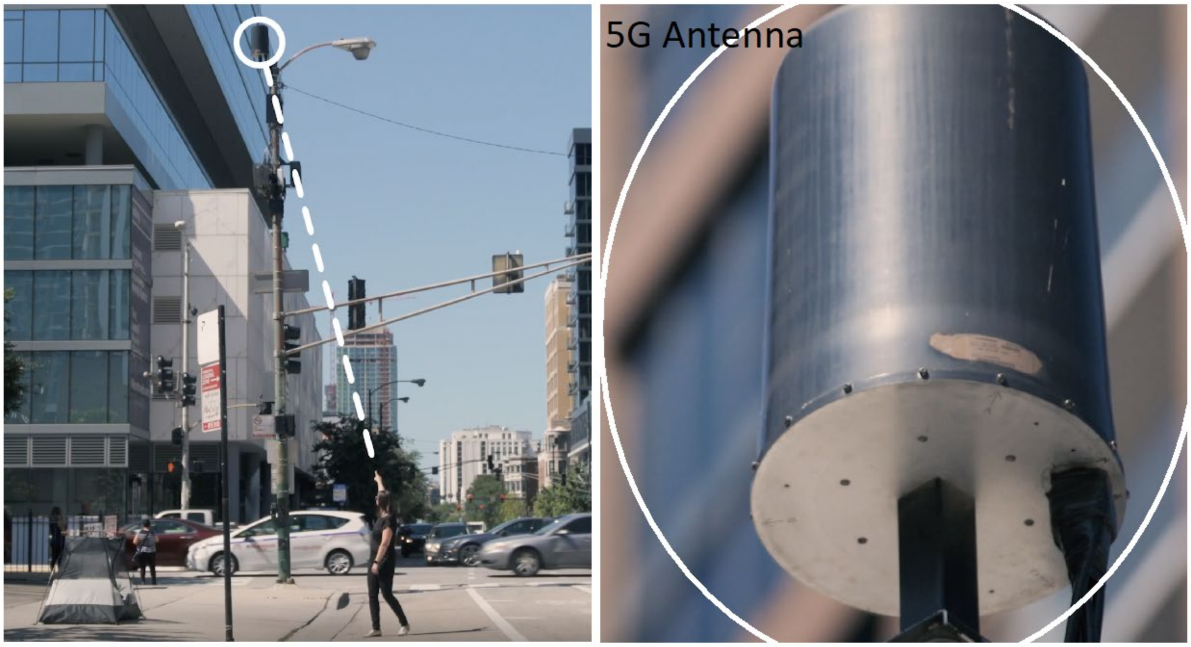


Figure 10 Radio Access Network showing 5G Cell Site with mmWave

The point to take away here is that the usage of Cable Edge Hub sites would be suitable to provide fiber connectivity between these new cell sites with RUs and edge sites with DUs/CUs.

<u>This is quite an opportunity for Cable Operators to position their distributed Hub Sites as the ideal places with CEC to host one or more virtual RAN functions for deeper and denser radio deployment.</u>

.

2.1.2 CBRS

The US Government/FCC has provided 3.5 GHz (3550-3700 MHz) for Citizen Broadband Radio Service (CBRS). It has three tiers such that tier 1 “incumbents,” including ship-borne Navy radars, fixed satellite stations, and wireless providers, are protected from lower tier users at all times.

CBRS offers an economical path for Cable MSOs to enter the wireless industry via an MVNO strategy can now deploy LTE network and minimize network expenses by offloading the traffic to its owned CBRS LTE network (instead of sending it to the host MNO). This means that Cable MSOs can offer not only in-building & outdoor wireless coverage, but also capacity expansion for their own benefits or for their B2B partners’ benefits.

<u>This is an upcoming opportunity for Cable Operators to leverage the Hub Sites as the suitable places with CEC to host CBRS functions.</u>

2.1.3 Subscriber Edge/User Plane Functions (e.g. CMTS, BNG, PGW)

Cable Operators have historically used CMTS / CCAP as the Subscriber Edge / Gateway function, which dictates the per-subscriber session and policy enforcement. Similarly, mobile Operators have used SGW/PGW (in 4G/LTE) as the Subscriber Edge / Gateway function. With the advent of Control Plane User Plane Separation (CUPS), Gateways could be decomposed, disaggregated and cloudified such that User Plane Function (UPF) could be deployed in a distributed manner, whereas CPF could be deployed in a centralized manner depending on the level of scale and aggregation needed.

Cable Operators can exploit the opportunity to embrace Fixed Mobile Convergence (FMC) by placing the Subscriber Edge Functions in virtualized manner at the transformed Hub sites. Furthermore, Control Plane User Plane Separation (CUPS) could enable a common converged UPF for cable access and mobile access distributed, while placing the CPF in a regional/centralized sites. This fits well with decomposed CCAP + RPHY approach many Operators are already pursuing.

The figure below illustrates the CMTS and SGW/PGW placement possibilities -

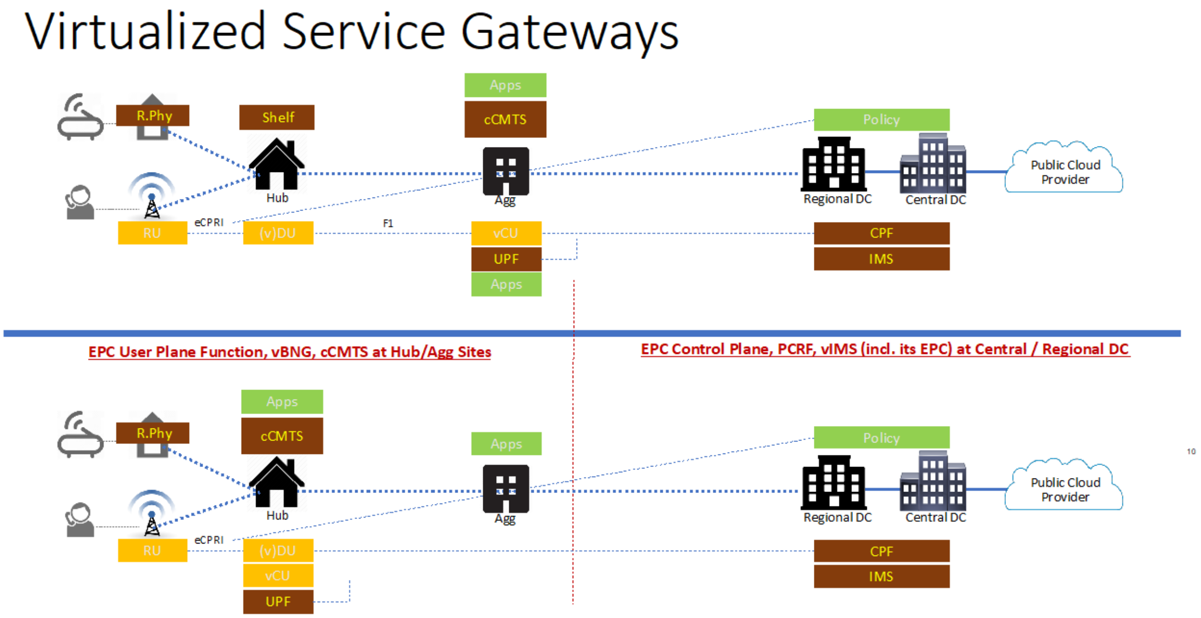


Figure 11 Subscriber Edge Functions, CUPS

Lastly, if certain traffic needs to be offloaded, then the CEC enabled Hub Sites could facilitate that, thanks to UPF.

2.2 B2C Service Use-Cases

Operators have been offering plethora of services to the consumers. Some are highlighted here in the context of CEC.

