EURASIP Journal on Wireless Communications and Networking

RESEARCH

Open Access

Dynamic service migration in ultra-dense multi-access edge computing network for high-mobility scenarios

Haowei Lin[1†], Xiaolong Xu[1*†], Juan Zhao[2] and Xinheng Wang[3]

*Correspondence:
xuxl@njupt.edu.cn
†Xiaolong Xu and Haowei Lin contributed equally to this work.
[1] Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, NanJing, China
Full list of author information is available at the end of the article

Abstract

The multi-access edge computing (MEC) has higher computing power and lower latency than user equipment and remote cloud computing, enabling the continuing emergence of new types of services and mobile application. However, the movement of users could induce service migration or interruption in the MEC network. Especially for highly mobile users, they accelerate the frequency of services' migration and handover, impacting on the stability of the total MEC network. In this paper, we propose a hierarchical multi-access edge computing architecture, setting up the infrastructure for dynamic service migration in the ultra-dense MEC networks. Moreover, we propose a new mechanism for users with high mobility in the ultra-dense MEC network, efficiently arranging service migrations for users with high-mobility and ordinary users together. Then, we propose an algorithm for evaluating migrated services to contribute to choose the suitable MEC servers for migrated services. The results show that the proposed mechanism can efficiently arrange service migrations and more quickly restore the services even in the blockage. On the other hand, the proposed algorithm is able to make a supplement to the existing algorithms for selecting MEC servers because it can better reflect the capability of migrated services.

Keywords: Dynamic service migration, Ultra-dense network, Multi-access edge computing, High-mobility scenarios

1 Introduction

With the advent of various mobile and Internet of Things (IoT) devices, new types of services and mobile applications are emerging that utilize machine learning (ML) and augmented reality (AR) technology [1]. These services performed computing resources' selection among MEC and centralized, cloud-based resources, towards efficient service orchestration [2]. They require high computing power and low latency due to recent advances in mobile network technology [3–7]. For example, the vehicular networks support more complex applications for both the vehicles and passengers nowadays, such as automatic driving, intelligent auxiliary driving for vehicles and augmented reality (AR), on-line interactive gaming, and other rich media applications for passengers [8], which require intensive communication and computation resources with low latency. Therefore, the fog computing (FG) and multi-access edge computing (MEC) are proposed to suit these requirements [9–12].

Although the cloud computing with MEC could provide a well-established distribution model and application platform [13, 14], the MEC network still has to deal with many problems on its own. The migration and handover of services are relatively frequent in MEC scenario considering users' movement, which may induce service interruption and additional cost [15]. Especially for highly mobile users, they accelerate the frequency of services' migration and handover, impacting on the stability of the total MEC network. Moreover, the architecture of ultra-dense network also causes more difficulties in the collection and delivery of information, easily leaving the backhaul network saturated and resulting in insufficient resources. The precise full system information is hard to be synchronized between base stations (BSs) and user equipment (UE) for mobility management decision making [16].

To solve these problems, various kinds of network architectures and methods are emerging. For example, the power-domain non-orthogonal multiple access (NOMA) [17–21] allows multiple MTs (mobile terminals) to reuse the same time slot and frequency resource blocks by dividing their transmit power and further exploiting the successive interference cancelation to mitigate the co-channel interference among them. Because of the advantages of NOMA, NOMA-based MEC system can provide large-scale access and complete the computing offloading services in large-scale access networks [22–24]. Service function chaining (SFC) is a promising mechanism to decompose a giant and resource-hungry network service into a chain of moderate, loosely connected virtual network functions (VNFs) in a specific order. By adopting SFC/NFV in 5G MEC networks, the operators could implement a variety of network services in a very flexible and efficient manner [25], but the reply of the post-migrated services still cost much time in the case of the saturated MEC network and the insufficient resources of MEC servers. Moreover, in terms of the selection to the suitable MEC servers for migrated services, many learning-based algorithms are established to evaluate various MEC servers for various migrated services, but lack further study on quality of service (QoS) of migrated services, which is the criteria of selecting the suitable MEC servers. Although many kinds of research for MEC are actively conducted, there are few studies on schemes related to service migration for users with high mobility in the ultra-dense network.

Based on the above observation, this paper is dedicated to solving the abovementioned problem. We propose a mechanism to minimize service interruption due to user mobility in high-mobility scenarios and ultra-dense network, and an algorithm to evaluate the services after being migrated to the target MEC servers in various aspects. The contributions of this paper are outlined as follows.

- We introduce the problems of the dynamic service migration in ultra-dense MEC network for high-mobility scenarios. To solve the problem, we propose a hierarchical MEC architecture, setting up the infrastructure for dynamic service migration in the ultra-dense MEC networks (Section 3).

- We propose a new mechanism for users with high mobility in the ultra-dense MEC network, called Chain Management with Valuation Adjustment Mechanism (CMVAM), efficiently arranging service migrations for users with high-mobility and ordinary users together. In the case of network saturation, the CMVAM can quickly restore the services in the blockage (Section 4).

- We propose an algorithm for evaluating migrated services, called Migration Effect Evaluation with QoS-aware (MEEQ), to make a supplement to the existing algorithms for selecting MEC servers, which is conducive to choose the suitable MEC servers for migrated services (Section 4).
- We implement a system prototype using the "networkx" to perform real service migration in an ultra-dense network. We saturate the ultra-dense network and test restoration of services in the blockage. Moreover, we test the effects of different algorithms for evaluating migrated services on services. Results from extensive experiments prove the capability of the proposed CMVAM and MEEQ (Section 5).

The rest of the paper is organized as follows. In Section 2, we present the related works. In Section 3, we propose a hierarchical multi-access edge computing architecture. In Section 4, we propose a new mechanism for users with high mobility in the ultra-dense MEC network. Moreover, we propose an algorithm for evaluating migrated services. Then, we test the effects of different algorithms for evaluating migrated services on services in Section 5. Finally, this paper is concluded in Section 6.

2 Related works

To solve the aforementioned problems of dynamic service migration in the MEC environment, there are studies related to service migration scheme in MEC environment. Kondo et al. [26] developed a MEC platform supporting service migration for MEC servers. Their proposed platform used IP mobility support gateway that applied the extended virtualized mobility support gateway (vMSG) to achieve mobility management requirement, but there was a lack of consideration for the movement of the mobile user requiring service migration. Ridhawi et al. [27] proposed the solution to decompose the data in the cloud into a set of files and services. They performed caching on the user mobile device to provide frequently requested files and services quickly. They considered the movement of the mobile user requiring task migration, but there was a lack of consideration for task migration situations that occur when users move frequently. Chen et al. [28] focused on the incorporation of virtual network function (VNF)/service function chaining (SFC) and multi-access edge computing (MEC). Based on this incorporation, they proposed an on-line algorithm called Follow-Me Chain to prevent unacceptably long service delay with inter-MEC handoffs for the VNFs of the corresponding SFC. However, they lacked further study on MEC server switching in the high-density environment to enable users with high-mobility and ordinary users to migrate services together without sufficient resources. Sun et al. [29] developed a novel user-centric energy-aware mobility management (EMM) scheme, in order to optimize the delay due to both radio access and computation, under the long-term energy consumption constraint of the user. Although the methods provided efficient solutions for migrating services in the ultra-dense networks, they lacked the consideration of high-mobility scenarios and details of the service migrated in the MEC servers. Aloqaily et al. [30] proposed the solution to provide fast service in situations where users demanded data or services at the same time in a high-density environment such as the stadium or the subway station. They used service-specific overlays to provide mobile devices with faster data access by using data replication methods from the cloud to the edge in the congested environment. Although the methods provided efficient solutions for managing data and services in congested situations, they did not provide efficient solutions to the use of MECS resources in situations where mobile users are frequently moving and generating tasks. Balasubramanian et al. [31] proposed the mobile device cloud system architecture to solve the buffering problem of mobile devices in the dense and congested environment. To this end, they utilized MPTCP (MultiPath-TCP) in their proposed system architecture. They also proposed an OS-side architecture that can manage the traffic coming from different owns. The OS-side architecture and MPTCP-based methods allowed efficient use of cloud resources, but lacked the consideration of the MEC environment.

