IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MOBILE EDGE COMPUTING AND MOBILE CLOUD COMPUTING: ADDRESSING HETEROGENEITY AND ENERGY ISSUES OF COMPUTE AND NETWORK RESOURCES

Received August 2, 2019, accepted August 26, 2019, date of publication August 30, 2019, date of current version September 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938534

Edge Computing in 5G: A Review

NAJMUL HASSAN[1], KOK-LIM ALVIN YAU[1], (Senior Member, IEEE), AND CELIMUGE WU[2], (Senior Member, IEEE)

[1]Department of Computing and Information Systems, Sunway University, Bandar Sunway 47500, Malaysia
[2]Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan

Corresponding author: Kok-Lim Alvin Yau (koklimy@sunway.edu.my)

ABSTRACT 5G is the next generation cellular network that aspires to achieve substantial improvement on quality of service, such as higher throughput and lower latency. Edge computing is an emerging technology that enables the evolution to 5G by bringing cloud capabilities near to the end users (or user equipment, UEs) in order to overcome the intrinsic problems of the traditional cloud, such as high latency and the lack of security. In this paper, we establish a taxonomy of edge computing in 5G, which gives an overview of existing state-of-the-art solutions of edge computing in 5G on the basis of objectives, computational platforms, attributes, 5G functions, performance measures, and roles. We also present other important aspects, including the key requirements for its successful deployment in 5G and the applications of edge computing in 5G. Then, we explore, highlight, and categorize recent advancements in edge computing for 5G. By doing so, we reveal the salient features of different edge computing paradigms for 5G. Finally, open research issues are outlined.

INDEX TERMS 5G, cloud computing, edge computing, fog computing.

I. INTRODUCTION

Edge computing is a computational paradigm that enables edge servers in mini clouds (or edge clouds) to extend cloud capabilities at the edge of the network to perform computationally-intensive tasks and store a massive amount of data at close proximity to user equipment (UEs) [1]–[3]. Traditional cloud computing, which is a centralized computing paradigm that provides continuous access to highly capable data centers, has been adopted to allow UEs to offload computation and storage to the data centers [4]. This is because UEs have limited processing, computational, and storage capabilities. Nevertheless, edge computing is preferred to cater for the wireless communication requirements of next generation applications, such as augmented reality and virtual reality, which are interactive in nature. These highly interactive applications are computationally-intensive and have high quality of service (QoS) requirements, including low latency and high throughput (e.g, ultra reliable low latency communication (URLLC), tactile internet) [5]–[7]. Most importantly, these applications are expected to generate a massive amount of data up to 30.6 exabytes per month [8]. The limited capabilities of UEs warrants the need for edge computing to: a) receive and store a massive amount of real-time data, b) process, compute, and analyze the data, and c) make and distribute decisions on mini clouds locally. Hence, edge servers in the mini clouds have the capabilities of a cloud but on a different scale, and they are located locally instead of remote data centers which may be far away from UEs [9].

The associate editor coordinating the review of this manuscript and approving it for publication was Junaid Shuja.

A. OUR CONTRIBUTIONS

This paper highlights recent advances of edge computing in 5G. Some analyses have been made on a particular computational platform of edge computing in 5G, particularly mobile edge computing (MEC) [10], [11]. In addition [12], [13] focuses on edge orchestration and its related issues in 5G environment and MEC architecture. In [14] a survey of edge computing, including its applications and key challenges, is presented from the perspective of vehicular networks. Our paper is first of its kind to present: a) a taxonomy of edge computing in 5G covering the objectives, computational platforms, attributes, the use of 5G functions, performance measures, and the role of edge computing; b) a review of state-of-the-art edge computing schemes in 5G; and c) open issues in this research topic. This topic is timely due to the recent advent of 5G, and the evolving roles of edge computing in the realization of 5G.

B. ORGANIZATION OF THIS PAPER

The rest of this paper is organized as follows: Section II presents an overview of 5G and edge computing, respectively,

and answers a host of questions on the use of edge computing in 5G, including the time characteristics of data, the key requirements, and the applications of edge computing in 5G. Section III presents a taxonomy of edge computing in 5G. Section IV presents the state-of-the-art schemes for edge computing in 5G. Section V presents open research issues. Section VI concludes this paper.

