Cognitive Service Architecture for 6G Core Network

Yuanzhe Li, Jie Huang, Qibo Sun, Tao Sun and Shangguang Wang, *Senior Member, IEEE*

***Abstract*—5G communication is making much progress in achieving the Internet of Things and improving the quality of user experience in large bandwidth scenarios. By introducing a variety of new technologies, the performance of 5G has been greatly improved. However, emerging applications put forward more stringent requirements in terms of latency, reliability, peak data rate, service continuity, etc. Communication technology still needs to be further developed. In this paper, the next generation of core network is conceptualize. Inspired by the nervous system of octopus, we propose a new cognitive service architecture. Cognitive service architecture is a new architecture designed for 6G core network. It is proposed to enhance the core network so that it is qualified for the increasingly high requirement for quality of service and complicated scenarios. We first give a short vision on the 6G core network. Then cognitive service architecture is demonstrated in detail. A case study is demonstrated to show how cognitive service architecture enhance the performance of system. Enabling technologies for 6G cognitive service architecture are discussed at last.**

***Index Terms*—6G, Core Network, Cognitive Service.**

I. Introduction

A. An Overview of 5G Network

THE past few years have witnessed the rapid development and accelerating deployment of the 5G communication technology. 5G is not a simple update of the 4G communication technology. It stands out in terms of three communication scenarios including enhanced Mobile BroadBand (eMBB), massive Machine Type Communications (mMTC) and Ultra Reliable Low Latency Communications (URLLC) [1]. In particular, 5G network provides user devices with 0.1 Gbps data rates in the uniform spatial distribution with 10-20 Gbps peak data rates in eMBB [2]. For mMTC, the number of connected devices supported in 5G increases 10 to 100 times [3]. For delay-sensitive applications, 5G provides uRLLC to achieve low latency service with reliability. When end to end latency is as low as 1 ms, the reliability is guaranteed as high as 99.99% [4].

In order to increase the user's access bandwidth, various emerging technologies have been proposed, such as millimeter wave communication, massive multiple-input multiple-output, ultra-dense network, etc. To achieve low latency in 5G core network, mobile edge computation is proposed [5]. Services can be deployed at proximity to prevent the unnecessary latency contributed by the transmission between radio access network and core network. Besides, a variety of network technologies, such as network slicing, software defined network and network function virtualization, are introduced to make the network architecture flexible [6]. For example, with the application of network function virtualization, traditional functions have realized the decoupling of software and hardware. Virtual network functions running on cloud servers take the place of dedicated devices. This reduces the equipment cost and makes the deployment of system more flexible.

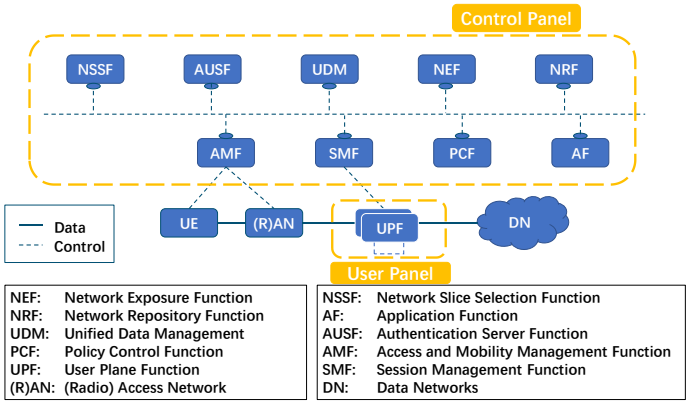

Fig. 1. 5G Service Based Architecture.

Based on these technologies, service based architecture [1] is adopted as the standard of 5G core network architecture. As is shown in Fig. 1, 5G core network is split into multiple network functions. Each network function is achieved by means of multiple microservices. By defining a unified interface specification, the network functions are decoupled from each other. The decoupled network functions can be expanded and deployed independently. As a result, the flexibility and efficiency of core network are improved. The same network function can be called by multiple network functions. This reduces the coupling degree of interface definition between network functions and bring advantages such as better load balance, better disaster tolerance, easy capacity expansion and upgrading. Finally, it realizes the customization of the whole network function on demand [7].

However, 5G still remains to be further developed in the following aspects.

1) Poor Coverage: Currently, 5G network is deployed mainly as terrestrial mobile communication. However, high deployment cost and technology limitations makes it impossible to cover everywhere. In fact, remote areas are hardly covered. At present, about 80% of the land area and more than 94% of the sea area on the earth are not within the coverage

Yuanzhe Li, Jie Huang, Qibo Sun and Shangguang Wang, are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. (Corresponding author: Shangguang Wang)
E-mail:{buptlyz;huangjie;qbsun;sgwang}@bupt.edu.cn
Tao Sun is with China Mobile Research Institute, Beijing, China.
E-mail:suntao@chinamobile.com

TABLE I
COMPARISON OF POSSIBLE KEY PERFORMANCE INDICATORS BETWEEN 6G AND 5G

Factors	6G	5G
Peak data rate	$> 1000\ Gbps$	$10\ Gbps$
Connection density	$> 10\ million/km^2$	$1\ million/km^2$
Mobility	$> 1000\ km/h$	$350\ km/h$
Delay	$< 0.1\ ms$	$Tens\ ms$
Reliability	$> 99.99999\%$	99.99%

of terrestrial mobile communication networks [6]. That means network is not available when users are in mountains or sailing across the sea where base stations are not deployed. Terrestrial mobile communication cannot provide network access for upper air, either.

Although non-3GPP access network has already been supported since 3G [8]. 5G core network has supported satellite communication [9], as well. The existing network architectures are mainly designed for terrestrial mobile communication. They cannot guarantee the quality and continuity of service when users switch between different access modes.

2) Not Capable of Internet of Everything: 5G has made much progress in achieving the goal of Internet of Things, especially in the three typical scenarios, i.e. eMBB, mMTC and uRLLC. However, it still has a long way to go. It still fails to provide service for data-rate intensive applications with ultra low latency [10]. For example, multi-sensory XR are among those applications which 5G network does not support well enough. Multi-sensory XR applications indicates virtual reality, augmented reality and mixed reality applications that provide not only immersive experience, but also strong real-time interaction. However, current network architecture is still not capable of satisfying both low latency and high data rate at the same time. Besides, how to guarantee the quality of service in highly dynamic environments, for example frequent service migration triggered by user movement, is still a challenge.

