DOCUMENT

234.10.01

Edge Computing and Small Cell Networks

June 2020

www.smallcellforum.org

Small Cell Forum develops the technical and commercial enablers to accelerate small cell adoption and drive wide-scale densification.

Broad roll-out of small cells will make high-grade mobile connectivity accessible and affordable for industries, enterprises and for rural and urban communities. That, in turn, will drive new business opportunities for a widening ecosystem of service providers.

Those service providers are central to our work program. Our operator members establish the requirements that drive the activities and outputs of our technical groups.

We have driven the standardization of key elements of small cell technology including Iuh, FAPI, nFAPI, SON, services APIs, TR-069 evolution and the enhancement of the X2 interface. These specifications enable an open, multivendor platform and lower barriers to densification for all stakeholders.

Today our members are driving solutions that include:

- 5G Components, Products, Networks
- Dis-aggregated 5G Small Cells
- Planning, Management and Automation
- 5G regulation & safety
- Neutral Hosts & Multi-operator
- Private and Public Network coexistence
- Edge compute with Small Cell Blueprint
- End to end orchestration

The Small Cell Forum Release Program has now established business cases and market drivers for all the main use cases, clarifying market needs and addressing barriers to deployment for residential, enterprise, rural & remote, and urban small cells. It has also established initiatives relating to both public and private (MNO) coordination. The Small Cell Forum Release Program website can be found here: www.scf.io

If you would like more information about Small Cell Forum or would like to be included on our mailing list, please contact:

Email info@smallcellforum.org

Post Small Cell Forum, PO Box 23, GL11 5WA UK

Member Services memberservices@smallcellforum.org

Credits

Contributors	Company
Prabhakar Chitrapu (Co-Editor)	AT&T
Santi Ibarz (Co-Editor)	Druid
K. Satish (Co-Editor)	Nokia
Zeki Keles	American Tower
Mark Reudink	Crown Castle
Hiren Surti	Crown Castle
Mark Gullett	Intel
Nick Ross	Intel
Srinivasa Addepalli	Intel
Andy Odgers	Quortus
Ravi Sinha	Reliance Jio
Glenn Morrow	Samsung
Hunter Kanai	Senko

Scope

This document provides a comprehensive overview of edge computing. In particular, it looks at how edge computing impacts the future of small cell networks. It also pinpoints existing gaps in the evolution and application of edge computing. Much still needs to be done in such areas as standardization and industry practices. How can these gaps be filled in order to allow edge computing and small cell networks to be deployed quickly and at scale?

The overview includes use cases, the architecture framework and implementation approaches. It discusses edge computing solutions based not only 4G and 5G cellular access but also on Wi-Fi access.

The target audience for this document includes technology developers, technology strategists and engineering managers as well as marketing and business professionals in related operator, infrastructure and application developer companies.

The paper's editors are Prabhakar Chitrapu, AT&T; Santi Ibarz, Druid; and Satish Kanugovi, Nokia. The paper's contributors are Srinivasa Addepalli, Intel; Mark Gullett, Intel; Hunter Kanai, Senko; Zeki Keles, American Tower; Glenn Morrow, Samsung; Andy Odgers, Quortus; Mark Reudink, Crown Castle; Nick Ross, Intel; Ravi Sinha, Reliance Jio; Hiren Surti, Crown Castle.

Executive summary

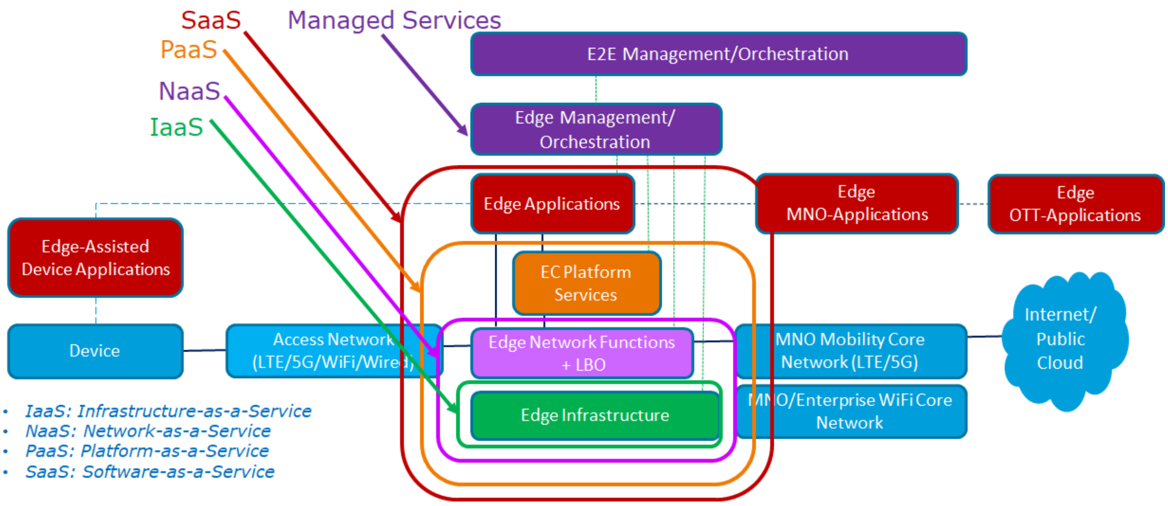
Edge compute is one of the most important developments for mobile operators (MNOs) and other service providers to consider in their next generation network strategies. There are significant opportunities to harness edge compute and combine localized data center (processing, storage and analytics) capabilities with connectivity.

This paper begins by outlining a range of benefits from improving quality of experience for existing services (e.g. delivering video close to the user); to enabling new services that require very low latency or precise contextual awareness, compute offload from devices, edge analytics and others. It also describes how edge computing can help various industry segments, such as automation and Industry 4.0, worksites, mission critical services, enterprises and public safety.

Our approach acknowledges that edge computing means different things to different audiences. To this end, we set out a comprehensive, end-to-end framework organized in five layers:

- edge computing infrastructure,
- edge network functions,
- edge platform services,
- edge applications and
- managed services.

Each of these layers can be delivered by different players in the edge ecosystem, which in turn implies a comprehensive edge services framework.



Functional framework for edge computing

The paper then studies various deployment scenarios for edge computing. In particular, it makes a distinction between network-based edge and premise-based edge use cases. Network-based use cases are typically relevant to an MNO network. Premises-based use cases tend to be relevant to enterprise, venue and campus environments.

We go on to describe why premises-based edge use cases are best understood and addressed as three different sub-use cases:

- fully private cellular networks (PCNs);
- PCNs that have a roaming relationship with MNO networks; *and*
- PCNS integrated with MNO networks.

These use cases are discussed in detail, highlighting the requirements and characteristics that are unique to each environment. This work is complemented by recent SCF work on private network market drivers. [1]

The centerpiece of the document considers the core synergies between edge computing (EC) and small cell networks (SCNs). It highlights how those synergies are present across multiple domains – technical, deployment, product and vendor ecosystems. We also make specific recommendations as to how EC and SCNs can enhance each other, at both technical and business levels.

The paper then continues into detailed exposition and analysis of each of the five layers mentioned above. Each section concludes with a set of recommendations, which identify either technical/business gaps or best practices/next steps for the industry. Detailed recommendations for each section appear on pages 21, 29, 40, 46, 52. The high-level recommendations are now listed here:

- Edge network recommendations: For edge computing solutions, the 'edge network' has to work in concert with the 'core network' via open interfaces and APIs to enable true multi-vendor ecosystem. We conclude that current specifications from various organizations are not completely adequate and need enhancements.
- Edge platform services and applications recommendations: Open and consistent APIs across multiple organizations need evolve and align to enable a broad ecosystem of edge platform and edge applications.
- EC platform solution recommendations: Blueprints/reference-designs/solutions for open-source edge computing platforms are urgently required for rapid growth of the EC ecosystem.
- EC and small cell recommendations: Small cell networks and edge computing platforms can leverage each other for advanced solutions benefiting at multiple levels: shared virtualized implementations leading to cost efficiencies; integrated network functions & mutually beneficial analytics (radio environment, RAN characteristics, location etc) leading to advanced functionalities to the edge computing platforms.
- EC infrastructure recommendations: For optimal edge computing infrastructure, it is recommended that the guidelines from TIA, BICSI and OCP be followed.

It is expected that these and other recommendations detailed will help the industry forge best practices. Through its work program and in collaboration with other SDOs, SCF will drive future activities to fill the gaps identified.

[1] Private Cellular Networks with Small Cells, April 2020, http://scf.io/doc/235

Contents

Figures

Tables

1. Introduction

Edge computing and 5G are together enabling technologies that have the potential to unleash a variety of advanced use cases resulting in new user experiences and new business opportunities. However, the 'edge' is not in one place. It can, in fact, be at one of several possible physical locations, each enabling a different set of edge applications. The edge can, for example, be at the customer premises, within the small cell or radio access network, or within regional aggregation networks.

Edge computing solutions are complex. They involve edge computing platforms, edge network functions, edge platform services, edge applications and edge orchestration and management solutions. Each of these functions can potentially involve different stakeholders, making for intricate business and technical interdependencies.

This document aims to provide a simple but comprehensive framework that ties all these components together. It will provide detailed descriptions of various components and identify gaps in various aspects of the edge computing ecosystem, such as standardization and operational approaches.

The document also focusses on various implementation aspects related to edge computing and their impact on small cell networks – specifically how small cell networks can be enhanced to support edge computing solutions and how edge computing platforms can be used to realize small cell network functions.

Of course, edge computing and small cells already enjoy a number of synergies. These include:

Technical

- Both edge computing and small cells address geographic areas with a small footprint
- Edge network functions include centralized SCN/RAN functions (as well as distributed CN functions)
- Edge computing applications make use of small cell analytics as well as edge computing services
- SCN/RAN and EC may be managed using the same management framework.

Deployment

- Both enterprise customers and mobile network operators may deploy small cell networks as well as edge computing platforms together in order to offer combined coverage/capacity and services solutions
- Fiber infrastructure may be put in place to jointly support 5G SC as well as EC.

Products

- Small cell networks can be designed to include edge computing functions
- Edge computing platforms can implement (centralized) parts of (disaggregated) small cell networks.

The vendor ecosystem

- Some vendors are beginning to offer solutions and services related to both small cells and edge computing. These vendors include traditional tower or cellular infrastructure companies and neutral host vendors as well as systems integrators.

2. Edge Computing – the Drivers

Mobile network data traffic is growing, driven by the vast number of end-user smart devices and IoT. The switch to data-centric services has also led to an increased demand for personalized end-user services and an improved user experience.

Mobile services have traditionally been targeted at mass-market consumers. Now, however, operators are starting to address the business enterprise market. This market requires enhanced and secure services. The result has been greater convergence between functions normally assigned to IT and those assigned to the operator.

Edge computing addresses these issues by enabling the use of what could be described as applications at the edge of the network – often within the IT domain – to provide flexible and configurable services. Applications located at the edge offer a number of benefits compared with traditional deployments. They include:

- Lower latency
- Localized, and, by extension, personalized or contextualized applications and services
- Improved quality of experience (QoE)
- Improved network operations

Use cases needing or benefiting from low latency include vehicle to everything (V2X), autonomous vehicles, immersive AR/VR, industrial control, real-time local video distribution and gaming.

Localization enables contextualization and personalization. Potential use cases here include local content distribution, network analytics and private LTE/5G networks.

Edge-based implementation can also be used to improve user QoE of traditional services such as video delivery. These services can be enhanced by such methods as acceleration enabled by edge-based analytics. As we have noted, low latency also improves QoE for applications such as gaming and augmented reality/virtual reality (AR/VR).

Edge-based implementation can also improve network operations in areas such as edge analytics, backhaul usage reduction and improved edge security.

Edge computing also provides business benefits to the operator, enabling the delivery of value-added services and, potentially, additional revenue streams. Processing traffic, or delivering services at the edge of the network, can also help to reduce backhaul bandwidth and improve network operations, reducing OPEX.

The following Table 2–1 provides a sample of potential business/deployment use cases:

Industry segment	Industry examples	Edge computing application
Automation and industry 4.0	Factory floor Logistics and warehousing	Edge network for internal communications. Automation. IoT edge applications (video, logistics, warehousing, automation, etc)
Worksite industries	Mining, agriculture, remote oil/gas	Extension of wireless IoT networks to remote/temporary locations. Removal of black spots in coverage Edge-enabled video, local communications, automation, alarms
Mission-critical services	Electricity, power plants	Critical communications. High availability Edge-enabled local communications, automation, alarms
Enterprise/venue services	Airports, stadiums, hospitals, ports	Private network for employees. Data privacy and public network for visitors. Edge-enabled local communications, videos, local applications, automation
Public safety	Patrol, first responders	Private network for emergency situations. Portable communications Edge-enabled communications, video, body camera feeds, drone video feeds

Table 2–1 Edge computing applications

Edge computing also provides architectural solutions that can address enterprise requirements, particularly around security. Such solutions might include allowing end-user data to remain on site and ensuring that sensor/IoT traffic is only accessible within the enterprise premises; or enabling end-users to access corporate LAN resources from the wireless network.

Finally, edge computing enables the transformation to 5G and related advances in communications technology including low latency communications, ultra-reliable communications and massively dense IoT. Edge computing also complements the 5G-led move towards network function virtualization (NFV) and network slicing.

For such use cases, a far more distributed edge is required, and small cells have a vital role to play. Securing locations for edge nodes will come with the same challenges as for small cells – challenges SCF has been instrumental in addressing.

There is strong correlation between operators which have advanced densification programs, and those with high interest in edge compute. These operators will help to drive edge compute forward and accelerate its progress in many industries.

The high level of intent to deploy edge compute in tandem with small cells was highlighted in a major operator survey which SCF conducted at the start of 2020. This found that almost 75% new indoor small cell deployments will be co-located with edge and/or private EPC.

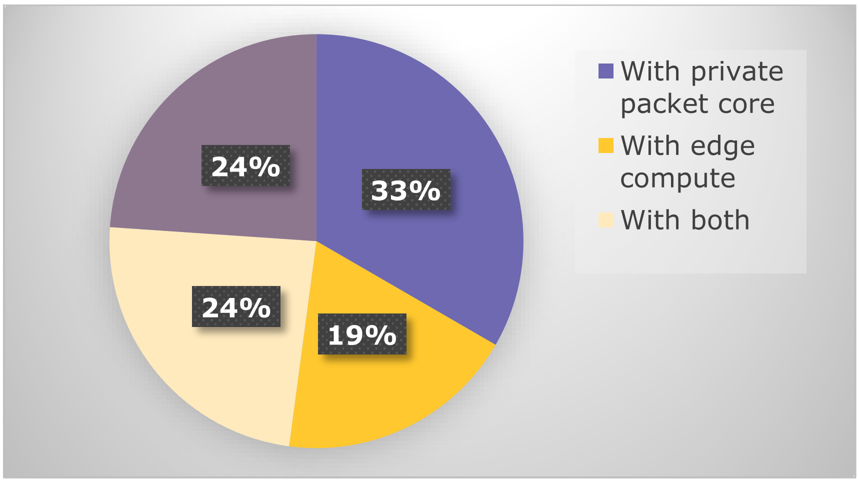


Figure 2–1 **Percentage of operators with small cell deployment plans, which also plan to deploy edge compute and/or private EPC by 2025 (SCF/Rethink operator survey 2020)**

It is not just traditional MNOs which will engage in parallel deployment of dense small cells and edge compute. New service providers are emerging which will provide neutral host platforms, especially for industrial or enterprise use cases; or which will build their own small cell networks, targeting a particular location (e.g. a city) or a particular industry (e.g. manufacturers). Our operator survey reveals a range of new business models by 2025 including edge as a service, direct services to subscribers and as support for other service providers.

The combination of small cells, shared spectrum (e.g. CBRS in the USA) and edge compute will enable new operators to support differentiated services for particular market requirements.

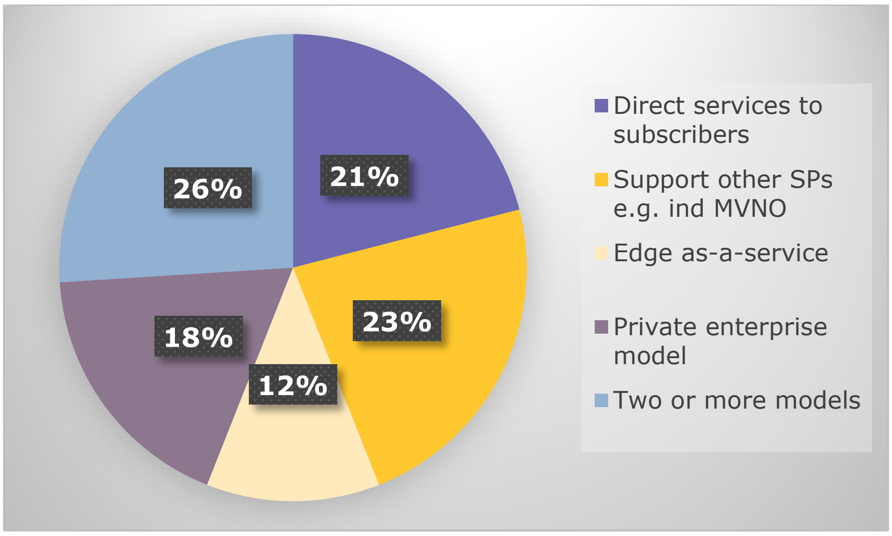


Figure 2–2 **Small cell and edge business models by 2025 (SCF/Rethink operator survey 2020)**

3. Edge Computing Framework

3.1 Defining the Edge

From a deployment perspective, edge computing solutions can be classified into two categories:

- **Network-based** – for example, where the MNO deploys edge computing assets as part of the overall MNO network infrastructure
- **Premises-based** – for example where the enterprise or venue owner associated with the premises deploys EC assets as part of its IT infrastructure

In both cases, the notion of the edge is flexible and is determined by the required closeness of the user to the EC assets. The various options are depicted in Figure 3–1 below.