2.2.1 Gaming

Online Video Gaming has attracted the attention of network engineers ever since the first multi-player online games appeared. Arguably, Online gaming has become one of the most profitable businesses on the Internet. The gamers do expect SLAs in both upstream and downstream direction notably in lowest latency and higher bandwidth with or without AR/VR. The game traffic is interactive and usually in Hub&Spoke pattern (as all client traffic is sent to the server(s) and sent back to the clients).

Google Stadia (recently announced) is making online gaming similar to that of content streaming that are hosted in the cloud allowing any type of endpoint devices. The below figure illustrates Gaming Node being hosted closer to the subscribers –

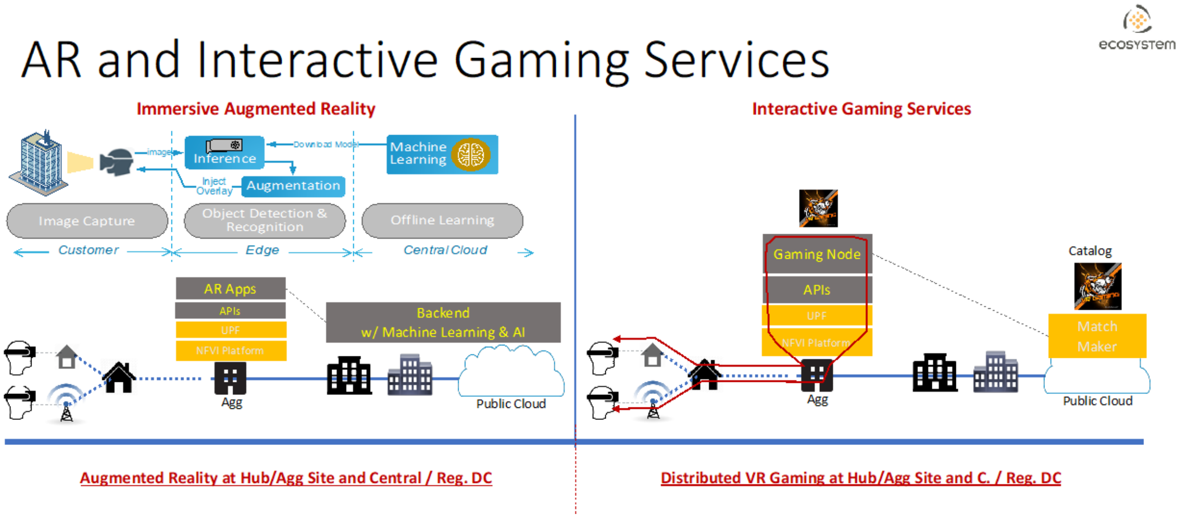

AR and Interactive Gaming Services
ecosystem
Immersive Augmented Reality
image
Inference
Download Model
Machine Learning
Inject Overlay
Augmentation
Image Capture
Object Detection & Recognition
Offline Learning
Customer
Edge
Central Cloud
AR Apps
APIs
UPF
NFVI Platform
Backend w/ Machine Learning & AI
Agg
Public Cloud
Augmented Reality at Hub/Agg Site and Central / Reg. DC
Interactive Gaming Services
Gaming Node
APIs
UPF
NFVI Platform
Catalog
Match Maker
Agg
Public Cloud
Distributed VR Gaming at Hub/Agg Site and C. / Reg. DC

Figure 12 Gaming Service

Note that Gaming Services may require GPUs in addition to CPUs on the x86 server nodes.

Interestingly, even if the games are played offline, the games get downloaded online and mere game download consumes a significant network (downstream) bandwidth. For example, Call of Duty: Black Ops 3 is 101GB, Grand Theft Auto V is 65GB. In contrast, an hour of 4K video on Netflix is about 7GB per hour, making a Call of Duty download equivalent to watching over 14 hours of 4K video!

It is worth highlighting that game downloads can affect ISP network utilization far worse than windows/iOS/macOS downloads, or even 4k movie download. Imagine the network contention that may arise if 20 users in a domain are downloading the latest game edition, while 20 users in the same domain are playing the game. In other words, caching such games (similar to that of on-demand videos) closer to the end customers could help to minimize the severe downstream bandwidth stress on rest of the network.

2.2.2 LiveTV

This one is an obvious one in which the Operators can move the content caches (live and on-demand) or translators closer to the subscribers in a distributed fashion in order to reduce the network load exponentially -

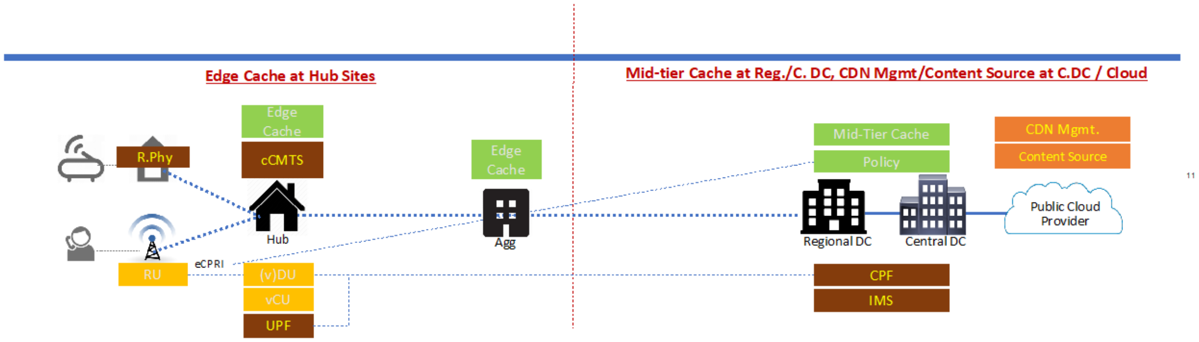


Figure 13 Managed Video/CDN

2.3 B2B Use-Cases

2.3.1 IoT & Public Cloud Hosting

One of the challenges of IoT is providing scalable connectivity to support a huge number of devices.

It is important to characterize IoT devices as – Heavy and Light. IoT Heavy devices are generally bandwidth intensive, require more power and sophisticated compute capabilities. Few examples are connected vehicles, autonomous control of large machinery, etc. IoT Light devices are highly constrained devices with minimal compute, memory, energy supply. Few examples are environmental sensors, under road monitors for smart parking, water meters etc. The IoT Light Devices require <200Kbps of data rate.

IoT Heavy devices such as the ones in factory automation etc. require lower latency treatment (say, few msec) and to do so, they would need to be hosted as close to the endpoints as possible for local processing.

