ETSI White Paper No. 24

MEC Deployments in 4G and Evolution Towards 5G

First edition – February 2018

ISBN No. 979-10-92620-18-4

Authors:

Fabio Giust, Gianluca Verin, Kiril Antevski, Joey Chou, Yonggang Fang, Walter Featherstone, Francisco Fontes, Danny Frydman, Alice Li, Antonio Manzalini, Debashish Purkayastha, Dario Sabella, Christof Wehner, Kuo-Wei Wen, Zheng Zhou

ETSI
06921 Sophia Antipolis CEDEX, France
Tel +33 4 92 94 42 00
info@etsi.org
www.etsi.org

About the authors

Fabio Giust
NEC - Editor

Gianluca Verin
Athonet - Editor

Kiril Antevski
UC3M

Joey Chou
Intel

Yonggang Fang
ZTE

Walter Featherstone
Viavi Solutions

Francisco Fontes
Altice Labs

Danny Frydman
Saguna

Alice Li
Vodafone

Antonio Manzalini
TIM

Debashish Purkayastha
InterDigital

Dario Sabella
Intel

Christof Wehner
Artesyn

Kuo-Wei Wen
ITRI

Zheng Zhou
Huawei

Contents

Introduction

Multi-access Edge Computing is regarded as a key technology to bring application-oriented capabilities into the heart of a carrier's network, in order to explore a wide range of new use cases, especially those with low latency requirements. When it comes to deploying MEC, there are many potential scenarios where MEC can fit in, and – as the name clearly spells out – these are not limited to 4G or 5G at all! As a universal access technology, MEC offers itself to any application that has locality requirements like a shopping mall or a sports arena, or wherever low latency is required such as 5G V2X or autonomous vehicle applications.

Nevertheless, starting from the fact that the MEC's original target was the mobile network, when it comes to its deployment, MEC is often considered as a 5G-only feature. However, 4G is expected to still be successful in the years to come, thus a large part of the industry is working towards running MEC in existing 4G networks. In fact, the MEC reference architecture, defined in ETSI GS MEC 003 [1], is agnostic to the mobile network evolution, so that a MEC host deployed in a 4G network can be reused to support 5G services as well.

Therefore, understanding the impact of deploying an ETSI MEC system into current 4G and future 5G systems is crucial for mobile network operators (MNOs) in order to carefully plan their network upgrades. This way, MEC can be not only a technology ready for 4G, but also a driver to motivate 5G adoption, as it can allow operators to retain the investment made in 4G deployment. Indeed, from a mobile evolution perspective, products based on current MEC specifications can be smoothly migrated to support 5G networks through software update. This way, flexibility in the deployment architecture allows planning for the introduction of MEC services as the milestone to build the edge cloud, which is key for the success of 5G services such as URLLC (Ultra Reliable Low Latency Communications).

In light of the considerations above, the purpose of this white paper is to show the compatibility of an ETSI MEC system with 3GPP 4G and 5G architectures, aiming at:

- shedding some light on the potential deployment options available for operational 4G systems;
- providing a technical insight of MEC operations under such scenarios;
- showing how the creation of the mobile edge infrastructure in 4G can pave the way for 5G deployment and therefore protect investments and smoothly transit to future and more advanced service offerings.

The present document will first showcase different options to cast the MEC system into a running 4G system, maintaining backward compatibility with the 3GPP-specified architecture. In other words, such options explore how the "MEC box" can be drawn into the 4G system architecture, showing the challenges associated to each choice.

Moreover, a section devoted to the transition to 5G will demonstrate how and why deploying MEC in 4G can accelerate network transformation, leveraging on compliance to 3GPP standards and use of the cloud computing paradigm to bring future-proof added value to service providers.

Deploying MEC in 4G networks: scenarios and challenges

As per the GS MEC 011 [2] specification, a key baseline functionality of the MEC platform is to route IP packets to MEC applications which are meant to handle the traffic in the following different ways:

- In **Breakout** mode, the session connection is redirected to a MEC application which is either hosted locally on the MEC platform or on a remote server. Typical breakout applications include local CDN, gaming and media content services, and enterprise LAN.
- In **In-line** mode, the session connectivity is maintained with the original (Internet) server, while all traffic traverses the MEC application. In-line MEC applications include transparent content caching and security applications.
- In **Tap** mode, specified traffic is duplicated and forwarded to the tap MEC application, for example, when deploying virtual network probes or security applications.
- In **Independent** mode, no traffic offloading function is needed, but still the MEC application is registered in the MEC platform and will receive other MEC services, such as DNS, Radio Network Information Service (RNIS), etc.

Steering traffic to/from MEC applications is achieved by configuring the MEC's local DNS and the MEC host's data plane accordingly. From the list above, it appears straightforward that the implementation-specific details of the data plane within the MEC host (as per the MEC architecture in GS MEC 003 [3]) and the MEC platform, which is meant to program the data plane through Mp2 interface, are impacted by the point where the MEC host is installed in the 4G architecture. Many choices are possible, but all in all they can be condensed down into some base scenarios discussed in the following sections.

Bump in the wire

The expression "bump in the wire" encompasses all the scenarios in which the MEC platform installation point ranges in locations between the base station itself and the mobile core network. These options were first proposed in the MEC Introductory White Paper [1] and reproduced in Figure 1.

When the eNB implementation bundles the MEC platform into a single implementation, this latter is able on the one hand to route plain IP packets to/from the MEC applications (i.e., local switching mode), and on the other hand to route GTP-encapsulated packets to/from the Serving Gateway (SGW) for regular traffic as per the operator-configured Packet Data Networks (PDNs - i.e., the legacy S1-U mode). This deployment is very convenient e.g., in enterprise scenarios to allow intranet traffic to breakout to local services (similar to Local IP Access - LIPA), and also when MEC is co-located with a CRAN deployment (see the ETSI white paper "Cloud RAN and MEC: a Perfect Pairing" [4].)

In all the other locations, either in proximity of the radio node or at an aggregation point, the MEC platform sits on the S1 interface of the 4G system architecture. In this scenario, the MEC host's data plane has to process user traffic encapsulated in GTP-U packets, thus requiring the appropriate handling of these tunnels. This non-trivial operation poses some challenges, as a portion of the data may be generated or manipulated internally in the MEC hosts or may come from a local breakout, without passing through the core of the network. For this traffic, a dedicated solution may be required (e.g., the MEC GW in Figure 1) to handle operations such as lawful interception and charging. This solution can support CUPS, which ensures a 3GPP-compliant solution (see sections below). Also, in this solution, low

latency is supported by installing the MEC platform all the way down to the eNB, or in locations that ensure minimal latency. Additionally, it offers the capability to steer traffic on a per session and/or packet granularity, with flexible filtering support.

