Survey on edge computing and its evolution

Hongmei Zhu*, Yilin Lin, Zeng Qin
China Telecom Research Institute, Guangzhou 510630, Guangdong, China

ABSTRACT

Edge computing is the key technology for 5G to achieve high bandwidth and low latency. It has been widely applied in vertical industries. However, with the fast development of various intelligent application services, operators put forward new requirements for MEC as an edge cloud. Operators are continuously exploring how to help users find edge nodes more quickly, how to smoothly migrate users within edge nodes, and how to better serve vertical industry customers. This paper first introduced the development history of MEC technology. Then application scenarios of MEC are described. Later functions and architecture of MEC are described with the most popular function. After that 5G System enhancements for edge computing are introduced with R17, R18 and even R19 assumptions. Then the latest Work item ETSI MEC 047 is introduced. Lastly, the conclusion is made about development of MEC in future.

Keywords: Edge, MEC, distributed network

1. INTRODUCTION

Multi access edge computing, formerly called Mobile edge computing (MEC), was first introduced in 2013. At the GSMA Summit held in Barcelona in 2013, IBM and Nokia Networks jointly demonstrated the world's first mobile edge computing platform, which provides data services at the base station side, greatly improving data transmission efficiency.

In 2014, the European Telecommunications Standards Institute (ETSI) established the MEC specification working group and officially announced the promotion of MEC standardization.

In 2016, ETSI expanded the access mode of MEC from cellular network to WLAN and other access modes, that is, expanded the concept of mobile edge computing to multi access edge computing.

In 2017, ETSI completed the MEC Phase1 standard, laid the foundation of MEC infrastructure, and launched a series of specifications, such as GS MEC-002 Requirements and Use Cases, GS MEC-003 Architecture, GS MEC 010-2 Lifecycle Management, GS MEC 014UE API, and GS MEC-021 Mobility Management, as well as specifications supporting business such as V2X.

This stage is also the commercialization period of 4G/LTE. At that time, many operators launched MEC pilot projects to explore its technology and business model, carrying different industry applications, such as CDN, face recognition, video surveillance, etc., by adding computing, storage, data processing and other capabilities resources at the edge of the mobile network. These attempts laid the foundation for the robust development of MEC in the 5G era. However, there were some technical problems in the MEC solution in the 4G era. The most critical reason was that the control plane/user plane of the 4G core network was not separated, which led to complex edge side shunting and docking solutions. There were also deficiencies in supervision, security and billing in the wireless side ToF (Traffic Offload) shunting solution, which led to commercial difficulties. Coupled with the imperfect application ecosystem of MEC and the low enthusiasm of enterprise users, MEC did not get widely used in the 4G era.

In 2018, ETSI MEC launched a project to study the integration of MEC and 5G. In addition, MEC, as one of the key technologies of 5G, is used to uphold low-latency business scenarios of 5G.

In 2020, ETSI released a report on the integration of 5G MEC and announced the expansion of the corresponding functions of the specification working group. Thus, it's always considered that the MEC standard is jointly developed by ETSI and 3GPP. ETSI focuses on defining the framework and architecture of the entire MEC, including application deployment environment, application scenarios, management software architecture, and API interfaces; 3GPP focuses on defining the support and implementation of MEC by 5G network, as well as how to provide quality of service guarantee.

*zhuhongm@chinatelecom.cn

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133954P · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049389

However, with the passage of time, the importance of MEC has further been highlighted. As the GSMA considers how to better support the development of MEC from the perspective of network in 3GPP, the research and standardization work on "5G System Enhancements for Edge Computing" will be carried out separately from R17. The standard for R18 phase will be frozen in 2023, and the work for R19 phase is still in progress. ETSI MEC evolves further.

2. APPLICATION SCENARIOS OF MEC

The application scenarios of MEC are very wide, including holographic communication, augmented reality, virtual reality, intelligent transportation, Internet of Things, Industry 4.0, government office and other fields. In the field of holographic communication, augmented reality and virtual reality. Figure 1 shows top ten application scenarios of MEC. And ETSI GR MEC038[2] comprehensively reviewed the application of MEC in the park MEC can provide faster response speed and lower latency, thereby improving the user experience. In the field of intelligent transportation, MEC can cache real-time data on vehicles, thereby improving traffic safety. In the field of Internet of things, MEC can perform local processing on sensors and devices, thereby reducing the requirement for cloud processing and latency. In the field of Industry 4.0, MEC can perform real-time data processing and optimization on industrial equipment, thereby improving production efficiency and quality. In the field of government office, MEC can provide efficient services for citizens and enhance the user experience.