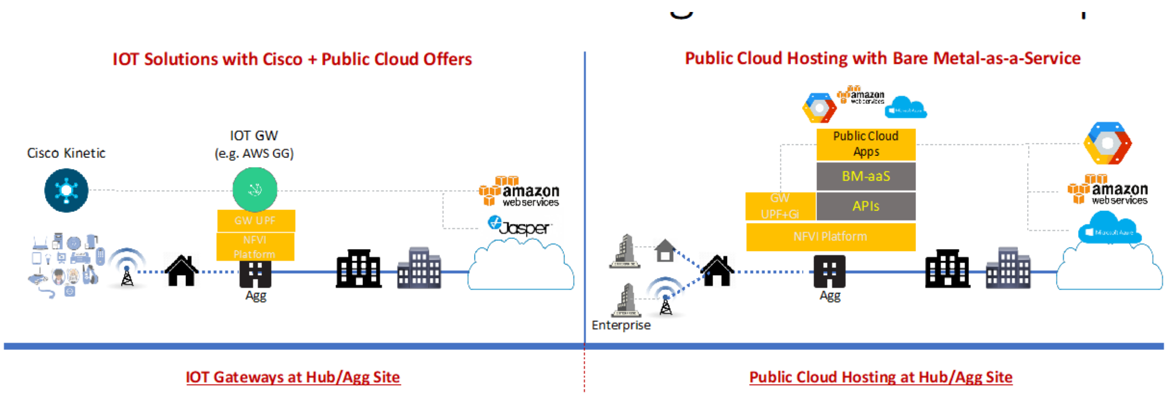


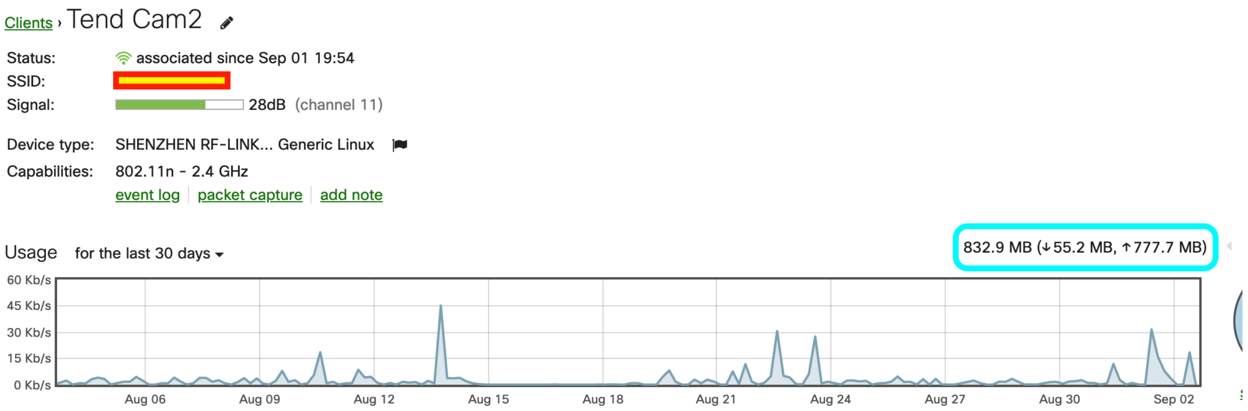
Figure 14 IOT and Public Cloud Hosting

2.3.2 3rd Party CDN

Operators can offer the transformed Hub sites to the 3rd party CDN operators and reduce the network bandwidth consumption while improving their customer experience.

2.3.3 Video Survilliance

The Video Surveillance / Cloud Monitoring services continue to gain traction in the context of Home/Business Security. The HD/UHD cameras could stress the network in upstream direction quite a bit. One HD camera could consume 1Mbps+, while streaming. The figure below illustrates the monthly consumption, mostly in the upstream direction -



Clients › Tend Cam2

Status: associated since Sep 01 19:54

SSID: [redacted]

Signal: 28dB (channel 11)

Device type: SHENZHEN RF-LINK... Generic Linux

Capabilities: 802.11n - 2.4 GHz

event log | packet capture | add note

Usage for the last 30 days

832.9 MB (↓55.2 MB, ↑777.7 MB)

Figure 15 Video Monitoring adds significant Upstream bandwidth Consumption

It is possible that such a service could be hosted in the Hub Sites, thereby significantly saving the network bandwidth in other domains.

2.3.4 Security

Security related workloads e.g. firewall, DPI, detection/prevention system etc. go hand in hand with the adjoining applications (of whichever category) in order to protect them, in addition to (cloud based) security related services that could be offered to the customers.

3 Cable Edge Compute – Architecture

It is important to highlight the possible Cable Edge Compute (CEC) locations in the perspective of E2E network architecture, as illustrated below, even though the focus is more on the Hub sites –

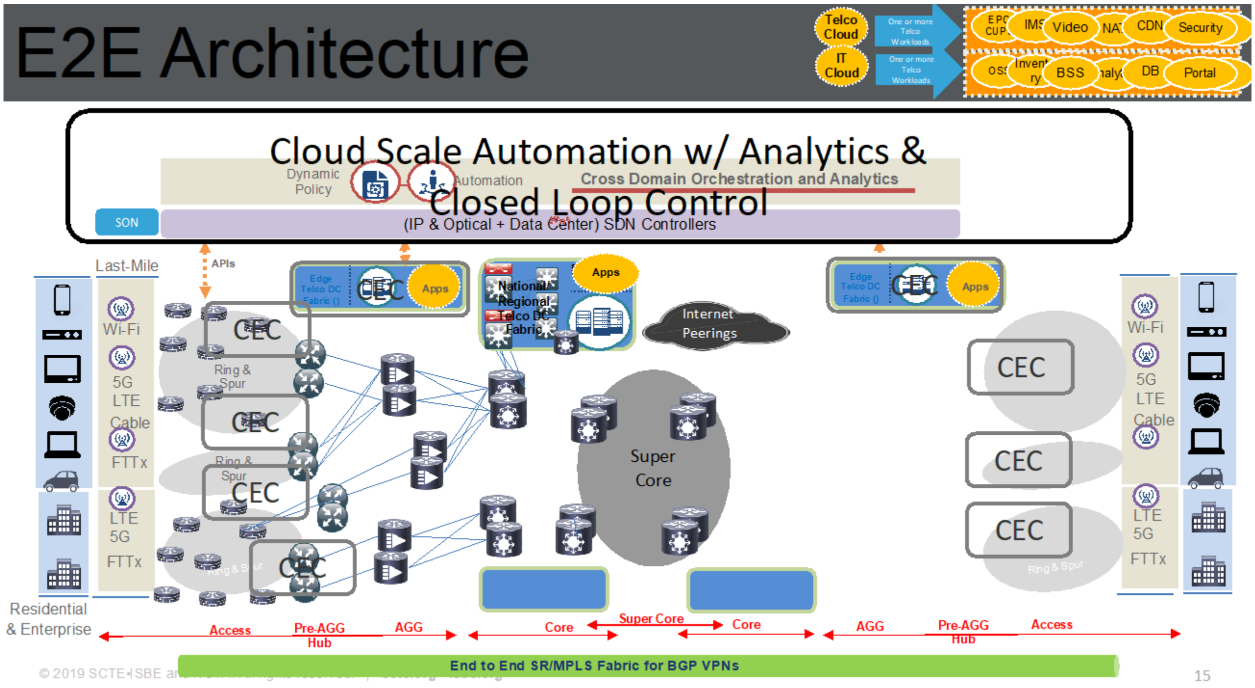


Figure 16 E2E Architecture Blueprint Showing Cable Edge Compute (CEC)

Given the wide variety of possible use-cases/applications (few are discussed in section 2) that could be hosted at the designated "Edge" locations, they require the presence of the same consistent cloud Platform, on top of which the application functions (mostly virtualized/containerized) can be deployed in a seamless manner (e.g. IaaS, PaaS, FaaS). This is possible with sufficient abstraction. A blueprint of such a platform is illustrated below -