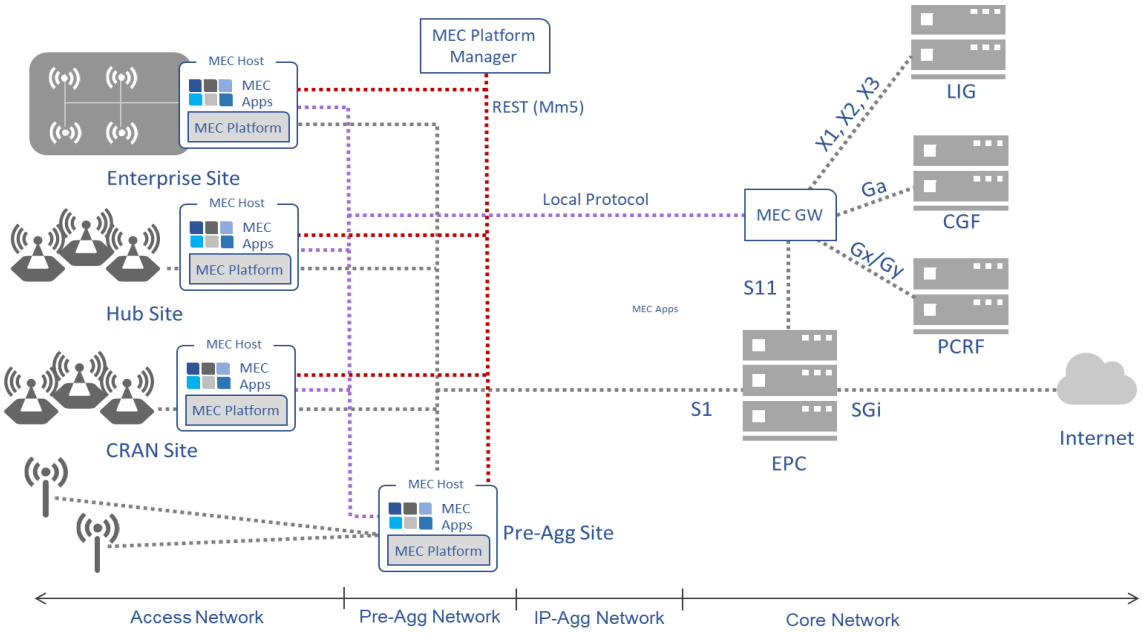

Figure 1: MEC deployment using the "Bump in the wire" approach.

Distributed EPC

Unlike the “bump in the wire” approach, in this deployment the MEC host logically includes all or part of the 3GPP Evolved Packet Core (EPC) components, as specified in the 4G system architecture in ETSI TS 123.401 [5], and the MEC data plane sits on the SGi interface. By doing so, in order to steer U-plane traffic towards the MEC system, two elements, the local DNS of MEC and the PDN Gateway (PGW) of a distributed EPC, play critical roles. In fact, as the UE subscribes to the distributed EPC co-located with the MEC host, the PGW thereupon terminates the PDN connection and assigns the IP address and local DNS information to resolve the MEC applications’ IP address. This scenario requires less changes to the operator’s network as standard 3GPP entities and interfaces are leveraged for operations such as session management, charging, etc.

This type of deployment can well serve Mission Critical Push to Talk (MCPTT), and M2M communications, where the communication with the operator’s core site is optional (see for example the upper diagram in Figure 2). In this case, the Home Subscriber Server (HSS) is co-located with the EPC as well, and there is no need for a working backhaul to keep the local service running. This type of deployment is typically used by first responders, public safety, and mission critical industrial sites.

In some other cases, the HSS is unique and centrally managed by the operator at the core site and the operator’s core site PGW can be used for some selected APNs (e.g. IMS or roaming). This allows the local management of the entire subscriber database and the use of the local EPC in the MEC to offload the entire APN traffic. Additionally, the distributed EPC offers the ability to deliver exactly the QoS and

configurability features that, e.g. an enterprise customer requests from the particular network service purchased (see the lower diagram in Figure 2).

The MEC applications can be co-located with the evolved packet core (EPC) functions in the same MEC host. This option can reduce costs as the EPC and its components can run e.g. as Virtual Network Functions (VNFs) on the same Network Functions Virtualisation (NFV) platform with the MEC components in order to improve scalability and better utilize network resources (see the example in Figure 3).