Figure 1. Top ten application scenarios of MEC.

3. FUNCTIONS AND STRUCTURE OF MEC

3.1 Architecture

In March 2016, ETSI GS MEC 003[1] issued a reference architecture for MEC in 5G, which clearly defined that MEC belongs to the 5G core network and uses the 3GPP user plane. At the same time, it defined the reference architecture variant of MEC in NFV, shown in Figure 2:

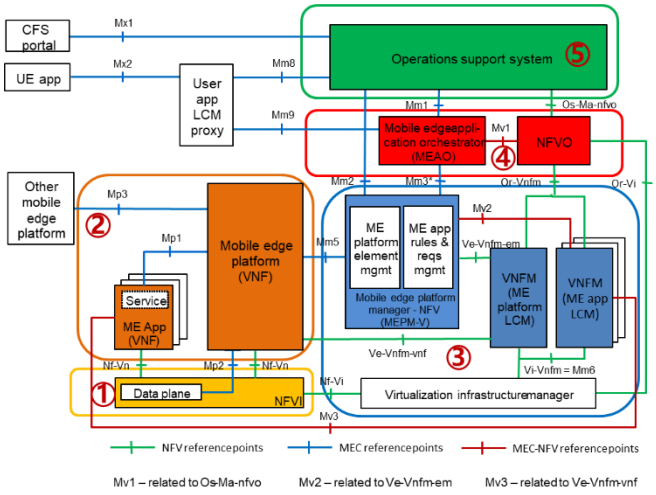

Figure 2. Reference architecture variant for MEC federation.

There are many entities in the architecture, which can be divided into five parts from a functional perspective, as shown by the numbers in the figure:

(1) NFVI: Based on the ETSI NFV framework, a virtualization platform that provides deployment environments for applications, services, MEPs, and more.

(2) MEP/MEC APP: MEP is a middleware capability for integrated deployment and network openness of MEC applications, which can host MEC services such as 5G network capabilities and business capabilities; MEC APP is an application that runs on the virtualized infrastructure of the MEC host and interacts with the MEC platform to use or provide MEC functions.

(3) MEPM/VIM: MEPM is the platform network management of MEC, which implements the monitoring, configuration, performance and other management of MEP, and manages the rules and requirements of edge computing applications; VIM stands for virtualization management, implementing lifecycle management for MEP and APP; The virtualization infrastructure manager is in charge of the allocation, management, and release of virtualization resources.

(4) MEAO/NFVO: It receives application requirements from OSS, decompose the requirements, and orchestrate the applications.

(5) OSS: Operator Service System.

Of course, ETSI GS MEC-003[2] also defines the reference architecture variant for MEC federation, which is proposed to support the Open Platform of GSMA and is also based on the standard work of ETSI GS MEC040: API Federation[3]. Later, the architecture specification will also introduce reference architectures for other functions, such as the MEC application slices that we will meet soon. Currently, the ETSI GR MEC 044[4] WI has been released, and the research on the MEC-based slice architecture will be written into the general architecture specification of ETSI GS MEC-003[1].

3.2 Functions

MEC is a technology that deploys computing and storage capabilities at the edge of mobile devices to improve application response speed and reduce network bandwidth consumption. This technology allows mobile devices to process tasks faster and more intelligently, thereby enhancing the user experience. Specifically, its functions can be summarized as follows:

(1) Cache and acceleration: MEC can cache static content to the edge of mobile devices, thereby improving response speed and reducing network bandwidth occupation.

(2) Real-time data processing: MEC can process real-time data on mobile devices, thereby reducing the need for and delay in cloud processing.

(3) Application optimization: MEC can optimize applications to improve their performance and response speed.