3) Lacking in Intelligence and Flexibility: Through the ubiquitous network connection, large number of computing requests converge to the edge of the network. Edge devices are faced with huge demand for computing resources. The constant emergence of new delay-sensitive and computation-intensive applications further aggravates the scarcity of resources. To cope with this situation, an on-demand real-time scheduling of communication resources and computing resources should be achieved efficiently. However, current 5G network architecture is not intelligent and flexible enough to be qualified. It lacks real-time perception and adaptive cognition of scenario changes. Although 5G introduces network slicing, it mainly provides services based on a small amount of semantic adaptation and simple rule matching for specific scenarios. It is difficult to deal with changeable network scenarios with high demand on short latency, such as multi-sensory XR applications and internet of vehicles.

B. Vision of 6G Core Network

In the 6G communication, new technologies, such as tera-hertz communications, visible light communications, advanced access-backhaul integration, etc., are introduced to realize ultra-high peak rate, ultra mass access, ultra-high reliability in communication network [11]. As is shown in Table I, compared with 5G, the performance of 6G will be greatly improved. However, only introducing new technologies is not enough. It is difficult to solve the aforementioned problems within current architecture. The core network architecture should be redesigned to achieve a powerful, flexible and intelligent network. In the following, we will give a brief vision of the 6G network.

1) Multiple Types of Mobile Communication: To access the Internet from anywhere, satellite communication, unmanned aerial vehicle (UAV) communication and maritime communication will be deeply integrated into 6G network. They will work as a supplement of terrestrial communication and greatly expand the coverage of 6G network. Note that, although satellite communication, UAV communication and maritime communication are all access network technologies, the expansion of the access network needs the support of the core network. The complex and changeable access scenarios set higher requirements for 6G core network. Take satellite communication as an example (shown in TABLE II), it has larger coverage and is not constrained by terrain. However, satellite communication also have problems such as long latency, higher path loss, Doppler shift and more frequent service migration. This presents great challenge to the 6G core network. First, unlike base stations located at one place, fast moving satellites and UAVs make network control more complicated. The core network has to deal with the frequent handover. To cope with the highly dynamic scenarios, the ability to cognize and predict changes in the access network environment is indispensable to core network. Second, latency of different communications varies a lot, making it hard for to provide integrated service with guaranteed quality. Third, handovers between different communications trigger not only communication handovers but also service migrations. Both communication topology and computation provision have to be rescheduled.

2) Reliable Low Latency with Mobile Broadband: As is mentioned, multi-sensory XR will be killer applications in the age of 6G. To satisfy the service quality requirements of these applications, both large bandwidth and low latency should be guaranteed to prevent deterioration of user experience such as black edge and frame dropping [12]. Considering XR are usually wearable devices, users may move around while wearing it. Communication hand-off and service migration may be frequently triggered, which will increase uncertainty of the system. In addition, multi-sensory XR applications are computation exhausted and latency sensitive. Its real-time rendering needs a lot of computing resources and a quick response. Both insufficiency of computing resources and network congestion will result in latency constraint violation. Providing a good network for these applications poses a huge challenge to the 6G core network. To guarantee the quality of

TABLE II
SATELLITE COMMUNICATION AND TERRESTRIAL COMMUNICATION

Factors	Satellite mobile communication	Terrestrial mobile communication
Link type	Service link & feeder link	Service link
Transmission distance	Above 600 km	About 1 km
Delay	Hundreds ms	Tens ms
Path loss	Above 180 dB	Within 140dB
Doppler shift	Up to several hundred kHz	Within kHz level
Cell Radius	hundreds km	300-500 m
Mobility	Inter-beams, inter-satellites	Inter-cells
Service Migration Trigger	Satellite mobility, user mobility	Only user mobility

service, 6G core network should have the ability to identify changes in the network environment. Then, communication and computation resources should be rescheduled quickly and fine-grained.

3) Communication Integrated with Artificial Intelligence: 6G network is obliged to fulfill the vision of Internet of Everything. Large number of heterogeneous devices are connected by 6G network and empowered by Artificial Intelligence (AI) applications. As AI applications are usually computation intensive, how to guarantee computation provision is a key challenge. Traditional cloud driven AI poses heavy load on backbone networks and suffers from long latency. Developing computation resources at network edge is also faced with resource limitation challenge. In previous generations of mobile communication technologies, the communication network only serve as a pipeline between user devices and cloud servers. As a result, communication and computation resources are scheduled individually. 6G will make a difference. It is not a simple pipeline system for information. Instead, it's more like a large field integrated with AI and computation resources. We call it the network empowered AI. All the devices in this field is empowered and scheduled by the network and thus form an Internet of Everything. On the other hand, to make the network work efficiently, the core network of 6G must have the ability to schedule all the devices within the coverage of the network. In this process, the controllers have to leverage the power of AI to deal with the fast changing heterogeneous environment. We call it the AI empowered network.

C. Related Work

Most existing works discuss the key performance indices and challenges [6] [10] [11] [13] [14] [15] [16] [17]. They analyze the technological trends and give potential solutions. Some work focus on discussing how to apply a specific technology in 6G network. T. Hewa et al. [18] give a vision on the role of blockchain in 6G and talk about the formidable challenges. Nei Kato et al. [19] analyze the challenges of applying machine learning technologies in 6G system. Rubayet Shafin et al. [20] give a possible roadmap on realizing artificial intelligence-enabled cellular network in 6G. Shuhang Zhang et al. [21] propose a UAV-to-Everything networking in 6G and put forward a reinforcement learning based framework for UAV-to-Everything communication. Wen Sun et al. [22] we present a new vision of Digital Twin Edge Networks in 6G scenario and propose a mobile offloading scheme to reduce offloading latency. Improving the performance of wireless communication at 6G base station is another hot topic. Shanzhi Chen et al. [23] study beam-space multiplexing in 6G system. Ertugrul Basar [24] introduces reconfigurable intelligent surface assisted communications to index modulation. It is a potential beyond MIMO solution that can be applied in 6G. Rony Kumer Saha [25] study the dynamic spectrum access and reuse in millimeter-wave spectrum and propose a hybrid technique involving interweave-underlay spectrum access and reuse. In Rony Kumer Saha's another work[26], he puts forward dynamic exclusive-use spectrum access method for 6G millimeter-wave communication. There is not much work on 6G network architecture. Guan Gui et al. [27] discuss key performance indices and propose a general architecture which demonstrate how different technologies are integrated. Xiuquan Qiao et al. [28] propose design principles for a distributed, decentralized, and intelligent application provisioning architecture for 6G.