In addition, there are also studies related to how to choose the suitable MEC servers to migrate services. Kim et al. [32] proposed using a vehicular cloud radio access network (vCRAN) in the automotive field with edge computing infrastructure where communication and networking resources were centrally controlled. The proposed vehicular network consisted of remote radio heads (RRHs) located at roadsides, MEC servers responsible for signal processing and service management, a cloud server managing MEC servers, and a software-defined network (SDN). Although this vehicular network provided a mobility support method for the allocation of services in the MEC servers, its centralized management was not suitable for the high-mobility users with low-latency requirements in the ultra-dense networks. Nasrin et al. [33] proposed a novel architecture, called Shared-MEC, to support service migration. They created a small cloud at the edge of the network to direct all MEC servers around, which could reduce communication costs and latency, but actually, this additional small cloud centers also required additional overhead. Ojima et al. [34] predicted user mobility with Kalman filter for estimation of the connectivity of MEC servers. They used mobility prediction to enable mobile users to select stable MECS during task requests and task collection, and the success rate of collecting results was improved. However, if the mobile users move frequently, they should select stable MECS at the new location each time. Moreover, if the predicted user's location is error, MEC servers will waste a lot of resources and need to spend more resources to remedy. Zhao et al. [35] proposed a vehicle mobility prediction module to estimate the future connected roadside units (RSUs) using data traces collected from a real-world vehicular ad hoc networks (VANET) deployed in the city of Porto, Portugal. Although wrong predictions will result in wasted resources in the MEC server, it is a good inspiration that they implemented a learning-based algorithm to constantly correct models for forecasting the locations of users. Wang et al. [16] proposed a Q-learning-based mobility management scheme to handle the system information uncertainties. Each user observed the task delay as an experience and automatically learnt the optimal mobility management strategy through trial and error. Peng et al. [36] considered the edge user allocation problem as an on-line decision-making and developed a mobility-aware and migration-enabled approach, called MobMig, for allocating users at real-time. Both schemes established the learning-based algorithms to choose the suitable MEC servers for the service to be migrated, but lacked further study on QoS of migrated services, leading to fail to choose the most suitable MEC servers.

3 The multi-access edge computing architecture and the dynamic service migration for high-mobility scenarios

In this section, we propose a hierarchical multi-access edge computing architecture, setting up the infrastructure for dynamic service migration in the ultra-dense MEC networks. The proposed architecture introduced a new mechanism to make the MEC network more adaptable to the ultra-dense state. Moreover, we propose a new mechanism for users with high mobility in the ultra-dense MEC network, called Chain Management with Valuation Adjustment Mechanism (CMVAM), which is extended by Follow-Me Chain (FMC) [28] and MEC architecture proposed in Section 3, efficiently arranging service migrations for users with high-mobility and ordinary users together. In the case of network saturation, the CMVAM can quickly restore the services in the blockage.

3.1 Architecture

In a conventional MEC network using MEC servers, as service requests from user equipment (UE) increase and the topology of a network changes dynamically because of high mobility, considerable increases occur in the complexity of establishing and maintaining connections between MEC servers and UE [32]. Hence, a new MEC network architecture that can support both of the flexible resource management and high mobility is necessary for improving QoS.

The proposed multi-access edge computing (MEC) architecture is shown in Fig. 1. In the architecture, MEC servers and BSs are integrated within the radio access network (RAN) and deployed on the user plane interface, i.e., between the core network (CN) and user equipment (UE). Each base station is equipped with a MEC server, which is also regarded as an independent network element complementing the existing cellular network without impacting the connectivity of other interfaces. These MEC servers are connected to corresponding base stations via dedicated reliable links to provide particular services for diverse end-user of such several nearby cells. MEC servers and centralized cloud centers are connected through the Internet backbone and share information with each other, so as to work in a cooperative manner [37].

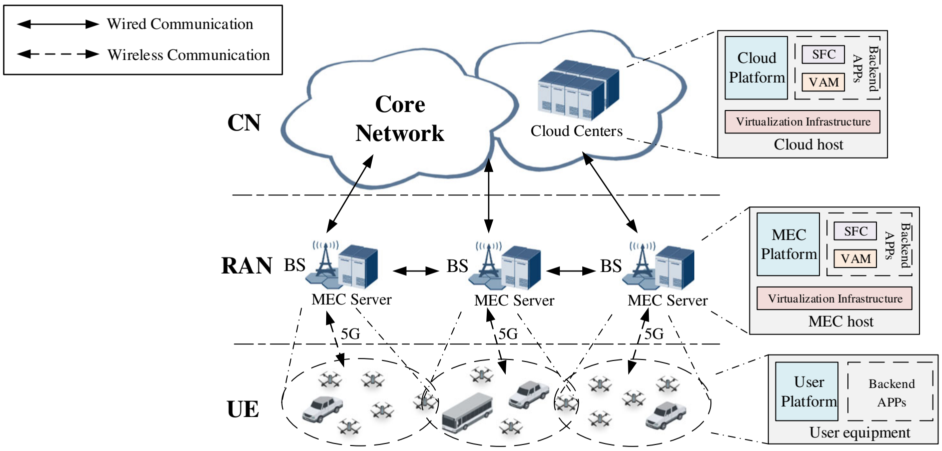


Fig. 1 The proposed MEC architecture. This is the proposed multi-access edge computing (MEC) architecture. In the architecture, MEC servers and BSs are integrated within the radio access network (RAN) and deployed on the user plane interface, i.e., between the core network (CN) and user equipment (UE). Each base station is equipped with a MEC server, which is also regarded as an independent network element complementing the existing cellular network without impacting the connectivity of other interfaces. These MEC servers are connected to corresponding base stations via dedicated reliable links to provide particular services for diverse end-user of such several nearby cells. MEC servers and centralized cloud centers are connected through the Internet backbone and share information with each other, so as to work in a cooperative manner

The MEC server enables user-related services and applications running as virtual machines (VM) and operating at the edge of the mobile network in a flexible and efficient way, based on a virtualization platform. Furthermore, particular network functions (NFs) are supposed to be tailored and instantiated by reusing the more abundant hardware resource of the MEC server, thereby enriching its service capability. Under the umbrella of the virtualization paradigm, it is feasible to integrate NFs and applications in the same virtualization infrastructure. Since the MEC servers have connection with both CN and UE, they could process flexible scheduling on user traffic according to the service demands. Although the incorporation of virtual network function (VNF)/service function chaining (SFC) and multi-access edge computing (MEC) allows 5G networks to deliver a variety of services and applications in a more flexible manner [28], the future cellular system calls for more intensive cell deployment and is designed to support more scenarios with high mobility [37]. Therefore, we have added the valuation adjustment mechanism (VAM) application to the MEC servers. This MEC networks can still provide stable services and dynamic service migration for common users and high-mobility users together in the ultra-dense network.