(4) Security and privacy: MEC can provide higher levels of security and privacy protection because data can be processed locally on mobile devices or kept out of designated areas. Park applications are also one of the key scenarios for MEC. Application scenarios in park is described in detail in ETSI MEC 038[2].

However, the above-mentioned excellent functions can only work when traffic is offloaded to the MEC.

If the traffic cannot detect the MEC, especially for 5G users, the terminal they hold is multi-functional, which can be used both under the MEC in the park and also on the internet that accesses the default UPF.

3.3 Traffic offloading

There are three ways: Uplink Classifier (UL CL), Local Area Data Network (LADN) and IPv6 multi-homed PDU session to enable local access to a DN/MEC, describes in ETSI GR MEC 031. Of course all these ways are defined in 3GPP TS 23.501[5].

UL CL is a feature supported by UPF, designed to transfer (local) traffic to match the traffic filters provided by SMF. Insertion and deletion of UL CL are determined by SMF and controlled by SMF using the general N4 and UPF functions. The entire process is not perceived by UE, so the user experience is good.

IPv6 multi-homed PDU session is also a type of offloading. A PDU session can be associated with multiple IPv6 prefixes, i.e., a multi-host PDU session. A multi-host PDU session is anchored by multiple PDU sessions. When the UE request type is “IPv4v6” or “IPv6”, the UE also provides the network with an indication of whether it supports multi-homing IPv6 over PDU sessions. Obviously, the UE needs to participate in this process.

LADN (Local Area Data Network) is another type of traffic offloading. LADN is a service provided by the serving PLMN and the LADN service area is a set of tracking areas. The UE is configured to know whether the DNN is a LADN DNN and the association between the application and the LADN DNN. Obviously, it's a restricted use.

Figure 3 shows special terminals in the park can access the MEC system and internal network through edge UPF without traffic leaving the park for secure purpose. However, the common users in the park can access the Internet. In this case, it requires special terminals in the park to sign LADN DNN, so as to trigger the establishment of LADN PDU sessions based on the user's location. Common users only sign up for Internet DNN, and Internet traffic is transmitted through central UPF/edge UPF. Of course, if the URSP technology is used, the same terminal can sign up for two DNNs at the same time, supporting simultaneous access to both the intranet and the Internet.

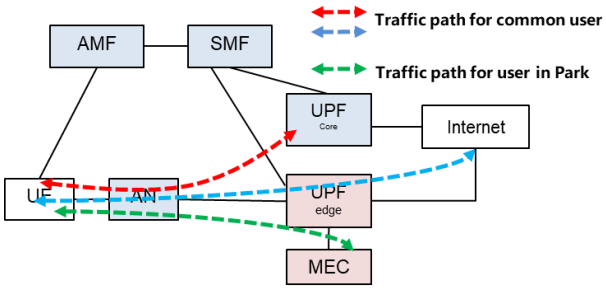

Figure 3. Traffic offloading of MEC in park area.

Then the question is, if one terminal only signs up for one DNN and access the intranet deployed on the edge MEC while also accessing the internet network?

This is a function defined by 3GPP. Below shows what 3GPP has done to support MEC applications.

4. 5G ENHANCEMENTS FOR MEC

The enhancement of 5G supporting for MEC is carried out in stages. The research work of 3GPP TR 23.748[6] was initiated from R17, and the relevant conclusions were finally written into 3GPP TS 23.548[7].

4.1 Newly introduced entity

(1) EASDF

EASDF (Edge Application Server Discovery Function): The main function of this NF is to process DNS messages according to the instructions of the SMF, including exchanging DNS messages from the UE ,then forwarding the DNS messages to C-DNS or L-DNS when DNS queries, adding EDNS Client Subnet (ECS) option into DNS Query for an FQDN, reporting to the SMF the information related to the received DNS messages, buffering/discarding DNS messages from the UE or DNS Server, also terminates the DNS security, if used.

The EASDF has direct user plane connectivity (without any NAT) with the UPF over N6 for the transmission of DNS signaling exchanged with the UE. The deployment of a NAT between EASDF and PSA UPF is not supported.

In fact, EASDF is extracted from the original SMF and UPF functions and introduced specifically for MEC, especially for EAS discovery and re-discovery for PDU Session with Session Breakout connectivity model, the details see in 3GPP 23.548[10] clause 6.2.3.