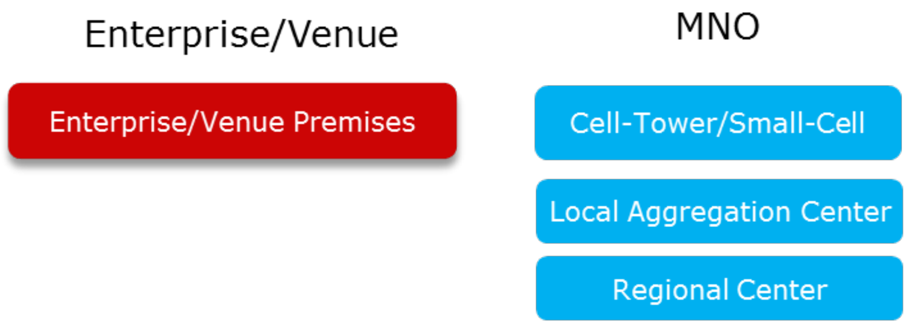
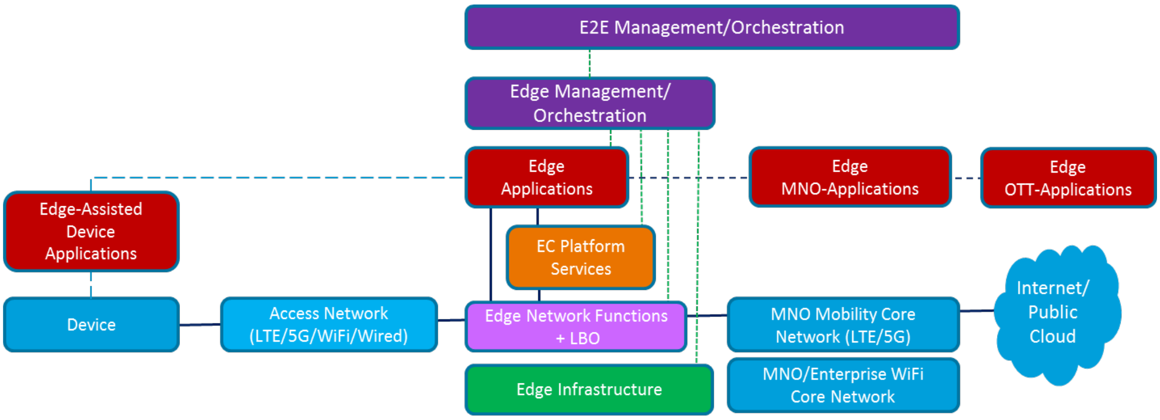


Figure 3–1 Edge locations

In the context of an MNO network, the edge location that is closest to the user would be a cell tower or a small cell location. Further into the network – and farther away from the end-user – would be a local aggregation center or a regional center (different operators may use different names for these). Latency is greater and location less precise as the edge moves farther away from the end-user.

In the context of an enterprises or venue (such as a stadium or event), the edge may be on the enterprise premises as part of the IT infrastructure. However, placing edge computing assets on the enterprise premises may not be cost-effective for small-to-medium enterprises. These may be better served by a network-based edge solution.

3.2 Functional Framework



LBO = local break out

Figure 3–2 Functional framework for edge computing

The above Figure 3–2 proposes a comprehensive framework for edge computing.

edge functions can be separated into five layers:

- Edge infrastructure
- Edge network functions
- Edge platform services
- Edge applications
- Edge management

Figure 3–3 below shows how the EC framework enables multiple service offerings.

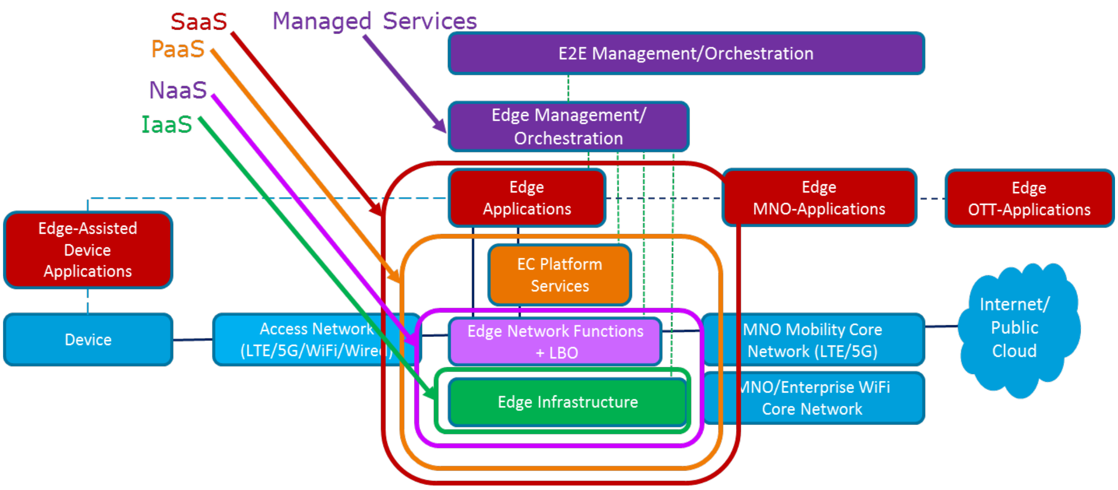


Figure 3–3 Edge computing-based service offerings

What does this mean in practice? A business entity, for example, may offer only edge infrastructure, making it an infrastructure-as-a-service (IaaS) provider. If a combination of edge infrastructure and edge network functions (including LBO) is offered as a service, the resulting offerings can be described as network-as-a-service (NaaS) offerings. Including edge platform services in this context enables platform-as-a-service (PaaS) provision. Including edge applications makes a business a software-as-a-service (SaaS) provider. Management and orchestration services may also be provided by certain business entities.

Edge computing makes multiple business opportunities possible. However, the overall framework highlights the technical and business complexities inherent in edge computing systems.

In the following sections, each of the components of the EC functional framework are briefly described. More detail can be found in Section 5 (EC infrastructure), Section 6 (Edge network) and Section 7 (EC platform and applications).

3.2.1 Edge Infrastructure

By edge infrastructure, we mean the hardware and software assets on which the edge network functions and on which edge platform services (and possibly edge applications) are implemented.

Although it is, in principle, possible to use dedicated hardware and software for edge infrastructure, the current trend is to use virtualized infrastructure. This consists of a generic commercial off-the-shelf (COTS) hardware platform, along with virtualization software layers or structures, such as hypervisors, virtual machines (VMs) and containers.

The infrastructure may also include hardware accelerators to implement compute-intensive functions such as encryption and decryption. The infrastructure may also include storage space. This can be leveraged by edge solutions for content caching or distribution applications.

Implementation details may differ. If the infrastructure is on enterprise or venue premises, there may be specific constraints or requirements from the enterprise or venue owner. If it is located on an operator's network, the MNO may dictate the nature of the edge infrastructure, so that it is aligned with the MNO's existing and evolving overall network infrastructure architecture strategies.

3.2.2 Edge Network Functions

Edge network functions are essentially connectivity functions. Let's look at MNO edge computing solutions first. Traditionally (i.e., for 4G and earlier generations, as defined by 3GPP), the MNO network consisted of the radio access network and core network (RAN and CN). These were quite separate and distinct in terms of their functions and implementations. However, with the onset of 5G and the advanced architectures of 5G networks, RAN functions and CN functions are beginning to overlap – notably at the edge, a position between the traditional RAN and CN (see Figure 3–4 below).

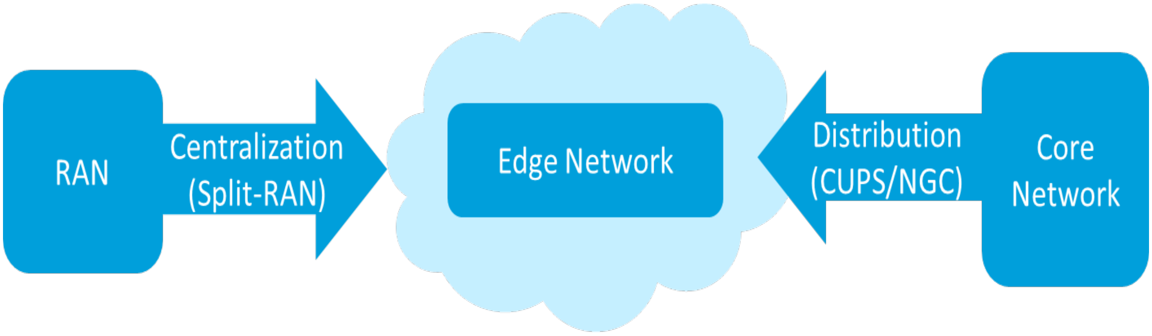


Figure 3–4 Concept of edge network

Traditionally, the RAN architecture consisted of distributed base stations and small cells (gNBs/eNBs/NBs). However, RAN architecture is evolving towards a centralized architecture (via various split architectures), whereby gNB/eNB/SC are split at specific protocol layers and implemented as a central unit connected to several remote units; 3GPP, ORAN and SCF are standardizing three different and complementary options for the split RAN architectures.

In an edge-enabled network, it follows that the centralized unit (CU) can now be implemented at an edge location based on edge infrastructure. If the CU and edge infrastructure are virtualized, then the virtualized CU can be implemented as virtualized network functions (VNFs) on the virtualized edge infrastructure.

By contrast, the core network (CN) traditionally had what could be called a centralized architecture, consisting of a few centralized locations in the MNO network implementing evolved packet core/core network (EPC/CN) functions. However, with the standardization of control and user plane separation (CUPS) and 5G next generation core (NGC), two fundamental changes happened: the separation of the control and user plane functions and the distribution of the user plane function. This allowed the user plane function to be located closer to the end-user (enabling, for example, low latency and location awareness). The edge became the location for such distributed user plane functions.

From an MNO network evolution perspective, therefore, the edge has the potential to become the confluence of centralized RAN functions and distributed CN functions. For Wi-Fi access, EC resources may also enable VNFs and Wi-Fi controllers.

Now let's consider enterprise or venue networks. These are essentially LAN/WAN IT networks, made up of routers, switches, firewalls and so on. The virtualized edge infrastructure can readily be used to implement software versions of such IT functions, thereby making the edge computing platform a shared asset for enterprise IT and edge functions.

Local breakout (LBO) function: Edge computing solutions can 'break out' certain traffic streams locally. The LBO traffic may be served locally at the edge, either by the edge applications or by the enterprise (or venue) application servers. The traffic streams to be broken out would be defined by LBO policy, relating for example to user (or device) identity, destination network, server type or application type.

3.2.3 Edge Platform Services

Edge platform services are typically provided by the edge platform to the edge applications. They may take inputs from the network functions but could also be standalone functions. Some examples, taken from ETSI-MEC publications and other sources[2] include:

- Edge domain name system (DNS) services (for resolving edge server names)
- Filtering rules and policy services (for implementing edge policy)
- Application discovery services (for discovering and accessing edge applications)
- Edge storage services (for content caching or distribution)
- Value-adding services including:
  - Radio network information services (for analytics of the underlying radio network)
  - Presence and location services (for detecting the presence of – and possibly tracking the location of – UEs)
  - Bandwidth and multi-RAT management services (for optimal integration of multiple radio access technologies (RATs), such as cellular and Wi-Fi)
  - Video compression services
  - Video analytics services (such as facial recognition)
  - Encryption services
  - Logging, statistics and analytics services (for edge monitoring applications)

The edge platform services are made available to the edge applications via a set of APIs. ETSI-MEC has defined a set of them[3]there are also proprietary sets of APIs.

3.2.4 Edge Applications

Edge applications are applications that are specific to – and/or tailored for – the users or devices at the edge. These may be, for example, augmented or virtual reality in a museum or robotic control in an industrial factory. The physical implementation of these applications may be at the edge or in the MNO network or over the top (OTT) in the internet. The choice would depend upon the requirements (notably latency) of the application.

It's worth noting that edge applications may or may not make use of edge platform services, depending upon the complexity of the edge solutions. When the edge use case is simple and consists of a small set of applications, it is possible, in principle, to realize the applications directly above the edge network functions layer. In this case the edge applications are connected both to the edge platform services layer and also directly to the edge network functions layer (see Figure 3–2).

Edge applications may be broadly characterized as enterprise or industrial applications; venue applications (stadiums, concert halls etc); and consumer applications. A more detailed classification appears in Section 8.2.

3.2.5 Edge Management and Orchestration

Clearly, the components of the entire edge solution need to be managed and orchestrated. Traditionally, network management consists of element management system (EMS) and network management system (NMS) layers, although common

---

[2] ETSI-MEC ISG Publications https://www.etsi.org/committee/1425-mec?jjj=1578244642448

[3] ETSI GS MEC 009 V2.1.1 (2019-01) "(MEC) General principles for MEC service APIs" http://www.etsi.org/deliver/etsi_gs/MEC/001_099/009/02.01.01_60/gs_MEC009v020101p.pdf

unified frameworks that manage the entire networks are in development. They include the open network automation platform or ONAP. These common frameworks are especially useful in networks that are virtualized. Other developments include the RAN being re-architected with its own controller (for example, Open RAN).

Based on these developments, the management and orchestration of edge assets (such as infrastructure) and functions, can, in principle, be performed as part of a single end-to-end framework or a layered framework, wherein the edge assets and functions are managed separately. This is then orchestrated at the end-to-end level. Figure 3–2 depicts the former case, with the individual dashed lines indicating interfaces from the edge infrastructure, and virtual network functions (VNFs) relating to network functions, platform services and applications to the management and orchestration layer.

3.2.6 Policy, Billing and LI

Policy for EC solutions is an important but complex topic. For edge infrastructure, policy solutions are needed to control access and partitioning of hardware resources, among other areas. For edge network functions, specifically for local breakout, policies are needed to configure and control the traffic that would be broken out – in terms of users, devices, destination networks, destination servers and application types.

For edge platform services, policies are needed to control which edge applications would be appropriate for which types of edge services. In addition, policy solutions should be designed to take into account the (potentially differing) requirements of the MNO and the enterprise or venue owner.

Of course, EC billing solutions are needed, not least so that locally broken out traffic may be monitored and billed based on usage. Approaches to billing will also depend on how EC applications are used. The consumption of EC platform services could also be monetized via appropriate billing mechanisms.

Legal interception (LI) also becomes important when traffic is locally broken out. Traditionally, the LI points are in the MNO's core network; when some traffic is locally broken out, the traffic is not seen by those LI points. Accordingly, mechanisms should be built into the EC solutions, such that traffic, which is under LI warrant, is made available to the LI points in the core network.

Another LI requirement is transparency: the party under LI warrant should not be able to detect that LI is in progress. This makes LI solutions for LBO traffic particularly challenging, as some mechanisms, such as disabling offloading of such traffic, or mirroring the traffic under warrant, could violate this transparency requirement.

3.3 Standardization

Edge computing solutions involve various stakeholders and distinct functional entities. It is critical that the interfaces needed for these distinct functional entities to interwork together are standardized. This would avoid fragmentation of the solution space that could be caused by the number of proprietary solutions that may not interwork with each other fully. So, what is the current situation? What gaps exist? And what should be done?

Firstly, edge infrastructure is generally assumed to be a network function virtualization infrastructure (NFVI), that can make use of generic COTS hardware and virtualization software layers. However, some of the edge applications are expected to be compute-intensive. Thus, special hardware accelerators may be needed; here standardization may be important.

Secondly, edge network functions are generally standardized to a great extent by 3GPP, so that architectures such as CUPS for LTE and 5G architecture can be readily used. However, simplified architectures – based on the so-called 'bump-in-wire' architectures – are also being developed by vendors. These do not have complete standards support and therefore need to be evaluated for standardization.

Edge computing is expected to work across all access technologies, including Wi-Fi and wired access networks. Therefore, these edge solutions should be developed in ways that maximize the use of standardization.

Thirdly, while there may be many platform services and many of these may be vendor-specific, it is useful to have a minimal set of services that is standardized.

Of course, the edge applications themselves do not need to be standardized. However, the interfaces (or APIs) to underlying edge platform services and network functions do. Here ETSI-MEC has done useful pioneering work by defining some APIs – for radio network information services (RNIS), location, and bandwidth, for example. These need to be harmonized with respect to other APIs that may be prevalent in the industry.

Other APIs not explicitly depicted in the framework figure (Figure 3–2) include APIs that may be needed for the edge network functions to communicate with the core network for the purposes of implementing edge functions. Several attributes related to various traffic flows are only known at the core network level and are not readily available at the edge. This makes certain edge tasks challenging. These tasks include mapping individual traffic flows to user or device identity, the ability to control the QoS of locally broken-out traffic flows by the core network, and being able to page only the edge devices that are connected to the edge network. Such tasks will require what could be called 'network APIs' between the edge and the core networks. Some of these are in place (notably via CUPS and 5G), but a thorough gap analysis needs to be performed to identify and fix any missing functionalities.

Finally, the management and orchestration of various edge functions needs to be standardized. Multiple solutions are available, such as ETSI-NFV and ONAP, but a global standard for these functions would be worth considering.

4. EC Deployment Use Cases

Edge computing is expected to cover a large variety of use cases where aspects of the core network are deployed at the edge of the network to enable access to one or more applications. This includes a wide range of network architectures – from fully private networks with all 3GPP nodes deployed at the edge, to hybrid architectures with user plane processing deployed at the edge and the remaining nodes deployed in central data centres.

The next section looks at a few relevant use cases.

4.1 Premise-based Edge Computing Use Cases

4.1.1 Fully Private Cellular network [4]

A private cellular network is deployed to provide service to employees, machines and devices, as authorized by the private network operator. It is characterized by the deployment of all relevant 3GPP nodes within the enterprise IT domain, with full user management capabilities. The offering includes data services with a local SGi interface to enterprise resources, and voice (if enabled).

User management (for example SIMs or IMSI) is managed by the enterprise, with no inbound or outbound roaming. This may also include optional support for location services to track IoT devices.

The mining industry offers an example of what a fully private network can offer. A local on-premise edge computing solution would provide full application support for a mining operation. The on-premise solution can provide support for necessary applications related to security, location, video surveillance, device monitoring and messaging – without the need to backhaul the traffic to a centralized EPC.

Figure 4–1 shows a private network (a fully autonomous edge core network) where all computing is done at the edge.