II. BACKGROUND

This section presents an overview of 5G, the time characteristics of data, edge computing, as well as the key requirements and the applications of edge computing in 5G.

A. REQUIREMENTS OF 5G SYSTEMS

5G is foreseen as the next generation wireless cellular network to cater for the needs of next generation networks. 5G possesses *three* main characteristics unseen in previous generation networks. Firstly, *a massive amount of data is generated*. According to the International Telecommunication Union (ITU), there are more than 7.5 billion mobile devices around the world in 2017 [15], and the number of mobile devices is expected to increase to 25 billion by 2020 [16], contributing to ultra-dense networks. Consequently, there is an explosive growth in the amount of data from 16.5 exabytes in 2014 to an estimate of 500 exabytes in 2020 [17], contributing to a growth rate of 30 times. Secondly, *stringent QoS requirements are imposed* to support highly interactive applications, requiring ultra-low latency and high throughput. *Thirdly*, *heterogeneous environment must be supported* to allow inter-operability of a diverse range of UEs (e.g., smart phones and tablets), QoS requirements (e.g., different levels of latency and throughput for multimedia applications), network types (e.g., IEEE 802.11 and Internet of things), and so on.

5G is comprised of *three* main new technologies to provide higher network capacity in order to support a higher number of UEs [18]. Firstly, *mmWave communication*, which uses high frequency bands (i.e., 30 GHz to 300 GHz [19]), provides high bandwidth (i.e., at least 11 Gbps [20]). Secondly, *small cells deployment* allows UEs to communicate using mmWave in order to reduce transmission range and interference. Thirdly, *massive MIMO (multiple-input multiple-output)* allows base stations (BSs) to use a large number of antennas (e.g., up to 16 antennas per sector) to provide directional transmission (or beamforming) in order to reduce interference, allowing neighboring nodes to communicate simultaneously.

B. MAIN CHARACTERISTICS OF 5G DATA

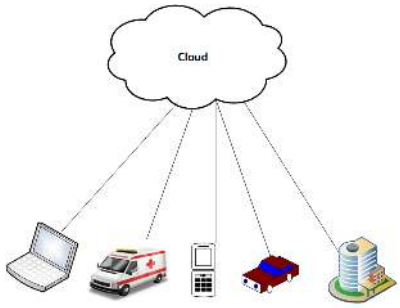
Data can be categorized into three main categories according to its time characteristics as follows:

- *Hard real-time* data has a strict predefined latency. Applications, such as video steaming, gaming, and healthcare services, generate this kind of data.
- *Soft real-time* data has a predefined latency, yet it can tolerate some pre-defined and bounded latency. Applications, such as intelligent traffic signal control system, generate this kind of data.
- *Non-real-time* data is not time-sensitive and can tolerate latency.

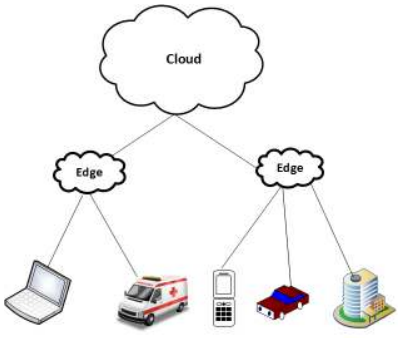
Edge computing is envisioned to handle applications and services with hard real-time requirement using edge servers due to their close proximity to UEs leading to significant reduction in latency. For applications and services with soft-real time requirement, or bounded end-to-end delay, tasks are handled by edge servers if the response delay between UEs and the cloud is higher than the requirement; otherwise, the tasks can be offloaded to the cloud. For applications and services with non-real-time requirement, tasks can be offloaded to the cloud for load balancing.