(2) EDC

EDC is the abbreviation of Edge DNS Client, which is a functionality in the UE that guarantees that DNS requests from applications are sent to the DNS Server's (e.g. EASDF/DNS resolver) IP address received from the SMF in the ePCO. Figure 5.2-1 of 3GPP TS 23.548[7] depicts the Edge DNS Client (EDC) functionality in the UE. The EDC functionality in the UE is also a UE capability that ensures the usage of the EAS discovery and re-discovery functionalities.

4.2 Connectivity models

3GPP SA2 has been focusing on its core work in R15 and R16 stages, such as the definition and standardization of 5G network architecture, network element functions, signaling processes, etc., and treating ETSI MEC as a common AF.

Although MEC has been used in different scenarios, 3GPP has defined three models of MEC from a rigorous perspective, see Figure 4, citing from 3GPP TS 23.548[7].

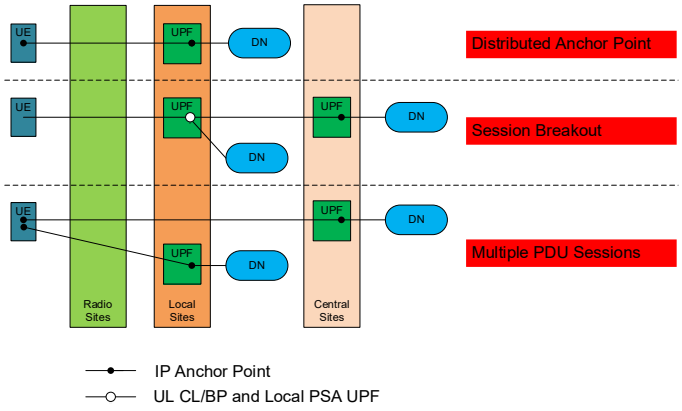

Figure 4. 5GC connectivity models for edge computing.

Based on this connectivity model, and combined with KI, 3GPP SA2 has proposed corresponding solutions to enhance various scenarios in which MEC may be applied.

4.3 Key issues

Based on the documents submitted by each company, as well as the ranking of importance, four KIs were finally determined.

(1) EAS Discovery and Re-discovery

The first KI is EAS Discovery and Re-discovery. That is, how to discover edge servers and offload traffic to it It's known that MEC can provide users with low-latency, high-bandwidth services, provided that traffic needs to be transferred to edge MEC. In traditional MEC systems, users can discover other edge servers through application layer messages be informed, but the application layer does not always guarantee QoS, and the application layer does not have enough perception of 5G networks.

Therefore, this is also the key reason why 3GPP SA2 wants to study and standardize this KI, that is, to solve the problem of user discovery of edge servers from the perspective of the network, providing a new path for users to use edge MEC.

(2) Edge Relocation

Edge relocation is actually caused by application mobility. Application mobility is defined in ETSI GS MEC021[8]. The first version was released in January 2020, and the third versions have been released so far. That document provides a specification for end-to-end MEC App mobility support in an edge system, describes the information flows, required information and the related operations and specifies the necessary API with the data model and data format.

Therefore, the MEC system side has already considered mobility, and the edge relocation here is still from the perspective of mobile networks, providing further guarantees for the mobility of Applications on the edge platform.

3GPP TS 23.548[7] clause 6.3 provides analysis and solutions for edge relocation due to UE mobility, but also considers relocation from the perspective of edge load balancing. If the relocation is caused by UE mobility, the trigger point is the network. If it is caused by load balancing, the trigger point is the edge platform itself. Both scenarios involve edge change. How to guarantee service continuity because packet loss may occur during edge change. Several solutions are provided:

The first is that Edge Relocation Using EAS IP Replacement. Details see clause 6.3.3 of 3GPP TS 23.548[7]. This function covers the scenarios that the UE moves from non-edge area to the edge area or the AF (Application Function) decides to enable the EAS IP replacement in the middle of a session, not only the mobility between two edge nodes. And the AF decides when and how to stop the Source EAS from serving the UE based on its local configuration. So, when 5G network support this new function, it will definitely enhance the service experience.