[4] For more private network use cases see Private Cellular Networks with Small Cells, April 2020, http://scf.io/doc/235

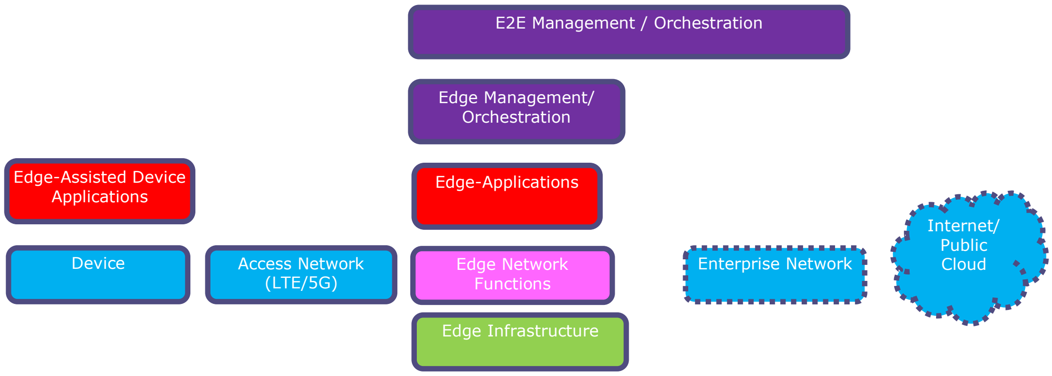


Figure 4–1 Fully autonomous private cellular network

The main characteristics of a fully autonomous private cellular network include:

- All core network components are deployed at the edge, as edge network functions
- The private cellular network provides communications to both employees and IoT devices
- User management is managed by enterprise (using private SIMs)
- There is no inbound or outbound roaming
- Location services may be used to track IoT devices
- Internet/public cloud access may be available (depending on business needs)

Figure 4–2 shows an extension of a previous use case where there is a single edge core network on a private network with multiple sites (for example, a mining company with several mining sites).

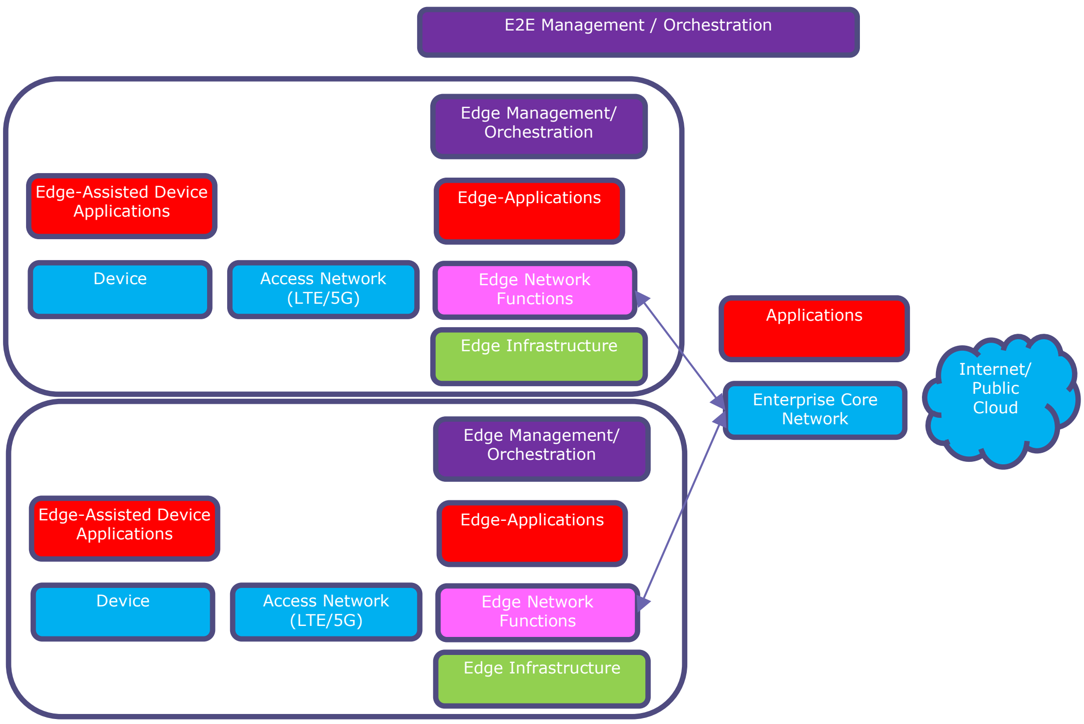


Figure 4–2 Fully autonomous private cellular network serving multiple sites

The main characteristics of this network are essentially the same as in a single-site fully autonomous private cellular network, with the addition of service and/or user mobility across the sites, as required by business needs.

4.1.2 Private Cellular Network with Roaming

A private cellular network with roaming is deployed to provide services to employees, machines and other devices, as authorised by the private network operator. It has inbound and outbound roaming agreements with other service providers and MNOs.

This network is characterised by the deployment of all relevant 3GPP nodes within the enterprise IT domain, with full user management capabilities and interconnectivity with standard IP exchange/GPRS roaming exchange (IPX/GRX) to enable 3GPP roaming.

The offering includes data services with home routed and visited routed traffic capabilities to allow for a configurable local SGi interface to enterprise resources, and voice (if enabled).

'Home user' management (that is, SIMs and IMSI) is managed by the enterprise. This may also include optional support for location services to track IoT devices.

Figure 4–3 maps this use case to the edge computing framework.

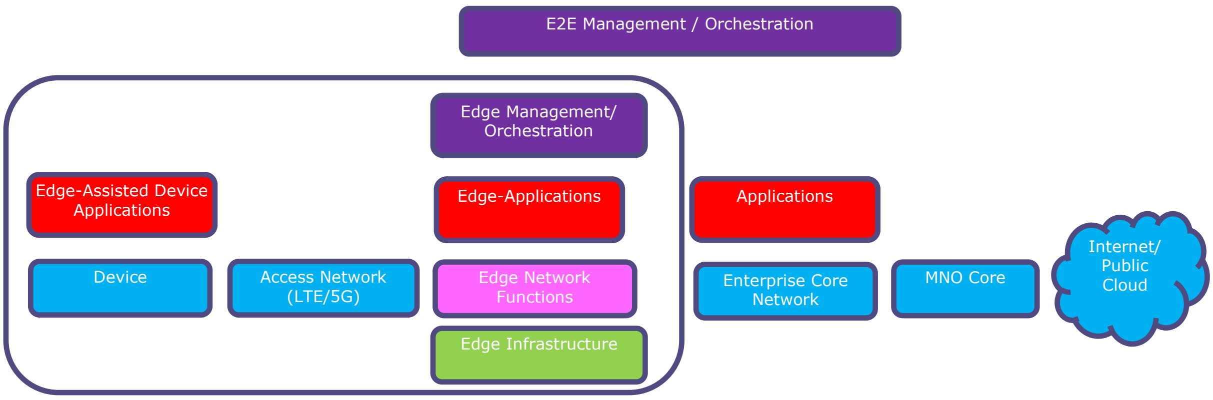


Figure 4–3 Fully autonomous private cellular network with roaming

The main characteristics of the private network with roaming are essentially the same as those of a fully autonomous private cellular network, with the addition of inbound and/or outbound roaming with other service providers and the MNO, and data services for home routed and visited routed traffic.

Deployment scenarios include:

- **Remote oil/gas, maritime** Edge applications with video, local communications automation, inbound roaming for employees. Reduced backhaul bandwidth needs
- **Hospital, ports, campuses** Edge application for local voice communications, automation, access to medical files/devices. Inbound roaming for patients and visitors

4.1.3 Private Cellular Network Integrated with MNO

An MNO-authorized private cellular network is deployed to provide service to employees, machines and devices, as authorized by the MNO, with inbound and outbound roaming agreements with other service providers and MNOs. It is characterized by the deployment of the relevant 3GPP nodes within the enterprise IT domain (for example mobility management entity, packet data network gateway and serving gateway (MME, PGW and SGW)). User management capabilities are provided by the MNO-managed home subscriber server (HSS). Interconnectivity with standard IPX/GRX enables 3GPP roaming.

This type of offering includes data services with home routed and visited routed traffic capabilities to allow for configurable local SGi interface to enterprise resources. Voice (if enabled) is provided using OTT applications and/or optionally voice over LTE (VoLTE) services, hosted by the MNO. Figure 4–4 maps this use case to the edge computing framework.

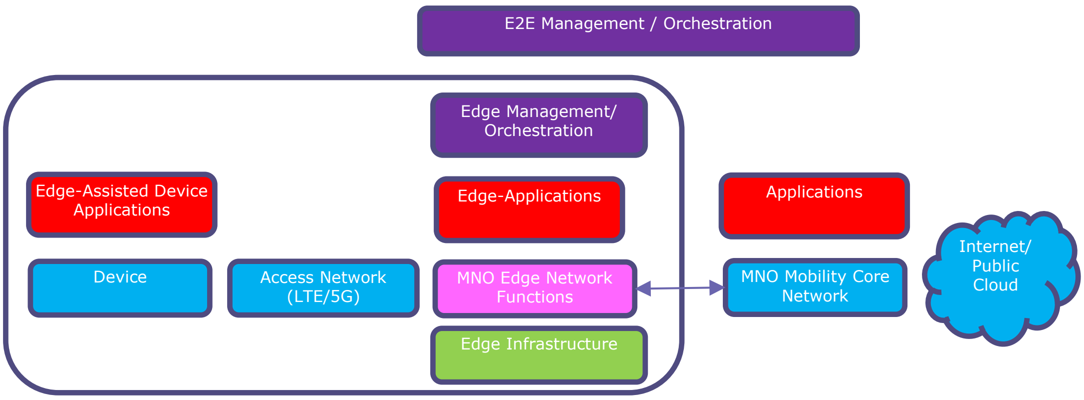


Figure 4–4 MNO-authorized private cellular network

The main characteristics of an MNO-authorised premises network are:

- MME, PGW, SGW components are deployed at the edge
- User management is managed by the MNO
- Local access control is managed by the enterprise
- The private network provides communications to employees and IoT devices
- There is inbound or outbound roaming with other service providers and the MNO
- Location services to track IoT devices
- Data services for home routed and visited routed traffic

Examples of deployment scenarios include:

- **Remote oil/gas, maritime.** Edge applications with video, local communications automation, inbound roaming for employees. Reduced backhaul bandwidth needs.
- **Hospital, ports, campuses**. Edge applications for local voice communications, automation, access to medical files/devices. Inbound roaming for patients and visitors.

4.2 A Network-based Edge Computing Use Case

The MNO network can be enhanced with distributed edge assets, with local breakout capabilities and edge computing functions. The architecture supporting such use cases is discussed in Section 5.

Such a distributed edge network may be used for providing edge services and applications to enterprises (via macro or on-premises small cell/DAS access networks) as well as to general public subscribers (via the macro or HetNet access networks). Corporate office buildings, autonomous vehicles and smart cities are just a few examples of enterprise applications.

Additionally, such a distributed edge network may be used as a content distribution network (CDN) for the MNO as well as to enterprise customers.

The network itself is characterized by the deployment of distributed 3GPP user-plane nodes (for example SGW/PGW) within the MNO network edge locations, with the remaining EPC functions located within the MNO core network. Relevant network architectures are in accordance with selected standards from 3GPP, ETSI MEC and other groups. The multi-access edge computing (MEC) initiative is an industry specification group within standards body ETSI

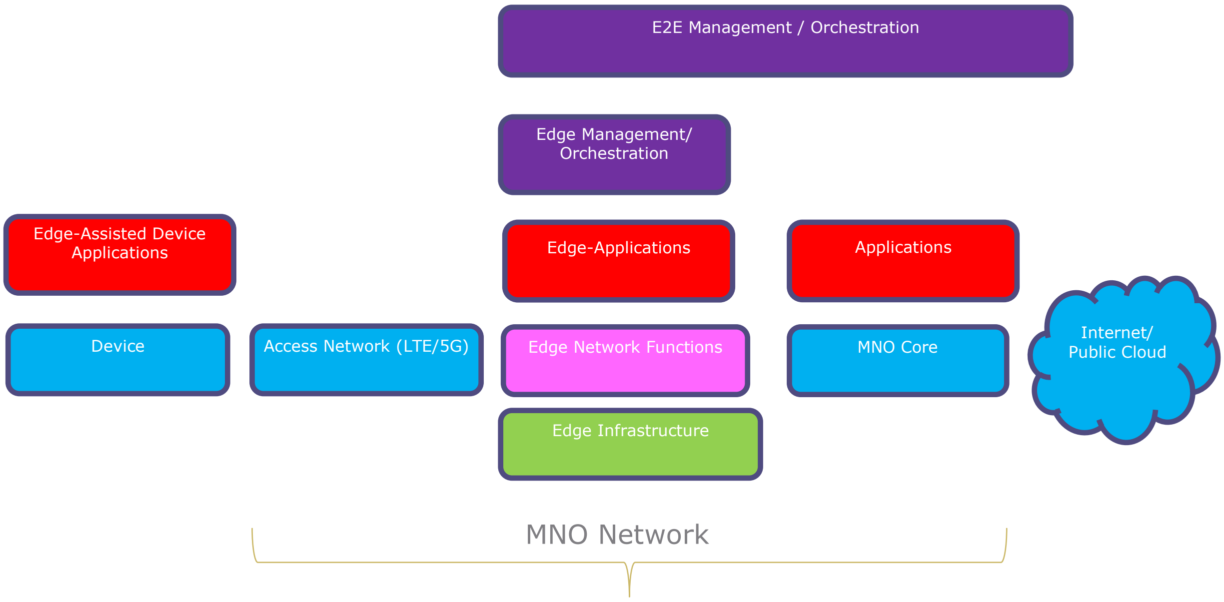


Figure 4–5 MNO distributed edge network

5. Edge Computing and Small Cell Networks

Small cells and small cell networks may be implemented in a variety of architectures – for example, small cell networks may be implemented as a network of integrated, all-in-one small cells. While such solutions would be suitable for small and medium-sized networks, disaggregated architectures would be more suitable for larger installations. The main reason for this is that by disaggregating the more expensive and functionally complex parts of the small cell network it might be centralized in a single element, whereas the less expensive and the functionally simpler parts could be replicated across the network.

This concept has led to two popular solutions in the industry: a two-way disaggregated architecture and a three-way disaggregated architecture. Generically, the elements of a disaggregated small cell network are referred to as RU (radio unit), DU (distributed unit) and CU (centralized unit). Depending upon the exact nature of the disaggregation, these elements may have different names: SCF-defined S-RU, S-DU & S-CU; O-RAN-defined O-RU, O-DU & O-CU; and 3GPP-defined gNB-DU & gNB-CU.

Figure 5–1 illustrates some of the possibilities.

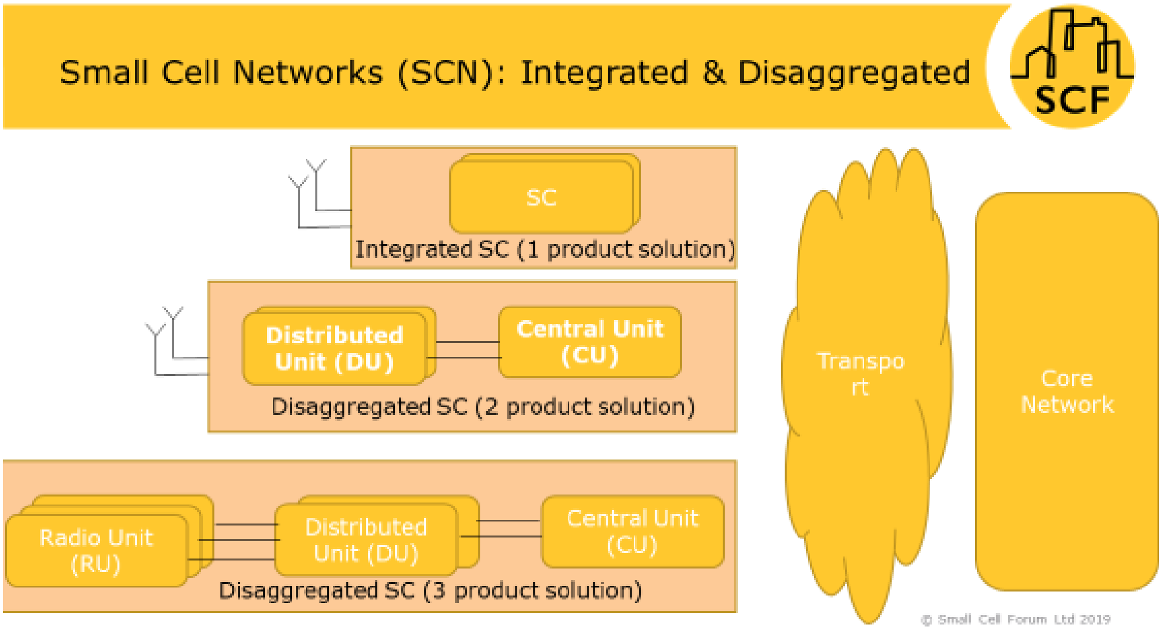


Figure 5–1 Integrated and disaggregated small cell network architectures

Small cells and small cell networks, due to their small footprints, and edge computing, due to its localization to the edge, are highly synergistic and can be integrated in more than one way, as we shall see.

5.1 SCN and EC Integration: Option One

As shown in Figure 5–2, the SC and MEC are completely integrated and create fully distributed MEC topology. This topology has the following challenges:

- Mobility complexity – that is, when a mobile subscriber (UE) crosses between SCs and should switch between application instances (for example, application instance #1 shown in Figure 5–2).

- Small cell cost – MEC and specific 4G or 5G core functions should be embedded as part of the small cell hardware, though it will raise the small cell cost as add-on processors, memory, and connectivity interfaces may be required. The operator should be very careful when specifying the MEC hardware resources (computing or memory, say) for the desired applications.
- MEC integration – The ideal aim of MEC integration would be the App Store model. The current state of the industry is that a combined small cell and MEC solution is custom-developed. To integrate the MEC application(s) onto the small cell, the MEC provider will be required to port its solutions to run onto the available small cell hardware. This development requires the parties to integrate interfaces, develop call flows, and ensure integration into the larger operator network.
- In most MEC-SC integrations, the solution will require three parties to interface: the operator, the small cell developer, and the MEC developer. With the increased costs and complexity of realizing the solutions, the operator should specify using distributed MEC with the small cell, or consider the value balance of doing so against increasing the cost. It's also worth noting that these partnerships and directives by operators for fully integrated MEC need to consider revenue share as well as ensuring a viable marketplace for all parties.