3.2 The VAM management application

We propose the VAM management application with the valuation adjustment mechanism (VAM) as the key, because there are the reasons for the difficulty in service migration for UE with high mobility in the ultra-dense network, i.e.,

1) The resource usage of the MEC servers cannot be synchronized and updated in time due to the frequent migration of UE with high mobility.

2) The frequent service migration of the UE with high mobility leads to an increase of the workload of the MEC servers.

3) A large number of UE applying for resources lead to resource crisis of MEC servers.

The concept of VAM originated from business, which was actually a form of an option. Through the design of the terms, the VAM can effectively protect the interests of investors. The VAM is that the acquirer (the investor) and the transferor (the financier) make an agreement on the future uncertainty when the merger or financing agreement is reached. If the agreed conditions arise, the financier can exercise a right; if the agreed terms do not appear, the investor can exercise a right.

If the MEC servers detect that the UE continuously moves among different areas in a short time, and the type of service that the UE is prepared to migrate includes an interaction type (low latency), the MEC servers use the VAM for the UE. Unlike ordinary UE, UE marked as high mobility needs to apply for resource usage time and hand in VAM code, which contains the credit level of the UE, while applying for resources. This credit level will change as the resource usage of the UE after the migration service conforms to the contract. If the UE migrates service again and returns resources near the agreed time, the VAM is completed, the credit level of UE as “the financer” is raised; If the UE does not migrate service again or return resources near the agreed time, the VAM is not completed, the credit level of UE as “the financer” is lowered. Moreover, the MEC servers as “the investor” determine the location of the migration service in the SFC and whether the service is transferred to other MEC servers based on the VAM code which contains credit level of the UE. However, since the VAM codes of UE are stored locally, the VAM codes may be modified, which will cause security issue. Therefore, the MEC servers must encrypt the VAM codes when modifying these codes and sending them to the UE again and verify the digital signature when accepting the VAM codes of the UE. In addition, the encryption parameters are periodically modified by the cloud servers, which will increase the security of the system without affecting the flexibility of the system.

The VAM application design, focusing on the functionalities located in the UE and MEC servers, is depicted in Fig. 2. In the MEC host, the VAM manager component manages the migration services of the UE with high mobility in the area, as well as the determination of the VAM and the modification of the VAM code. This VAM manager component, which is connected to the SFC application, has the right to determine the location of the migration task in the SFC and even decide whether to transfer the migration service to other MEC servers. This component facilitates SFC formation and SFC disengaging, as well as all UE joining/leaving the SFC. Furthermore, other UE that are not yet in the communication range of BSs can use the local VAM management component to start communicating with nearby MEC servers during the formation process. Other functionalities, residing in the cloud, take care of matching migration services with existing or new SFC and delegates tactical decisions to the correct MEC servers in case of sufficient resources. The encryption and decryption of the VAM code is the responsibility of the VAM encryption component, which is connected to the cloud servers for modifying regularly encryption parameters.

In a dynamic situation, the introduction of the VAM facilitates the MEC network to migrate services of UE with high mobility in the ultra-dense network, i.e.,

1) The mobility management of the MEC servers is implicitly undertaken by massive UE. The frequent service migration of UE with high mobility causes huge overhead for

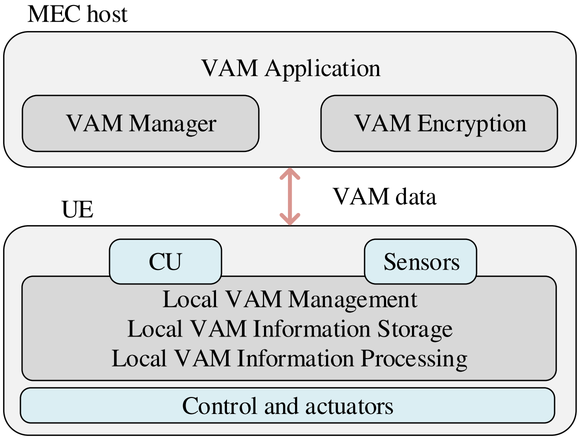


Fig. 2 The VAM management application architecture. This is the VAM application design focusing on the functionalities located in the UE and MEC servers. In the MEC host, the VAM manager component manages the migration services of the UE with high mobility in the area, as well as the determination of the VAM and the modification of the VAM code. This VAM manager component, which is connected to the SFC application, has the right to determine the location of the migration task in the SFC and even decide whether to transfer the migration service to other MEC servers. This component facilitates SFC formation and SFC disengaging, as well as all UE joining/leaving the SFC. Furthermore, other UE that are not yet in the communication range of BSs can use the local VAM management component to start communicating with nearby MEC servers during the formation process. Other functionalities, residing in the cloud, take care of matching migration services with existing or new SFC and delegates tactical decisions to the correct MEC servers in case of sufficient resources. The encryption and decryption of the VAM code is the responsibility of the VAM encryption component, which is connected to the cloud servers for modifying regularly encryption parameters

the MEC network, which also causes the resource changes of the MEC server to fluctuate so much that certain information cannot be obtained in time. The fundamental purpose of introducing VAM is to enable the MEC servers not to make complex judgments on how to give UE with high-mobility resources by massive UE submitting the VAM codes, the concentration of past performance of this UE. The proposed VAM code is one of the solutions to this type of problem, which can effectively alleviate the complexity of the MEC servers' judgments in high-mobility scenarios.

2) Special labels are added to all UE with high mobility at low cost. Based on the virtualization of network resources, a network slice forms an end-to-end logical network, which provides one or more network functions to cater to the demand side of the slice (e.g., vertical industry users, virtual operators, and enterprise users) [37]. However, network slices only identify the type of service and do not identify the subject of the service, as this would be a huge task. The introduction of VAM can be seen as adding different labels to all UE with high mobility to distinguish them. On the premise of network slicing, VAM aligns SFCs more accurately by identifying labels, which would not be a huge task because it is shared by massive UE.

3) The introduction of credit level in VAM makes the arrangement of service sequences in SFC more reasonable. Yu et al. [28] regarded the mobile SFC embedding problem as an NP problem and had made a series of studies on the placement and migration of SFC, but this would undoubtedly increase the workload of the MEC servers in the high-mobility scenarios. However, the credit level is a simple and effective basis for determining the order of services in SFC, which relies on the past performance of UE with high mobility.

3.3 The CMVAM

Each VNF of an SFC requires a certain amount of storage capacity and is connected to adjacent VNFs with virtual bandwidth. The VNFs are then chained in a specific order to realize the service [28]. For VNF n connecting to VNF m, we denote its storage, virtual link bandwidth requirements, and usage time as a three-dimensional vector of capacity-bandwidth-time, i.e., $(vc_l^n, vb_l^{mn}, vt_l^n)$, and regard each MEC server i as a bin, as shown in Fig. 3. Since each SFC includes multiple VNFs, the SFC management then becomes the bin-packing problem that the set of VNFs assigned to a set of bins connected by links characterized by the link bandwidth and propagation delay.