Most existing works mentioned above focus on either giving a vision of 6G network or studying a specific technology that may be adopted in 6G. Very few work studies how to upgrade the 6G network architecture. [27] [28] give a general 6G architecture, but lack specific design.

D. Contribution

In this paper, we focus on upgrading the core network architecture of 6G. Although 5G core network architecture has made much progress by adopting service based architecture, it is still not qualified for the emerging scenarios such as space-air-ground multiple access network, internet of everything and providing intelligent and flexible network service. The existing centralized architecture of core network lacks real-time perception and adaptability to the complex and changeable scenarios in 6G. In this paper, we focus on the core network architecture of 6G. The contributions are as follows:

1) We propose a cognitive service architecture for 6G core network. As far as we know, this is the first 6G core network architecture that have ever been put forward.
2) Basic conceptions of cognitive service architecture are presented. Possible enabling technologies are analyzed and discussed.
3) A case study is conducted to evaluate the performance of cognitive service architecture. In the process of session establishment, cognitive service architecture can save more than 55.11% time than non-cognitive service architecture.

The remainder of this paper is organized as follows. Section II demonstrates the design of cognitive service architecture. In specific, Section II-A presents features of octopus's nervous system. Section II-B describes cognitive service and

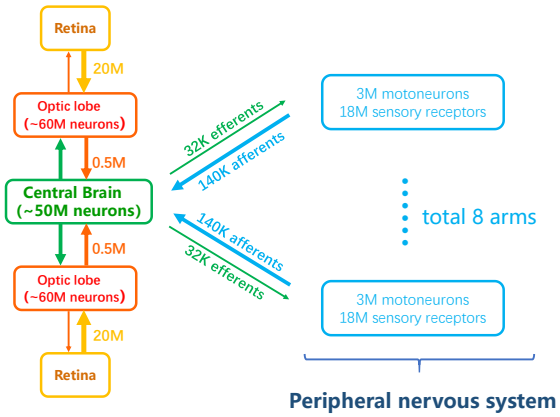

Fig. 2. The Nervous System of an Octopus.

other key concepts of cognitive service architecture. Section III summarizes enabling technologies for cognitive service architecture. Section V concludes the paper.

II. 6G Core Network Architecture

Diversified objectives, changeable service scenarios and personalized user requirements not only require 6G network to have large capacity, ultra-low delay, but also a remarkable degree of plasticity. In the face of the ever-changing requirements in the distributed scenario, 6G network architecture should have enough flexibility and scalability and is able to make very fine-grained adjustments to the network in the control layer. In this section, we dive into the design of 6G core network and try to propose an improved 6G core network architecture.

A. Inspiration

To satisfy requirements of new applications in 6G era, 6G core network must be powerful and flexible with high efficiency. It should meet following requirements: 1) ultra-large network capacity, 2) computation resources provided in the proximity, 3) strong cognition ability to recognize state change of the environment, and 4) efficient and compatible control layer.

It is quite difficult to design and deploy a system meeting all the above requirements. To find out a proper architecture, we try to get inspiration from animals. After a lot of investigations, octopus stands out because of its unique nervous system. Octopus is known for its flexibility and agility. Its nervous system is impressive in the following three aspects.

First, it has a strong and sensitive perception ability. The arms of octopus are excellent in terms of the sense of touch. The suction cups of arms are equipped with chemical receptors by which octopus has the ability to taste what it touches. Because the powerful cognitive ability brought by chemical receptors, octopus can recognize its skin and prevent arms from tangled or stuck to each other [29].

Second, it has an excellent control of its soft and complex body. The octopus's body is not limited by joints and skeletons. The long and flexible arms can be extended, shortened and bent at any point in any direction and length. The nervous system must deal with infinitely large degrees of freedom when it tries to control the arms [30]. To simplify the process of control, the octopus reduces the arms' degrees of freedom by

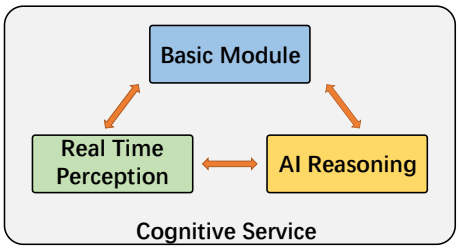

Fig. 3. Cognitive Service.

adopting special motion control strategies. For example, in a grasping action, the octopus's arm is divided into three parts by pseudo-joints. The middle section is as long as the near section. The far section is like a hand to grab food with a set of suction cups. Pseudo-joints are reshaped for each grasping action, and can be adjusted according to the position of the target and arm. The octopus calculates the position of the pseudo-joints from the arm itself: after touching the target, two muscle activation waves propagate toward each other. One propagates along the contact position to the proximal end and one propagates along the base of the arm to the far end, and the elbow is where the two waves collide [31].

Third, the structure of the octopus's nervous system is quite unique. As is shown in Fig. 2, the octopus' nervous system is divided into three parts: a central brain, two large optic lobes connected to eyes and axial nerve cords along the arms [30]. Most nerve cells are located in their nerve cords of its arms. Only part of nerve cells are localized in the brain. This structure is dramatically different from skeletal animals. Besides, the number of afferent and efferent fibers are relatively few. This means most complex motor skills are organized in peripheral nervous system in their arms [30][32]. The arm neuron network produces neuron activation patterns that specify all the space-time details of the basic mode of motion. The brain sends global commands to the arm neural network to activate and scale program variables [33].

Overall, octopuses achieve a good control of their soft body by leveraging a distinguished nervous system. Massive receptors bring excellent cognitive ability. Special motion strategies simplify the control. And most importantly, most decisions are made in the arms and the central nervous system only serve as a coordinator. All these are good examples on how to cope with control problems in a complicated system.

B. Cognitive Service Architecture for 6G Core Network

Inspired by the nervous system of octopus, we put forward cognitive service architecture for 6G core network. Cognitive service architecture is an upgrade of 5G service based architecture. In the following, we will introduce what is cognitive service, how it works and what is cognitive service architecture in details.