The second is that AF request for simultaneous connectivity over source and Target PSA at edge relocation. This is to enhance the transmission of data through a two-channel approach to ensure data is not lost.

Packet buffering for low packet loss, which is also to enhance the reliability of data transmission.

(3) Network Exposure to Edge Application Server

Applications deployed on MECs often have strict latency requirements, so if some real-time information from the network, such as entities' load and user path, can be made available to edge applications timely, it will greatly enhance the flexibility of the business, which is also the original intention of this KI.

The UPF will be instructed to report information for a PDU Session directly i.e., bypassing the SMF and the PCF. This reporting may target an Edge Application Server (EAS) or a local AF that itself interfaces the EAS. At the same time Local NEF deployed at the edge may be used to support network exposure with low latency to local AF. So, two options were provided for this KI: Usage of Nupf_EventExposure to Report QoS Monitoring and with a Local NEF serving the edge server.

5. R18 ENHANCEMENT FOR MEC

The above-mentioned 5G enhancements for MEC are mainly the content of 3GPP R17. Here is a brief summary of the content of 3GPP R18. The entire content of R18 has been frozen in June 2023.

Referring to 3GPP TS 23.548[7], it will be found that clause 6.7 and 6.8 are completely new sections.

5.1 Support of the local traffic routing in VPLMN (HR-SBO)

As we all know, the edge applications of R17 are still within the local network, including moving. However, with the development of business, especially the globalization of applications, applications deployed on the edge platform also have the requirement for roaming. This is what the content t in Section 6.7. This content also meets the requirements of the GSMA open platform initiative which is to achieve data roaming on the edge platform.

When roaming, the UE establishes a Home Routed Session that is to support session breakout in V-PLMN based on the subscription. After the HR-SBO PDU Session is established, the UE is able to access EAS deployed in EHE in VPLMN and the UE can also access the data network in the Home PLMN.

5.2 Support for mapping between EAS address and DNAI

Section 6.8 is to support for mapping between EAS address Information and DNA. This is to consider the future development and flexibility (scaling and expansion) of the network, changes in the IP address of the edge platform and DNAI (including mapping relationships). Therefore, the original pre-configuration method in AF cannot meet the requirements. What is proposed here is to write any changes of the network to the UDR, and AF implements subscriptions to the above dynamic change information through NEF, thereby quickly deciding the optimal access path of edge UEs.

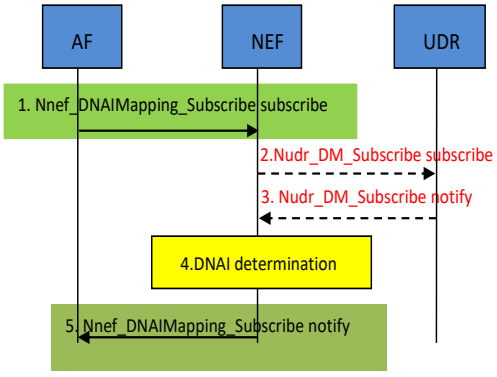

Figure 5. AF request for DNAI through NEF.

There are five steps for the process, shown in Figure 5. Step 1 is that AF invokes Nnef_DNAI Mapping_Subscribe service to subscribe to the DNAI information; Step 2 is that NEF invokes the Nudr_DM_Subscribe service to request DNAI mapping information for this DNN and/or S-NSSAI; Step 3 is that UDR notifies the NEF with all the DNAI mapping information for the requested DNN/S-NSSAI; Step 4 is that NEF determines the suitable DNAI(s) utilizing the DNAI

mapping information. And the final step is that NEF notifies the DNAI(s) or the updated DNAI information to AF in accordance with the request in Step 1.

In addition to the new content added in the above two clauses, R18 also updates other existing parts of R17. For details, please refer to 3GPP TS 23.548 V18.2.0[8].