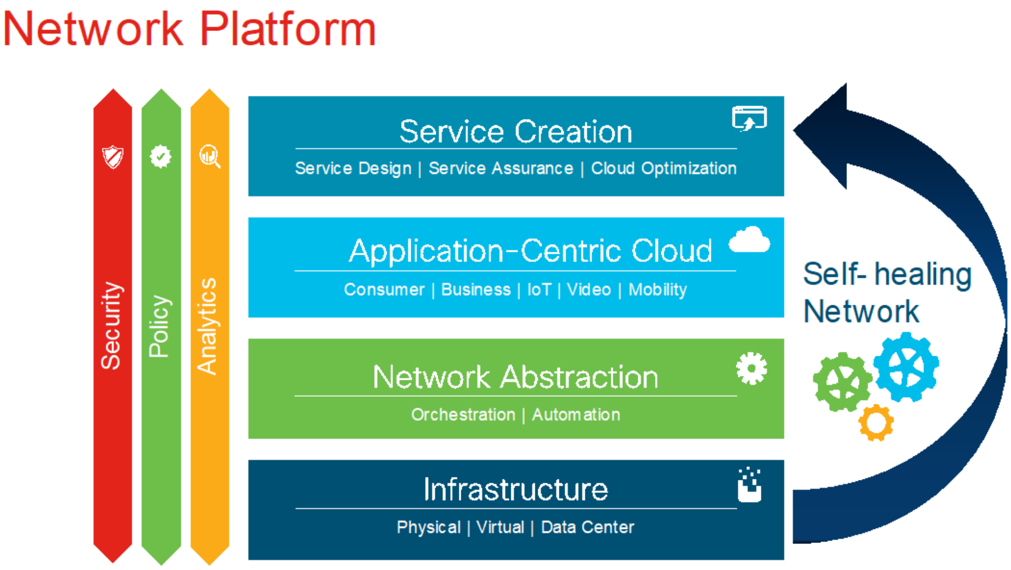

Network Platform
Security
Policy
Analytics
Service Creation
Service Design | Service Assurance | Cloud Optimization
Application-Centric Cloud
Consumer | Business | IoT | Video | Mobility
Network Abstraction
Orchestration | Automation
Infrastructure
Physical | Virtual | Data Center
Self- healing
Network

Figure 17 Network Platform - Abstraction is KEY

There are 4 building blocks that constitute "Edge" architecture to facilitate the Application Centric Cloud construct, each with certain unique properties –

1. Edge Infrastructure – Hardware
2. Edge Infrastructure – Software Platform (NFVI)
3. Network Fabric (SR)
4. Automation (SDN), Orchestration and Assurance

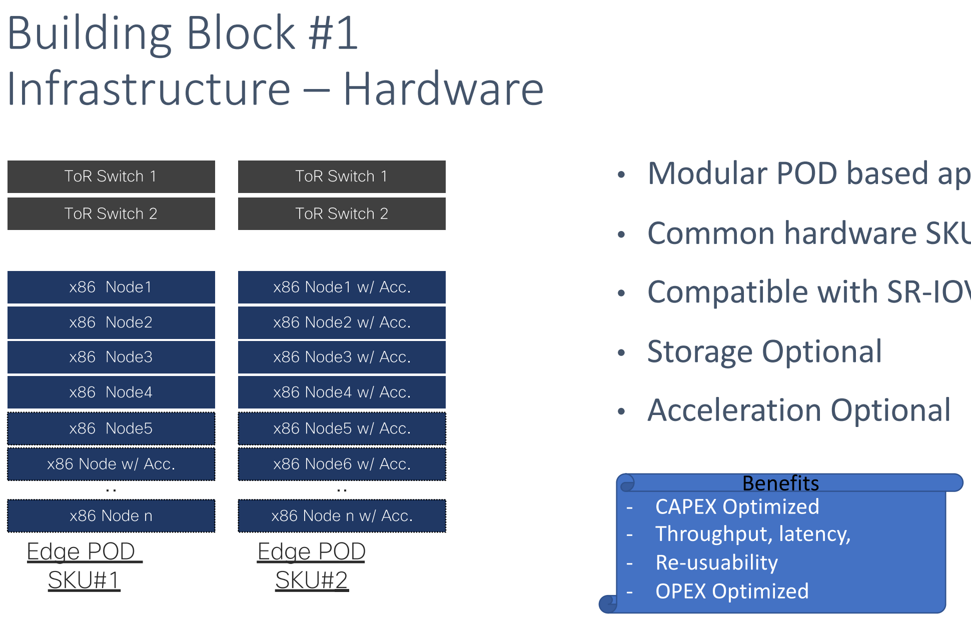
3.1 Infrastructure – Common Hardware

To morph MSO Hub Sites into the Application Centric Cloud that can accommodate wide variety of possible applications (mostly virtualized/containerized) for whatever time-period, the infrastructure should be flexible enough to not only provide higher throughput, lower latency, multi-tenancy, VM/Container workloads, security etc, but also enable scale-out in a policy-driven manner. Consistency is KEY for cost-effectiveness.

It is likely that many Hub sites may have smaller footprint (as compared to the traditional Data Centers) in terms of space, power, cooling, rack depth etc., whereas a few Hub Sites may have slightly larger footprint. Hence, the hardware infrastructure should be built while striking the balance among footprint optimization, performance and cost.

> In other words, whether a Hub site has space for only 2 racks or 6 racks or more, the hardware infrastructure should be expandable *without requiring any/much architectural changes*.

This flexible expandability requires a **modular POD based approach** that can provide consistency using a set of common hardware configurations (SKUs) for x86 server, storage and network. We refer to them as Edge POD, as illustrated in the figure below –



Building Block #1
Infrastructure – Hardware

ToR Switch 1
ToR Switch 2
x86 Node1
x86 Node2
x86 Node3
x86 Node4
x86 Node5
x86 Node w/ Acc.
..
x86 Node n
Edge POD SKU#1

ToR Switch 1
ToR Switch 2
x86 Node1 w/ Acc.
x86 Node2 w/ Acc.
x86 Node3 w/ Acc.
x86 Node4 w/ Acc.
x86 Node5 w/ Acc.
x86 Node6 w/ Acc.
..
x86 Node n w/ Acc.
Edge POD SKU#2

- Modular POD based approach
- Common hardware SKUs
- Compatible with SR-IOV, VPP etc.
- Storage Optional
- Acceleration Optional

Benefits
- CAPEX Optimized
- Throughput, latency,
- Re-usability
- OPEX Optimized

Figure 18 Hardware Infrastructure - Edge PoD SKUs

It is worth noting that recent innovations have helped to decouple storage and move the storage nodes to the locations such as conventional Data Centers that don't have similar constraints as Hub Sites do. This allows for simplicity and cost efficiencies, if/when application workloads don't require storage. This is further covered in the section 3.2.

The hardware infrastructure could also optionally include Accelerated Units (e.g. GPU, FPGA) on compute node(s), if required by any application workloads. For example, vRAN's vDU application workload may require eCPRI radio signal processing to be done on those x86 compute nodes housing specific Accelerated Units (e.g. FPGA), or Gaming's Application workload may require certain processing to be done on only those x86 compute nodes housing specific Accelerated Units (e.g. GPUs).

Additionally, It is very important for the hardware to be compatible with more than one virtual forwarding innovations whether user-space forwarder such as VPP/fd.io etc., or kernel-space forwarder such as OVS etc. or something that bypasses host kernel and user-space altogether such as SR-IOV etc. (as illustrated below) to ensure optimized forwarding behavior as expected by the application functions.