1) Cognitive Service: Cognitive service is the key concept of cognitive service architecture, which is an upgrade of traditional network functions. As is shown in Fig. 3, the upgrade is conducted by means of introducing two abilities. The first is the real time perception ability. Traditional network function interfaces are transformed into polymorphic interfaces with

cognitive ability, which endow cognitive services with multi-dimensional real-time perception ability, including request flow, resource and topology status, operation and maintenance events, etc. The second is AI reasoning ability. AI operators, rule matching units and approximate reasoning units are built in the network functions to realize online feature matching and local reasoning. Thus, basic module of traditional network function, real time perception and AI reasoning work as a whole to form a cognitive service.

However, only introducing cognitive service is not enough. The real time perception ability of different cognitive services should coordinate with each other to conduct complex perception. Besides, the AI reasoning ability should be updated regularly to achieve better performance. Therefore, at the system level, cognitive scheduling and knowledge graph management are introduced. Cognitive scheduling arranges, prunes and merges the core network system and dynamically adjusts the system capability according to the cognitive information of the network function and the polymorphic interface. Knowledge graph management works based on the theory of knowledge graph. It gathers information from network functions and interfaces and update the knowledge graph of cognitive service. Then the cognitive ability of network functions and polymorphic interfaces are upgraded based on the updated knowledge graph.

Knowledge graph is the core component of system's cognitive ability. It is in charge of analyzing and modeling of service quality demand and scene context. As is shown in Fig. 4, the knowledge graph building system consists of two modules: Knowledge mining module and knowledge generation module. Knowledge mining module has three sub-modules, which are in charge of extracting entity, relationship and attribute from unstructured data, respectively. Knowledge generation module makes use of structured data and information extracted by knowledge mining module to generate knowledge graph. It has several sub-modules. Knowledge Representation represents entities and relations as the structure of the three-tuple. Entity alignment points synonymous entities with different name to the same objective object. Ontology construction use top-down method to construct a hierarchical relationship between a set of conceptual definitions and concepts. Knowledge reasoning further explore the hidden knowledge based on the existing knowledge base to expand the knowledge base and obtain reasoning results. The process of constructing knowledge graph works as follows. First, user requirements for communication and computation resources are extracted from massive service instances based on data mining. The structured data in it is directly integrated with data in the third-party database. The unstructured data first enters the knowledge mining module. It is processed through entity extraction, relationship extraction and attribute extraction. Then, the integrated and mined data goes through knowledge representation and entity alignment operation. Next, if the represented knowledge does not exist on the ontology, the ontology is constructed first. After that, the represented knowledge is updated to the new ontology. Finally, all knowledge ontology forms knowledge graph after the process of quality assessment. The knowledge graph generated in this scenario can be imported into a model library for new scenarios.

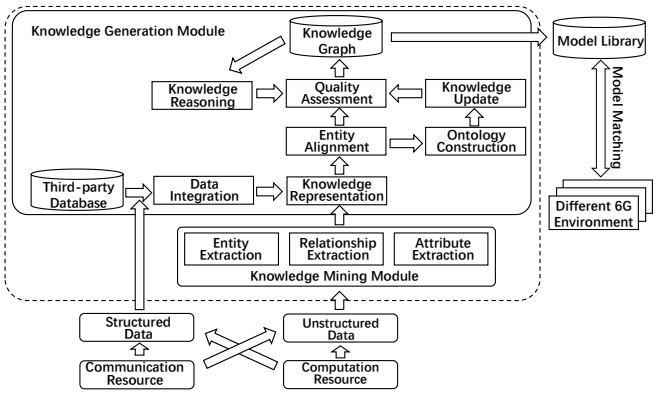

Fig. 4. Knowledge Graph for 6G Communication.

In traditional network architecture, it is difficult for the applications to accurately get the network performance in real time. In turn, the network cannot make an adjustment in real time according to the needs of the applications, either. The inability to adaptively match computing resources and communication resources leads to a decline in service experience in highly dynamic environments, especially for delay-sensitive and resource-exhausted applications. Therefore, the cognitive ability is of great importance. This is just like an octopus needs to know where the food is and which arm is best for getting food. Cognitive service architecture helps a lot to solve this problem. It makes the core network have the ability to perceive real-time computing demand and computing resource distribution. The first step is to build the unified modeling of heterogeneous computing and express it by means of ontology modeling and semantic description. Then, based on the age of information theory[34], the status of communication and computation resources are jointly evaluated.

2) Intelligent Scheduling: One of the most important features that distinguish 6G from previous generations of communication is that it does not only play a role of pipeline connecting difference devices. In other words, 6G communication is not just communication itself. Instead, it is more like a field which interact with all the devices in it, and vise versa. That is, just connecting network devices together is not enough. 6G network should provide intelligent scheduling for both communication resources and computation resources. For one thing, lacking any kind of resource will inevitably lead to deterioration of user experience. For another, to prevent resource waste and improve the overall performance of the system, communication resources and computation resources should be jointly scheduled at a fine-grained level. Although network slicing makes great progress in terms of managing communication resources conveniently leveraging network function virtualization, it is still in a coarse-grained level and lacks flexibility. To fulfill intelligent scheduling, new measures should be taken and relevant technologies should be upgraded.

Based on the distributed edge core network, we introduce AI as a Function in 6G core network to act as an intelligent resource scheduler. AI as a Function is different from AI as a Service. AI as a Service is a service provided by service providers which play a role of enhancing application perfor-

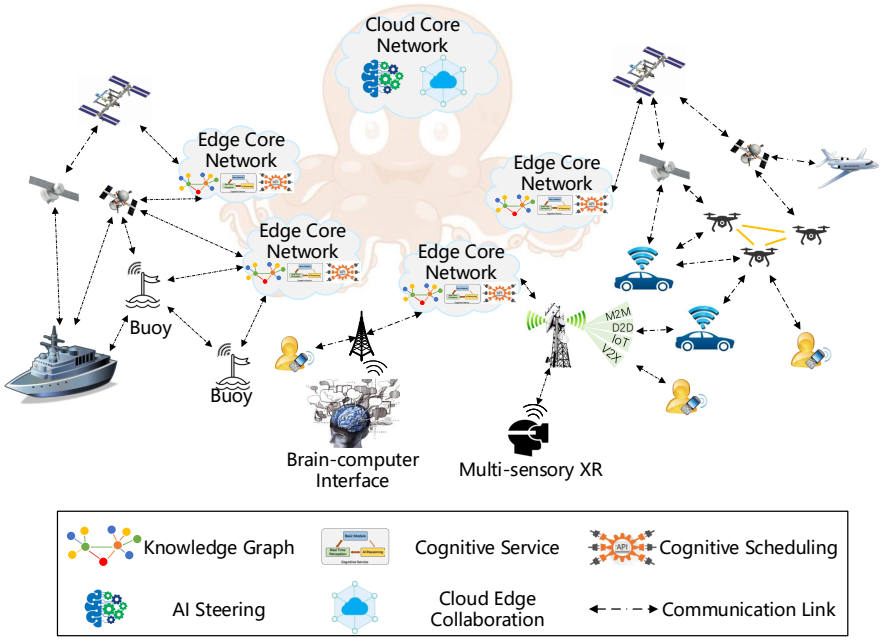

Fig. 5. Cognitive Service Architecture for 6G.