6. R19 ENHANCEMENT STUDY FOR MEC

After the completion of the R18 work, 3GPP launched the R19 study. Three KIs for MEC enhancement were identified in 3GPP TR 23700-49[13], as follows:

6.1 EAS (re)discovery and UPF (re)selection

The title of this key issue occurs in R17, here is the enhancement of R17 existing solutions. This KI is to investigate whether and how to reduce impact on the central NFs when supporting EAS (re)discovery and UPF (re)selection. Current edge computing design has central 5GC Control Plane NFs involved deeply. For example, the EAS Deployment Information and local UPF information (deployment information, or load) may be changed dynamically, as a consequence, they should be updated at the central SMF and to influence the decision of the subsequent action. Furthermore, central SMF can also be a bottleneck point for the DNS message handling since all decision is made by the central SMF. Now couples of solutions are provided within which is to introduce I-SMF or local SMF for the edge.

6.2 EAS and local UPF (re)selection

The second KI is the enhancement of EAS and local UPF (re)selection. In the existing design as described above, 5GS provides support for means to determine, to report and to expose UE on-path congestions status, date rate information and round-trip delay between UE and PSA UPF. However, no means are defined to also consider above metrics on data network (e.g. N6 delay) when multiple EAS instance(s) are available for selection to provide best possible E2E user experience. N6 is like the end of the 3GPP closed pipeline, with no information going outwards, which does cause inconvenience in selecting edge nodes. The purpose of this KI(Key Issue) is to investigate whether and how to enhance EAS and local UPF (re)selection considering dynamic information related to EAS with EAS load and especially N6 delay between the local PSA and EAS.

6.3 EC traffic routing between local part of DN and central part of DN

Sometimes the application traffic may need to be first steered to Edge and processed there for low delay. After that, the application traffic may still need to be further forwarded to the Application Server in the central part of DN for further processing for combining or some other purpose. The application traffic may not be able to be routed directly between the EAS in the local DN and the Server in the central DN in case there is no direct connectivity between the local DN and the central DN. Similarly, several solutions have been proposed, such as IP replacement of EAS, over session breakout model, or via PDU session. Which will be identified and written into TS, we will wait and see.

In fact, besides the enhancement of MEC in 3GPP SA2, 3GPP SA6 has also carried out Architecture for enabling Edge Applications. See 3GPP TS 23.558[9] for details.

7. ETSI EVOLUTIONS FOR MEC

ETSI MEC is accelerating its integration into the 3GPP 5G mobile network and promoting commercial use, while also stepping up its own evolution, enhancing its features from architecture and capabilities.

7.1 API federation

In 2020, GSMA launched the Operator Platform project to accelerate and simplify the deployment of MEC apps across operators and clouds, facilitating MEC development.

App providers or operators can promote edge applications more conveniently, which not only defines a reference architecture, but also defines four interfaces: northbound interface, southbound interface, user network interface and the east-west platform interface. As a platform, ETSI MEC actively undertakes the task of east-west interface and initiated research in 2020. The research results are shown in ETSI GR MEC-0035[10]. And the research content has been written into the specification GS MEC040[3].The second version, v3.2.1, is about to be released.

7.2 MEC application slices

It's known that network slice has got widespread attention. However, as described earlier about the N6 interface, the N6 interface does not provide any information to the edge, so MECs must manually configure it when interfacing with network slice, which cause greatly inconveniences operations. In addition, the existing ETSI MEC uses AppD as the benchmark during the instantiation process, but in reality, as a slice element, it requires the support of infrastructure. Therefore, it must carry the information of NSD (Network Service Descriptor). Driven by the above two factors, ETSI MEC has launched MEC application slice to achieve automatic adaptation and docking with 5G network slice. The WI is stable and about to be released.

7.3 Distributed edge network

Distributed network architecture is a network architecture that distributes computing, storage, and control functions across multiple nodes in the network. In distributed network architecture, each node can independently perform specific tasks and interact with other nodes through communication and collaboration. This distributed approach allows the system to work with higher efficiency, while also increasing the flexibility and fault tolerance of the system.

With the development of business, especially the vertical industries, the mobile network is also about to move from a closed, inefficient centralized network to a distributed one. IMT-2030(6G) Promotion Group released a white paper on the application scenarios and requirements of 6G distributed network technology[11]. And much more voices about distributed network have been heard from the many SDOs (Standards Developing/Development Organization).

MEC is distributed-native. As early as 2022, ETSI GS MEC-002[12] proposed distributed use cases and related functional requirements based on HTC (Holographic Type Communication) business. As shown in Figure 6, also refer to Figure A.41.2 of ETSI GS MEC002.