C. SIGNIFICANCE OF EDGE COMPUTING

5G is foreseen to support highly interactive applications with low latency and high throughput requirements [21]. Edge computing adopts a decentralized model that brings cloud computing capabilities closer to UEs in order to reduce latency. Fig. 1 shows the cloud computing and edge computing models. Edge computing can either operate as a single computing platform, or a collaborative platform together with other components, including the cloud [22]. Edge computing is necessary as the traditional cloud computing model is not suitable for highly interactive applications that are computationally-intensive and have high QoS requirements, including low latency and high throughput. This is because cloud may be far away from UEs, which also increases



(a) Cloud computing.



(b) Edge computing.

FIGURE 1. Cloud computing and edge computing models.

energy consumption. In other words, cloud servers are typically located at the core network, and edge servers of the mini clouds are located at the edge of the network [23].

To understand the need of an edge computing, consider real-time packet delivery among self-driving cars that requires an end-to-end delay of less than 10 ms [24]. The minimum end-to-end delay for an access to cloud is greater than 80 ms [25], [26], which is intolerable. Edge computing fulfills the sub-millisecond requirement of 5G applications, and reduces energy consumption by around 30% to 40% [27], which attributes to up to five times lesser energy consumption as compared to accessing the cloud [28].

D. KEY REQUIREMENTS OF EDGE COMPUTING IN 5G

There are four key requirements for the successful deployment and operation of edge computing in 5G. While all the four key requirements are important, achieving a balanced trade-off among them must be considered depending on the applications.

Firstly, *real-time interaction*, which is the fundamental motivation for the use of edge computing over cloud computing, ensures low latency to support delay-sensitive applications and services (e.g. remote surgery, tactile internet, URLLC, unmanned vehicles [29], [30] and vehicle accident prevention) in order to improve QoS. A diverse range of services, including decision making and data analysis, can be provided by edge servers in a real-time manner.

Secondly, *local processing* is feasible since data and user requests can be processed by edge servers, rather than the cloud. This means that, by reducing the traffic amount across the connection between a small cell and the core network: a) the bandwidth of the connection can be increased to prevent bottleneck; and b) the traffic amount in the core network is reduced.

Thirdly, *high data rate* is necessary to transmit the massive amount of data generated by a diverse range of applications (e.g., virtual reality and remote surgery) to edge clouds [31]. Edge servers, which can be embedded in the BSs, allow easy access to edge clouds without the need to access the core network. The use of mmWave frequency bands in a small cell provides a high data rate transmission.

Fourthly, *high availability* ensures the availability of the cloud services at the edge. Since edge computing pushes data and application logic to the edge clouds, the availability of the edge clouds is important.

E. APPLICATIONS OF EDGE COMPUTING IN 5G

Many applications of 5G are relying on edge computing for real-time interaction, local processing, high data rate, and high availability, including:

- *Healthcare*, such as remote surgery and diagnostics, as well as monitoring of patient vital signs and data. Doctors can use a remote platform to operate surgical tools in order to save life from a distance where they feel safe and comfortable.
- *Entertainment and multimedia applications*, such as streaming HDTV or 3D TV.
- *Virtual reality, augmented reality, and mixed reality*, such as streaming video contents to virtual reality glasses. The size of the glasses can be reduced by offloading computation from the glasses to edge servers.
- *Tactile internet*, which is the next evolution of Internet of things, provides an ultra-responsive and ultra-reliable network connectivity to ensure successful delivery of real-time control messages and physical tactile experiences remotely [32], [33].
- *URLLC*, which ensures high reliability between UEs specifically in M2M communications, supports low-latency transmissions of small payloads with very high reliability from a limited set of UEs, such as fire alarms [34].
- *Internet of things*, such as smart appliances that connect devices (e.g., household appliances) to the internet.
- *Factories of the future*, such as smart machines, to improve safety and productivity. Operators can use a remote platform to operate heavy machines, particularly those located at hard-to-reach and unsafe places, from a safe and comfortable place.
- *Emergency response*, whereby different kinds of data and information about an event or incident are gathered from different sources at different times. The partially available data and information are used to make critical decisions, and they provide a more complete picture of the event as time goes by. Decisions made are shared with emergency response team (e.g., firefighters) in real time, even prior to their arrivals at the location of the event.
- *Intelligent transportation system*, whereby drivers can share or gather information from traffic information centers to avoid vehicles that are in danger, or stop abruptly, in a real-time manner in order to avoid accidents. In addition, unmanned vehicles can sense their surroundings and move safely in an autonomous manner.