mance. It includes a large variety of intelligent services such as voice assistant, computer vision, intelligent recommendation and so on. They are all user-oriented intelligent services. However, AI as a Function is a network function of 6G core network. It works for the network and serve as a scheduler. It is transparent to users but help to improve the overall performance of the network.

Intelligent scheduling is non-trivial in resource scheduling. Although mobile edge computing is introduced to equip access network with edge servers, the computation resource of a single edge node is still limited. Generally, the arrival of the task is non-uniform. The computation resource needed per task is uncertain, either. Poor scheduling will lead to computation resource shortage or network congestion locally. To guarantee the quality of experience, computation resource of different edge nodes and the status of network should be jointly arranged. That is, the access control of computing tasks and network routing is decided at the same time. For example, the routing protocol used in 6G network could be redesigned. The routing process considers not only the status information of the network, but also the distribution of available computation resources.

Just resource scheduling is not enough. Even with ample computation resources and good network status, users may still suffer from service quality degradation if services coordination relationship is not carefully scheduled. To deal with this problem, multiple measures should be taken. First, it is essential to realize the intelligent monitoring of service operation status and establish a service operation quality prediction model. Then, the service collaboration trigger mechanism should take user requirements, equipment failures and system iteration upgrades into consideration. To achieve the coordination of the services, the collaboration strategy has to be carefully designed. Deep reinforcement learning will be leveraged to cope with massive factors, such as user movement and scene changes that may result in service quality degradation. Then the coordination scheme is fulfilled by means of service rescheduling, resource redistribution and service migration. The key is to dynamically adjust the service chain and keep it adapted to the ever-changing user behavior.

3) Cognitive Service Architecture: In 5G network, edge computing is introduced to provided computation resources in the proximity. Although the 5G core network adopt service based architecture, it still works as a whole logically. As we mentioned above, communication resources and computation resources should be scheduled jointly to cope with different resource-exhausted situations. However, in previous generations of communications, the network only serve as a pipeline and has no cognitive ability to computation resources. Besides, core network and edge nodes are located at different level of the network. The division of network and computation results in much difficulty in the joint scheduling.

Inspired by the octopus's distributed nervous system structure, core network will be split into edge core network and cloud core network in 6G cognitive service architecture. Edge core network will further sink to the edge of the network. As is shown in Fig. 5, cognitive service architecture will leverage edge computing to form a multi-center architecture to provide efficient, flexible, ultra-low delay and ultra large capacity network services. Edge core networks act like peripheral nervous system in octopus's arms, while the cloud core network plays a role of central brain. Most of the cognitive services will be deployed at the edge core networks. Cognitive scheduling and knowledge graph management are mainly deployed at edge core network as well. Cloud core network will no longer directly participate in the communication. It is in charge of AI steering and helps edge core networks to coordinate with each other.

With core network sinking to the edge, the control panel of communication resources and computation resources are finally laid at the same tier of the system, which paves the way for realizing the visions in Section I-B. First, the cognitive service architecture works like an octopus. The flexible architecture makes it adapt to various access scenarios. Besides, the cognitive service can be optimized for different mobile communications. Second, as control panel of communication resources and computation resources are all deployed at the proximity of users, low latency is guaranteed while bandwidth pressure is relieved. Due to the sinking deployment of edge core network, the cognitive service architecture will realize the whole network coverage from the core network to the user devices. On this basis, service continuity between different communication modes could be guaranteed. When service scenario and requirement changes, service link and communication link can all switch seamlessly. Third, through AI as a function running in edge core network, edge core network supports service adaptation, service migration, service collaboration and service evolution in terrestrial communication, satellite communication, UAV communication and maritime communication.

III. Enabling Technologies for 6G Cognitive Service Architecture

A. Unified Semantics for Network

6G network will attach great importance to the satisfaction of users' personalized needs, especially when it comes to multi-sensory XR. In cognitive service architecture, unified semantics should be defined to describe the cognitive objects of the system.

First, it helps to recognize the personalized needs of users. It is necessary to conduct quantitative modeling and analysis of human subjective experience to meet the information processing and transmission needs of differences. Besides, it is important to explore the service cognitive mechanism and the coupling of business logic and cognitive function based on user quality of experience. Building a quantitative model of human subjective experience is the first step to give core network ability of cognition.

Second, a unified quantitative model of network states should be defined. As mention above, 6G core network must jointly schedule communication and computation resources. Therefore, the model does not only contain communication states of the network such as network delay, bandwidth, jitter, packet loss rate, etc., it also contains computation resource state including the distribution and the usage of different kind of computing resources.

Third, the mapping relationship between system resource status and user requirements is indispensable. As edge core network is introduced to 6G cognitive service architecture, network state and resource provision should match with user demands. Therefore, it is not enough if 6G core network only has the ability to recognize user demands and network state. Finding out the gap and the effect of interaction between them is the key to make a decision.

B. Polymorphic Interface Supporting Cognitive Services

Traditional adaptive systems are mostly used for resource allocation in data centers, such as load balancing, container arrangement, etc. Communication resources mainly rely on manual configuration, lacking in flexible adaptive mechanism. However, this is not suitable for distributed edge core network. For one thing, the edge core networks are located at different places, which is not convenient for unified management. For another, traditional measures are unable to meet the requirements for real-time and rapid adjustment. Therefore, polymorphic cognitive service interface technology is introduced. By introducing rule matching, approximate reasoning and other technologies, polymorphic interface has the functions of situational awareness, demand identification, state statistics and so on. The introduction of polymorphic interface endows 6G core network with fine-grained perception ability. The polymorphic interfaces not only complete the function of traditional network interfaces, but also have the ability of network state perception. It serves as a unified semantic description of the whole system and acts like the chemical receptors on an octopus's arm. Although the cognitive ability of one polymorphic interface is simple, large quantities of polymorphic interfaces work together to form the basis of cognitive capabilities in cognitive service architecture.