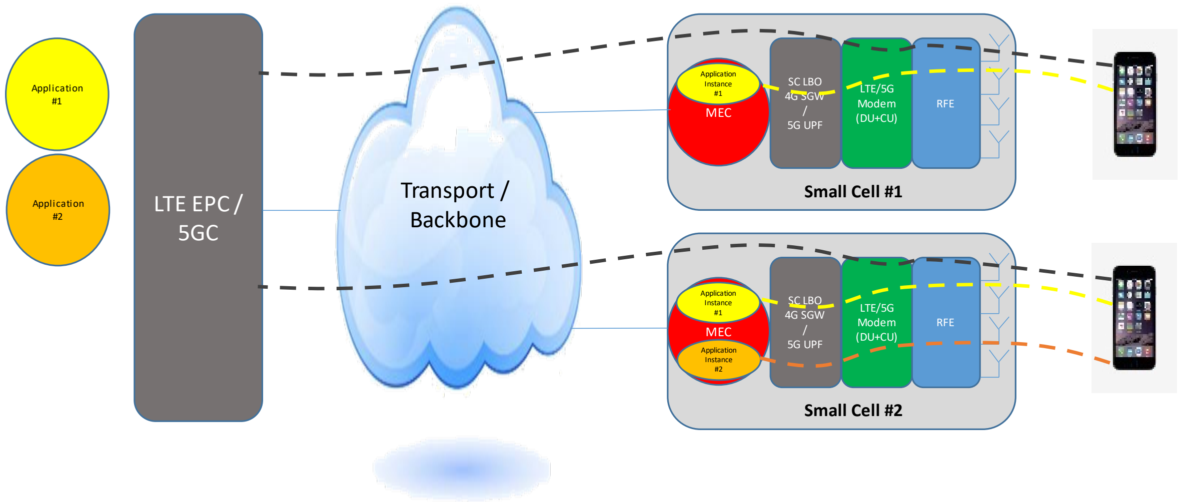


Figure 5–2 Fully integrated EC & SC topology option

In spite of these challenges, this topology can be useful for residential and SOHO locations, where the operator may be interested in enabling localized applications, which may serve the subscribers attached to the specific small cell. It can be applicable – and integrated – as part of a virtual customer premises equipment (vCPE) solution, where vCPE, SC and MEC can be integrated into one box and provide local service for the users connected over cellular, WiFi or LAN.

5.2 SCN and EC Integration: Option Two

As shown in Figure 5–3 and Figure 5–4, in this case the MEC is localized behind a central controller, which may comprise only LBO functionality (4G SGW or 5 UPF) or additionally VRAN BBU functions (with various eCPRI split options).

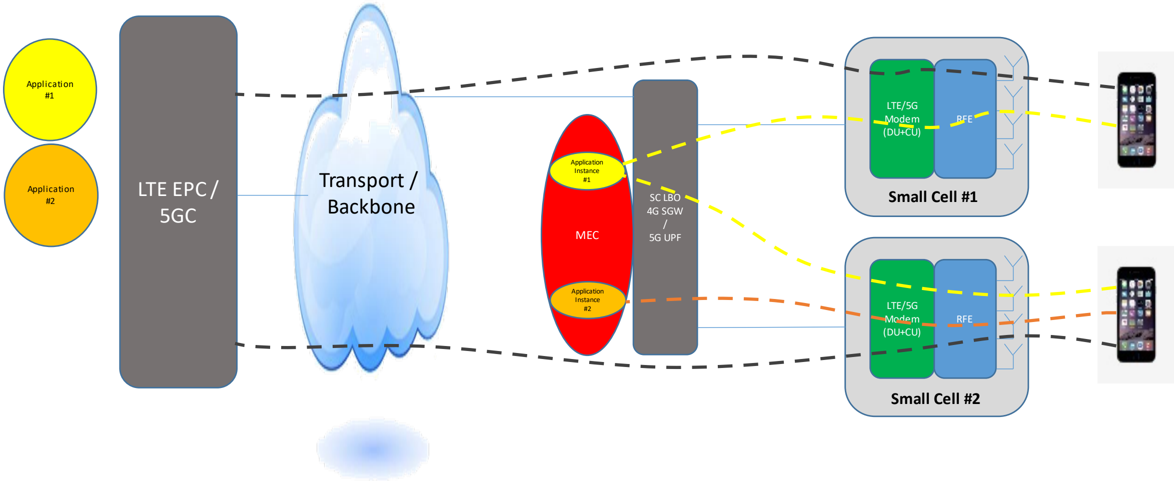


Figure 5–3 Partially Integrated EC & SC Topology-1

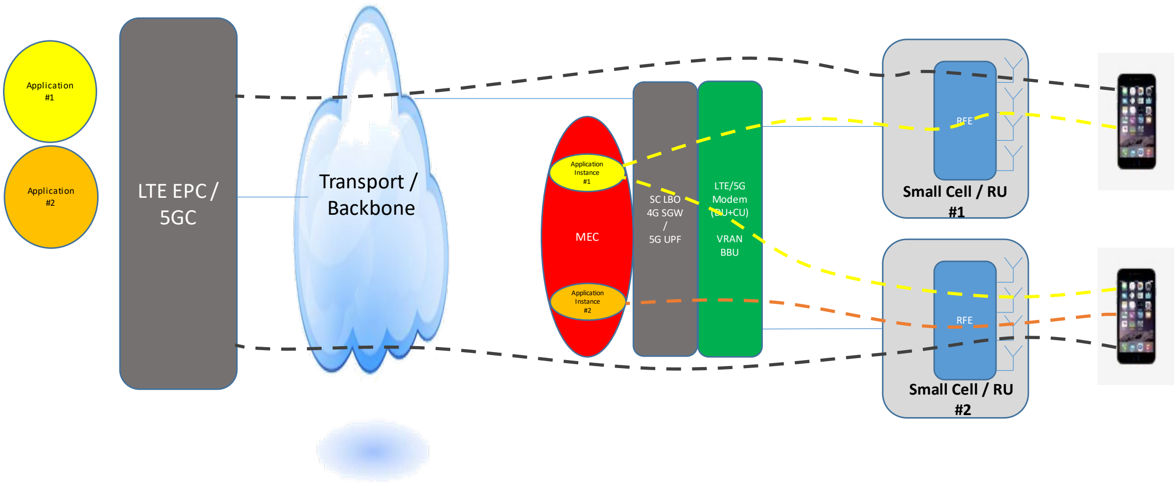


Figure 5–4 Partially Integrated EC & SC Topology-2

This topology is in fact very common with MEC architecture proposed for the macro deployment, where MEC infrastructure is located in the POP/aggregation sites.

In the localized MEC solution topology, the small cell manufacturer and operator do not need to do extensive integration work on the small cell. This method is therefore by far the easiest to develop and should have the shortest time to market. The downside is that this model works best for venues, campuses, and multi-dwelling units (MDU) where networks can have a localized aggregation point in a common IP network. For this reason, this method will not be cost-effective to service standalone deployments, like home or SMB-type small cells.

5.3 Recommendations

The recommendations regarding small cells and edge computing can be split into three categories. They are:

- **Computing infrastructure should be shared between edge computing and virtualized parts of disaggregated small cell networks:** – With the disaggregation and virtualization of small cells, the computing infrastructure can be beneficially shared between edge computing functions and the virtualized parts of the disaggregated small cell network. This can produce considerable cost savings and increased functionalities as well as make deployments easier, especially in indoor enterprise scenarios. The virtualized small cell functions may be DU and/or CU of a disaggregated small cell network.

- **Centralized functions of small cell networks and distributed functions of the core network should be realized as common edge computing network functions** – Small cell network functions, such as DU and/or CU in a disaggregated SCN, and edge computing network functions, such as distributed user plane functions and local break out functions – along with associated local policy engines – may be beneficially co-implemented, resulting in small cell network solutions with integrated EC capabilities.
- **Small cells networks should be enhanced to provide analytics as edge computing platform services** – Small cell network products, either integrated SCs or RU/DU/CU in disaggregated SCNs, may incorporate edge computing services, such as radio network information services (RNIS) and location services. This will make for more efficient implementations of combined small cell network and edge computing solutions.

6. Edge Computing Infrastructure

Edge computing infrastructure can be located at a variety of possible edge locations (see Figure 6–1).

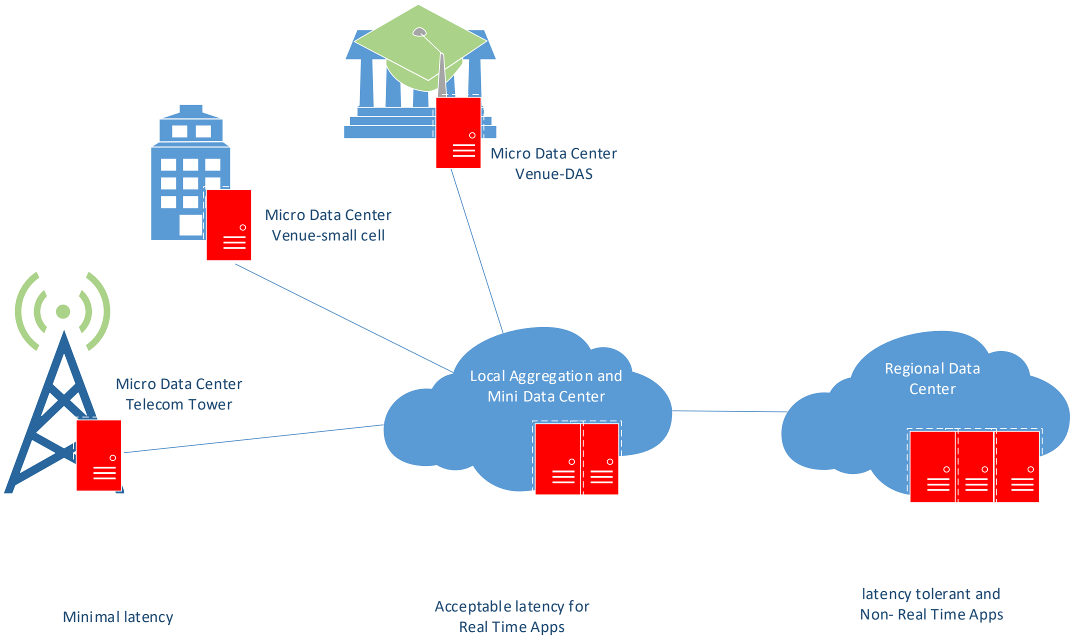


Figure 6–1 Edge infrastructure location options

Various locations are suitable for a number of edge applications, including:

1. Service provider wireless and wired use cases for flexible indoor and outdoor on-premise deployment such as 4G/5G/Wi-Fi base band deployment at micro data centers (or deep edge locations).
2. Enterprise service use cases for indoor and outdoor on-premises deployment such as IIoT, factory automation, private LTE and private 5G use cases (such as critical and remote control) at micro data centers (or deep edge locations).
3. Metro deployments, usually at a higher level of aggregation (such as mini data centers) for vRAN, vEPC, CDN and 5G NW slicing use cases.
4. Regional data center deployments. These are a good fit for vEPC 5G network slicing; data analytics engine; AI as a common infrastructure; SoN; and ML engine for a regional state-level cluster.
5. Centralized data center for end-to-end NSO orchestration, cloud-based services, CDN, vEPC and 5G NSA and SA core services.

Note: In this white paper, the two locations labelled as micro data centers for venue small cell and venue DAS, are referred to as premise-based EC infrastructure. The others (labelled as micro data center telecom tower, mini data center and regional data center) are referred to as network-based EC-infrastructure.

6.1 Premises-based EC Infrastructure

Premises-based edge computing infrastructure essentially consists of COTS hardware, with a virtualization environment for other edge-related functions. The COTS hardware should have sufficient compute-storage-networking resources. There may also be dedicated accelerators for compute-intensive functions, such as video and graphics or encryption. The edge-related functions – edge network functions, edge platform services and edge applications – may be executed as virtual machines or containers. See Figure 6–2.

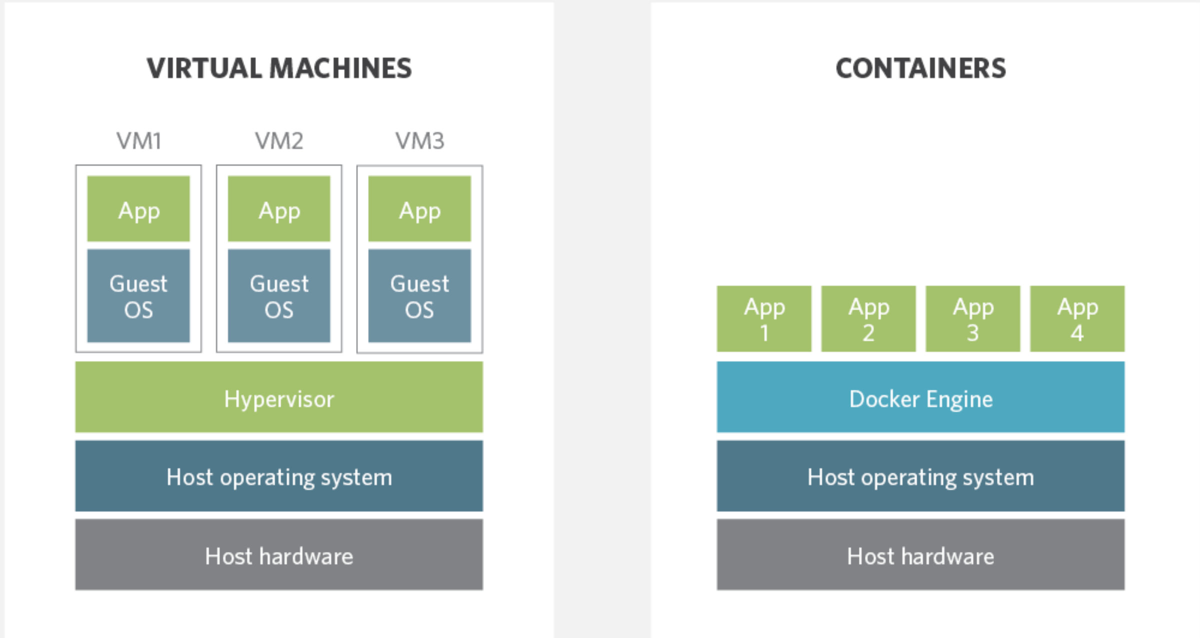


Figure 6–2 Virtualized Edge Infrastructure

Linux Foundation's edge project, Akraino[5], is developing infrastructure frameworks for edge computing use cases.

Premises-based edge computing can take advantage of collocation space, an on-site data center, a server room and/or a wiring closet. Depending on the facility, all of these on-premise locations typically have power, fiber, space, and the necessary infrastructure to enable edge computing.

6.2 Network-based EC Infrastructure

Network-based edge infrastructure consists of edge computing assets distributed across the network. The edge locations may be at one or more of the following:

1. Cell tower/small cell
2. Local aggregation center
3. Regional center

Cell tower and small cell sites allow for edge computing opportunities as they provide wireless service to a variety of end-users but are only one hop away from the end-user. They are close to mobile and nomadic users, so can provide ultra-low latency.

As for positioning, the ground space at tower sites would be ideal for edge data centers (EDCs), which may be in a purpose-built facility, an existing facility or in a cabinet or enclosure and operate within the 10-150 kilowatt range. AC/DC power is readily available at tower sites and in the majority of cases there are backup generators for redundancy. For fronthaul, backhaul and the various edge computing

[5] https://www.lfedge.org/projects/akraino/

network architectures, fiber connectivity is important to support the small cell densification that is required for 5G with edge computing.

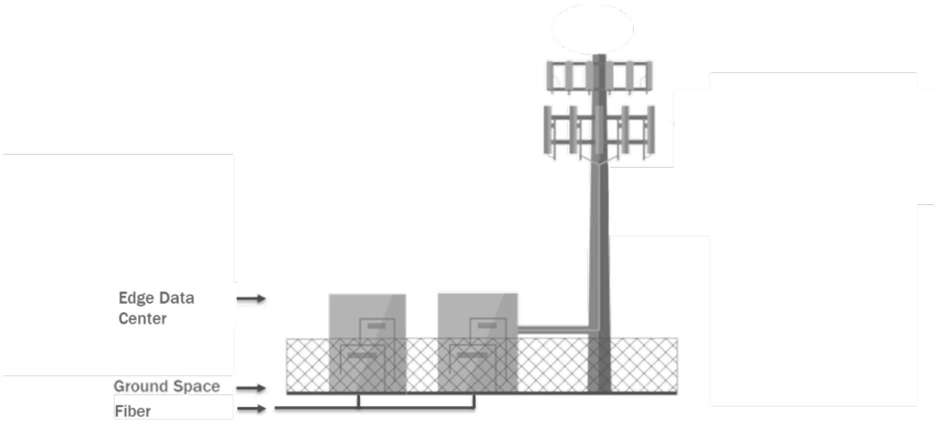


Figure 6–3 **Cell tower-based EC infrastructure**

Local aggregation and regional centers provide additional opportunities for edge computing.

The local aggregation network comprises multiple access networks, typically connected via a fiber network, with traffic travelling to regional points of presence (POPs). Local aggregation sites are where switching and routing equipment aggregates access traffic and are suitable locations for deploying edge computing hardware. The locations of these sites vary, which could affect latency, but they provide another resilient tier of possible locations for edge computing.

The regional centers are mobile network facilities that are typically centralized and have some core network equipment. These sites generally have fiber, power, cooling and space, making them high-value edge compute locations. The locations are usually manned and easily accessible. However, the latency at these locations tends to be greater than at the radio sites and local aggregation sites.

6.2.1 Deployment Considerations for Edge Data Centers

Data processing that is required at the edge of the RAN can be achieved by the deployment of edge data centers (EDCs). This approach enables a natural transition to 5G by providing network architectural solutions to support network function virtualization (NFV) and network slicing.

Edge Data Centers

The deployment of EDCs involves a significant upfront investment in an extensive network point of presence anticipating applications and customers that may or may not arrive. Without a guaranteed revenue stream, this can be a huge investment risk.

The co-location of EDCs and C-RANs is a logical bridging step through which the EDC provides additional processing power for a collective of BBUs that will support a gradual transition from 4G to 5G. Deployment of management software like software defined networks (SDN) on virtual machine (VM) clusters – to consolidate processing workload onto active servers – can further optimize equipment utilization. This can help to reduce both upfront investment and power consumption from idle servers.

A number of pieces of equipment need to be installed in EDCs; these include servers, storage, edge routers, and switches. All this equipment needs physical connectivity – further increasing the need for rack mounting space, which comes at a premium. The use of CS connectors to replace standard LC connectors (see Figure 6–4) can enable an increase in space efficiency of up to 50 per cent, helped by a smaller connector footprint. The use of SN connectors further improves rack space utilization by allowing individual lane management when terminated into a QSFP or QSFP-DD transceiver by removing the need for MPO connectors and fan-out modules.