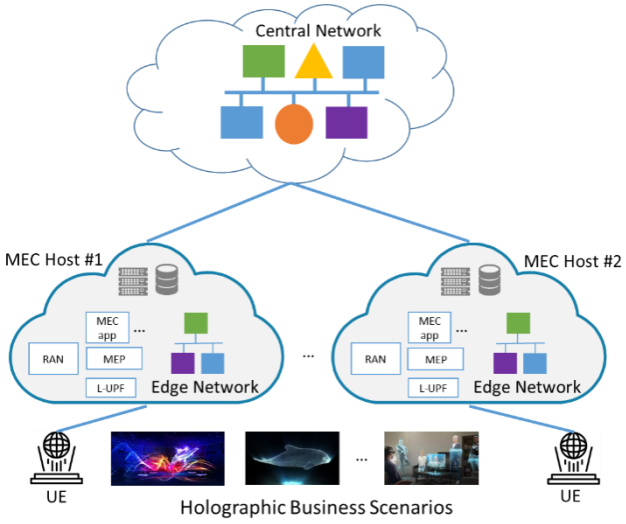

Figure 6. Example diagram of holographic service deployment based on MEC.

Assuming that the holography-enabled terminal devices are able to connect to MEC applications and maintain service continuity as the users move. The MEC hosts use general-purpose hardware, where specific network functions (e.g. on-premises UPF and RAN) can be instantiated to form an edge network to serve the UEs with extreme connection requirements (such as real-time high-definition holographic communication). So, the MEC hosts is able to have various forms to support holographic services with large bandwidth and low latency combined with on-premises UPF and RAN. They are co-located and share the infrastructure and even with different signaling path regards the time delay.

ETSI GR MEC 047[13] was born under this situation. But it's just ongoing and will be frozen in the end of 2024.

8. CONCLUSIONS

Edge computing is the product of network evolution and cloud computing technology development. It provides computing, storage, network and other infrastructure near users, and provides users with edge cloud services by deploying and running applications on this infrastructure

With the increase of per capita income, consumers' willingness to consume has undergone a fundamental change. Edge computing will gradually transform its consumption attributes from optional consumer goods to mandatory consumer goods. At the same time, with the advancement of XR and AI, the space of edge computing is unlimited.

Of course, with the evolution of the network, especially the introduction of the concept of 6G distributed network and the concept of network reconfiguration, edge computing may not be limited as the core network in the future. It can make more use of the resources of the wireless network to realize the integration of computing and network, which will greatly improve the user experience.

REFERENCES

[1] ETSI GS MEC 003: "Multi-access Edge Computing (MEC); Framework and Reference Architecture," (2023).
[2] ETSI GR MEC 038: "Multi-access Edge Computing (MEC); MEC in Park enterprises deployment scenario," (2022).
[3] ETSI GS MEC 040: "Multi-access Edge Computing (MEC); Federation Enablement APIs," (2023).
[4] ETSI GR MEC024 V2.1.1, "Multi-access Edge Computing (MEC); Support for network slicing," (2019).
[5] 3GPP TS 23.501 V17.1.1, "System architecture for the 5G System; Stage 2," (2021).
[6] 3GPP TR 23.748: "Study on 5G System Enhancements for Edge Computing," (2024).
[7] 3GPP TS 23.548: "5G System Enhancements for Edge Computing," (2023).
[8] 3GPP TS 23.548 V18.2.0: "5G System Enhancements for Edge Computing," (2023).
[9] 3GPP TS 23.558: "Architecture for enabling Edge Applications (EA)," (2023).
[10] ETSI GR MEC 035: "Multi-access Edge Computing (MEC); Study on Inter-MEC systems and MEC-Cloud systems coordination," (2022).
[11] ETSI GR MEC044 V3.1.1, "Multi-access Edge Computing (MEC); MEC Application Slices," (2024).
[12] ETSI GS MEC 002: "Multi-access Edge Computing (MEC); Use Cases and Requirements," (2024).
[13] ETSI GR MEC047 V4.1.1, "Multi-access Edge Computing (MEC); Distributed Edge Network," (2024).