The choice of a set of candidate MEC servers in the proximity for the service migration is similar to a multi-bin packing problem. The difference is that there is no relation among items in the set of bins for the multi-bin packing problem; however, in the service migration, there is a specific order among the VNFs in the SFC. In general, the candidate bins to place a VNF include the bins with largest residual areas in the highest vt_l^n (i.e., the product of residual bandwidth and residual storage capacity). Let $rc_i(t)$ and $rb_{ij}(t)$ denote the residual capacity of MEC server i and the residual bandwidth of e_{ij}, respectively. Then, the candidate bin is determined by the weighted inner product of $(rc_i(t), rb_{ij}(t))$ and $(1,\omega)$, i.e.,

$$\max_{i,j \in M} \left\{ (rc_i(t), rb_{ij}(t)) \times (1, \omega) \right\} \tag{1}$$

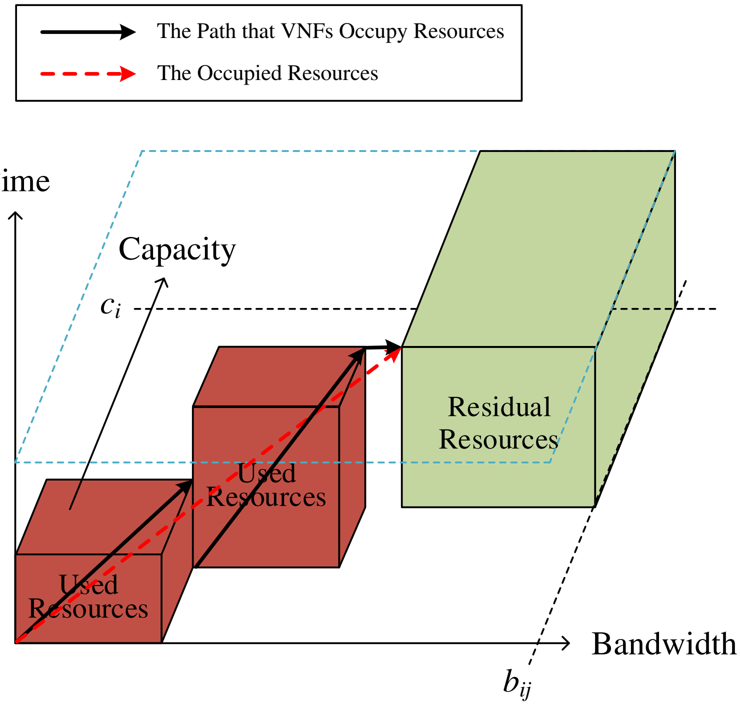


Fig. 3 The bin. This is the bin which each MEC server is regarded. In this architecture, we denote its storage, virtual link bandwidth requirements, and usage time as a three-dimensional vector of capacity-bandwidth-time

where

$$rc_i(t) = c_i - \sum_{l \in L, n \in S_l} x_i^{(n,l)}(t) \times vc_l^n, \forall i, t \tag{2}$$

$$rb_{ij}(t) = b_{ij} - \sum_{l \in L, (m,n) \in S_l} x_i^{(n,l)}(t) \times x_j^{(m,l)}(t) \times vb_l^{nm}, \forall i, j, t \tag{3}$$

where c_i is the total storage resource capacity of MEC server i; b_{ij} is the transmission bandwidth between MEC i and j; L is the total number of users in the system; S_l is the service function chaining (SFC) of user l; $x_i^{(n,l)}(t)$ is the decision variable indicating if VNF n of S_l is placed at MEC i at time t; vc_l^n is the capacity of MEC server i; vb_l^{mn} is the bandwidth of MEC server i.

$$\omega = \frac{ts_l}{tr_l}, \forall l \tag{4}$$

where ω characterizes difference in importance between $rb_{ij}(t)$ and $rc_i(t)$ under the same size of residual area, ts_l is the connection time of service, and tr_l is the residence time of service.

For VNFs from users with high mobility (i.e., shorter residence time in cells), traditional algorithm preferred to choose a candidate bin with a larger inner product because such bins tended to consume more bandwidth partly for VNF migration (handoff user) and partly for normal network access to other SFC owns (other normal access users). For more static users (i.e., higher residence time in cells), their VNFs were placed in a bin with a smaller inner product so that the bandwidth of the bin was saved for more frequently moving users [28].

Before all UE SFC is integrated, CMVAM will adjust the location of the service migration in advance. After the adjustment, a new chain will be generated on the basis of several original SFC. In the new chain, the service migrations requested by UE whose credit levels are positive in the VAM code will be all placed at the front end of the chain. Among them, the higher the UE's credit level, the higher the order of the service migrations that the UE applies for. The service migration requested by UE whose credit levels are negative in the VAM code will be all placed behind the ones requested by UE whose credit levels are positive. Among them, the lower the UE's credit level, the lower the order of the service migrations that the UE applies for. The UE that does not have a VAM code means that the UE is not determined to be without high mobility, and the service migrations requested by this UE will be placed behind the ones requested by UE with high mobility. Only a new queue will be generated for a while. Only when all service migrations in the chain are completed within the limited time, the service migrations in the next chain will continue. If there are still unresolved service migrations in the chain after a limited time due to the error of these services themselves, these service migrations are suspended until all the service migrations in the next chain are completed. This ensures that only wrong services will be suspended, and other service migrations will be performed successfully in the situation of having sufficient resources.

So far, CMVAM has integrated the SFCs of each UE in the current time period into one chain and performs service migration according to the VNF order in the chain. Upon handoff, CMVAM needs to perform service migration from the current MEC server to the target MEC server. However, all the VNFs of the chain may not be able to be placed in time to the new MEC server, due to the limited bandwidth on the link between MEC servers. To reduce service interruption, the VNFs of the chain could be placed in a set of nearby MEC servers as long as such a set of MEC servers suit migration requirements of these services. However, the current algorithms of selecting MEC servers only suit the single or multiple requirements, such as bandwidth, propagation delay, workload, and energy consumption and thus lead to two potential dangers, i.e.,

1) The greedy algorithm always makes the best choice at the moment and does not consider what might happen in the future. For example, interactive services have strict requirements for latency, but not all interactive services have the same tolerance interval for latency. If there is no optional MEC server for a service that requires extremely low latency because another interactive service with the relatively lower requirement of latency has previously selected this MEC server, the current service migration will fail. Moreover, the resources of MEC server, which are able to suit the requirements for previous service with the relatively lower requirement of latency, are wasted.

2) Uniform selection criteria cannot accommodate a variety of different types of service migrations. In addition to interactive tasks, cloud services can be divided into computational tasks and data-based tasks. The uniform selection metrics cannot be used for all types of service migrations because the resources required for these service migrations are different. Even for the same type of migration services, there are subtle differences in the migration requirements for MEC servers.

In view of the above situation, we propose an algorithm for evaluating migrated services, called migration effect evaluation with QoS-aware (MEEQ), to make a supplement to the existing algorithms for selecting MEC servers, which is conducive to choose the suitable MEC servers for migrated services. The MEEQ will be shown in detail in Section 4.