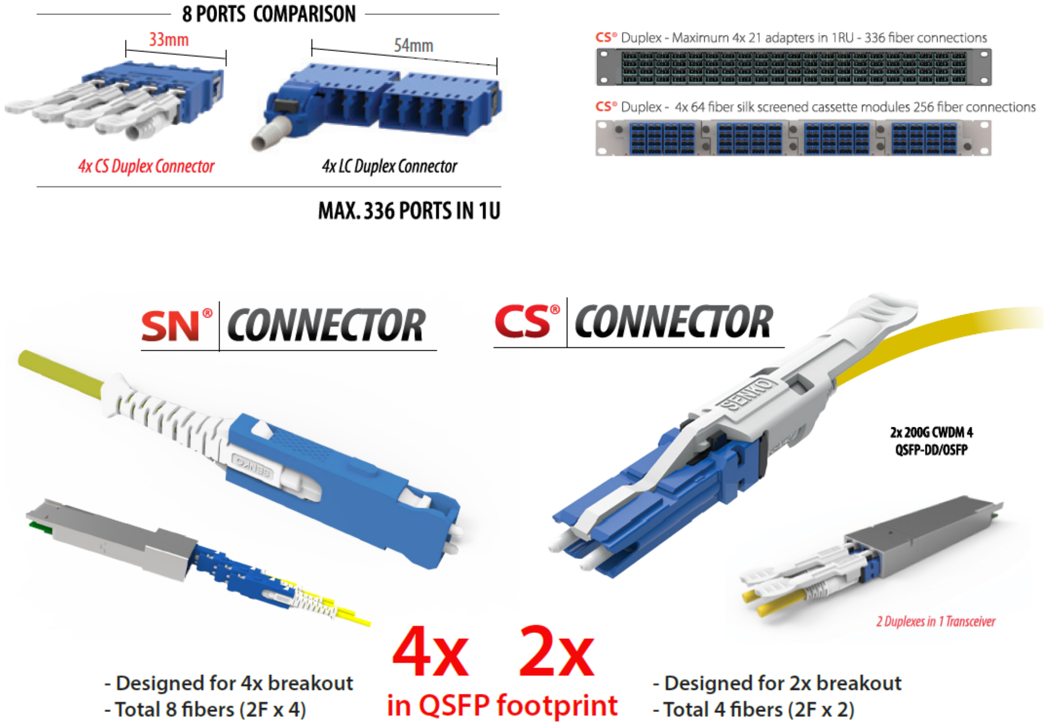


Figure 6–4 Types of connectors for edge data centers

Fiber Connectivity

The centralization of EDCs and C-RANs will require a fronthaul network with a large bandwidth capacity to support high-speed data transmission with low latency. The only logical choice for a mobile fronthaul network is optical fiber, which ticks all the boxes for low latency, high reliability, scalability, and high bandwidth. A simple fronthaul network can be designed as a point-to-point (P2P) architecture; this essentially means a dedicated fiber to each remote radio head (RRH). Although this architecture requires the least planning, it needs the most fiber runs from the central office to each RRH.

Another way of deploying a high bandwidth fronthaul network with less optical fiber is by implementing a wavelength division multiplexing (WDM) passive optical network (PON), which allocates dedicated wavelengths to each RRH over a shared fiber network.

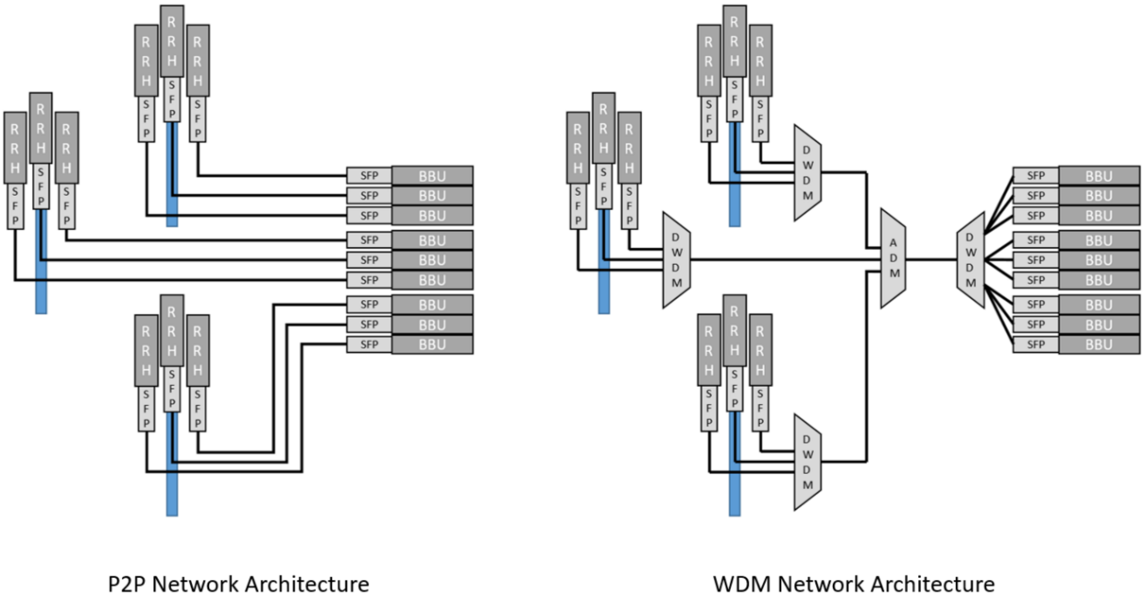


Figure 6–5 Fiber network architectures

A wavelength division multiplexing (WDM) coupler is a mature optical component that is widely deployed in access and regional long-haul networks.

The most common type is the silica-on-silicon arrayed waveguide grating (AWG) that has a close modal field match with optical fiber. This means it has excellent heat dissipation properties that allow it to be deployable in a wide temperature range while maintaining good channel uniformity and low attenuation. This, in turn, allows the flexibility of installing WDM filters in both indoor and outdoor environments. Outdoor-rated WDM filters with a wide working temperature between -40°C to +75°C are available.

Figure 6–6 AWG and WDM couplers

Central Offices as EDCs

Depending on network security requirements, the central network can be housed in mobile containers, office floors, data centers, or in telecom central offices. There are many considerations to take into account when deploying infrastructure to house a cloud or centralized base band unit (C-RAN BBU). Among the requirements are a rack mounting system, power supply, battery backup, access security controls, fiber management and termination, and, most importantly, backhaul connectivity to the core network.

Of all potential locations that can house the C-RAN BBU, telecom central offices provide all the required infrastructure, network connectivity, and access controls. In fact, there is a move underway to deploy data center facilities within central offices through what are called central office re-architected as a datacenter (CORD) deployments. CORD replaces the traditional central office infrastructure with hardware building blocks.

Existing central offices are already fitted with racking infrastructure, large power capacity, battery backup systems (including generators for additional equipment), environment control systems, building fire protections, network installation and maintenance procedures, disaster recovery protocols, and many other systems to guarantee network reliability. All existing outside plant networks – such as optical fiber cables, pits and conduits, and aerial poles – lead back to central offices. The deployment of a fronthaul network will be either an overbuild over the existing network or will access the existing network, perhaps by leasing dark fiber or overlaying wavelengths into the live access fiber network using WDM couplers.

Power Considerations

Other than the need for a large fronthaul bandwidth connectivity, 5G base stations are extremely power hungry. A typical 4G base station uses a 4x4 array, while a 5G base station is expected to use a 32x32 array or even up to 64x64 array, depending on the carrier's MIMO requirements. Based on current technology, each 5G base station is expected to require 10-15 kW of power.

Not only will each 5G base station need more power, but the coverage area of a 5G base station will be much smaller compared to 4G, due to the usage of higher radio frequencies. The short transmission range of 5G means it will need at least three to four times more base stations to have the same coverage area as the current 4G LTE network. This adds to the already high-power requirement for each base station. The only practical way of providing constant and reliable high-power supply to 5G base stations is by connecting them to the local power grid.

One of the biggest challenges in a large-scale deployment is finding technicians skilled in both power and fiber connectivity. There is, however, the option of combining power and fiber connectivity in a single cable and connector. This can help to reduce the cost of cable installation. A pre-terminated hybrid power and fiber plug-and play solution enables both power and fiber connection to be done within a single IP68 rated connector. With the ability to also integrate MPO connectors into a hybrid cable and connector, multi-fiber connections can also be made in a single cable and connector assembly. This removes the need to manage two separate cables which may require two separate truck rolls for technicians for fiber and power.

Figure 6–7 Hybrid power and fiber connectors

Challenges and Solutions in Deployment

In a densely populated city, space comes at a premium and, with multiple mobile carriers, everyone is vying to use for the same spaces – unless an infrastructure-sharing agreement can be reached. Network operators can not only deploy standalone mobile towers, but can also locate them on existing infrastructure such as power poles, lamp posts, traffic lights, road signage, and even on rooftops or the sides of buildings.

Getting the consent or rights of way (ROW) to deploy base stations and antennas on existing infrastructure is critical. Base station installation on infrastructure belonging to councils, local government, government boards, or utilities provides access to a large network of infrastructure as well as established deployment rules. An agreement with a single body could enable the use of a wide range of infrastructure over a wide coverage area in strategic locations. The alternative could be gaining access to individual buildings on a deal-by-deal basis.

Figure 6–8 Tower/pole-mounted deployments (standalone mobile tower, top/side of lamp post)

Figure 6–9 Rooftop mounted deployments

The next challenge is getting the fronthaul optical fiber network to the base stations. Optical fiber network deployment can be simplified if there is existing infrastructure such as pits and conduits or even dark fiber that can be leased. However, if there is insufficient capacity in the existing infrastructure, a new fiber network will be required.

There are many deployment methods – from aerial networks, direct buried, air-blown network and cable hauling through to conduits. The installation method selection depends on the level of network reliability, network capacity, future expansion flexibility, ease of network fault restoration, and, most important of all, cost of installation.

The cheapest method is the aerial network, but this network has the highest fault risk. An underground network has a higher level of reliability but is more expensive. The least invasive underground cable installation method is directly burying cable via the slot-trenching method. This is a much faster and cheaper install method, but the fiber count of the direct buried cable must be sufficient for immediate and potential future use as it is nearly impossible to deploy additional cable into the same slot trench. In addition, if there are any cable faults from a third-party excavation, network restoration will be much harder as there is no slack cable in pits at either end of the fault location where cable can be easily pulled and a restoration joint made.

Existing utilities – such as power cables, water pipes, sewage pipes, and existing copper and fiber cables – may also be along the planned route. Some countries have strict rules regarding the separation of utilities and minimum bury depths. The positioning of antennas on infrastructures can also cause installation and maintenance problems. If power cables are present, health and safety regulations may require minimum distance separation or the use of conduits. Technicians may even be prevented from doing any installation work near power cables if they do not have relevant certification. The higher radio power output from 5G antenna arrays will also create a wider no-approach zone. Proper training, visual signage, and radio antenna shutdown procedures must be in place to provide a safe working environment.

Some references for data center best practices can be found in specialized documentation that provides data center tier system and classification.[6]

6.3 Recommendations

TIA & BICSI Recommendations

There are various different regulatory and engineering requirements, depending on the location of the site. Installers should refer to the local requirements before engaging any engineering and civil work. Nevertheless, there are useful recommendations and best practice guidelines from organizations such as the Telecommunications Industry Association (TIA)[7] and the Building Industry Consulting Service International (BICSI)[8]. They include:

- TIA position paper: edge Data Centers - *https://www.tiaonline.org/wp-content/uploads/2018/10/TIA_Position_Paper_edge_Data_Centers-18Oct18.pdf*
- BICSI: Telecommunications Distribution Method Manual Ed. 14
- BICSI: Outside Plant Design Reference Manual Ed. 6
- BICSI: Information Technology Systems Installation Method Manual Ed. 7

[6] Information technology – Generic cabling for customer premises – Part 5: Data Centers ISO/IEC 11801-5: https://bit.ly/3gJxXl8
Uptime Institute - data center tier system and classification: https://bit.ly/2yIC6V8

[7] TIA https://bit.ly/36FePA9

[8] BICSI https://www.bicsi.org/

The guidelines may not be applicable to all sites. Nevertheless, they provide a generally acceptable best practice in the industry for site deployment and infrastructure design.

OCP Recommendations

Another useful set of recommendations for building EC infrastructures comes from the Open Compute Project (OCP).[9]

OCP was founded by a consortium of companies from the hardware, silicon, entertainment, banking and social media sectors. The OCP is a collaborative community focused on redesigning hardware technology to efficiently support the growing demands on compute infrastructure. Its goal is to build one of the most efficient computing infrastructures at the lowest possible cost without the limitation of vendor lock-in. OCP focusses on four pillars of infrastructure products. They are racks, power, chassis PCB and software. OCP deliverables include rack design specifications, requirements and recommendations for improving power consumption, blueprints for various use cases, and frameworks for open source software.

[9] https://bit.ly/2B90xMd

7. Edge Network Architecture

There are three main architecture approaches for the edge network. They are bump-in-wire (BIW) LTE architecture, CUPS-based LTE architecture, and 5G-NGC architecture. These are described briefly below.

7.1 Bump-In-Wire (BIW) EC Architecture for LTE

Bump-in-wire architecture essentially places a multi-access edge computing (MEC) infrastructure somewhere between the base station and the core interface where the traffic is either unencrypted or the MEC solutions have access to the keying material for the encrypted traffic. The latter approach poses significant challenges, due to key distribution. From a security perspective, this architecture can be described as a man-in-the-middle approach.

Depending upon where the MEC infrastructure is placed, there are three deployment options depicted below; options 1 and 2 are typically considered the most advantageous.

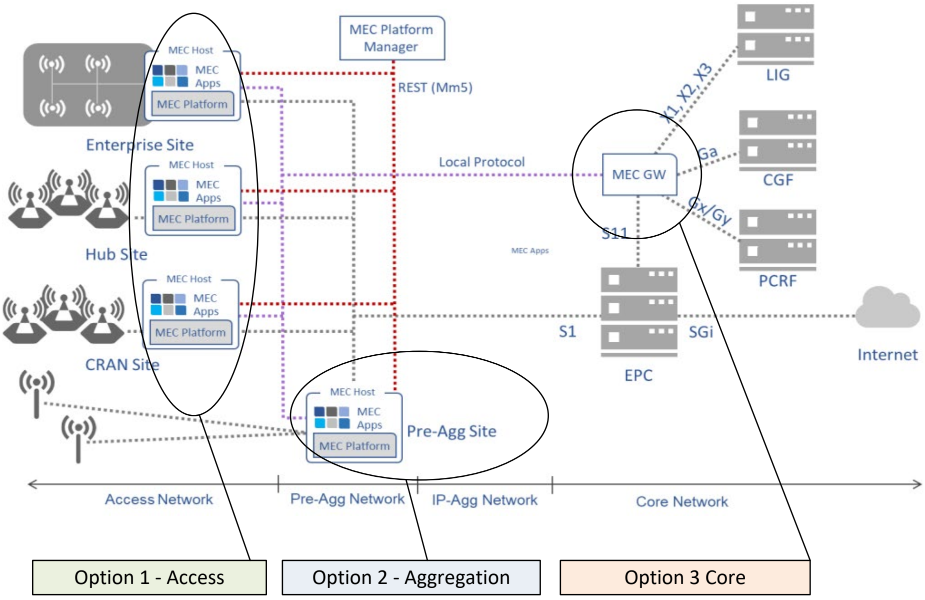


Figure 7–1 BIW Deployment Options[10]

- **Option 1: Access deployment** – In this option, MEC infrastructure is located in the access (RAN/SCN) network, either at an enterprise location, a cellular hub site or the BBU or CU of a CRAN/disaggregated deployment. In all these cases, the MEC infrastructure could, in principle, be integrated into the RAN baseband units. Such deployment architectures may have difficulty meeting legal intercept requirements for some services.
- **Option 2: Aggregation point deployment** – In this option, the MEC infrastructure is placed in a RAN/SCN aggregation point. This option makes sense from a legal intercept point of view if the aggregation point is secure.
- **Option 3: Core deployment** – In this option, the MEC infrastructure is located in the core network. While it can provide some latency-tolerant

[10] ETSI-MEC white paper #24 "MEC Deployments in 4G and Evolution Towards 5G" https://bit.ly/3gBeaUI

services, it has less value as the edge location is far from the end-users or devices. There is no latency improvement, no backhaul cost reduction and the applications could, quite possibly, be delivered by the core network itself.

In any of these options, the solutions are likely to have proprietary components. From a security perspective, additional interfaces to exchange use identity, flow mappings and security credentials are needed. The good news is that the industry is moving towards more standard interfaces favouring other approaches.

In summary, bump-in-wire solutions are feasible for some services but will always be hampered by challenges in obtaining security credentials. The solution is likely to be very fragile in terms of support from release to release. Longevity is doubtful. A proprietary solution is inevitable with the associated vendor-lock-in limitations. In addition, the industry has moved on to solutions that are not based on man-in-the-middle paradigms.

7.2 CUPS-based EC Architecture for LTE

Control and user plane separation (CUPS) is a powerful concept that has been successfully deployed in IT networks, and is now finding its way into mobile networks. Coupled with distribution of the separated control and user planes, three edge computing architectures, based on the CUPS principle, have emerged.

- **Distributed data plane** – This is the MEC model being standardised by 3GPP and ETSI. It distributes the user traffic plane to the edge site, allowing user data to be connected with edge-hosted applications.
- **Distributed control plane** – In this model, along with the data plane, control plane signalling is also distributed to the edge. Additional benefits of this architecture are reduced signalling requirements over the backhaul interface (for example transition from idle to active mode), and increased resilience (local data services can continue during backhaul outage).
- **Fully autonomous edge core network** – In this model, the edge network is capable of operating in a completely standalone manner, without access to the centralised core network.

Distributed Edge User Plane Architecture

This is the MEC model being standardized by ETSI.[11] As shown in Figure 7–2, it distributes the user traffic plane to the edge site, allowing user data to be connected with edge-hosted applications.