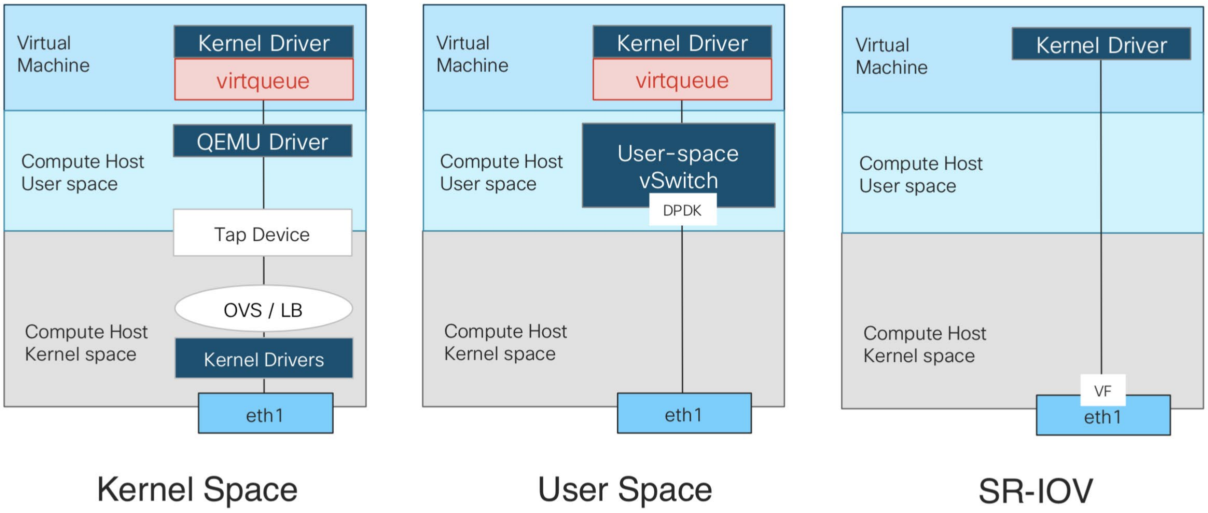


Figure 19 Virtual Forwarding Options

The Edge POD **consistency** ensures not only repurpose-ability, but also operational simplicity.

An example is shown for 4 NFV Infrastructure hardware SKUs that could be consistently deployed in the Hub sites depending on the requirements –

Table 2 Infrastructure Hardware - POD SKUs

	Hardware SKU1	Hardware SKU2	Hardware SKU3	Hardware SKU4
CPU / GPU	2 sockets [2.4GHz, 48 Cores, …]	2 sockets [2.4GHz, 48 Cores, …]	2 sockets [2.4GHz, 48 Cores, …]	4 sockets [2.4GHz, 48 Cores, …]
Memory/RAM	256GB	256GB	256GB	1TB
Memory/Disk	2x1TB SSD	2x1TB SSD	2x1TB SSD	2x1TB SSD
HW RAID	Yes	Yes	Yes	Yes
NIC	4 NICs: 2x10Gbps	4 NICs: 2x10Gbps	4 NICs: 2x10Gbps	4 NICs: 2x40Gbps
Acceleration	Yes/FPGA	- NA-	- NA -	- NA -

The smaller set of SKUs help to simplify operational and budgeting.

While the current Edge POD design assumes dedicated x86 server nodes, in the future, it may be possible to leverage the compute capacity of routers to be able to host containers or functions.

3.2 Infrastructure – Software (NFVI)

Software Infrastructure comprises of the Operating System (e.g. Linux/KVM) that facilitate reliable and deterministic "Virtualization" environment on top of the Hardware Infrastructure (i.e. Edge POD) that would be deployed during the Hub Site transformation, as well as Cloud Orchestration Platform that facilitate virtualized Infrastructure management.

Given the varying set of Hub sites constraints (mentioned in section 3.1), the Infrastructure Software stack must allow for maximizing the usage of Edge POD hardware resources for the designated Cable/Telco/IT applications workloads. This is KEY for superior cost efficiencies.

> This means a typical NFVI software stack inc. Virtual Infra Manager (VIM) must take as least overhead as possible to keep most of the resources available for application workloads related to the use-cases.
>
> Put it other way, VIM such as Openstack that may require at least 7 server hardware nodes (3 nodes for control, 3 nodes for storage and at least 1 node for management; see [6] and [7] for more details) either in the same rack or in adjacent racks in the site) for non-service purposes might NOT be acceptable. Consider a space/power constrained Hub site that can accommodate only 10 server nodes, then if the VIM software takes a large percentage (e.g., 70%) of the server nodes dedicated for cloud management, and not for hosting application workloads, then it would yield "Poor cost efficiencies".

The Infrastructure Software Stack should have the following attributes in order to minimize the Edge POD hardware footprint suitable for CEC cloud in Hub Sites –

1. Common Cloud Orchestration for VNF or CNF
2. Cloud Orchestration Control Nodes combined with Compute Nodes
3. Deterministic NFVI performance
4. Remote Storage with Optimization and Security
5. Remote Management & Monitoring

Each of these are now further detailed below.

3.2.1 Common Cloud Orchestration Platform

The infrastructure software stack should allow for orchestrating both VM based Application Functions aka Virtual Network Functions (VNF), Containers based Application Functions aka CNF as well as Physical Functions (e.g. Bare Metals) on the chosen Infrastructure Hardware i.e. x86 by appropriately leveraging Virtualized Infrastructure Managers (VIM) independent of where the functions are instantiated (e.g. Serverless or not). This allows for future compatibility.

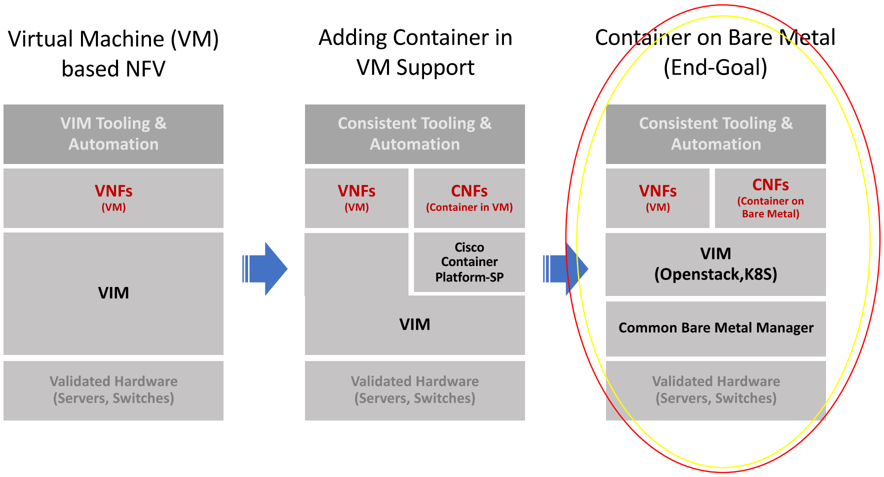

Virtual Machine (VM) based NFV
VIM Tooling & Automation
VNFs (VM)
VIM
Validated Hardware (Servers, Switches)
Adding Container in VM Support
Consistent Tooling & Automation
VNFs (VM)
CNFs (Container in VM)
Cisco Container Platform-SP
VIM
Validated Hardware (Servers, Switches)
Container on Bare Metal (End-Goal)
Consistent Tooling & Automation
VNFs (VM)
CNFs (Container on Bare Metal)
VIM (Openstack,K8S)
Common Bare Metal Manager
Validated Hardware (Servers, Switches)

Figure 20 Converged Cloud Platform

Containers may be hosted on Bare Metal or inside a VM. Few Examples of VIM are Openstack, K8S, vSphere etc.

VNF Manager (VNFM) and NFV Orchestrator (NFVO) would likely live in the centralized / regionalized data centers, though a lite version of VNFM could be hosted on each Edge PoD for VNF/CNF monitoring.

3.2.2 Orchestration Control Nodes combined with Compute Nodes

Instead of dedicating 3 Nodes for VIM/Cloud Controller purposes (e.g. Openstack Controller), they could be configured to run both VIM/OpenStack controller and compute functions. More specifically, the compute nodes run the host operating system (e.g. Linux/KVM) along with VIM, as well as the application workloads (IT/Cable/Telco Cloud).