After the MEEQ, CMVAM calculates migration time of the VNFs in the chain according to the chaining order, among the $N_{\max}$ highest ranking migrating candidates $\{p_1, p_2, ..., p_{N_{\max}}\}$. Each migration path may consist of multiple links. The longer the migration path, the more the bandwidth is consumed, and the higher the chance in the occurrence of network congestion. Thus, we take the most residual resources first principle to decide migration path for each VNF in the chain, in case the migration time is met. (The quality of services has been guaranteed before by MEEQ.) After the service migration, the migrated services should not violate the constraints on the new MEC servers. Otherwise, the migrated services will be viewed as service down [28].

3.4 The total service migration procedure

Figure 4 illustrates the total handover (HO) and migration signaling based on dynamic service migration with the CMVAM. In the procedure, handover decision and some related configuration work are performed by the MEC server instead of the involved UE or BSs (as LTE generally does). The advantage of this approach is to avoid the redundant signaling interaction between the UE and the BSs, hence the handover latency is effectively reduced. This is the advantage of MEC, which can execute and migrate services with lower latency, while the estimated mean value of such latency in the existing LTE system is 12ms (for contention-free access) [38].

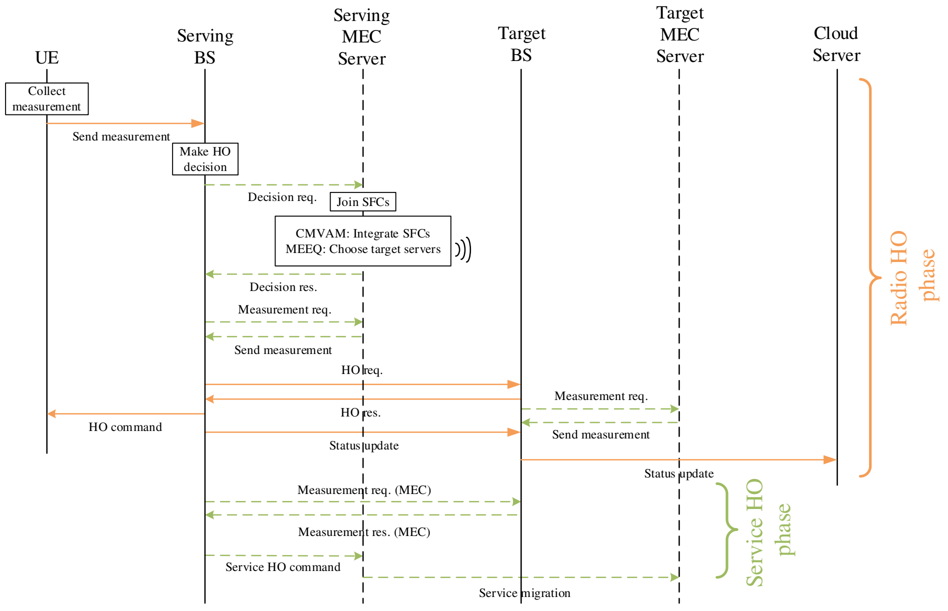


Fig. 4 The total HO and migration signaling. This is the total handover (HO) and migration signaling based on dynamic service migration with the CMVAM. In the procedure, handover decision and some related configuration work are performed by the MEC server instead of the involved UE or BSs (as LTE generally does). The advantage of this approach is to avoid the redundant signaling interaction between the UE and the BSs, hence the handover latency is effectively reduced. This is the advantage of MEC, which can execute and migrate services with lower latency, while the estimated mean value of such latency in the existing LTE system is 12 ms (for contention-free access)

4 Migration effect evaluation with QoS-aware

In this section, we propose an algorithm for evaluating migrated services, called migration effect evaluation with QoS-aware (MEEQ), to make a supplement to the existing algorithms for selecting MEC servers, which is conducive to choose the suitable MEC servers for migrated services. The MEEQ focuses on the impact of service experience in the different QoS-aware areas when interacting with humans. Moreover, MEEQ can evaluate the actual performance of the services (e.g., interactive services and computational services), which have different requirements for MEC servers. The MEEQ has been used in Section 4 to select the appropriate MEC servers for the service migration, as shown in detail in this section.

The three areas, which users' experience of QoS can be divided into, are the best area, the sensitive area, and the unusable area. The migrated services in the best area, which users have already obtained the best experience from, are not able to provide users with a better experience as the performances of these services continue to improve. The different performances of services only in the sensitive area can be clearly perceived by users, and will obviously affect the users' experience of migrated service. The migrated services in the unusable area cannot be used normally by users. The notations which we define in this section are summarized in Table 1.

The area determination s_k is used to determine the area where QoS parameters are located, i.e.,

$$s_k = \begin{cases} +\infty & QoS \text{ in the best area} \\ \frac{B_k - W_k}{B_k - q_k} & QoS \text{ in the sensitive area} \\ 1 & QoS \text{ in the unusable area} \end{cases} \tag{5}$$

Table 1 The notations of MEEQ

Notation	Description
q_k	The actual QoS parameter of service *k*
B_k	The boundary value of the best area and the sensitive area of service *k* (i.e., the best QoS value)
w_k	The boundary value of the sensitive area and the unusable area of service *k* (i.e., the worst QoS value)
s_k	The parameter used to determine the area where QoS parameters are located (i.e., the area determination)
$d(q_k, B_k)$	The value of difference between the best QoS value and the actual QoS parameters after adjusted (i.e., the difference)
dir_k	The nature of the QoS parameter, which is used to make the formula constant for positive values by assigned 1 or (– 1) (i.e., the directional feature value)
v_k	The satisfaction with service *k*, which is a real number between 0 and 1 (i.e., the user satisfaction)
f_k	The satisfaction with service *k* when $d(q_k, B_k) \in (0, +\infty)$
v_k^i	The user satisfaction for different QoS parameters *i* of service *k*
a_k^i	The weight for different QoS parameters *i* of service *k*
$VQoS_k$	The final values for evaluating the total user experience of the migrated service *k* in their MEC servers

where B_k is the best QoS value, W_k is the worst QoS value, and q_k is the actual QoS parameter of service k.

According to the area determination s_k, we calculate the value of difference between the best service experience and the actual QoS parameters and adjust this value to obtain the difference $d(q_k, B_k)$, i.e.,

$$d(q_k, B_k) = \begin{cases} +\infty & s_k = 1 \\ 0 & s_k = +\infty \\ (B_k - q_k) \times dir_k \div s_k & s_k \in (1, +\infty) \end{cases} \tag{6}$$

where q_k is the actual QoS parameter of service q, B_k is the best QoS value, dir_k is the directional feature value, and s_k is the area determination.

Next, the user satisfaction v_k that represents the user's satisfaction with the migrated services in the form of a numerical value can acquired based on the difference $d(q_k, B_k)$, i.e.,

$$v_k = \begin{cases} 0\infty & d(q_k, B_k) = +\infty \\ 1 & d(q_k, B_k) = 0 \\ f_k(d(q_k, B_k)) & d(q_k, B_k) \in (0, +\infty) \end{cases} \tag{7}$$

where $d(q_k, B_k)$ is the difference.

Due to the trend of the function Sigmoid [39] in line with our scaling requirements for QoS parameters in sensitive areas, we use this function as the basis function, take the defining domain of the function into $[0, (B_k\text{-}W_k) \times dir_k]$ and adjust the result range of the function into [0,1] to obtain the user satisfaction function $f_k(d(q_k, B_k))$, i.e.,

$$f_k\left(d(q_k, B_k)\right) = 1 - \frac{1}{1 + e^{-\left\{\frac{5 \times \left[d(q_k, B_k) - z\right]}{z}\right\}}} \tag{8}$$

$$z = \frac{(B_k - W_k) \times dir_k}{2} \tag{9}$$

where $d(q_k, B_k)$ is the difference, B_k is the best QoS value, W_k is the worst QoS value, and dir_k is the directional feature value.