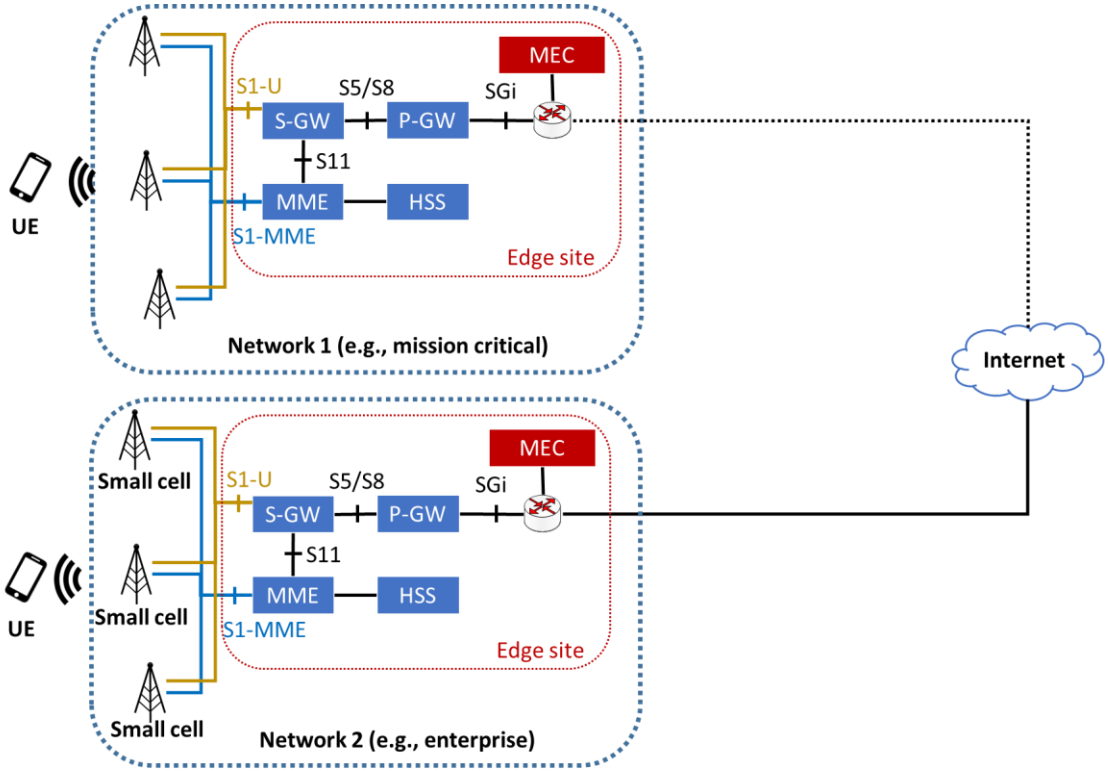

Figure 2: MEC deployment with distributed EPC

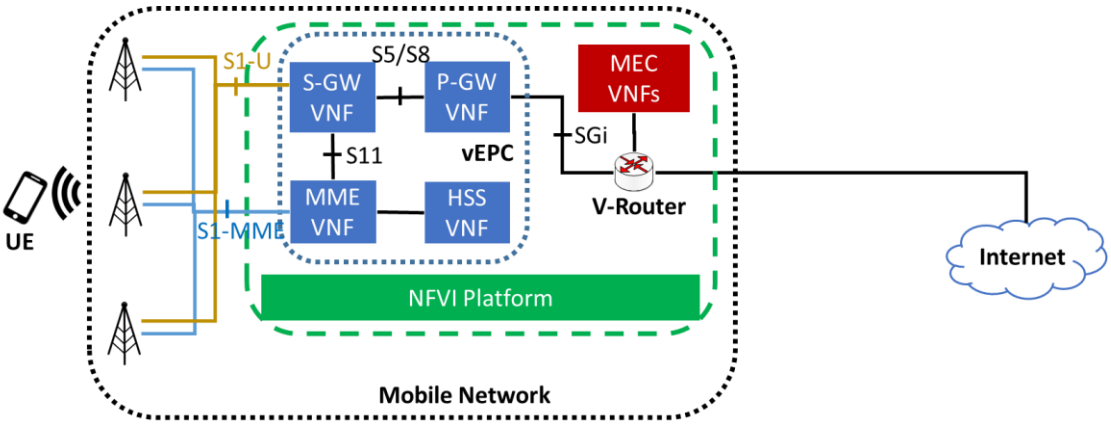

Figure 3: MEC deployment with EPC and MEC application on the same NFV platform (same MEC host).

Distributed S/PGW

The distributed S/PGW deployment option is similar to the previous one, except that only SGW and PGW entities are deployed at the edge site, whereas the control plane functions such as the Mobility Management Entity (MME) and HSS are located at the operator's core site. Still, the MEC host's data plane connects to the PGW over the SGi interface.

Similarly to the previous option with the whole distributed EPC, the SGW and PGW can also run as VNFs together with the MEC application on the NFV platform as part of the same MEC host. The local SGW selection is performed by the central MME according to the 3GPP standard DNS procedures and based on the Tracking Area Code (TAC) of the radio where the UE attaches to. This architecture allows offloading the traffic based on the APN, which means, for example, that the IMS for VoLTE APN and roaming APNs may not be offloaded.

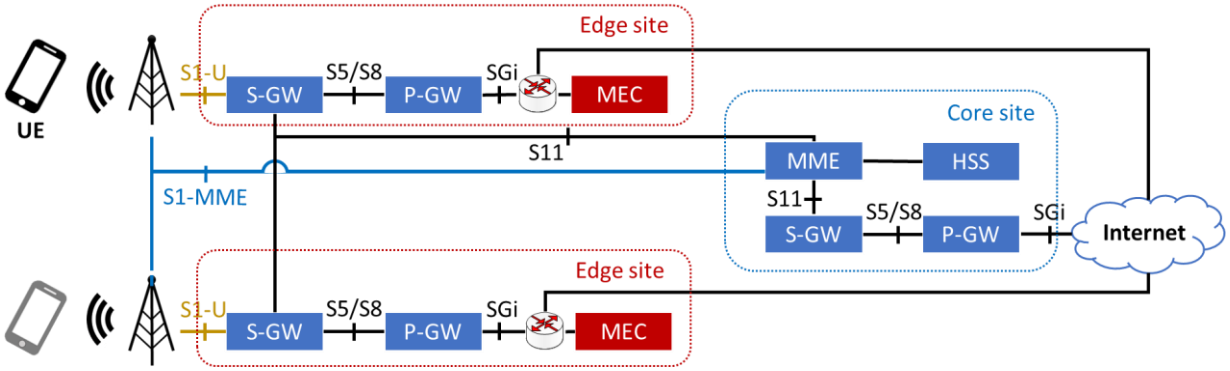

Figure 4: S-GW and P-GW MEC deployment

The diagram above shows the deployment with the SGW and PGW co-located at the network edge, which requires the operator to extend the S5 interface to the MEC site. This type of deployment allows the operator to retain full control over the MME.

Distributed SGW with Local Breakout (SGW-LBO)

Local breakout at the SGWs is a new architecture for MEC that originates from operators' desire to have a greater control on the granularity of the traffic that needs to be steered. This principle is dictated by the need to have the users able to reach both the MEC applications and the operator's core site application in a selective manner over the same APN.

With the Distributed SGW deployment, one of the optional MEC deployment scenarios is to co-locate MEC hosts with the SGW. Both the SGW-LBO and the MEC application may be hosted as VNFs in the same MEC platform. The following figure describes co-locating MEC hosts with the SGW in a mobile network where the MEC system and the distributed SGW are co-located at the edge.

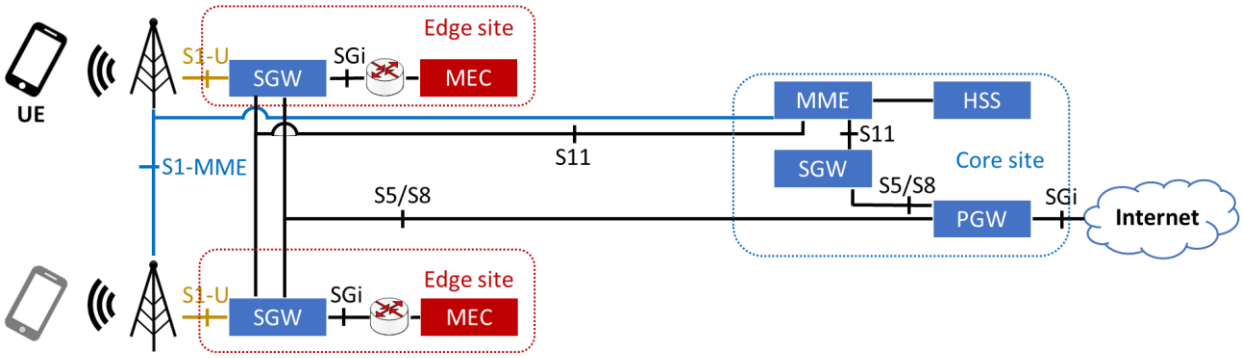

Figure 5: SGW-LBO MEC deployment

The traffic steering uses the SGi - Local Break Out interface which supports traffic separation and allows the same level of security as the operator expects from a 3GPP-compliant solution. This solution allows the operator to specify traffic filters similar to the uplink classifiers in 5G, which are used for traffic steering. This architecture also supports MEC host mobility, extension to the edge of CDN, push applications that requires paging and ultra-low latency use cases.