Behind the polymorphic interface is a lightweight learning agent which runs real-time AI reasoning decision-making unit. The agent has the ability of fast reasoning and learning based service adaptation. The function of the agent is similar to the peripheral nervous system in the octopus arm. Due to the large number of deployments on various edge devices, the learning agent must be lightweight and able to work in a resource-constrained environment. At the same time, these agents do not handle complex tasks. They only match the extraction of information recognized by the polymorphic interface and perform predefined operations. Agents with the same function in the system upgrades through the knowledge graph, federal learning and transfer learning.

C. Service Continuity Guarantee

The introduction of edge core network architecture and other communications (satellite, UAV and maritime communication) in 6G networks has brought great challenges to the continuity of services. When users move across two edge core networks, it is necessary to solve the cross-domain hand-off of the network and the seamless migration of services [35]. On the other hand, the movement of satellites and drones may also trigger network hand-off and service migration [6]. To guarantee service continuity, the following technologies are required.

The seamless service migration technology makes the service migration process completely transparent to users through the unified scheduling of communication resources and computing resources. The main purpose is to overcome the link interruption caused by network re-establishment, service state preservation, data transmission, and state restoration during service migration. Such interruptions will greatly affect the user experience of applications [36] that are sensitive to delay and require high reliability.

By learning and matching the tracks of users, satellites and UAVs, the migration prediction technology can predict the occurrence time and destination node of service migration. As a result, the system can migrate some state data of services in advance to reduce the downtime in the process of service migration. Generally, the reliability of the migration process is guaranteed by increasing the calculation redundancy(multiple edge nodes working together) and network link redundancy(setting up multiple links simultaneously). Based on the migration prediction, the cost brought by redundancy can be reduced. The higher the accuracy of prediction, the higher the utilization rate of system resources.

In addition, a lightweight migration carrier technology is needed. The existing technologies are not suitable for service migration. Virtual machine is too cumbersome, which will bring too much additional data transmission. Virtual machine is not suitable for microservice, either. Although container technology is light enough, it is not conducive to the preservation of context state and remote recovery. Therefore, a lightweight migration carrier technology combining advantages of virtual machines and containers is needed. It should not only meet requirements of microservice deployment, but also can quickly encapsulate the current runtime context state to meet the rapid migration and recovery of services.

D. Universal Platform for Computation Network Integration

One of the main design principles of 6G core network is to achieve the integration of communication and computation resources. In order to achieve this goal, we first need to build a universal platform to achieve unified dynamic scheduling. From the perspective of the platform, the application instances and network functions are all services. The only difference lies in who they serve. Therefore, the platform is conducive to a more flexible scheduling of computation and communication resources.

The cross-domain orchestration management technology supports collaboration between different edge core networks. Since the edge core network is deployed on a universal platform, the number and location of edge core network deployments can be very flexible. The cross-domain orchestration management technology flexibly adjusts the deployment scheme of edge core networks according to the dynamic requirements of communication and computing resources in the city. The edge networks can merge or split to adjust the number of edge nodes. The coverage area size of a single edge core network changes as well according to the distribution of user requests. In addition, cross-domain orchestration management technology supports grayscale deployment of core network functions and smooth upgrade of network functions.

In cognitive service architecture, network functions and computing resources are abstracted as services. To achieve the unified management of communication resources and computing resources, a distributed service discovery mechanism is need in the distributed core network scenario to achieve efficient cross domain service discovery, service registration and service analysis, etc. The mechanism is event driven, has standard control interface and supports iterative optimization of service composition.

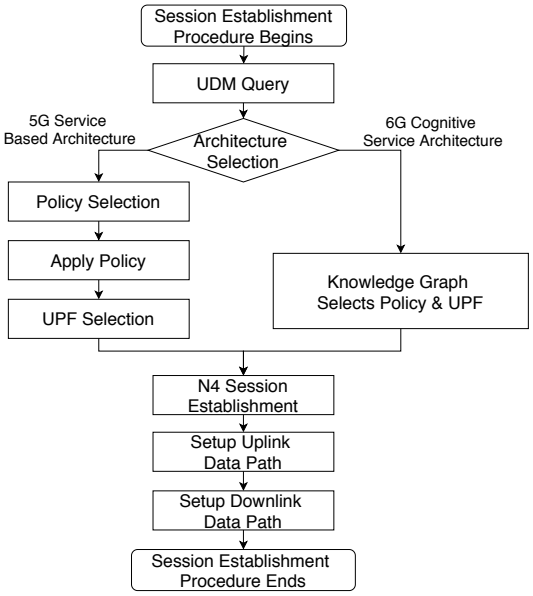

Fig. 6. Session establishment procedure

E. Next Generation of Network Slicing

Although network slicing in 5G realizes the flexible allocation of communication resources [37], it is still a coarse-grained network management method. 5G network slicing provides services for fixed process scenarios based on a few semantic adaptation and simple rule combinations. It is difficult to meet the low delay service requirements of 6G network caused by the changeable scenarios and the complex service combination. Therefore, 6G network slicing technology should be upgraded to meet the following requirements: 1) It should support more fine-grained and flexible resource allocation. 2) It supports the cooperation of multiple distributed edge core networks. 3) Network functions and computation resources are jointly managed and scheduled in the form of services. 4) In the 6G high mobility and high reliability scenario, service migration are guaranteed with high reliability. 5) Each network slice is deployed and released rapidly to achieve high efficiency. To meet these requirements, The upgrade mainly comes in two folds. On the one hand, the next generation of network slicing should leverage cognitive service and intelligent scheduling to meet the above requirements. On the other hand, fine-grained and flexible network slicing is one of the key technologies that enable cognitive service architecture to integrate computing resources and communication resources in a unified architecture.

More specifically, the upgrade of network slicing relies on the upgrade of its enabler technologies. That is, Network Functions Virtualization (NFV) and Software Defined Networking (SDN) should also be upgraded to meet the high requirement of flexibility and scalability in 6G network. The controller of NFV and SDN should be upgraded leveraging cognitive service. This enables the scheduling of network and network functions to be integrated into cognitive service architecture. Besides, the virtualization of NFV and SDN should be realized with finer granularity. This can guarantee the flexibility in resource scheduling.