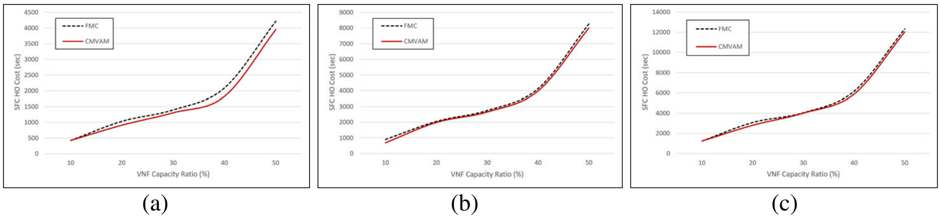
Moreover, different services have different requirements for different MEC servers, depending on the types of these services (e.g., interactive services and computational services). Therefore, we introduce special weight matrix, which reflects the relative importance of several QoS parameters for their services [40], to gain the final values $VQoS_k$ for evaluating the total user experience of the migrated services in their MEC servers, i.e.,

$$VQoS_k = \begin{bmatrix} v_k^1 & v_k^2 & \ldots & v_k^n \end{bmatrix} \times \begin{bmatrix} a_k^1 & a_k^2 & \ldots & a_k^n \end{bmatrix} \top \tag{10}$$

where v_k^i is the user satisfaction for different QoS parameters i of service k, $i \in [1, n]$; a_k^i is the weight for different QoS parameters i of service k, $i \in [1, n]$.

5 Results and discussion

In this section, we implemented a system prototype using the "networkx" to perform real service migration in an ultra-dense network. We saturated the ultra-dense network and test restoration of services in the blockage. Moreover, we tested the effects of different algorithms for evaluating migrated services on services. Results from extensive experiments prove the capability of the CMVAM and MEEQ.



(a) (b) (c)

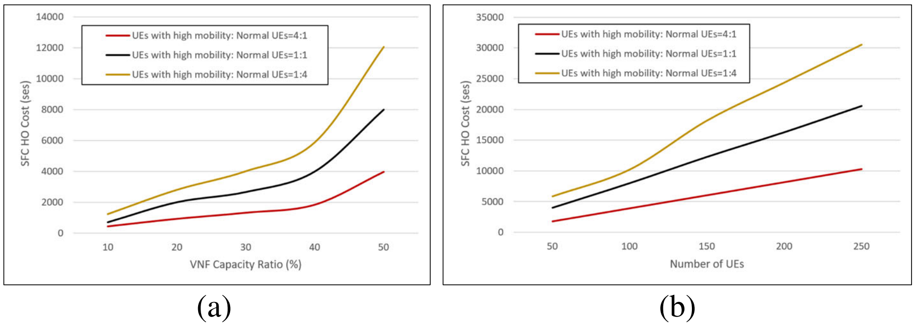
Fig. 5 The impacts of the VNF capacity ratio on the SFC HO costs. This is the results of the experiments about the impacts of the VNF capacity ratio on the SFC HO costs. The SFC HO cost meant the time that it took for all services to restore in the blockage, while the VNF capacity ratio meant the proportion of the storage of each VNF in the storage of total MEC server. In these experiments, we set the proportion of the UE of high mobility and the ordinary UE as 4:1, 1:1, and 1:4 in each cell

5.1 Experiment setup

We followed the ETSI MEC framework to develop the MEC network environment based on “networkx,” which was a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. The network that we considered consisted of 49 BSs in a hexagonal grid, where each BS was collocated with one MEC server. We set each slot time to be 1 s. We set the link bandwidth between any two MEC servers to 10 Gbps. We set the storage capacity of each MEC to 20 GB. The connection time of Each SFC owns service was ranged over 10 to 50 s. For all SFC owns, the bandwidth requirement between two VNFs was over 2 to 4 Gbps, and the latency requirement between two VNFs followed a uniform distribution with a mean over 1 to 3 ms. The residence time of ordinary UE followed a stochastic distribution within 300 to 1800 s, and the residence time of UE with high mobility followed a stochastic distribution within 5 to 50 s. Users were roaming with a random walk mobility model so that users might move to any one of the neighbors with an equal probability. We performed 10 experiments on each method and averaged them as a result.

5.2 Experimental results and analysis

The FMC had proven to perform better than the local optimal approach, random migration approach, and bandwidth-oriented migration approach [27]. In Figs. 5 and 6, the



(a) (b)

Fig. 6 The impacts of the different proportion of the UE on the SFC HO costs. This is the results of the experiments about the impacts of the different proportion of the UE on the SFC HO costs. The SFC HO cost meant the time that it took for all services to restore in the blockage, while the VNF capacity ratio meant the proportion of the storage of each VNF in the storage of total MEC server. In these experiments, we set the proportion of the UE of high mobility and the ordinary UE as 4:1, 1:1 and 1:4

Table 2 The QoS parameters and their weights

FP (%)	ARSD (s)	SPR (%)	MTR (s)	Weight of FP	Weight of ARSD	Weight of SPR	Weight of MTR
1	1.5	10	10	0.2	0.3	0.2	0.3
1.5	2	20	10	0.2	0.3	0.2	0.3
2	2.3	40	12	0.25	0.25	0.3	0.3
1.3	3	60	12	0.25	0.25	0.3	0.3
1.6	2.7	50	11	0.3	0.2	0.3	0.2

SFC HO costs were shown with regard to the VNF capacity ratio. The SFC HO cost meant the time that it took for all services to restore in the blockage, while the VNF capacity ratio meant the proportion of the storage of each VNF in the storage of total MEC server. We saturated the ultra-dense network and tested restoration of services in the blockage. We set the proportion of the UE of high mobility and the ordinary UE as 4:1, 1:1, and 1:4 in each cell, and the results were given in Fig. 5a–c, respectively. Figure 5a showed that the CMVAM can more quickly restore the services than the algorithm "Follow-Me Chain (FMC)" in the blockage. Thus, the CMVAM proved to performs better in the case of changes of the VNF capacity ratio. Since the proposed scheme managed the SFCs in the MEC servers by VAM code, it was possible to provide a fast service even when the UE frequently moved among different cells. In addition, the management advantage of CMVAM for SFC was best reflected especially when the VNF capacity ratio was 40%.

However, this advantage of performance was not explicitly reflected when the proportion of the UE of high mobility and the ordinary UE was 1: 1 or 1:4, as shown in Fig. 5b, c, because CMVAM was a management method for the UE of high mobility. The smaller the number of the UE with high mobility, the less intervention CMVAM had for services. Therefore, CMVAM can only play a better role only when the UE of high mobility accounted for the majority of all UE, as shown in Fig. 6a, b.