The SGW selection process performed by MMEs is according to the 3GPP standard and based on the geographical location of UEs (Tracking Areas) as provisioned in the operator's DNS.

The SGW-LBO offers the possibility to steer traffic based on any operator-chosen combination of the policy sets, such as APN and user identifier, packet's 5-tuple, and other IP level parameters including IP version and DSCP marking.

Control/User Plane Separation (CUPS)

The deployment options above which distribute the EPC gateways at the edge, either co-located with or within the MEC host, can also be built using the CUPS paradigm standardized in 3GPP Release 14 and have the new User Plane built in the MEC host.

Local User Plane (UP) distribution allows the use of the CUPS architecture to locally steer the traffic. SGW-C and PGW-C are the end points that populates the UP routing tables.

Challenges in the different approaches

From the deployment scenarios outlined above, a clear distinction emerges of two major categories, depending on whether the MEC host leverages the EPC packet gateways' functionalities or not. This section examines the impact of the different types of deployment scenario with respect to session and mobility management, security, charging and Lawful Interception. As expected, approaches that use standard 3GPP NFs to support MEC show the least impact.

Session management

In the bump in the wire scenario, MEC is located on the S1-U reference point. The eNB and SGW are not MEC-aware as MEC components are not involved in the standard 3GPP procedures of session management, including PDN connection setup, deletion and paging. However, it is necessary for MEC to get the UE context for the right traffic routing, which makes it more challenging. There are at least two feasible approaches to manage the UE context for MEC:

1. User plane packet inspection: MEC creates the UE context according to the S1-U tunnel IP addresses and Tunnel Endpoint Identifiers (TEID-Us) learned from the user plane packets (see also

the section below “Identifying specific subscribers at the MEC platform”). For the traffic that needs to be offloaded, MEC routes specific packets to specific applications by a traffic offload function. For traffic flows that do not need to be offloaded, MEC behaves as a transparent device. In addition, a dedicated yet not standard mechanism is necessary to trigger paging from the MEC application, e.g., for push notifications.

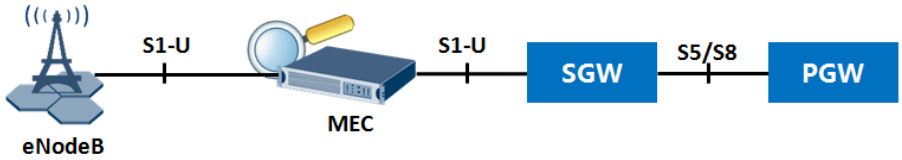

Figure 6: User plane packets inspection

2. Controlled by the PGW: The enhanced PGW controls the session management of MEC to create, update, delete the UE context and delivers the charging characteristics/LI information. Although a new reference point needs to be created between PGW and MEC, charging and LI are supported and it can be easily upgraded to the CUPS deployment mode and then evolved smoothly to 5G. The reference point between PGW and MEC will be Sx in CUPS deployment mode and N4 in 5G.

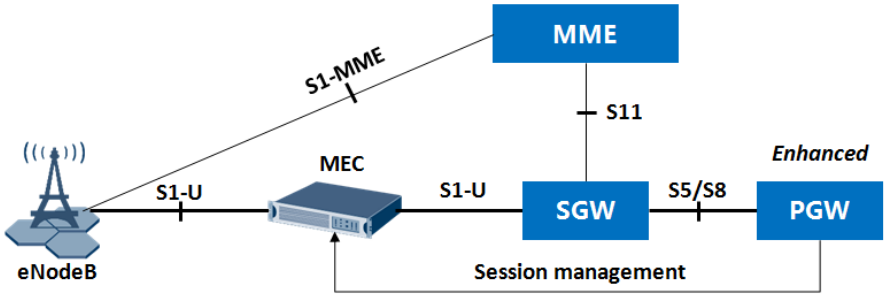

Figure 7: PGW-Controlled MEC

In the other MEC deployment options which make use of **EPC co-location,** there is much less impact on session management, as it is handled by the EPC functions installed along with the MEC host. In particular we observe the following

- EPC MEC, SGW+PGW MEC: Session management is not impacted, even for inter-MEC handover since the standard 3GPP procedures are used to keep the original PGW as anchor. This assures session continuity as well as paging idle UEs. Application level mobility is achieved by reassigning the IP address to the user or enforcing a breakout policy into the target SGW. Charging and lawful interception are supported natively by the solution.
- SGW-LBO MEC: The connectivity to the standard PGW assures the creation and deletion of the UE context similarly to the approach above.
- CUPS MEC: All the MEC options can incorporate the CUPS solution which requires a UP capable of performing traffic offload in order to steer traffic to/from the MEC applications.

Mobility management

Mobility is concerned with service continuity when the UE is moving intra or inter MEC. MEC needs to be aware of the handover of the UE in the underlying network and updates the UE context to keep the service continuity. Two scenarios appear relevant:

1. The UE moves from one eNodeB to another, but is still in the coverage of the same serving MEC host, i.e., intra MEC mobility. The MEC system should be able to route the traffic to the UE via the correct eNodeB and tunnel.
2. The UE moves out of the coverage area of the source MEC host to enter the coverage area of target MEC, i.e. inter MEC mobility or MEC Handover. This scenario may result in interruption of service to the UE. In order to provide service continuity to the UE, the MEC system needs to relocate the service to the UE from the source to the target MEC.

Depending on the selected solution, MEC Handover is handled in different ways. In the bump in the wire approach, mobility is not natively supported. One solution is to have the MEC implementation to detect the UE handover and act accordingly. An alternative solution is to update the UE context in MEC by the PGW as described in the session management part.