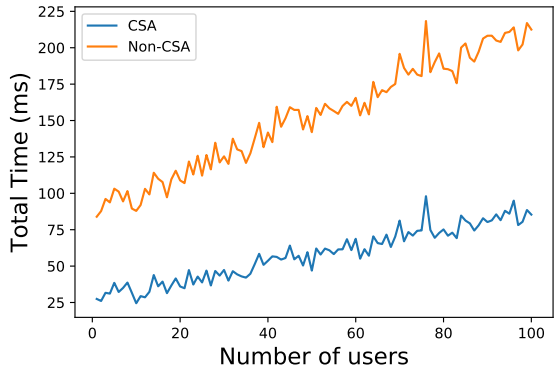

Fig. 7. Total time of session establishment

IV. Case Study on Cognitive Service Architecture

Cognitive service architecture improve the overall performance and flexibility of core network by means of introducing cognitive service and upgrade the network architecture. However, as 6G networks are still in a very early stage of research and standards of 6G network haven't been released yet, it's very hard to test the overall performance of the system architecture. To demonstrate the performance improvement brought by cognitive service architecture, the process of session establishment is chosen to conduct an evaluation.

A. Experiment Setup

The test environment is set up based on free5gc [38], which is an open source project that implement the 5G core network. It is used to emulate the process of session establishment of 5G network. In this experiment, a knowledge graph is added to the system to optimize the procedure of session establishment. As is shown in Fig. 6, the first step of establishing a session is querying data from Unified Data Management (UDM). Then, if the system works in 5G Service Based Architecture, policy selection, applying policy and User Plane Function(UPF) selection are conducted in sequence. However, in 6G cognitive service architecture, these works are conducted by the knowledge graph. After finishing the selection policy and UPF, N4 session is established. Then follows the setup of uplink and downlink path.

B. Total Time of Session Establishment

We measure the total time of establishing sessions with different number of users. In Fig. 7, CSA denotes cognitive service architecture while Non-CSA denotes Non-cognitive service architecture. Generally, CSA uses less time than Non-CSA. When there is only one user, the total time of CSA is 56.47 ms less than that of Non-CSA. As the user number increases, the gap becomes larger. When user number is 98, the gap even reaches 128.43 ms. With different user number, CSA saves more than 55.11% time than Non-CSA. This is because CSA leverages knowledge graph to simplify part of session establishment workflow.

V. Conclusion

6G network is destined to realize full coverage, full scene and Internet of Everything. In order to cope with the complex and changing scenarios, the 6G core network must implement a flexible and highly efficient architecture. This requires the 6G core network to have cognitive capabilities and to achieve integrated scheduling of communication resources and computing resources. Inspired by the nervous system of octopus, we propose cognitive service architecture for 6G core network. In this architecture, network functions are upgraded to cognitive services by adding real time perception and AI reasoning abilities. Besides, the previous core network is divided into edge core networks and cloud core network. Edge core network is in charge of resource management and scheduling of a specific area. Most network management and control are conducted at edge core network. The cloud core network only serve as a coordinator. The key idea of cognitive service architecture is to achieve a flexible network with cognitive ability by imitating the nervous system of an octopus. A case study is conducted and the result show that by applying knowledge graph, cognitive service architecture can simplify the workflow of network function interaction and improve the performance of system. Possible enabling technologies are also discussed in this paper.

In future work, we will concentrate on enhancing the coordination between different edge core networks. Typical real world scenarios, such as dynamic service deployment among different edge core networks and the migration of computation-intensive application will be tested. Besides, we will upgrade current 5G core network functions to make it applicable in 6G cognitive service architecture.

Acknowledgment

This work was supported in part by the National Key R&D Program of China (2018YFE0205503), NSFC (61922017, 62032003, and 61921003) and BUPT Excellent Ph.D. Students Foundation(CX2019133).

References

[1] 3GPP, "System architecture for the 5G System," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.501. [Online]. Available: ftp://ftp.3gpp.org/Specs/archive/23_series/23.501/

[2] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, Third quarter 2016.

[3] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[4] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3098–3130, Fourth quarter 2018.

[5] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, oct 2016.

[6] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, "Vision, Requirements, and Technology Trend of 6G: How to Tackle the Challenges of System Coverage, Capacity, User Data-Rate and Movement Speed," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 218–228, apr 2020.

[7] M. Condoluci, S. H. Johnson, V. Ayadurai, M. A. Lema, M. A. Cuevas, M. Dohler, and T. Mahmoodi, "Fixed-Mobile Convergence in the 5G Era: From Hybrid Access to Converged Core," *IEEE Network*, vol. 33, no. 2, pp. 138–145, mar 2019.

[8] 3GPP, "Architecture enhancements for non-3gpp accesses," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.402. [Online]. Available: ftp://ftp.3gpp.org/Specs/archive/23_series/23.402/

[9] G. Giambene, S. Kota, and P. Pillai, "Satellite-5G Integration: A Network Perspective," *IEEE Network*, vol. 32, no. 5, pp. 25–31, sep 2018.