The VIM/OpenStack controllers on the chosen 3 nodes continue to be in an active-active-active cluster configuration (with load sharing) for redundancy. Any additional nodes may be used as a pure compute node to scale the cloud on an as-needed basis. This is shown in the picture below –

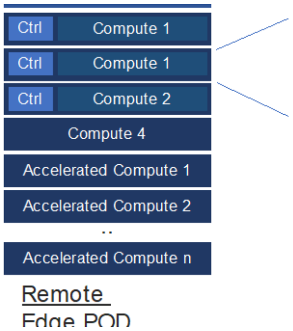


Figure 21 Collapsed VIM Controller and Compute Nodes

Additional compute nodes could potentially be reconfigured to have VIM controller function, if an existing compute+controller node failed.

3.2.3 Deterministic NFVI Performance

The collapsed control and compute functions are known to hamper the overall NFV performance. For example, latency may increase time to time.

To dramatically improve the overall NFV system performance in a deterministic manner, the software stack must be designed such that the VIM and VNF/CNF related task(s) run only on the specified CPU core(s), and the specified CPU core(s) are allowed to only run the chosen task(s), whether shared or not. This could be done by appropriately fencing the CPU cores from two distinct angles, as illustrated in the figure below -

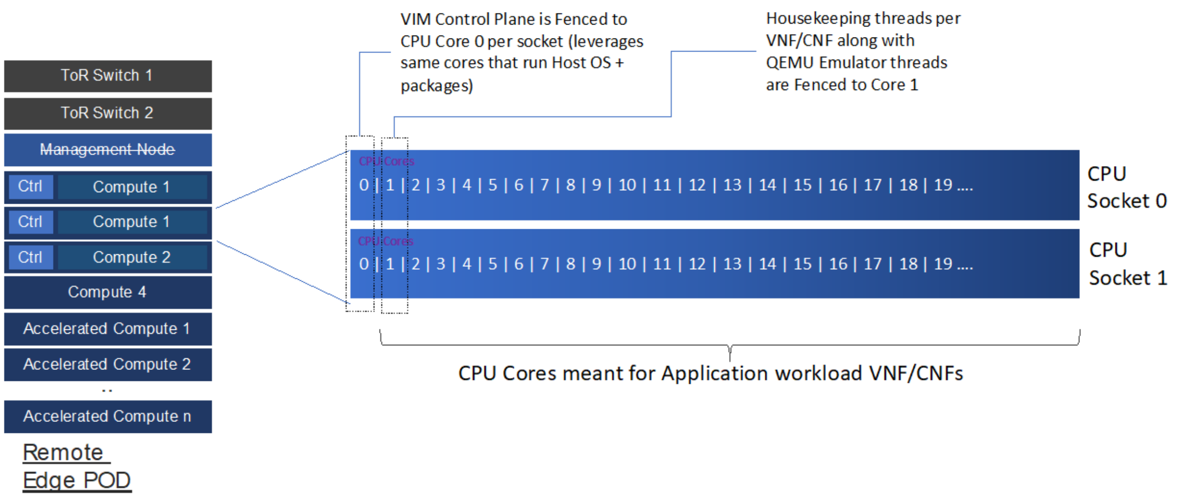


Figure 22 Deterministic NFVI Performance Logic

At least one of the CPU cores per socket can be used to receive interrupts and cannot be used for guest workloads. The logic employed is that the application VNFs can share the CPU core(s) on a particular socket with other VNFs' to run non-real-time tasks. This logic is followed whether or not hyper-threading is enabled.

In addition to specific CPU reservations, the software stack should allow for leveraging the virtual forwarding innovations such as SR-IOV, VPP etc. for optimized forwarding behavior as expected by the application functions. For ex, VNF1 may require SR-IOV, whereas CNF2 may require VPP based forwarding on the same x86 host.

3.2.4 Remote Storage with Optimization & Security

To avoid having to dedicate any nodes for Storage purposes (e.g. persistent storage, object storage and similar, unaffected by latency variations), Storage can be moved to remote location (e.g. centralized or regional Data Centers that don't have similar constraints) and accessed by the CEC Hub sites over the network.

> For example, the Ceph service for glance image services (including bulk transfer such as software image download) is no longer available locally inside the CEC Hub site. This is due to the assumption that Image-based (and object-based) storage is infrequently fetched and can be cached at the expense of latency (and would be costly if done at the edge). Thankfully, Latency isn't an issue, given the time it takes to download a VNF image is a lot more than WAN latency. Avoid using block-based storage because it may not be feasible to centralize it due to slow responsiveness (and would be costly at the edge).

This is illustrated in the figure below -

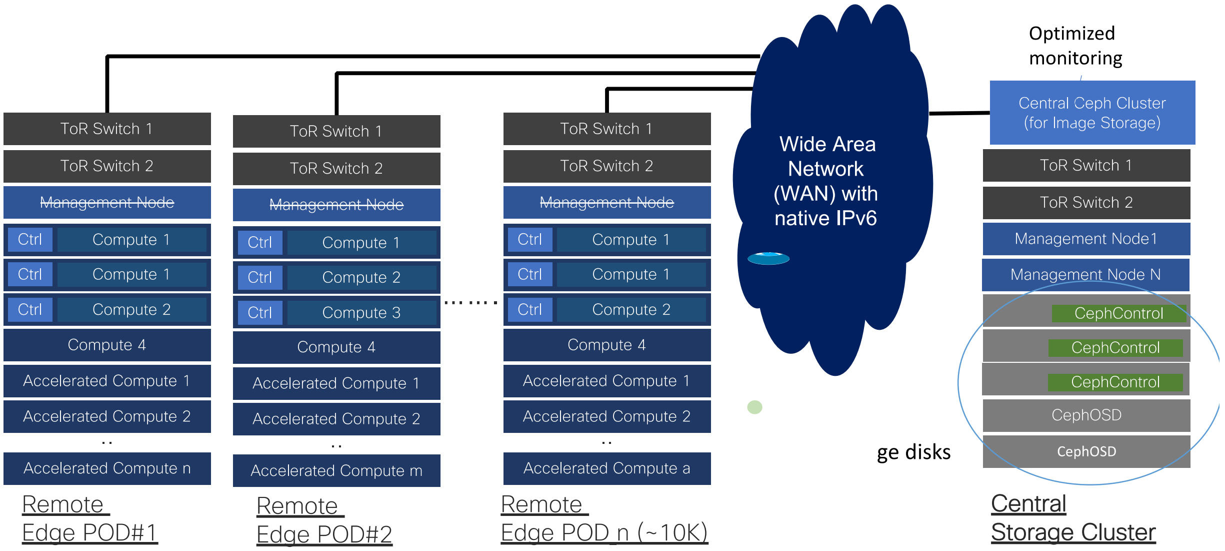


Figure 23 No more Local Storage for VIM

It is important to limit the potential bottleneck if multiple CEC edge PODs simultaneously interact with a single central Ceph cluster. Also, to ensure tight secured access between Edge PODs and Storage cluster, proper authentication and encryption (if necessary) of their communications (REST o TLS) is enforced.