[11] ibid

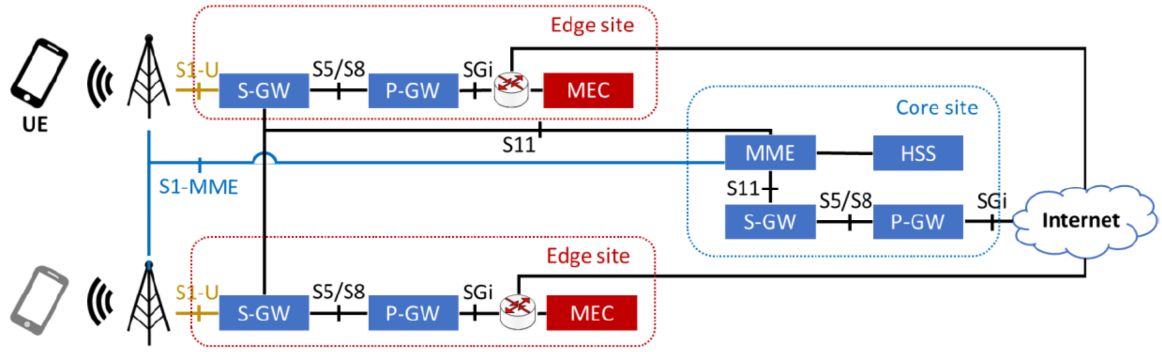


Figure 7–2 Distributed user plane EC architecture Option 1[12]

Note that the MEC box in Figure 7–2 (from ETSI-MEC) refers to EC platforms and EC application in the SCF EC framework introduced in Section 3.

As described by ETSI-MEC[13] in this architecture, the core network functions, and the SGW and PGW entities are deployed at the edge site, whereas the control plane functions, such as the mobility management entity (MME) and home subscriber server (HSS), are located at the operator's core site. The MEC host's data plane connects to the PGW over the SGi interface.

Another distributed deployment proposed by ETSI is a local breakout solution. This gives operators greater control of the granularity of the traffic that needs to be steered, as shown in Figure 7–3, where the PGW is now in the operator's core network, allowing greater control of traffic flows.

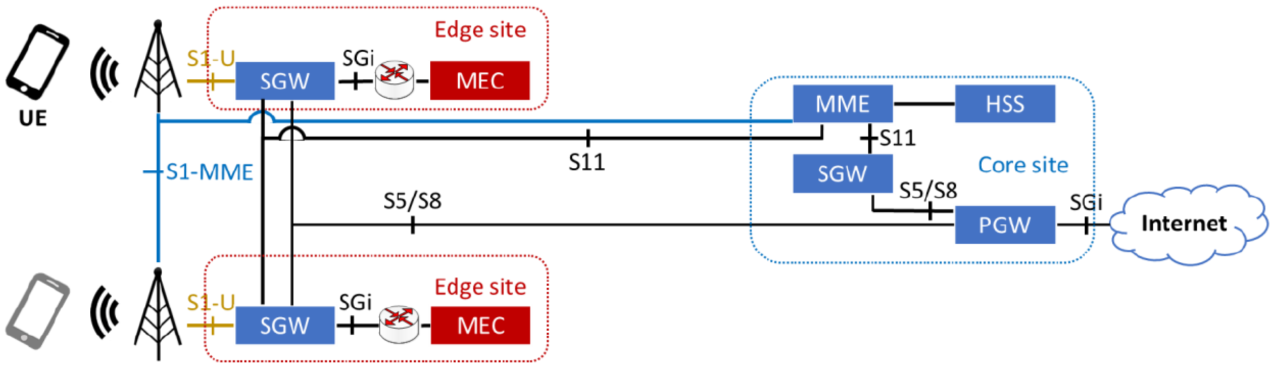


Figure 7–3 Distributed User Plane EC Architecture Option-2[14]

Note that the MEC box in the above figure (from ETSI-MEC) refers to EC platforms and EC application in the SCF EC framework introduced in Section 3.

[12] ibid

[13] ibid

[14] ibid

Finally, 3GPP has standardized a CUPS-based architecture outline in an overview[15] and specifications[16] which can be utilized for EC purposes, as shown in Figure 7–4.

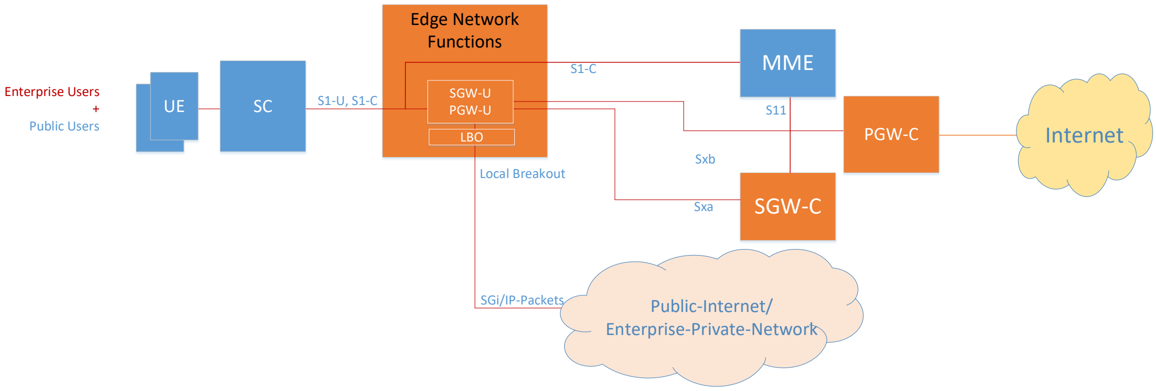


Figure 7–4 3GPP CUPS based EC Architecture[17]

Here the user plane functions, SGW-U & PGW-U, are realized in the edge location, whereas the control plane functions, SGW-C and PGW-C, are realized in the operator core network. The interfaces are Sxa (between SGW-U & SGW-C) and Sxb (between PGW-U & PGW-C), standardized by 3GPP in its Release 14 specifications [6-2 & 6-3].

Distributed Edge Control and User Plane Architecture

In this architecture, the edge includes the control plane functions, such as MME and HSS, in addition to the user plane.

[15] 3GPP CUPS Overview https://www.3gpp.org/news-events/1882-cups

[16] 3GPP CUPS Specifications TS 23.214 "Architecture enhancements for control and user plane separation of EPC nodes"

[17] ibid

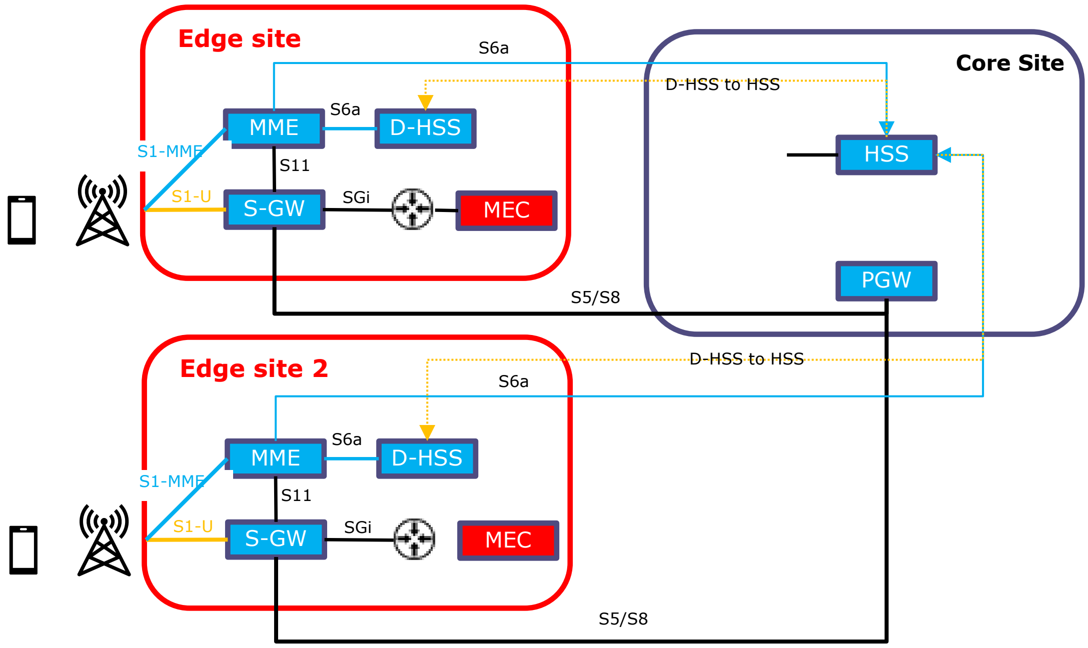


Figure 7–5 Distributed user and control plane EC architecture[18]

Note that the MEC box in the above figure (from ETSI-MEC) refers to EC platforms and EC application in the SCF EC framework introduced in Section 3.

This architecture leads to improvements in:

- **Resilience** – Local UEs can continue uninterrupted service during backhaul failure. Additional resilience is also achieved as the edge core has a local copy of MME and HSS. However, the D-HSS to HSS interface is not defined by ETSI/3GPP and therefore distributing HSS may be considered as being optional.
- **Latency** – Faster switch to/from idle mode – important over high delay/jitter links (such as satellite or wireless)
- **Backhaul bandwidth** – No bandwidth is used for local data signalling

[18] ETSI-MEC white paper #24 "MEC Deployments in 4G and Evolution Towards 5G" https://bit.ly/3gBeaUI

Fully Autonomous EC Architecture

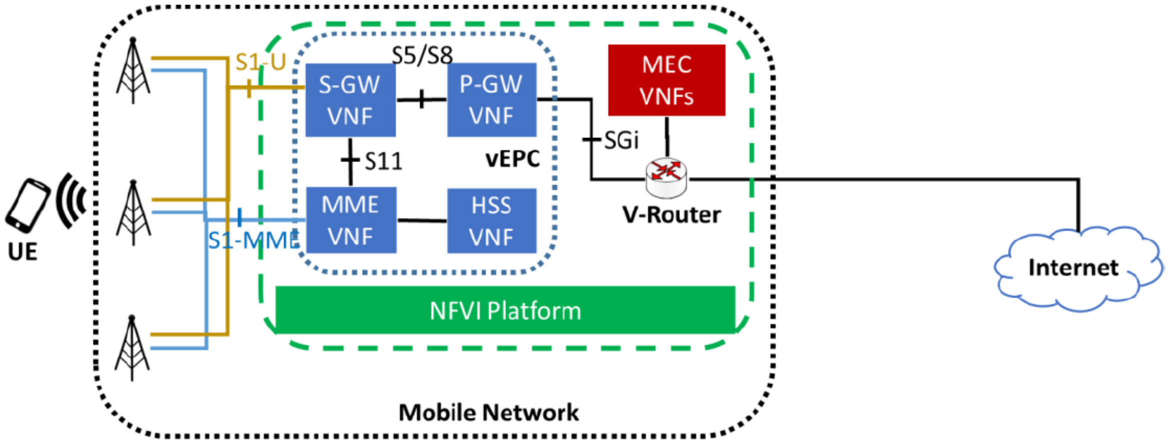


Figure 7–6 Fully autonomous EC architecture[19]

Note that the MEC box in Figure 7–6 (from ETSI-MEC) refers to EC platforms and EC application in the SCF EC framework introduced in Section 3.

In this architecture, all the core network functions are implemented at the edge, resulting in full edge autonomy. This type of deployment serves mission-critical push-to-talk or remote industrial use cases like mining sites, where communication with an MNO core is optional. It is typically used by first responders, public safety, and mission-critical industrial sites.

Some important characteristics of this architecture are:

- Support for edge SIMs which operate without reliance on backhaul
- Support for local IMS services – including VoLTE
- Local management of entire subscriber database
- Ability to deliver QoS and configurable features to enterprise customer requests.

[19] ibid

7.3 5G NGC-based EC Architecture

The high-level architecture of NGC support for MEC is detailed below as defined by 3GPP TS 23.501:[20]

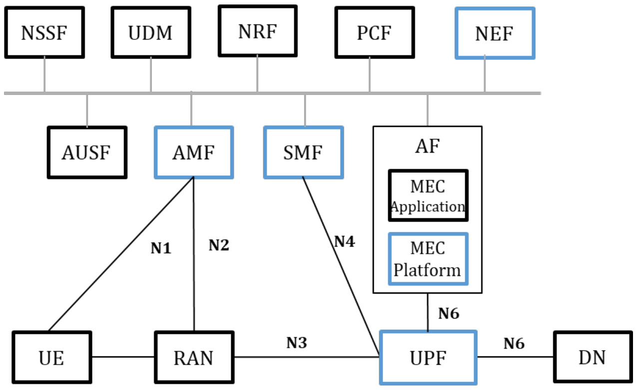


Figure 7–7 3GPP 5G NGC-based EC Architecture

3GPP has many functions that facilitate extremely powerful and flexible MEC deployments and applications. They include:

- **User plane (re)selection** – The 5G core network can select and reselect the user plane function (UPF) to route the user traffic to the local data network (DN).
- Provisions for application functions (including MEC applications) to influence UPF (re)selection and traffic routing via policy charging function (PCF) or network exposure function (NEF)-related interfaces.
- **Application triggering services** – These allow the application function (AF) to contact the UE or the UE to contact the AF – for example, push services in both directions. (NB: a UE could also be an AF.)
- **Local routing and traffic steering** – The 5G core network can explicitly select traffic sets to be routed to specific applications in the local data network, including the use of a single PDU session with multiple PDU session anchor(s) (for example UL CL/IPv6 multi-homing). This includes the ability for an application function to influence routing and steering internally via NEF for external AFs.
- **Three sets of session and service continuity** – These are relevant to UE and application mobility options.
- **Network capability exposure** – A 5G core network and application function allowing them to provide information to each other – via NEF for external AFs and directly for internal AFs.
- **QoS and charging** – Here the PCF provides rules for QoS control and charging for the traffic routed to the local data network.
- **Support of local area data network (LADN)** – Here the 5G core network provides support to connect to the LADN in a certain area where the applications are deployed. LADNs in Release 16 are defined via a set of TAIs.

[20] [6-4] 3GPP Specifications System architecture for the 5G System (5GS) TS 23.501

- **Support of time-sensitive networking** – In this case, the 5G core network provides support to connect to the LADN in a certain area where the applications are deployed. LADNs in release 16 are defined via a set of TAIs.
- **Extended set of PDU session types** – IPv4, IPv6, IPv4v6, Ethernet and unstructured. Ethernet and unstructured are new.

7.4 Wi-Fi-related Architectural Considerations

Edge computing platform benefits can be made available to applications over the Wi-Fi access network as well. This is particularly relevant in enterprise or venue deployments with Wi-Fi installations, where enhanced audio-visual content (such as augmented reality) needs to be delivered to an audience. The edge computing platform provides a low-latency, high-bandwidth environment which is a key enabler for such applications.

The edge computing platform introduces requirements – and offers unique design choices – for the Wi-Fi network. Along with the applications, the resources on the edge computing platform can be shared with network VNFs such as the Wi-Fi controllers.

The edge computing platform needs to interface with the underlying Wi-Fi radio access network (via Wi-Fi access points or controllers) to collect information on the network state and parameters. The WLAN information, either in the native form or the insights derived by analytics on these inputs, can be exposed to the application layer to tailor behavior (for example, adapting content) for improving network and application performance.

ETSI MEC is working on a standardizing the interface to expose the WLAN information to the application[21] from the Wi-Fi network elements.

Wi-Fi-based EC solutions can be provided by the MNO community or jointly with MNOs and public network operators (PNOs). These solutions are described in the next section.

7.4.1 Public (MNO) Wi-Fi Edge NW

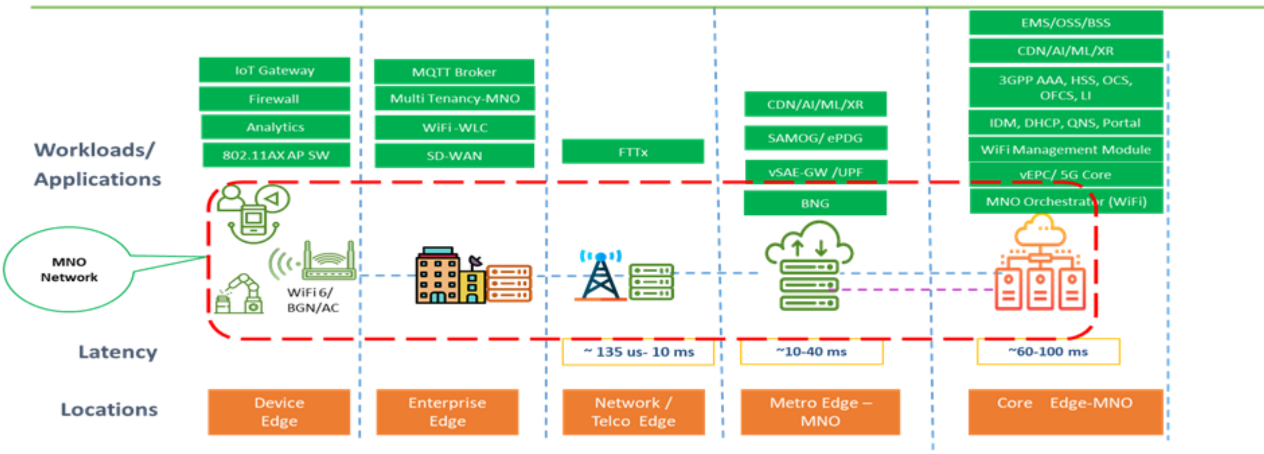


Figure 7–8 MNO-offered Wi-Fi EC solutions

[21] ETSI-MEC paper, "MEC in 5G networks"

Figure 7–9 describes the functional architecture of the edge functions in an MNO-offered end-to-end Wi-Fi network. Depending on the nature of the applications and services needed, the edge computing functions can be distributed across the different network nodes, including the Wi-Fi access point (AP).

The standalone Wi-Fi AP device offers the compute platform to host device edge supporting functionality such as AP data and control and management plane functions, as well as value-added functions like IOT gateway and security or firewall services.

In addition to hosting the wireless LAN controller (WLC) functions, the enterprise edge provides services like multi tenancy, message queuing telemetry transport (MQTT) broker and SD WAN, creating a platform for B2B use cases.

A metro edge owned by an MNO supports BNG, SAE GW, UPF, SAMOG, ePDG, CDN, and AI/ML modules.

The MNO core edge platform provides a common platform that supports the 4G/5G core network functions and functions specific to Wi-Fi network management and security (IDM, DHCP, QNS, portal). It offers an ideal platform to implement functions for Wi-Fi interworking with 3GPP, including authentication and charging (3GPP AAA, HSS, OCS, OFCS, LI) as well as OSS and BSS services.

A common distributed edge architecture for 4G/5G and Wi-Fi network with a common orchestrator simplifies enterprise service deployment and maintenance across access networks.