[10] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *arXiv:1902.10265 [cs, math]*, jul 2019.
[11] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 55–61, mar 2020.
[12] iLab, "Cloud VR Black edge and Network Delay Relationship White Paper," Shenzhen, China, Huawei, White Paper, 2019.
[13] J. M. C. Brito, L. L. Mendes, and J. G. S. Gontijo, "Brazil 6G Project - An Approach to Build a National-wise Framework for 6G Networks," in *Proceedings of 2nd 6G Wireless Summit (6G SUMMIT 2020)*, mar 2020, pp. 1–5.
[14] M. H. Alsharif, A. H. Kelechi, M. A. Albreem, S. A. Chaudhry, M. S. Zia, and S. Kim, "Sixth Generation (6G) Wireless Networks: Vision, Research Activities, Challenges and Potential Solutions," *Symmetry*, vol. 12, no. 4, p. 676, apr 2020.
[15] A. K. Yerrapragada and B. Kelley, "On the Application of K-User MIMO for 6G Enhanced Mobile Broadband," *Sensors*, vol. 20, no. 21, p. 6252, jan 2020.
[16] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, and C. S. Hong, "6G Wireless Systems: A Vision, Architectural Elements, and Future Directions," *IEEE Access*, vol. 8, pp. 147 029–147 044, 2020.
[17] S. Nayak and R. Patgiri, "6G Communications: Envisioning the Key Issues and Challenges," *arXiv:2004.04024 [cs, eess]*, apr 2020.
[18] T. Hewa, G. Gür, A. Kalla, M. Ylianttila, A. Bracken, and M. Liyanage, "The Role of Blockchain in 6G: Challenges, Opportunities and Research Directions," in *Proceedings of 2nd 6G Wireless Summit (6G SUMMIT 2020)*, mar 2020, pp. 1–5.
[19] N. Kato, B. Mao, F. Tang, Y. Kawamoto, and J. Liu, "Ten Challenges in Advancing Machine Learning Technologies toward 6G," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 96–103, jun 2020.
[20] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, "Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 212–217, apr 2020.
[21] S. Zhang, H. Zhang, and L. Song, "Beyond D2D: Full Dimension UAV-to-Everything Communications in 6G," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6592–6602, jun 2020.
[22] W. Sun, H. Zhang, R. Wang, and Y. Zhang, "Reducing Offloading Latency for Digital Twin Edge Networks in 6G," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12 240–12 251, oct 2020.
[23] S. Chen, S. Sun, G. Xu, X. Su, and Y. Cai, "Beam-Space Multiplexing: Practice, Theory, and Trends, From 4G TD-LTE, 5G, to 6G and Beyond," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 162–172, apr 2020.
[24] E. Basar, "Reconfigurable Intelligent Surface-Based Index Modulation: A New Beyond MIMO Paradigm for 6G," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 3187–3196, may 2020.
[25] R. K. Saha, "A Hybrid Interweave–Underlay Countrywide Millimeter-Wave Spectrum Access and Reuse Technique for CR Indoor Small Cells in 5G/6G Era," *Sensors*, vol. 20, no. 14, p. 3979, jan 2020.
[26] ——, "On Exploiting Millimeter-Wave Spectrum Trading in Country-wide Mobile Network Operators for High Spectral and Energy Efficiencies in 5G/6G Era," *Sensors*, vol. 20, no. 12, p. 3495, jan 2020.
[27] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening New Horizons for Integration of Comfort, Security, and Intelligence," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 126–132, oct 2020.
[28] X. Qiao, Y. Huang, S. Dustdar, and J. Chen, "6G Vision: An AI-Driven Decentralized Network and Service Architecture," *IEEE Internet Computing*, vol. 24, no. 4, pp. 33–40, jul 2020.
[29] P. Graziadei, "Receptors in the Suckers of Octopus," *Nature*, vol. 195, no. 4836, pp. 57–59, jul 1962.
[30] B. Hochner, "An Embodied View of Octopus Neurobiology," *Current Biology*, vol. 22, no. 20, pp. R887–R892, oct 2012.
[31] G. Levy and B. Hochner, "Embodied Organization of Octopus vulgaris Morphology, Vision, and Locomotion," *Frontiers in Physiology*, vol. 8, mar 2017.
[32] L. Zullo, G. Sumbre, C. Agnisola, T. Flash, and B. Hochner, "Nonsomatotopic Organization of the Higher Motor Centers in Octopus," *Current Biology*, vol. 19, no. 19, pp. 1632–1636, oct 2009.
[33] G. Sumbre, Y. Gutfreund, G. Fiorito, T. Flash, and B. Hochner, "Control of octopus arm extension by a peripheral motor program," *Science (New York, N.Y.)*, vol. 293, no. 5536, pp. 1845–1848, sep 2001.
[34] C. Li, S. Li, and Y. T. Hou, "A General Model for Minimizing Age of Information at Network Edge," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM 2019)*, apr 2019, pp. 118–126.
[35] S. Wang, R. Urgaonkar, M. Zafer, T. He, K. Chan, and K. K. Leung, "Dynamic Service Migration in Mobile Edge Computing Based on Markov Decision Process," *IEEE/ACM Transactions on Networking*, vol. 27, no. 3, pp. 1272–1288, jun 2019.
[36] L. Ma, S. Yi, N. Carter, and Q. Li, "Efficient Live Migration of Edge Services Leveraging Container Layered Storage," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 2020–2033, sep 2019.
[37] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, p. 106984, feb 2020.
[38] free5gc, "free5gc/free5gc," 2020. [Online]. Available: https://github.com/free5gc/free5gc

Yuanzhe Li received a Bachelor degree of Engineering degree in communication engineering from Beijing University of Posts and Telecommunications(BUPT), in 2016. Currently, he is a Ph.D. candidate at the State Key Laboratory of Networking and Switching Technology, BUPT. His research interests include mobile edge computing, cloud computing and service computing.

Jie Huang is currently a Ph.D. candidate at the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT). Her research interests include mobile edge computing and service computing.

Qibo Sun received his Ph.D. degree in communication and electronic system from the Beijing University of Posts and Telecommunication in 2002. He is currently an associate professor at the Beijing University of Posts and Telecommunication in China. He is a member of the China computer federation. His research interests include services computing, Internet of things, and network security.

Tao Sun got the B.S. degree on Automation in 2003, and Ph.D. degree on Control Science and Engineering in 2008, both from Tsinghua University, China. He currently is leading the research and standardization of the mobile network architecture and service evolution in China Mobile Research Institute. He worked on directions such as network slicing, 5G capabilities enablement for verticals, introducing AI to network and new IP technologies. He has more than 10 years of experience on Mobile network standardization work. He currently serve as Vice Chair of 3GPP SA2.

Shangguang Wang is a Professor at the School of Computing, Beijing University of Posts and Telecommunications, China. He is a Vice-Director of the State Key Laboratory of Networking and Switching Technology. He has published more than 150 papers, and his research interests include service computing, cloud computing, and mobile edge computing. He served as General Chairs or TPC Chairs of IEEE EDGE 2020, IEEE CLOUD 2020, IEEE SAGC 2020, IEEE EDGE 2018, and IEEE ICFCE 2017, etc., and Vice-Chair of IEEE Technical Committee on Services Computing (2015-2018). He is currently serving as Executive Vice-Chair of IEEE Technical Committee on Services Computing (2021-), and Vice-Chair of IEEE Technical Committee on Cloud Computing (2020-). He is a senior member of the IEEE.