3.2.5 Remote Management & Monitoring

Management Nodes perform number of important tasks – VIM deployment, monitoring, operations (e.g. node addition, replacement), version control, software version changes etc. To avoid having to dedicate any nodes for Management purposes in CEC Hub sites, management functions can be moved to remote location (e.g. centralized or regional Data Centers that don't have similar constraints) and accessed by the CEC Hub sites over the network. This could be quite challenging, but can be achieved, as illustrated in the figure below -

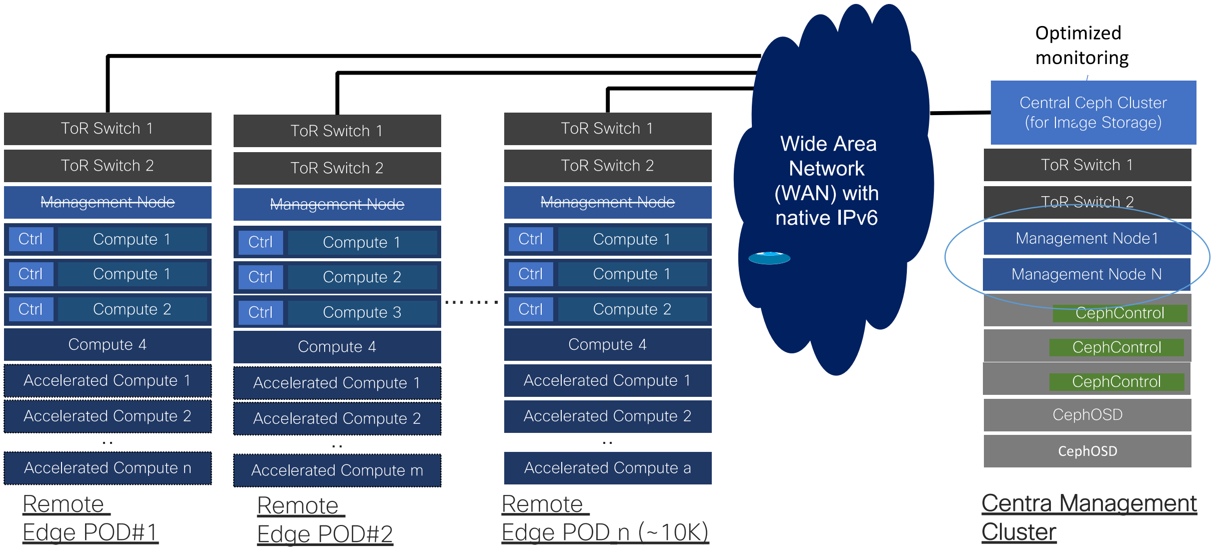


Figure 24 No more local Management Nodes

For example, the installation as well as monitoring of CEC nodes can be completely automated via a single intent file.

3.3 Network Transport Fabric

In order for the traffic to enter and exit the Hub Site, the Compute complex must appropriately connect to the network transport. There are 4 different options (4-tier, 3-tier, 2-tier, 1-tier) to design Network Fabric, as illustrated in the figure below –

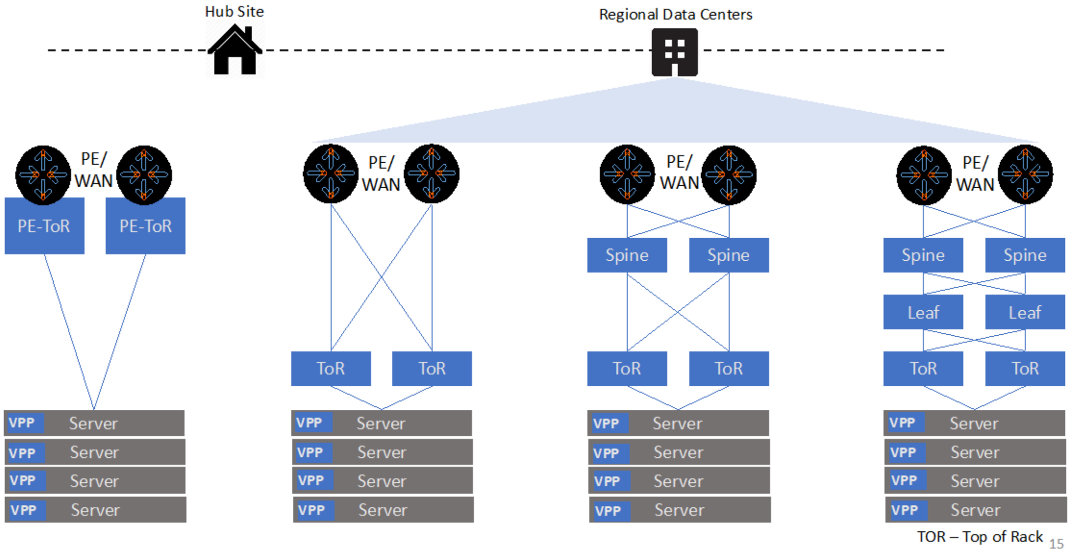


Figure 25 Network Transport Fabric

For the Hub Sites, 1-tier or 2-tier design is more appropriate to keep the footprint minimal while ensuring SLAs. Network Transport Fabric should have the following attributes -

- 10/40/100Gbps+ Ethernet Transport
- Any-to-Any Reachability
- E2E IPv6 with Segment Routing
- BGP based VPN
- Programmable WAN and DC Fabric
- IP+Optical WAN

3.4 Automation, Orchestration and Assurance

The architectural framework must have two foundational elements to ensure the Application Centric Cloud transformation of Hub Sites becomes viable.

Firstly, the Hardware Infrastructure (e.g. x86 nodes, FPGAs etc.), Software Infrastructure (e.g. VIM) and Network transport (e.g. TORs) must be managed with stringent automation and orchestration for its entire life-cycle. Zero-touch instantiation of Infra management is also required. Consider FPGA installation or firmware changes should be handled in an automated manner.

Secondly, each service instantiated at the Edge may comprise of one or more number of application functions that would need to be instantiated, provisioned and configured in the right sequence with the right service level agreements (SLAs) in a zero-touch manner. Assurance would be foundational to monitor the SLAs with closed-loop.

> In order to appropriately enforce the policies (such as access control, inspection, prioritization, optimization etc.) on the traffic passing through any one or more of the (service) functions deployed at the edge, it is important to maintain the sequence in which the traffic must pass through those functions. This is referred to as service-chaining.

For example, Service chaining can help to regulate the traffic flow for service A to pass through application functions 1, 2 and 4 in a sequence, whereas traffic flow for service B to pass through functions 2 and 5.

Application Functions can be inserted or deleted into the flow by modifying the chain, as controlled by the orchestrator, as explained in [5].

The architectural framework is based on the concept of hierarchical management and orchestration, and consists of domain-level orchestration systems, where the domain would correspond to Network Transport, Edge POD, Data Centers etc. The domain-level orchestration systems would cater to domain specific management. For example, Edge POD in CEC Hub site would be managed via the MANO stack (VIM, VNFM, NFVO etc.). These orchestration systems are glued together in a modular architecture framework with an end-to-end service orchestration that would interact northbound with OSS and BSS systems, which offer a comprehensive set of service instantiation, service lifecycle management and operational workflows.