7.4.2 Public (MNO) Private (PNO) WiFi Edge Network

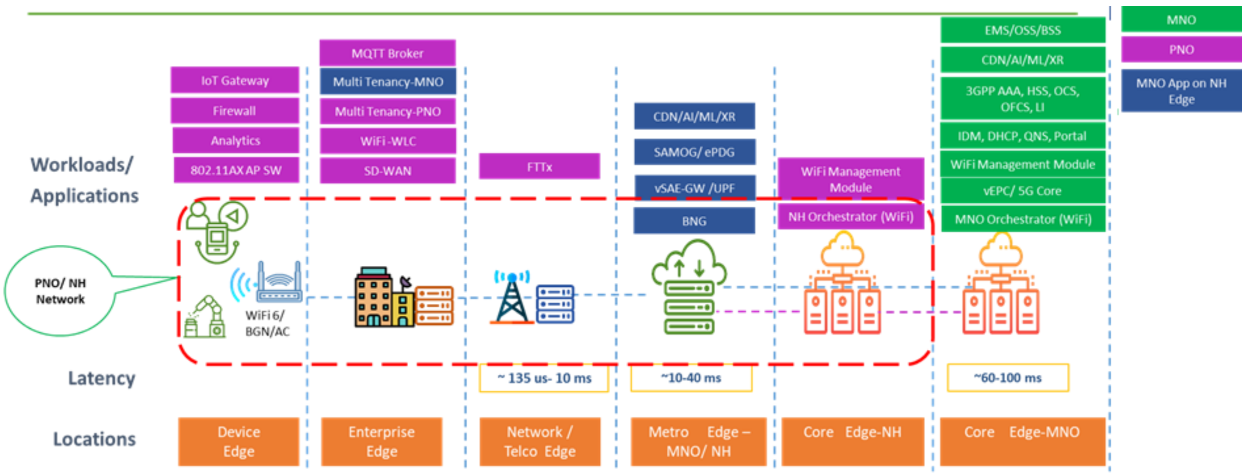


Figure 7–9 Joint MNO and PNO-offered Wi-Fi EC solutions

The Wi-Fi edge architecture for the joint MNO and PNO-offered Wi-Fi network is described above.

The metro edge can be owned by the neutral host (NH) and/or the public or private operator. Based on the partnership between the 5G network and the NH service provider, the service orchestrators of the two providers will need to interwork. When an NH partner owns a multiple cluster of networks with different business modules, it will own the unified orchestrator for functions like end-to-end SLA management and billing as well overall Wi-Fi network management functionality.

7.5 Recommendations

This chapter has presented the potential edge network architectures being standardized and deployed: distributed data plane, distributed control plane and distributed control plane with distributed subscriber information.

Edge network functions must be based on standardized 3GPP interfaces: CUPS architecture was defined in 3GPP Release14, but vendors already had proprietary solutions for control/data splits to provide the required throughput, performance and scalability. Vendors were reluctant to move their efficient *de facto* internal architecture to a *de jure* standard for CUPS. However, for a healthy growth of the edge computing industry, the solutions should move towards standardized CUPS based architectures.

Open edge-core interfaces should be encouraged via necessary interface enhancements and plugfests to enable multi-vendor edge computing eco-system: The edge-core interfaces, standardized by 3GPP, are Sx and N4 interfaces for LTE and 5G respectively. An efficient, verified interoperability for these interfaces between vendors requires very close cooperation. Furthermore, the performance of core network implementation of CUPS architecture, throughput and scalability are also closely linked to tight integration of the edge and the core via these interfaces. This interoperability had not been in huge demand by the industry or MNOs for some time, but the trend is now changing as more and more edge solutions are being deployed.

Accordingly, the edge-core interfaces must be enhanced as required to enable true multi-vendor operation and organizations such as SCF should organize plugfests to validate such interoperability. Additionally, there will be significant advantages for the control/data plane split solutions if dedicated hardware /software 4G/5G data switches are developed by data switch companies, which should also be encouraged.

Security of the CP-UP interfaces is of paramount importance: An unauthenticated and unencrypted control plane channel to the user plane should provide, at a minimum, authentication vectors between elements, particularly when those elements are distributed and could come from different vendors. Any control interfaces should also employ some form of cryptographic authentication, and include cryptographic transport (if not for privacy, then for integrity). While these aspects are tackled in the context of the 5G-N4 interface, there are gaps for the LTE-Sx interfaces which must be addressed.

Solutions resilient to backhaul disruption should be standardized and developed: CUPS architecture, with a split where the data plane is at the edge, provides less resilience than when the control plane is shared at the edge. Standards need to be defined to have control plane information shared or distributed between edge and core NF.

Statefulness for EC mobility will be crucial in evolving EC systems: The current edge architectures cover UE mobility from the edge to the core network, but have not yet addressed mobility from one edge to another edge when the control plane is at the edge.

8. Edge Platform Services and Applications

The edge computing platform brings together the network and application layer domains that enable a whole range of new applications. The platform is an enabler for a new class of use cases with support to host latency-sensitive, bandwidth and compute-demanding application functions close to the end-user device. It also enables high performance by exposing network awareness and configuration as a service to the applications.

8.1 Platform Services

The platform services obtain information from different sources – both network functions and the devices (for instance, by using signalling between the UE and edge app). Figure 8–1 below illustrates the interactions through which the edge platform services layer is realized in different deployments.

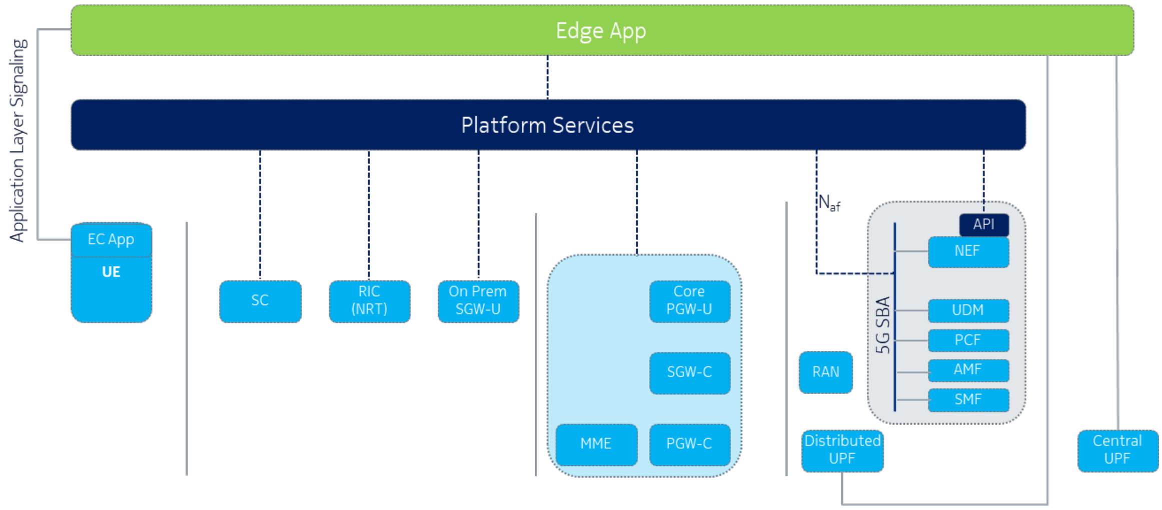


Figure 8–1 Realizing edge platform services

Information Source	Interface to platform services layer	Remarks
UE	Application layer signalling	Client app on UE conveys information to corresponding server application on the edge platform
SC, RIC, on-premises SGW	Direct	The information sources and network functions are also at the edge
MME	API gateway/3GPP network exposure function	
5G SBA elements	Naf	
Other 5G network functions	NEF APIs	

8.1.1 Radio Network Analytics

The edge computing platform obtains information through its interfaces with the underlying radio network and applies network analytics to derive insights related to network conditions. These insights are used to tailor application behaviour and transport options. The availability of real-time network information is key for the viable support of performance-sensitive applications over wireless networks. As a potential implementation option in the ORAN architecture, the radio network analytics function can be implemented as an NRT RIC workload.[22] ETSI MEC has developed a specification for exposure of APIs from cellular LTE[23] and WLAN networks[24].

8.1.2 Real-time Network-based Localization and User/Device Tracking

Reliable localization and device identification are key functional requirements for factory automation and enterprise use cases involving precise control and content delivery. This is particularly important given the limitations of the global navigation satellite system (GNSS) in indoor deployments.

Edge computing apps can communicate with the application on the user equipment to obtain UE-measured location information. Therefore, the edge computing platform can support high-accuracy hybrid localization information exposure by leveraging information collected from the user and interfaces to the underlying network-based localization functions. This platform service can build on work already done in this area – for example, ETSI MEC location[25] and user identity[26] APIs.

8.1.3 Application and Function Placement

The edge computing platform provides a secure application life-cycle management framework to enable easy onboarding, configuration and monitoring of applications. Given the dynamic nature of the network conditions and user location, the platform should support application relocation and distribution of functions between the edge clouds and/or the central cloud infrastructure.

In addition, edge computing resources are usually high value compared to centralized cloud resources. Therefore, the edge cloud platform needs to expose services to enable dynamic relocation of functions not only between the edge and central clouds but also between the device and edge clouds. ETSI's UE application interface work[27] specifies APIs that enable client-side applications to access edge compute life-cycle management services to get a list of supported applications (that is, applications that can be offloaded from a device to the edge), as well as instantiation and migration of applications on the edge.

As an example, the key requirements for an immersive gaming experience are the support of continuous interaction between the user and the gaming server (in other words the key/mouse/joystick inputs to the game server), and real-time feedback –

[22] O-RAN Alliance "O-RAN Non-real-time RAN Intelligent Controller & A1 interface Use Case Requirements" Version 1.0 - June 2019 (ORAN-WG2.Use Case Requirements v01.00) O-RAN Alliance "O-RAN Non-real-time RAN Intelligent Controller & A1 interface Use Case Requirements" Version 1.0 - June 2019 (ORAN-WG2.Use Case Requirements v01.00)

[23] ETSI-MEC paper: ETSI GS MEC 012, "Multi-access edge computing (MEC); Radio network Information API".

[24] ETSI-MEC paper: ETSI GS MEC 028, "Multi-access edge computing (MEC); WLAN Information API" (Work In Progress)

[25] ETSI-MEC paper: ETSI GS MEC 013, "Multi-access edge computing (MEC); Location API"

[26] ETSI-MEC paper: "Mobile edge computing (MEC); UE Identity API"

[27] ETSI-MEC paper: "Multi-access edge computing (MEC); UE application interface"

for example, the other player's position or new objects in the game environment. Edge computing is an ideal platform to onboard such functions, because of its ability to support the compute needed within the stringent delay requirements. Also, given the increasingly dynamic nature of the AR and VR components in the game (Pokemons inserted into a gamer's real-world view, for example), in-game data-processing computing needs and complexity have grown. This situation creates a compelling need not only to offload some of the device-hosted functions to the cloud, but also to ensure they are accessible within latency boundaries.

The edge computing platform should enable cloud infrastructure for the entire ecosystem of the distributing gaming service – network operators, users, gaming companies and mediators.

8.1.4 Multi Access Network and Traffic Management

Many edge computing use cases, such as factory automation, are expected to be deployed with multiple access technologies. The edge computing platform not only has to abstract but to make the best use of multiple access network paths. The edge computing framework should leverage what is known about different application traffic streams in order to maintain desired performance levels and maximize network resource utilization. This platform service should work across a number of different multi-connectivity technologies such as ATSSS (3GPPATSSS)[28], MPTCP (IETFMPTCP)[29], MPQUIC (IETFMPQUIC)[30] and MAMS (IETFMAMS)[31] to enable its application across a wide variety of deployments.

8.1.5 Integration with Network Slicing

Integration with a network slicing framework allows the edge computing platform to provide services to the applications for on-demand creation and configuration of network slices.

The end-to-end network slice consists of access, transport and core segments. The edge computing platform can access functions that manage and control network infrastructure segments, as well as managing the edge network functions. The ability to configure slices can be leveraged by applications to get the desired network configuration for best performance. Implementation of such a platform service will require cross-industry collaboration – which is already under way. For example, the ETSI MEC standards group is working on identification of necessary enhancements to the API framework for support of network slicing, during which it is considering the findings from SDOs, industry collaborations and regional projects (e.g., ETSI NFV, 3GPP, NGMN, 5G-Transformer, 5G-Coral).

8.1.6 Value-adding services

The platform services discussed above can be seen as network-type services. The EC platform may provide value-adding services, such as machine learning functions,

[28] 3GPP TR 23.793 - Study on access traffic steering, switch and splitting support in the 5G System (5GS) architecture

[29] IETF RFC 6824, TCP Extensions for Multipath Operation with Multiple Addresses

[30] IETF draft Multipath Extensions for QUIC (MP-QUIC), draft-deconinck-quic-multipath-03

[31] IETF RFC 8743, Multi-Access Management services (MAMS)

video analytics functions, graphics functions and security functions. These can assist and greatly enhance the performance of edge applications.

8.2 Applications

There are many edge applications and the list is growing. There is no established way to classify or organize them. However, this is a rough guide.

8.2.1 Application Categories

1. Enterprise, venue and consumer applications

- Examples of enterprise edge applications are enterprise intranet access via LBO, AR/VR and IOT gateways.
- Potential venue applications include instant video replay in sports venues and security applications.
- Potential consumer applications include AR/VR, and autonomous vehicles.

2. Revenue-generating and cost-saving applications

- Potential revenue-generating applications include IOT gateways for smart cities and local content applications for hospitals.
- Potential cost-saving applications include applications resulting in backhaul savings by local offload and/or compression or edge analytics for improved network troubleshooting.

3. User experience-enhancing, network optimization and local applications

- Examples of user experience-enhancing applications are improved video quality by end-to-end optimization and improved gaming application experience due to low latency.
- Potential network optimization applications include backhaul bandwidth optimization applications.
- Potential local applications include content applications in museums.

4. Pass-through, locally terminating and. locally originating applications

- This classification is based on architecture of the traffic flow and is shown in Figure 8–2 below.

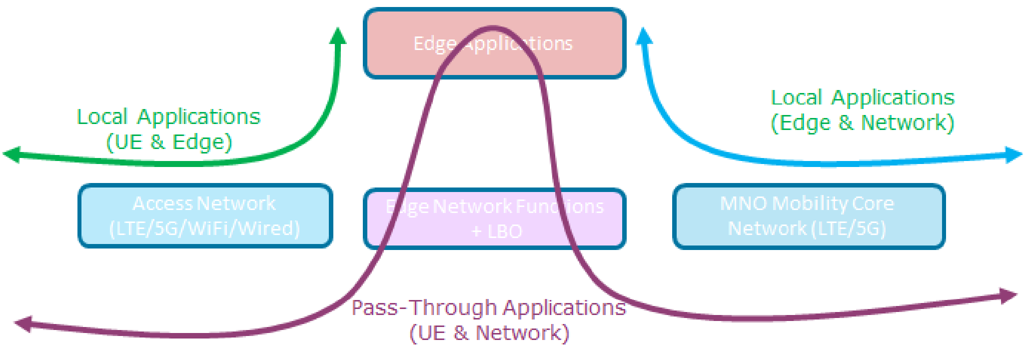


Figure 8–2 Traffic flow-based application types

8.2.2 EC application examples

Some examples of edge computing applications are:

LBO to enterprise

- This is possibly the simplest EC use case. Here, locally broken out data is made available to the enterprise IT network for accessing enterprise servers.

AR/VR

- These applications are latency sensitive. EC can be used to meet these low-latency requirements.
- These applications are also compute intensive. Some computing can be offloaded to the EC platform. Only essential computing needs to be performed at the UE or device.

Autonomous vehicles

- These applications are latency sensitive. EC can be used to meet the low latency requirements.
- These applications are also compute intensive. Some computing can be offloaded to the EC platform. Only the essential computing needs to be performed at the vehicle.
- EC can also provide network-based road vision to the vehicle, which can be used in conjunction with its own autonomous vision systems, leading to safer implementations.

Video compression/video optimization

- Video compression using compute-heavy algorithms can be implemented in the EC platform without burdening the UE or device
- Video stream transmission can be optimized by the EC application based on radio and backhaul transport qualities.

Presence determination and personalized adverts

- Using small cells and/or Wi-Fi APs, the presence of users or devices can be determined by the EC application and personalized ads, among other notifications, can be targeted.

Location tracking

- Using small cells and EC resources, indoor location can be accurately estimated and tracked. This can be useful in industrial scenarios, where both objects and personnel can be tracked for productivity, security and safety purposes.

Security

- Especially in enterprise settings, enterprise security policies and firewalls can be implemented as EC applications for mobile traffic as well. (Note that Wi-Fi traffic can generally be controlled by enterprise policy servers but that mobile traffic tends to be transparent to the enterprise networks, in the absence of EC.)

Robotic control

- In industrial environments, latency-sensitive LTE-based robotic and machine control can be achieved using LBO and EC-based control applications.

Content caching/distribution

- Locally generated content (for example in sports or music events) can be distributed locally to edge users (such as visitors to the events).
- Using the storage capabilities of the EC platform, content that is relevant to the edge users (training material in enterprises, say, or medical images in a hospital setting) can be made available to edge users via the EC platform.
- Using edge storage capabilities, popular content can be either transparently cached or distributed via a CDN framework and made available to edge users more efficiently, reducing backhaul costs.

IOT gateways

- In smart cities, for example, EC solutions can be used to implement IOT gateways, in order to perform functions such as message aggregation/compression, message store and forward and protocol conversion.

8.3 Recommendations

In collaboration with appropriate SDOs, SCF should drive the development of a set of harmonized and consistent application and network APIs to enable small cell networks to facilitate EC services and applications, via the following steps:

- Explore open edge network APIs for the small cell network to offer radio network analytics as part of the EC applications and/or to the EC platform services.
- Develop a common API framework to facilitate a microservices environment – a service registry that allows apps to communicate and exchange services with HTTP RESTful-based APIs.
- Explore the development of edge computing platforms that support a 'bring your own service' concept with API guidelines that allow description of any service or any new API.
- Explore collaborating with appropriate industry-forums for application APIs, system-level APIs and edge infrastructure APIs etc.

9. Edge Computing Platform Solutions

Section 3 introduced the functional framework for edge computing. This section addresses the implementation of solutions for the EC framework: platform, network, services and applications.

As the industry is transitioning from a virtual machine (NFVI)-based infrastructure to a complete container-based infrastructure, open source initiatives like Akraino are beginning to play a major role by defining the framework and blueprints for software platform packaging.