In the EPC MEC, SGW + PGW MEC, and CUPS MEC, the MEC handover is supported using 3GPP standard S1 Handover with SGW relocation by maintaining the original PGW as anchor. The same considerations apply for the SGW-LBO MEC deployment. In the latter case, the target SGW enforces the same policy towards the local MEC application. It is the MEC application's responsibility to synchronize at application level and maintain the session in the case of a stateful application.

Lawful interception

Lawful Interception (LI) and Retained Data (RD) play a crucial role in helping law enforcement agencies to combat terrorism and serious criminal activity. Providers of public telecommunications networks and services are legally required to make available to law enforcement authorities information from their retained data which is necessary for the authorities to be able to monitor telecommunications traffic as part of criminal investigations.

Typically, this functionality is supported at nodes within the core network. However, traffic carried from the UE to an application at the network edge is currently designed to avoid the core, and hence would avoid the usual intercept points. In the context of MEC, it is recommended that LI and RD collection functions are implemented at the edge of the network, alongside or as part of the functionality being intercepted. Any edge node including LI/RD collection features must support strong physical security requirements similar to core network sites. In these regards, ETSI MEC is collecting informative and normative aspects in GS MEC 026 (work in progress), as this has strong impact on the MEC entities especially under the bump in the wire approach.

On the contrary, the solutions that include an EPC gateway, such as EPC MEC, SGW+PGW MEC, SGW-LBO MEC, and CUPS MEC are compliant with LI requirements as they natively support the usage of X1, X2, X3 interfaces towards the operators' Mediation Device, responsible to connect to the agency and transfer the requested data for the target.

Security

MEC offers an IT service environment and cloud-computing capabilities for hosting applications at the edge of the mobile network. As illustrated above in some deployment models the MEC applications run on the same physical platforms as some network functions. The third party applications are not controlled by the operator directly, so there are risks of these applications exhausting resources that are needed by the network functions. There are also risks of poorly designed applications allowing hackers to infiltrate the platform and hence affect the network functions running on the platform – or even of malicious

applications doing the same thing themselves. One solution to tackle these issues is to run both the MEC applications and the network function(s) in robustly segregated virtual machines, providing an assurance of confidentiality for sensitive data/information between VMs running on the same physical platform, and between a hypervisor and the host operating system. In addition, there is an opportunity for the MEC system to provide security/assurance services for the hosted applications. One example is to perform integrity assurance checks on applications at installation and upgrade, or after a server restart. Another is to expose security services APIs to sufficiently trusted third party MEC applications, e.g. for user identification.

Another option to enforce security is to allow deployment types that run applications on segregated hardware. This is particularly relevant for CDN, which usually have strict hardware requirements for copyright and privacy issues. Security is also enforced with an appropriate network design of the edge site where the MEC platform is connected with the use of L2/L3 traffic separation and firewalls.

Moreover, IPSec is envisaged to protect packets on the S1 interface. This creates additional challenges for bump in the wire approaches, which may require special, not yet standardized network design and more complex approval process by operators.

One last observation concerns the Distributed EPC deployment when including the HSS at the edge: this scenario requires special care on handling the confidentiality of subscribers' data and extending standard 3GPP core network interfaces at the edge of the network.

Charging

MEC needs to support off-line and on-line charging:

- Off-line charging: MEC periodically collects and reports the data records to the off-line charging function for aggregation and correlation. Billing systems use the aggregated/correlated event records to charge the consumer at the end of the billing cycle.
- On-line charging: Upon the first chargeable event of a consumer, MEC triggers an on-line charging request towards the on-line charging function to get a quota granted. When the allocated quota is almost fully used, MEC reports the usage of the resource and requests for an additional quota from the on-line charging function. The charging function may allocate a new quota or deny it. In case of denial, MEC will reject the resource usage request.

When it comes to charging, the bump in the wire approach natively supports the case of traffic passing through MEC application and then further to the CN, for which charging is taken care of by the 3GPP functions. Conversely, for traffic that is either terminated at MEC applications or breaking out to an external network, alternative solutions need to be considered to provide the necessary charging support. For instance, an alternative, yet not standardized solution needs the cooperation of MEC and CN functionalities, whereby MEC reports charging data to the PGW (as in Figure 7), or to the MEC Gateway based on the charging policies from the PGW (see Figure 1). Then the PGW aggregates and reports them to the billing system.

The other deployment options leverage the EPC data plane functionalities, so that both offline and online charging are supported natively as in the standard EPC, for all packets terminated locally or forwarded to external APNs configured in the core site's EPC for home and roaming traffic. This applies also to solutions like SGW + PGW MEC, SGW-LBO MEC, and CUPS MEC. In the latter case, the User Plane component employed for traffic offloading is able to forward usage statistics to the SGW-C and PGW-C according to

the standards. However the PGW-C needs to be customized to support both the PGW-U and the customized MEC-UP.

Identifying specific subscribers at the MEC platform

Traffic routing is part of the MEC platform's (MEP) essential functionality [1] and is enabled by applying configurable traffic rules. The functionality supports use cases such as breakout of encrypted user-plane traffic to a local network (e.g. enterprise network) by the MEP. Such traffic routing enables e.g. employees using authorized smartphones and tablet PCs to enjoy a fast broadband connection directly to their enterprise LAN, rather than such traffic having to traverse the mobile core network via a latency-inducing transport network. A key aspect of the routing is the identification of the packets to be filtered, where several filtering options are provided, including source/destination IP, port and tunnel addresses.

In order to filter based on UE identity, ETSI MEC has specified the UEIdentity feature [7]. This includes a dedicated API to trigger the MEP to route specific UE traffic flows to specified end points, without having to route the traffic via a MEC application. For local breakout, for example, the process would begin with the UE entering the serving area of the MEP within the enterprise zone. Detection could be provided by a BYOD client application on the UE, which would then initiate a connection to the BYOD server MEC application hosted by the MEP. Once the connection to the BYOD server had been established, it would be responsible for invoking the traffic routing at the MEP.

A challenge with this approach is that identification of individual traffic flows for a specific UE at the MEP can be problematic since the necessary information to do so may be obfuscated due to the MEP location within the mobile network architecture. For instance, considering a CRAN or bump in the wire deployment, identifiers, such as the UE's International Mobile Subscriber Identity (IMSI), are generally not exposed over the S1 interface. Therefore, the MEP must be provided with alternate identification information that it has direct access to, such as the temporary connection identifiers used on the S1 interface. Considering the user-plane only, such temporary identifiers include the pair of S1 GTP-U Tunnel Endpoint Identifiers (TEIDs). These are dynamically allocated and may change even while the connection is active due to factors such as UE mobility. The consequence is that the EPC sourced mapping information must be provided in near real time. One solution is to deploy probe agents within the EPC to capture temporary identifier information for a given UE identity, e.g. that are provided by the BYOD system. For example, by monitoring the S11 interface, the TEID assignments per IMSI can be recovered by the probe and then forwarded to the MEP or to the BYOD server to invoke at the MEP.