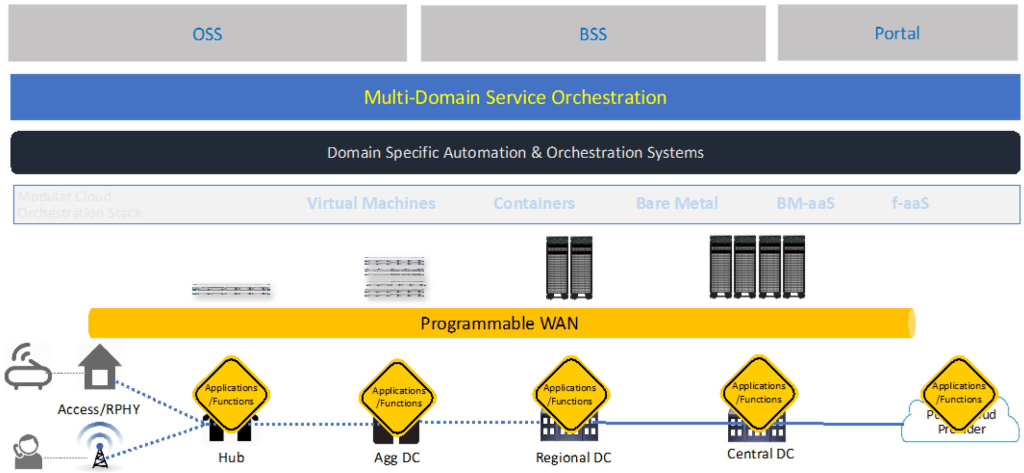


Figure 26 Automation, Orchestration and Assurance

The architectural framework should have the following attributes –

1. Infrastructure Software - lifecycle management with CI/CD pipeline
2. Automated Service Validation
3. Self-Service Portal
4. Closed Loop Assurance
5. Single Pane of Glass

4 Summary

Cable Operators can transform their Hub sites into Next Gen Application-Centric Cloud sites (similar to how Telecom Operators could transform their Central Offices (COs)) and take advantage of next swath of revenue generating services that can be hosted closer to the subscribers in order to ensure superior customer experience.

They should ensure that the Hub transformation approach adheres to key architectural building blocks – IH (Infrastructure Hardware), IS (Infrastructure Software Stack), NF (Network Fabric) and AOA (Assurance, Orchestration, Automation) along with key attributes.

4.1 Opportunities

Many Operators have 1000s of Hub Sites deployed across the country and they are not currently fully utilized to host applications pertaining to B2B or B2C services and offer superior customer experience.

If these Hub sites can be transformed into hosting qualified applications such as IoT, bandwidth hungry applications e.g. CDN/cache edge, lower latency applications e.g. gaming, then these applications (and resulting services) would get a lot closer to the customer.

4.2 Challenges

There are number of challenges in sufficiently utilizing Edge Computing in MSO environment. Few are captured below –

1. IT Cloud and Cable Cloud separation – Diverse application types (in IT and Cable/Telco space) would demand different SLAs (some may be more stringent than the others), hence, a common cloud could be deemed difficult. However, maintaining separate Clouds for Cable/Telco and IT applications / workloads could result in many clouds distributed on the network and less efficient usage of infrastructure.

   In other words, one hand, the desire to push for high performant, efficient use of Edge resources have to be balanced with a cookie cutter way of deploying a common cloud.

2. People/Skills/Culture – People get less excited about making drastic changes or get entrenched with identifying tons of issues that will make things not work. Sometimes, it is because of lack of appropriate skillsets, sometimes, it is because of the organization culture. This can single-handedly make or break the Hub transformation. Also, Network Centric teams tend to downplay the IT Centric teams and vice versa.

3. Stringent Assurance Needs – While one of the most important, Assurance topic doesn’t surface until later in the conversation. That could hamper the overall efficacy. Programmatic connectivity to ensure the “Service Assurance” would be difficult if Cloud and Network together are managed with intense Automation.

4. Hub Site Physical constraints – Power availability (AC vs DC) and Space availability (e.g. raised floor, rack size etc.) to install x86 platforms would become important. Hub sites conventionally

may not be built according to the cloud needs. Also, the failure rate with x86/storage nodes may be higher than the routing/switching equipments, so frequent swapping may require people (if not robots) with suitable skills.

5. Lawful Interception (LI) – It may not be feasible to comply with LI if applications are designed and deployed where there is no way to place taps.

6. Multi-vendor and 3rd Party/Partner workloads – B2C vs B2B business approach may drive application workloads from different vendors / partners with their prerequisites that may break the feasibility. For ex, vendor-proprietary virtualization manager co-located with the workload instance, or VNF/CNF not fully compliant with APIs etc. If they are not able to take advantage of the APIs for Easy Consumption of the Platform, then it could jeopardize the overall efficacy, and may even derail the Edge Computing insertion.

7. Challenging Software Lifecycle Management – In virtualization paradigm, the agility is key. So, the number of software changes is a lot more than typical operations expect. If lacking 100% automation, then it would require human intervention.

8. Cloud-Native Applications – Not all applications are fully containerized yet. This means they would not be able to take advantage of Edge Computing resources that tend to assume disaggregated applications for utmost efficiency.

9. Too many Standards and Industry Groups – This makes Cable MSOs choosing particular group(s) a bit difficult and also hampers the overall progress. For ex, 3GPP, MEC WG, CORD, TIP, OpenFOG, OPNFV, ONAP, etc.

4.3 Recommendations

This paper highlights several recommendations that Cable Operators could consider as they look to transform their Hub Sites with a focus on Application-Centric -

1. Develop a roadmap for what (IT and/or Telco) applications could run on the common CEC platform initially vs later on. Carefully Analyze and Select the Application type that would yield better bang for the buck (e.g. bandwidth savings, latency reduction etc.) on a common CEC platform.
2. Train and retrain the workforce and push them to drive the changes (not just accept the changes). Take risk, fail fast approach should be advocated.
3. Mandate "Assurance-first" approach while evaluating the CEC offerings, even if the Assurance systems are different for Applications, CEC infra and Network. They must be programmatically linked to ensure AI/ML powered Operations.
4. Choose Hub sites that have limited or no physical constraints early on. Ensure CEC PODs with highly optimized hardware and software stack in fewer variations (e.g. 8 server nodes, 12 server nodes, 24 server nodes) that can be well standardized for different Hub sites.
5. Applications that run on CEC must be first cleared for the LI compliance.
6. Specify a common set of requirements so as to accommodate application workloads from different vendors / partners
7. Develop 100% automated software lifecycle management with CI/CD pipeline and ensure that it works in as many failure scenarios as possible.

8. Push the applications owners to provide decomposed and/or disaggregated workloads that can run on Kubernetes etc.
9. Choose fewer standards groups that are not limited to mobility only. For ex, Edge Computing Consortium.

Abbreviations

CEC	Cable Edge Computing
VNF	Virtual Network Functions
CNF	Cloud Native Functions
HFC	hybrid fiber-coax
MEC	Multi-Access Edge Computing
B2B or B2C	Business to Business, or Business to Consumer
ISBE	International Society of Broadband Experts
SCTE	Society of Cable Telecommunications Engineers
CUPS	Control Plane and User Plane Function
CPF	Control Plane Function
UPF	User Plane Function
BNG	Broadband Network Gateway
CCAP	Converged Cable Access Platform
cCMTS	Containerized Cable Modem Termination System
VIM	Virtual Infrastructure Manager
MANO	Management and Network Orchestration

Bibliography & References

1 Managing the 5G Telco Cloud – https://blogs.cisco.com/sp/managing-5g-at-the-data-center-level-first-steps

2 Reimagining the End-to-End Mobile Network in the 5G Era - https://www.cisco.com/c/en/us/solutions/service-provider/mobile-internet/reimagining-mobile-network.html

3 Real-World 4G/5G Use Cases https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/0522-mobility-ckn.pdf

4 Open vRAN Ecosystem https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/0920-mobility-ckn.pdf

5 Using SDN Controller in Telco DC https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/application-centric-infrastructure/white-paper-c11-740717.pdf

6 Openstack Example Architecture https://docs.openstack.org/install-guide/overview.html#example-architecture

7 https://www.mirantis.com/blog/making-openstack-production-ready-kubernetes-openstack-salt-part-1/

8 Towards MEC Edge Compute https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/0417-DC-CKN-PDF.pdf

9 Latency matters more than Bandwidth https://www.igvita.com/2012/07/19/latency-the-new-web-performance-bottleneck/