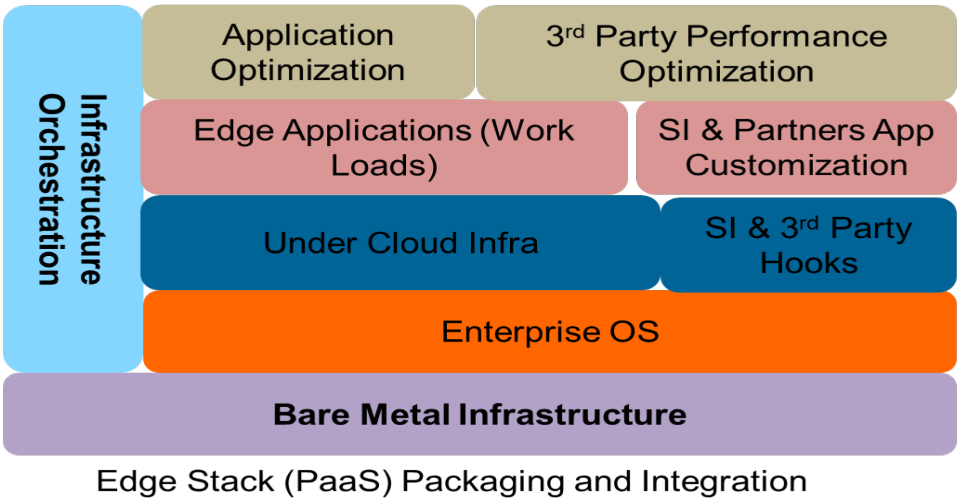


Figure 9–1 Software packaging for edge platforms

The industry faces a number of challenges to building an edge software stack based on open stack offerings. They include:

- Major open source framework and blueprints have chosen their host OS based on their past experience with the OS vendor. However, blueprints and frameworks packaged with the open source OS (like Ubuntu or CentOs) may not be a good fit for a specific application, which needs an enterprise-grade OS.
- Open source under cloud platforms is not a complete package yet, and still has to rely on a lot of integration.
- Most open source frameworks are working on the MANO (LCM, domain NSO) and automation tools. The packaging still heavily depends on open source tools like Calico, Ceph, Atomic Host, Runtime Docker Container package, Kubernetes, and Helm Chart integration.
- The infrastructure may also need the support of hosting VMs. Openstack Helm is not fully tuned and customized.
- Applications/workload integration and performance optimization are major challenges for an open source-based solution.
- Domain orchestration API hooks are not mature enough to integrate with ONAP E2E orchestration and need a lot of system integration.

Despite all the challenges, an open source framework and blueprint initiatives and projects are attractive, since the PaaS offering is getting hardened, with multiple projects being built around it, many of which are close to initial trials. A very cost-effective solution can be built complying with Apache 2.0 licensing.

The various MEC platforms all across the industry (wired, wireless, and enterprise) – from the centralized data centre to customer premises – may also have a variety of real-time and non-real-time requirements. Hence it is important to have a very flexible architecture (see Figure 9–2 below).

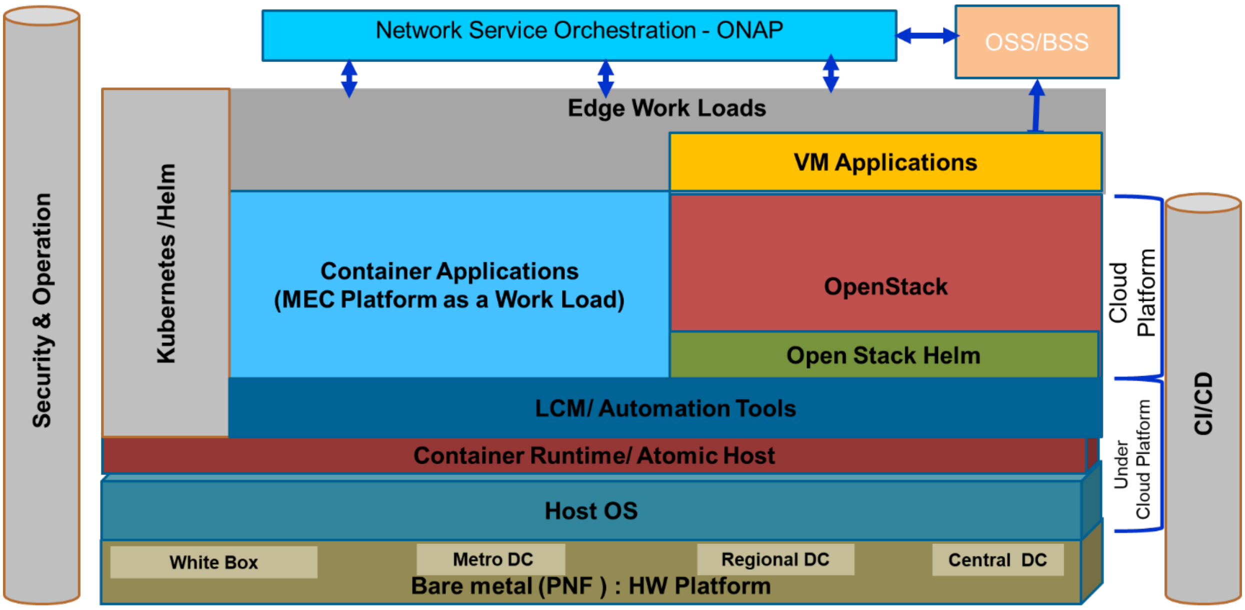


Figure 9–2 Generic EC solution framework

9.1 Opensource Framework and Edge Stack

There are multiple open source projects that provide the component capabilities required for edge computing. However, there is no holistic solution to address the need for fully integrated and deployable edge infrastructure.

Akraino Edge Stack, a Linux Foundation project initiated by AT&T and Intel, intends to develop a fully integrated end-to-end edge infrastructure solution with different recipes targeting multiple sizes of edges, deployments and use cases; the project is completely focused on edge computing.

This open source software stack provides critical infrastructure to enable high performance, reduced latency, improved availability and lower operational overhead. It also aims to provide scalability, address security needs, and improve fault management. The Akraino community will address multiple edge use cases and every segment of the wired and wireless telecommunications industry. The Akraino community has developed two releases so far with multiple blueprints. Each blueprint addresses various use cases and edge deployments.

AT&T has contributed a seed code to enable carrier-scale edge computing applications to run in virtual machines and containers. The combination of Intel's Wind River Titanium Cloud components with Open Stack provides Kubernetes infra controller (k8s) support with Docker runtime life cycle management (LCM) to the Akraino framework and relevant blueprints.

The Akraino project is a complementary open source project, and interfaces with the existing projects: Acumos AI, Airship, Ceph, DANOS, edgeX Foundry, Kubernetes, LF networking, ONAP, OpenStack, Homeedge, Project EVE and StarlingX. Project Eve and

Homeedge – along with Akraino – are part of the newly formed Linux Foundation edge (LFE) umbrella project. See Figure 9–3 below.

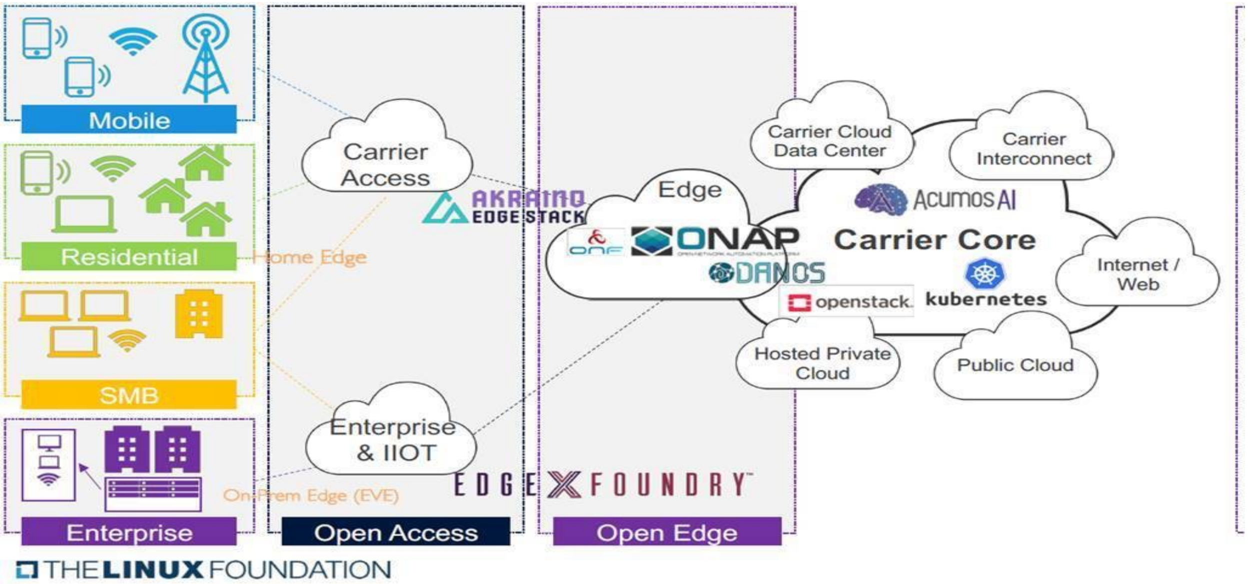


Figure 9–3 Ecosystem of open source projects in support of EC

Of the components making up the ecosystem of open source software, blueprints are developed as solutions for various use cases.

Examples of blueprints include:

- Radio edge cloud (REC) – Akraino radio edge cloud (REC) provides an appliance tuned to support the O-RAN Alliance and O-RAN software community's radio access network intelligent controller (RIC). It is the first example of the telco appliance blueprint family.
- Network edge cloud (NEC)
- Integrated edge cloud (IEC)
- Provider access edge
- Far edge distributed cloud
- SDN-enabled broadband access (SEBA)
- Connected vehicle blueprint – Connected vehicle blueprint focuses on establishing an open source MEC platform, which is the backbone for the V2X application.
- Edge lightweight and IoT (ELIOT) blueprint family – This blueprint provides edge computing solutions for smart office scenarios. It manages intelligent devices, delivers AI training models and configures a rules engine through cloud/edge collaboration.
- The AI edge blueprint family – The AI edge blueprint mainly focuses on establishing an open source MEC platform combined with AI capacities at the edge, which could be used for safety, security, and surveillance sectors.
- Integrated cloud native (ICN) family – This family is introduced in Akraino 2.0. The ICN family of blueprints focuses on cloud native-based edge deployments. They are meant for edges that require the co-existence of network functions and applications. They support containerized and virtualized network functions. This family of blueprints is generic and supports multiple use cases such as disaggregated 5G RAN, 5GC, AI and visual computing. The ICN family of blueprints includes multiple features to

support a wide variety of edge use cases. Some notable edge features that are also part of the ICN family are:

- Multi-cluster application orchestration (ONAP4K8s, a cloud native edge profile of ONAP). This is meant to deploy geo-distributed network functions and applications across multiple edges and clouds. It also enables inter-application connectivity by automating the service meshes in edges.
- Software-defined edge WAN. The ICN family includes a next generation cloud-native SD-WAN – called SD-EWAN opensource – to provide secure overlays among edges. This is needed to provide traffic routing across multiple WAN links, firewall security, IPsec-based overlays, anti ddos protection via traffic shaping and policing, and source and destination network address translation (NAT) for addressing edges with overlapping network addresses.
- SmartNIC-friendly OVN-based networking controller. The ICN family also includes a next-generation OVS-based network controller (OVN-for-K8s-NFV, an open source project in OPNFV) to provide support for provider networks, multiple virtual networks and service function chaining that is required for any MEC applications.
- Accelerator and security services. The ICN family includes multiple platform acceleration and platform functions from the OpenNESS open source project that help in improving the performance and security of both edge network and edge applications. These functions are critical as many edges are resource constrained. Some of the critical functions include CPU manager for core dedication and affinity, NUMA-aware scheduling, SRIOV-NIC, SRIOV-Crypto, SRIOV-FPGA and AI inferencing acceleration services.
- Distributed UPF (dUPF). The ICN family also plans to include dUPF open source as it is felt that UPF is needed for local breakout and traffic steering to local MEC applications.

Figure 9–4 below shows a network cloud blueprint, which provides a base PaaS to build and customize specific infrastructure and flexibly embrace the structure and framework for multiple variations of edge platforms.

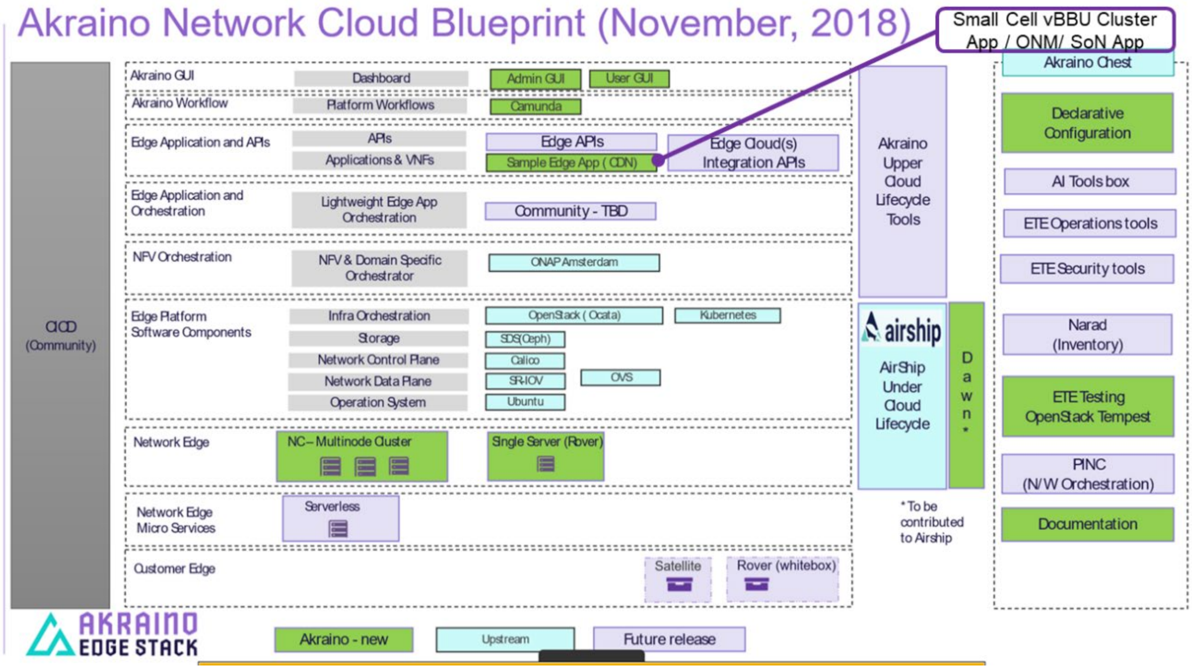


Figure 9–4 Network cloud blueprint

A point of delivery (POD) is the method through which a blueprint is deployed to an edge site. For example, an edge location could have a single server or multiple servers contained in one or more racks. As the blueprint uses the declarative configuration, point of delivery (POD) allows organization for the deployment. POD also allows repeatable methods to deploy on a larger scale (10,000-plus locations) at a reduced cost.

9.2 User Plane Acceleration and Customization

For carriers as well as enterprise service providers, there is a common challenge of building an edge fabric which is a good fit for all possible deployment scenarios. Ultimately the industry is looking for edge PaaS, where the automation tools offload the customization of user plane acceleration needs for real-time and non-real-time applications and, based on the use case, the connectivity requirements are managed by dynamic plug-and-play-based hardware (GPU, Smart NIC) platform as well as user plane control modules.

The P4Runtime Project from ONF is an initiative that brings vendor/silicon-independent fabric and VNF/container offloading. The project takes care of programmable switches (portable switch architecture) and data planes.

9.3 E2E Orchestrator Integration

ONAP (Casablanca or Dublin releases) is a good starting point for network service creation and orchestration of all network enabler or user application workloads in VNFs and container forms. A centralized or distributed ONAP model should be supported.

Every edge application needs to be packaged to ensure ONAP compliance and support local agents for LCM notification and services to ONAP (for example, a data collection alarm and events module or an application controller).

9.4 Recommendations

These recommendations are based on the overview of EC infrastructure (both premises-based and network-based) that appear above.

1. **Security** – Security between enterprise network and premises-based EC infrastructure should be carefully considered and implemented. This may require negotiation between the enterprise (for example the IT department) and the entity installing and managing the EC infrastructure (for example the MNO or neutral host company). Security considerations would include firewall rules and access control schemes among others. SCF to establish a work item to create a best practice blueprint.
2. **VMs and containers** – As a matter of best practice, infrastructure options (in both premises-based and network-based contexts) must include white box solutions, either as simple bare-metal solutions (with no software running on them) or virtualization platforms (with virtualization software supporting, for example, VMs and/or containers).
3. **Virtualization** – Virtualization framework options include VMs and containers; VMs tend to pre-date containers. However, VMs can sometimes (depending upon the EC use case) be unnecessarily heavy and may also not have the dedicated computing, storage or networking resources, when needed by the application. In such cases, containers (realized and orchestrated by Dockers and Kubernetes respectively) are useful alternatives. SCF recommends the use of containers depending on the characteristics of the applications, such as:

    - Bandwidth-intensive applications
    - Latency-sensitive applications
    - High-reliability applications
    - High device density applications

4. **Slicing** – Vendors must ensure EC infrastructure (both premises-based and network-based) that needs to support different types of applications (such as those listed above) also supports slicing capabilities wherever possible.
5. **Openness** – Currently, there are initiatives in the industry toward specification of open hardware. Examples include the Facebook Open Compute project, which is suitable mainly for large data centers. There are other initiatives, such as those started by ARM, for developing open hardware specifications for smaller footprint computing platforms. We recommend the development of such open specs suitable for EC infrastructure. SCF to work on this in collaboration with the appropriate SDOs.

10. Conclusion

The aim of this edge overview has been to address edge computing, small cell networks and the synergies between them.

It started by providing an extensive overview of edge computing, by first putting forth an end-to-end EC framework. The EC framework consists of a four-layer taxonomy for edge computing: EC infrastructure, edge network functions, EC platform services, and EC applications. The role of edge management and orchestration was also addressed. This taxonomy leads to various types of EC-based service offering possibilities, including IaaS, NaaS, PaaS, SaaS and managed services. Network-as-a-service (NaaS) in particular is a unique service offering that is made possible by edge computing technology.

The paper further developed the details of each of the layers in the EC framework. Each of these detailed discussions includes a set of recommendations for the small cell and edge computing industry.

Some of these recommendations will be the basis of future work items in the Small Cell Forum, collaborating with appropriate standards development organizations and/or industry forums.