MEC as driver to 5G adoption

Multi-access Edge Computing makes no assumptions on the underlying radio infrastructure, which makes it a highly flexible element in the communications networks. As the delivery technology, together with the underlying hardware of the MEC platform, remains open, this enables new levels of adaptability to the chosen deployment scenario. Therefore, service providers (SPs) can use MEC as a revenue generator and application test bed (including service producing applications) without being forced to wait for full ratification of the 5G standard and the associated capital investment. This approach enables SPs to offer third parties a cost effective way to trial their applications. Using an "edge cloud", the SP can host applications in a virtual retail space, test the revenue return, and scale up or remove as appropriate. So, starting out as a 4G edge test bed with limited deployments at first, MEC allows a smooth transition into the 5G network rollout, removing the need for major upgrades when the time for transition arrives.

Another focus area for transitioning from today's 4G to 5G networks is re-using the existing deployed systems in the process. Due to the virtualized characteristics of MEC, it has never been easier to monitor performance and resource needs of an application, which, in turn, enables more accurate pricing for operators towards application providers for hosting the applications, as well as dimensioning the edge equipment exactly as required for the application set proposed.

The common feature set of providing much-improved capabilities at the edge of the network, improved intelligence about resources needed at the edge, and the ability to charge for service delivered by cycles, memory, storage and bandwidth delivered, makes it very attractive to start the deployment now in early test sites, roll out to sites that show promise and need for MEC based applications, and then roll out as part of the 5G transition without losing any upfront investment from the earlier test deployments.

Taking into account the above considerations, in the next sections we illustrate how MEC compatibility towards 5G networks may involve:

- Integrating the MEC data plane with the 5G system's one for routing traffic to the local data network and steering to an application;
- An Application Function (AF) interacting with 5G control plane functions to influence traffic routing and steering, acquire 5G network capability information, and support application instance mobility;
- The possibility of reusing the edge computing resources and managing/orchestrating applications and/or 5G network functions, while MEC still orchestrates the application services (chaining).

Deploying MEC in the 5G system architecture

The 5G Service Based Architecture (SBA) specified by 3GPP TS 23.501 [6] contains multiple control plane functional entities, like the Policy Control Function (PCF), the Session Management Function (SMF), the Application Function (AF), etc., and data plane functional entities like the User Plane Function (UPF).

In contrast to the current mobile network architecture, the 5G system was conceived to allow a more flexible deployment of the data plane, aiming to natively support edge computing. As a consequence, the MEC architecture can easily be integrated into that defined for 5G. Figure 8 illustrates an example MEC mapping to the 5G system architecture, where for example the MEC host's data plane can be mapped to 5G's UPF element.

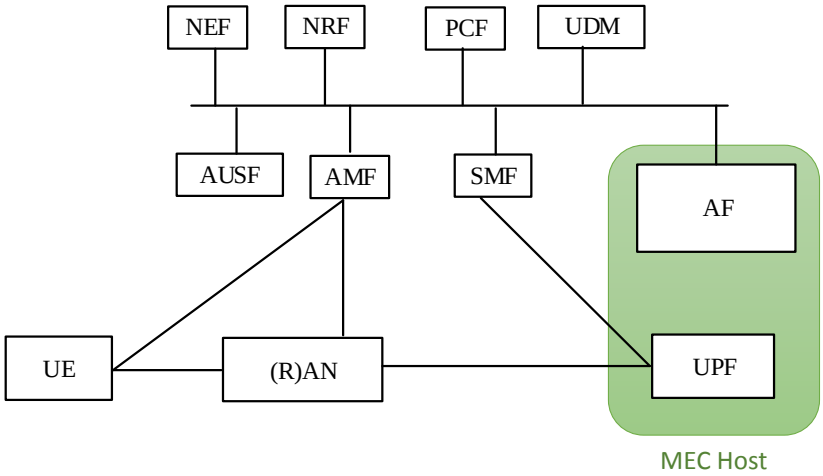

Figure 8: An example of MEC mapping with 5G system architecture

In the example above, the MEC platform would leverage the 5G network architecture and performs the traffic routing and steering function in the UPF. For example, a UL Classifier of UPF is used to divert to the local data plane the user traffic matching traffic filters controlled by the SMF, and further steer to the application. The PCF and the SMF can set the policy to influence such traffic routing in the UPF. Also the AF via the PCF can influence the traffic routing and steering. Therefore MEC in 5G is able to influence the UPF through the standardized control plane interface in SMF similarly to some of the EPC MEC deployment scenarios that we examined in 4G.

Although the position of MEC at the edge site is left to the operators' choice, similarly to what we have done for the 4G MEC deployment, here are a few migration examples to 5G selected architectures. The pictures below show how the MEC host, which includes the 4G core network functions, can be transformed to support 5G by software upgrading the relevant network functions. In the transition to 5G the MEC functionalities introduced with the 4G technology are preserved, fulfilling key requirements such as:

- reusing the edge computing resources;
- interaction with 5G control plane;
- integration with the 5G network.