III. TAXONOMY

Fig. 2 shows a taxonomy of edge computing in 5G, covering objectives, computational platforms, attributes, the use of 5G functions, performance measures, and the roles of edge computing in 5G.

A. OBJECTIVES

There are *five* main objectives of edge computing in 5G as follows:

O.1 *Improving data management* to handle a large amount of delay-sensitive data, which are generated by UEs, that needs to be handled locally in a real-time manner. For instance, the local UEs in a smart factory is expected to generate up to 1 petabyte of data daily [35]. Since accessing to cloud incurs high latency [36], the data can be handled locally by edge servers. Such efficient

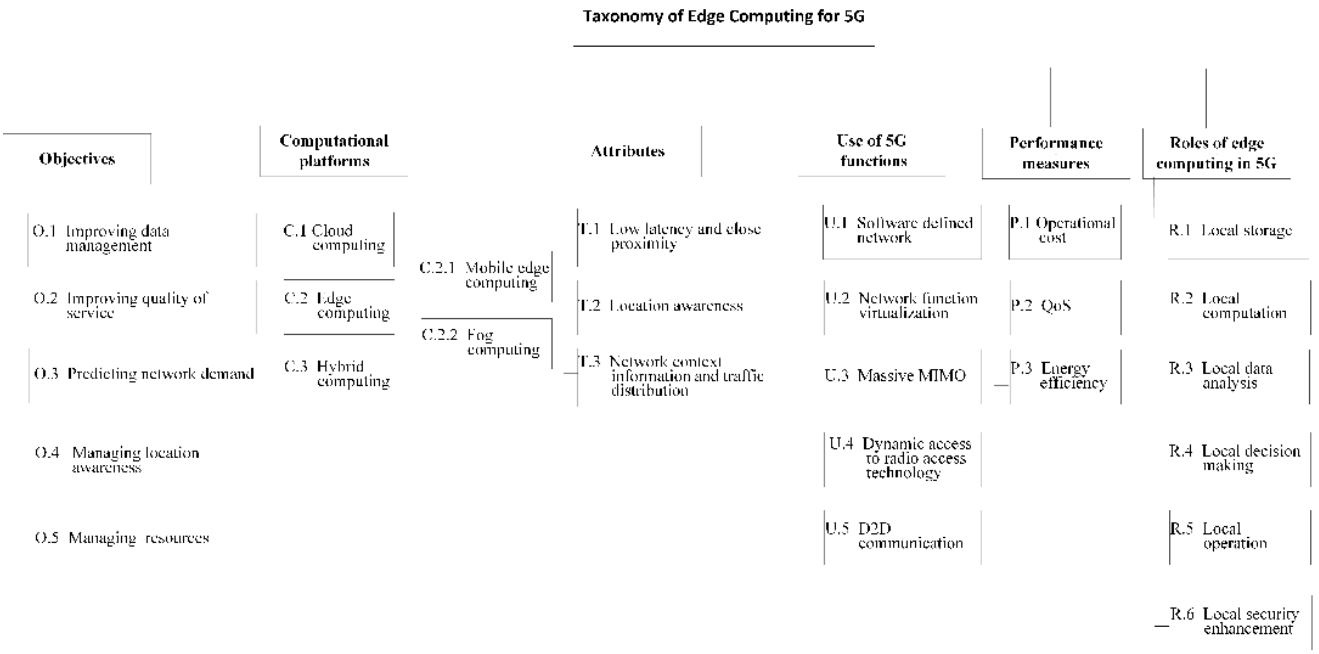


FIGURE 2. Taxonomy of edge computing in 5G.

data management is needed to support local functions (e.g., D2D) and real-time applications (e.g., remote surgery).