On the other hand, the MEEQ was an evaluation for the service effect essentially, which focused on the impact of service experience in the different QoS-aware areas when interacting with humans. To evaluate the importance of the QoS-aware areas, we set other

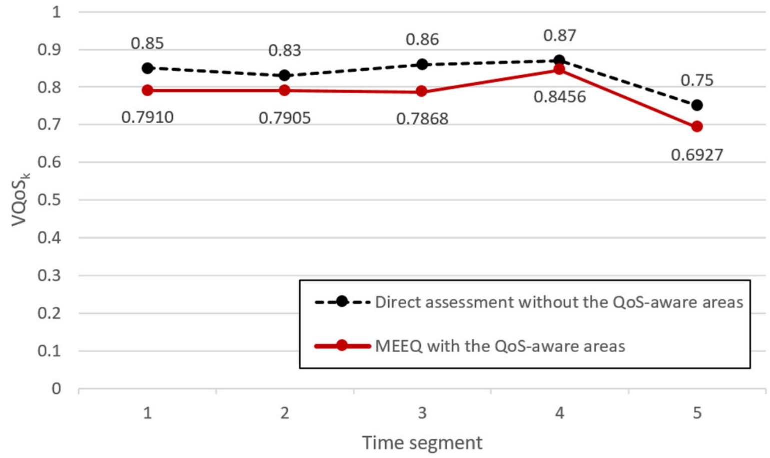


Fig. 7 The QoS evaluation of migrated services. This is the evaluation values with regard to the different time segment. This figure showed that the MEEQ proposed in this paper can better reflect the capability of migrated services, compared with direct assessment without the different QoS-aware areas

Table 3 The best QoS values and the worst QoS values

	FP (%)	ARSD (s)	SPR (%)	MTR (s)
The best QoS values	0	0.4	100	1
The worst QoS values	10	10	50	30

method without the QoS-aware areas as control group [40], in which the QoS parameters were failure rate (FP), average request submission delay (ARSD), stability of processing request (SPR), and mean time to repair (MTR). The QoS parameters and their weights were shown in Table 2.

In Fig. 7, the evaluation values $VQoS_k$ were shown with regard to the different time segment. The best QoS values and the worst QoS values that we set were shown in Table 3. Figure 7 showed that the MEEQ proposed in this paper can better reflect the capability of migrated services, compared with direct assessment without the different QoS-aware areas. In Fig. 8, since there were three times that the SPR appeared in the unusable area and two times that the SPR appeared near the boundary of the unusable area and the sensitive area, the final values $VQoS_k$ for evaluating the total user experience of the migrated services in their MEC servers were reduced. Therefore, it was also reasonable that the evaluation values of MEEQ were lower than other methods, which confirms the MEEQ can better reflect the quality of migrated services indeed.

6 Conclusions

To solve the problem of dynamic service migration in ultra-dense MEC network for high-mobility scenarios, we first propose a hierarchical MEC architecture, setting up the infrastructure for dynamic service migration in the ultra-dense MEC networks. Then, we propose a new mechanism "CMVAM" for users with high mobility in the ultra-dense MEC network, efficiently arranging service migrations for users with high-mobility and ordinary users together. Moreover, we propose an algorithm "MEEQ" for evaluating migrated services to contribute to choose the suitable MEC servers for migrated services.

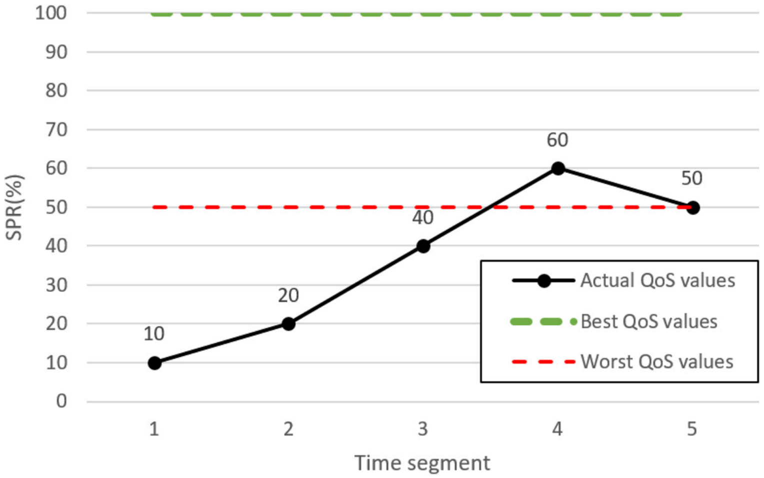


Fig. 8 The changes in SPR. This is the changes of the stability of processing request in the different time segment. There were three times that the SPR appeared in the unusable area and two times that the SPR appeared near the boundary of the unusable area and the sensitive area

The results show that the mechanism “CMVAM” can efficiently arrange service migrations and more quickly restore the services even in the blockage. On the other hand, the algorithm “MEEQ” is able to make a supplement to the existing algorithms for selecting MEC servers because it can better reflect the capability of migrated services.

Abbreviations

MEC: Multi-access edge computing; IoT: Internet of Things; ML: Machine learning; AR: Augmented reality; FG: Fog computing; BS: Base station; UE: User equipment; NOMA: Non-orthogonal multiple access; SFC: Service function chaining; VNF: Virtual network function; QoS: Quality of service; CMVAM: Chain management with valuation adjustment mechanism; MEEQ: Migration effect evaluation with QoS-aware; vMSG: Virtualized mobility support gateway; EMM: Energy-aware mobility management; MPTCP: MultiPath-TCP; vCRAN: Vehicular cloud radio access network; RRH: Remote radio head; SDN: Software-defined network; RSU: Roadside unit; VANET: Vehicular ad hoc networks; FMC: Follow-Me Chain; RAN: Radio access network; CN: Core network; VM: Virtual machines; NF: Network function; VAM: Valuation adjustment mechanism; HO: Handover; FP: Failure rate; ARSD: Average request submission delay; SPR: Stability of processing request; MTR: Mean time to repair

Acknowledgements

We would like to thank the reviewers for their comments helping us improve the quality of this paper.

Authors’ contributions

Haowei Lin conceived of the study and participated in its design and coordination and helped to draft the manuscript. Xiaolong Xu carried out the modification of the architecture in ultra-dense multi-access edge computing network. Juan Zhao carried out the experimental correction and screening. Xinheng Wang carried out the provision of information and design of the dynamic service migration. The authors have read and approved the final manuscript.

Funding

This work is sponsored by the National Key Research and Development Program of China (2018YFB1003700), etc.

Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests

The authors declare that they have no competing financial interests.

Author details

[1] Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, NanJing, China. [2] School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, NanJing, China. [3] Department of Electrical and Electronic Engineering, Xi’an Jiaotong-Liverpool Univesity, Suzhou, China.