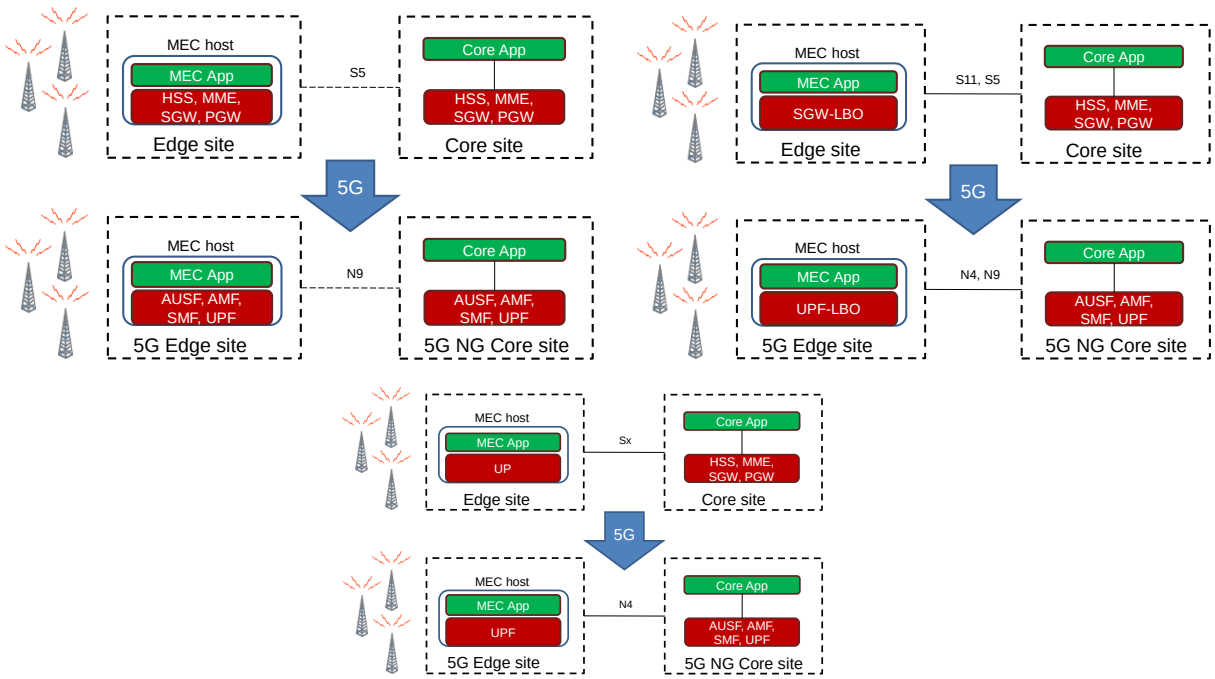

Figure 9: Migration patterns for MEC deployments from 4G to 5G. In the top left diagram, MME, SGW, PGW and HSS migration e.g. to support private networks and mission critical applications. At the top right, SGW-LBO MEC migration to 5G for selective traffic offloading. In the bottom diagram, CUPS migration to 5G.

Management and Orchestration of Cloud vs Edge resources

There is a growing consensus that in the long term, 5G deployments will increasingly integrate fixed-mobile networks infrastructures with cloud computing and MEC. In these future scenarios, the borders between cloud and MEC virtual resources will blur, paving the way towards a sort of "continuum" of logical resources and functions, offering flexibility and programmability through global automated operations. This will require that the orchestration capabilities, which are already a key element for exploiting cloud computing capabilities, become an essential part of the operation of future 5G infrastructure.

In cloud computing, orchestration is a mature concept and is generally referred to as the automation of tasks involved with arranging, managing and coordinating services deployed across different applications and enterprises, i.e., administrative domains, with the purpose of exposing them as single service instances. In 5G service scenarios integrating cloud and MEC, orchestration will have to span across the different service levels: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

On the other hand, in 5G future service infrastructures, IaaS will assume a new meaning, as the CPU, storage, and network resources are not only provided by a collection of data centres, but also by the CPU, storage, and network resources deployed into the Points of Presence (PoPs) of the telecommunications network (in its core, edge and access segments). The PaaS layer will feature a pool of software appliances that facilitate the end-to-end lifecycle of developing, testing, deploying, and hosting services and applications. Some examples of these appliances are databases, web servers, application servers, Apache Hadoop, Apache Storm, and load balancers, each of which will be integrated with other network

appliances (e.g., telecom/internet middle-boxes) to design complex service chains, functions and applications. SaaS will integrate multiple, interoperable PaaS and IaaS resources to deliver services and applications to the end users.

In this broader perspective, orchestration should satisfy both horizontal and vertical interoperability requirements: horizontal, when considering the interoperability between the same tiers in different infrastructure stacks (such as cross-SaaS, cross-PaaS, or cross-IaaS); and vertical when addressing downstream-compatible infrastructure tiers in different stacks.

The integration of 5G management, control and orchestration processes is expected to facilitate applications/services development by providing controlled access to high-level abstractions of 5G resources (e.g. abstractions of computing, memory/storage and networking) thus enabling any vertical application. Moreover just like a true operating system, it should provide automated resource management, scheduling process placement, facilitating inter-process communication, and simplifying installation and management of distributed functions and services, spanning from cloud computing to MEC. This implies a shared data structure capable of supporting multi-vendor systems and applications, which together enable sharing of common data amongst different protocols. Data structures include network state information, i.e., data about system and interface state, forwarding information base state, neighbours table and routing information base and policies. Also, standardized data models are required using, for example, a data-modelling language such as YANG.

The key to a successful management integration of MEC platforms, regardless of their deployment site, is to support standard modes of management, be it IPMI-based management at a very low level, SNMP, modern REST-based protocols or DMTF's Redfish, all of which connect into the operators' network management solutions. Whether Open Source MANO, ONAP or other open- or closed-source solutions, they all offer interfaces into these standard technologies. As MEC moves closer to standard data centre practices, wherever there is a non-standard management approach, typically the integration will become tedious, and, in some cases, even fail.

MEC and NFV

A consolidated vision of the MEC system is about deploying it as part of an NFV environment where MEC applications would be deployed as Virtual Network Functions (VNFs). In this deployment scenario, we have already illustrated some examples of how the MEC and EPC functions may co-operate to create the end-to-end network service.

In order to fully understand the implication of deploying MEC in an NFV environment, ETSI MEC has already started looking at how the two technologies can blend together, resulting in a proposed architecture available in the GR MEC 017 [8]. The MEC system would be virtualized as well and offered as a Network Service which introduces additional challenges in all life-cycle and enablement procedures for the MEC application (VNFs). Also, the management and orchestration systems from both MEC and NFV are meant to co-operate in order to carry out their respective functions.

Support to third party service providers

Third party service providers may own, deploy and manage compute and storage resources to provide MEC service. In order to manage the service, they require a control interface to the mobile network's OAM system. However, no single standardized interface or open API specification exists to interconnect a 3rd party MEC service with a specific MNO network. Each third-party cloud service provider must work

independently with each operator's network, where they intend to provide service, conforming to the operator's or vendor's interfaces, when available. As an alternative to offering edge services, third party service providers may utilize the Service Capability Exposure Function (SCEF) in a 4G system to monitor, gather network information and create innovative services in the cloud. But these are not edge services in true sense.

In order to avoid this ugly scenario, providers of MEC platforms may support standardized and non-standardized interfaces as they feel necessary, but with a focus on the standards-based management functions delivering access to the complete set of capabilities.

Management of MEC applications

The MEC application life-cycle consists of procedures such as: on-boarding, enablement, instantiation, termination, query, disablement and deletion. The Mm3 interface (connecting MEC orchestrator and MEC Platform Manager) is a key component for MEC application on-boarding and enablement. A MEC Orchestrator is the brain, makes placement decisions, whereas a MEC Platform Manager is the executor, allocating resources through VIMs and instantiating applications through a MEC Platform on each MEC host. For each scenario the placement decision is based on the demands of a MEC application and real-time monitoring capabilities of a MEC host.

The MEC platform specification assumes a completely virtualized environment. This is a key requirement in order to enable seamless application lifecycle management paired with seamless platform management. Some applications, however, require hardware acceleration in order to perform certain tasks that are too difficult to achieve in a fully virtualized regime. A resulting requirement for this is the possibility to add access to the acceleration function as part of the virtualization platform. It would be even better if these requirements can be fulfilled in a single box, and can be configured upon start to allow communality of units across multiple deployments, while matching the local requirements when the unit is started.

A dynamic start-up and shutdown of applications across multiple machines, selecting the best-cost solution that matches the application's requirements, enables telecoms operators to select the best match between application, performance needed and delivered without adding unnecessary overhead and upfront investment until it is really required.

Conclusions

Multi-access Edge Computing brings a network technology featuring a whole set of application-oriented functionalities, such as: policy-based traffic forwarding control, DNS policy management, application enablement and orchestration, and, optional services, like RNIS, location and bandwidth management. The key element in the MEC architecture is the MEC host, a general purpose edge computing facility that provides the computing, storage and other resources required by applications such as IoT data pre-processing, VR/AR, video streaming and distribution, V2X, etc.

In this document, we have explored how the MEC system can be deployed in existing 4G networks, by showing different options to install the MEC host along with the 4G system architecture components, and observing how such installation choices impact on the running system and architecture.

Moreover, we have demonstrated how the MEC system deployed for 4G networks could be migrated to future 5G networks, looking at the problem from different angles, including compliance with 5G system architecture, adoption of cloud computing and NFV paradigm, protection of the investment during network upgrade.

It is clear that in order for MEC platforms to be widely adopted by Mobile Network Operators as bridge to 5G, the MEC approach used needs to

1. create value for the customers with real business justifications
2. have minimal impact on existing 4G architecture and network processes
3. use standard 3GPP interfaces to the largest possible extent
4. provide a seamless software-only upgrade to 5G user plane functionality.

Whereas individual use cases that are deployed independently of mobile operators may allow deployment of any of the solutions described above, it appears that the family of solutions that push EPC functionality to the edge and are fully softwarized (i.e., cloud-ready) provide the most effective bridge to 5G.

List of abbreviations

3GPP	3rd Generation Partnership Project
4G, 5G	4th, 5th generation of mobile networks
AF	Application Function
API	Application Programming Interface
APN	Access Point Name
AR	Augmented Reality
BYOD	Bring Your Own Device
CDN	Content Delivery Network
CRAN	Cloud RAN
CUPS	Control/User Plane Separation
DMTF	Distributed Management Task Force
DNS	Domain Name System
DSCP	Differentiated Service Code Point
eNB	Evolved Node B
ETSI	European Telecommunications Standards Institute
EPC	Evolved Packet Core
GTP, GTP-U	GPRS Tunnelling Protocol, GTP-User plane
HSS	Home Subscriber Server
IaaS	Infrastructure as a Service
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IoT	Internet of Things
IP	Internet Protocol
IPMI	Intelligent Platform Management Interface
IPsec	Internet Protocol Security
LAN	Local Area Network
LI	Lawful Interception
LIPA	Local IP Access
M2M	Machine to Machine
MANO	Management and Orchestration
MCPTT	Mission Critical Push to Talk
MEC	Multi-access Edge Computing
MEP	MEC Platform
MME	Mobility Management Entity
MNO	Mobile Network Operator
NFV	Network Functions Virtualisation
OAM	Operations, Administration and Management
ONAP	Open Network Automation Platform
PaaS	Platform as a Service
PCF	Policy Control Function
PDN	Packet Data Network
PGW, PGW-C	PDN Gateway, PGW Control plane
PoP	Point of Presence

QoS	Quality of Service
RD	Retained Data
REST	Representational State Transfer
RNIS	Radio Network Information Service
SaaS	Software as a Service
SBA	Service Based Architecture
SCEF	Service Capability Exposure Function
SGW, SGW-C	Serving Gateway, SGW Control plane
SGW-LBO	SGW with Local Breakout
SMF	Session Management Function
SNMP	Simple Network Management Protocol
SP	Service Provider
TAC	Tracking Area Code
TEID	Tunnel Endpoint Identifiers
UE	User Equipment
UL	Uplink
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
V2X	Vehicle to Everything
VoLTE	Voice over LTE
VIM	Virtualised Infrastructure Manager
VM	Virtual Machine
VNF	Virtual Network Functions
VR	Virtual Reality
YANG	Yet Another Next Generation

References

[1] Several authors, “Mobile-Edge Computing – Introductory Technical White Paper,” Sept., 2014. https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing_-_introductory_technical_white_paper_v1%2018-09-14.pdf

[2] ETSI GS MEC 011 V1.1.1, “Mobile Edge Computing (MEC); Mobile Edge Platform Application Enablement” (2017-07). http://www.etsi.org/deliver/etsi_gs/MEC/001_099/011/01.01.01_60/gs_mec011v010101p.pdf

[3] ETSI GS MEC 003 V1.1.1, “Mobile Edge Computing (MEC); Framework and Reference Architecture” (2016-03). http://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/01.01.01_60/gs_mec003v010101p.pdf

[4] ETSI White Paper No. 23, “Cloud RAN and MEC: A Perfect Pairing”, First Edition, February 2018, http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp23_MEC_and_CRAN_ed1_FINAL.pdf

[5] ETSI TS 123 401, “LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access,” 2008-2017. http://www.etsi.org/deliver/etsi_ts/123400_123499/123401/

[6] 3GPP TS 23.501 V15.0.0, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15)” (2017-12) http://www.3gpp.org/ftp//Specs/archive/23_series/23.501/23501-f00.zip

[7] ETSI GS MEC 014 V1.1.1, “Mobile Edge Computing (MEC); UE Identity API” (2018-02) http://www.etsi.org/deliver/etsi_gs/MEC/001_099/014/01.01.01_60/gs_mec014v010101p.pdf

[8] ETSI GR MEC 017 V1.1.1, “Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NFV environment” (2018-02) http://www.etsi.org/deliver/etsi_gr/MEC/001_099/017/01.01.01_60/gr_MEC017v010101p.pdf

ETSI
06921 Sophia Antipolis CEDEX, France
Tel +33 4 92 94 42 00
info@etsi.org
www.etsi.org