Received: 12 May 2020 Accepted: 17 September 2020
Published online: 06 October 2020

References

1. J. Lee, D. Kim, J. Lee, Zone-based multi-access edge computing scheme for user device mobility management. Appl. Sciences-Basel. **9**(11), 2308–2323 (2019)
2. S. Barmpounakis, G. Tsiatsios, M. Papadakis, E. Mitsianis, Collision avoidance in 5G using MEC and NFV: the vulnerable road user safety use case. Comput. Netw. **172** (2020)
3. A. Baydin, B. Pearlmutter, A. Radu, J. Siskind, Automatic differentiation in machine learning: a survey. J. Mach. Learn. Res. **18**, 1–43 (2018)
4. R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, R. Gao, Deep learning and its applications to machine health monitoring: a survey. Mech. Syst. Signal Process. **115**, 213–237 (2019)
5. H. Ye, L. Liang, G. Li, G. Kim, L. Lu, M. Wu, Machine learning for vehicular networks: recent advances and application examples. IEEE Veh. Technol. Mag. **13**, 94–101 (2018)
6. R. Palmarini, J. Erkoyuncu, R. Roy, H. Torabmostaedi, A systematic review of augmented reality applications in maintenance. Robot. Comput. Integr. Manuf. **49**, 215–228 (2018)
7. J.Lee, J.Lee, Preallocated duplicate name prefix detection mechanism using naming pool in CCN based mobile IOT networks. Mob. Inf. Syst. **2016**, 1–9 (2016)
8. E. Ahmedand, H. Gharavi, Cooperative vehicular networking: a survey. IEEE Trans. Intell. Transp. Syst. **19**(3), 996–1014 (2018)
9. Y. Mao, I. Zhang, K. Letaief, Dynamic computation offloading for mobile-edge computing with energy harvesting devices. IEEE J. Sel. Areas Commun. **34**, 3590–3605 (2016)
10. P. Mach, Z. Becvar, Mobile edge computing: a survey on architecture and computation offloading. IEEE Commun. Surv. Tutorials. **5**, 450–465 (2018)
11. H. Li, G. Shou, Y. Hu, Z. Guo, in *2016 IEEE 4th International Conference on Mobile Cloud Computing, Services, and Engineering*, Mobile edge computing: progress and challenges, vol. 19, (Oxford, 2016), pp. 83–84
12. J. Lee, J.Lee, Hierarchical mobile edge computing architecture based on context awareness. Appl. Sci. **8**, 1160 (2018)
13. H. Peng, J. Wang, A multicriteria group decision-making method based on the normal cloud model with zadeh'sz-numbers. IEEE Trans. Fuzzy Syst. **26**, 3246–3260 (2018)

14. C. Stergiou, K. Psannis, B. Kim, B. Gupta, Secure integration of IOT and cloud computing. Futur. Gener. Comput. Syst. **78**, 964–975 (2016)
15. X. Guan, X. Wan, F. Ye, B.Choi, in *2018 IEEE International Smart Cities Conference*, Handover minimized service region partition for mobile edge computing in wireless metropolitan area networks, (Trento, 2018), pp. 1–6
16. J. Wang, K. Liu, M. Ni, J. Pan, in *2018 IEEE Global Communications Conference*, Learning based mobility management under uncertainties for mobile edge computing, (Taiwan, 2018), pp. 1–6
17. Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, H. Poor, Application of non-orthogonal multiple access in LTE and 5G networks. IEEE Commun. Mag. **55**(2), 185–191 (2017)
18. H. Huang, J. Xiong, J. Yang, G. Gui, H. Sari, Rate region analysis in a full-duplex-aided cooperative nonorthogonal multiple-access system. IEEE Access. **5**, 17869–17880 (2017)
19. Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, L. Hanzo, Nonorthogonal multiple access for 5G and beyond. Proc. IEEE. **105**(12), 2347–2381 (2017)
20. G. Gui, H. Huang, Y. Song, H. Sari, Deep learning for an effective nonorthogonal multiple access scheme. IEEE Trans. Veh. Technol. **67**(9), 8440–8450 (2018)
21. Y. Wu, L. Qian, H. Mao, X. Yang, X. Shen, Optimal power allocation and scheduling for non-orthogonal multiple access relay-assisted networks. IEEE Trans. Mobile Comput. **17**(11), 2591–2606 (2018)
22. P. Corcoran, S. K. Datta, Mobile-edge computing and the internet of things for consumers: extending cloud computing and services to the edge of the network. IEEE Consum. Electron. Mag. **5**(4), 73–74 (2016)
23. X. Chen, Q. Shi, L. Yang, J. Xu, Thriftyedge: resource-efficient edge computing for intelligent IOT applications. IEEE Netw. **32**(1), 61–65 (2018)
24. Q. Fan, N. Ansari, Application aware workload allocation for edge computing-based IOT. IEEE Internet Things J. **5**(3), 2146–2153 (2018)
25. Y. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile edge computing: a key technology towards 5G. ETSI white paper. **11**(11), 1–16 (2015)
26. T. Kondo, K. Isawaki, K. Maeda, in *2018 IEEE 42nd Annual Computer Software and Applications Conference, vol. 2*, Development and evaluation of the MEC platform supporting the edge instance mobility, (Tokyo, 2018), pp. 193–198
27. I. A. Ridhawi, M. Aloqaily, Y. Kotb, Y. A. Ridhawi, Y. Jararweh, A collaborative mobile edge computing and user solution for service composition in 5G systems. Trans. Emerg. Telecommun. Technol. **29**, 3446 (2018)
28. Y. Chen, W. Liao, in *2019 IEEE International Conference on Communications*, Mobility-aware service function chaining in 5G wireless networks with mobile edge computing, (Beijing, 2019), pp. 1–6
29. Y. Sun, S. Zhou, J. Xu, EMM: energy-aware mobility management for mobile edge computing in ultra dense networks. IEEE J. Sel. Areas Commun. **35**(11), 2637–2646 (2017)
30. M. Aloqaily, I. A. Ridhawi, H. Salameh, Y. Jararweh, Data and service management in densely crowded environments: challenges, opportunities, and recent developments. IEEE Commun. Mag. **57**, 81–87 (2019)
31. V. Balasubramanian, M. Aloqaily, F. Zaman, Y. Jararweh, in *2018 IEEE 7th International Conference on Cloud Networking*, Exploring computing at the edge: a multi-interface system architecture enabled mobile device cloud, (Tokyo, 2018), pp. 1–4
32. K. Yonggang, A. Namwon, P. Jaehyoung, L. Hyuk, in *2018 IEEE 7th International Conference on Cloud Networking*, Mobility support for vehicular cloud radio-access-networks with edge computing, (Tokyo, 2018), pp. 1–4
33. W. Nasrin, J. Xie, in *2018 IEEE International Conference on Communications*, SharedMEC: sharing clouds to support user mobility in mobile edge computing, (Kansas City, 2018), pp. 1–6
34. T. Ojima, T. Fujii, in *2018 International Conference on Information Networking*, Resource management for mobile edge computing using user mobility prediction, (Chiang Mai, 2018), pp. 718–720
35. Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. Braun, S. Sargento, Mobility prediction-assisted over-the-top edge prefetching for hierarchical vanets. IEEE J. Sel. Areas Commun. **35**(8), 1786–1801 (2018)
36. Q. Peng, Y. Xia, Z. Feng, J. Lee, in *2019 IEEE International Conference on Web Services*, Mobility-aware and migration-enabled online edge user allocation in mobile edge computing, (Beijing, 2019), pp. 91–98
37. L. Li, Y. Li, R. Hou, in *2017 IEEE Wireless Communications and Networking Conference*, A novel mobile edge computing-based architecture for future cellular vehicular networks, (San Francisc, 2017), pp. 1–6
38. European Telecommunications Standards Institute (ETSI), Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN), V13.0.0, 3GPP (2018)
39. F. Sheng, Y. Yin, C. Qin, K. Zhang, Research and implementation based on transcendental function coprocessor sigmoid function. Microelectron. Comput. **35**(2), 11–14 (2018)
40. X. Cai, X. Zhang, An energy efficiency evaluation model based on QoS sources reduction in cloud computing environments. Comput. Eng. Sci. **36**(12), 2305–2311 (2014)

Publisher's Note