O.2 *Improving QoS* to meet a diverse range of stringent QoS requirements in order to improve quality of experience (QoE) [37]. This helps to support next generation applications, including highly interactive applications and on-demand services. For instance, over-the-top (OTT) services enable online delivery of multimedia contents, which generally require low latency and high bandwidth, without the service providers being actively involved in the control and distribution of the content [38]. This service can promote new and personalized applications that allow service providers to customize QoS [39]. Service providers must have a holistic view of subscribers and customers, covering their contextual information, such as their preferences and interests. Subsequently, the information can be personalized for attracting potential customers and enhancing their QoE.

O.3 *Predicting network demand* to estimate the required network resources to cater for the network (or user) demand in a local proximity, and subsequently to provide optimal resource allocation to handle the local network demand. An accurate prediction of network demand helps to decide whether a network demand should be handled locally at the edge or at the cloud, and so it provides an efficient allocation of resources (e.g., bandwidth).

O.4 *Managing location awareness* to enable the geographically distributed edge servers to infer their own locations and track the location of UEs to support location-based services. This enables location-based service providers to outsource services and data to edge clouds. For instance, mobile UEs can query and search for information about points of interest in local proximities given their geographical locations. The number of queries can be high, such as queries related to *hospitals* and *medical advices* during emergency response.

O.5 *Improving resource management* to optimize network resource utilization for network performance enhancement due to the limited network resources available in the edge cloud as compared to the cloud. This is challenging as it is a multi-objective function that must cater to a diverse range of applications, as well as user requirements and demands, which vary as time goes by.

B. COMPUTATIONAL PLATFORMS

Different computational platforms provide varying computing capabilities (e.g., in terms of processing loads) with different characteristics (e.g., in terms of availability, the proximity from UEs, and the complexity of network infrastructure) to process data at different geographical locations. The computational platform can be used either individually or in combination based on the network scenarios and application/service requirements. As an example for the computational platform used individually, applications and services with strict QoS requirements can use edge servers to process real-time data. As an example for the computational platform used in combination, healthcare applications and services with both real-time and non-real-time data can use edge servers to process real-time and lightweight data, and cloud to process heavyweight data. There are *three* main computational platforms in 5G as follows:

TABLE 1. Comparison between fog and MEC.

Characteristics	Fog	MEC
Device (examples)	UE, edge-cloud, cloud server	BS
Number of hops	Single or multiple hops	Single
Context awareness	Medium	High
Access technologies (examples)	Wi-Fi, bluetooth, mobile networks	Mobile networks
Inter-node communication	Supported	Partial

C.1 *Cloud computing* gathers, processes, and stores a massive amount of network-wide data and information from UEs in the network. Subsequently, it sends back the data and information, or decisions, back to the UEs. While cloud can empower UEs with low computational and storage capabilities, it is not suitable to provide real-time services because cloud may be far away from UEs.

C.2 *Edge computing* gathers, processes, and stores a massive amount of local data and information from UEs in a local area. Edge computing has close proximity to UEs, while cloud may be far away from UEs. Table 1 presents a comparison of three types of edge computing platforms as follows:

C.2.1 *Fog computing* deploys local fog nodes which are local hardware devices, such as switches and routers, to provide local computation. According to the OpenFog Consortium, fog computing is ''a system-level horizontal architecture that distributes resources and services of computing, storage, control, and networking anywhere along the continuum from cloud to things'' [40]. Fog computing shares similar benefits to other edge computing variants (e.g., MEC) to provide low latency and real-time analytics; however, it has low storage capacity.

C.2.2 *MEC* provides storage and computational capacities at the edge of the network, such as the radio access networks (RANs) and BSs, to improve context awareness and reduce latency. The MEC servers, which are usually co-located with multiple hosts (e.g., BSs), use a virtualized interface to access storage and computation facilities. A MEC orchestrator overlooks the MEC hosts by gathering and providing real-time information regarding the services offered by each host, the available resources (e.g., network capacity and load), the network topology (e.g., UEs connected to the servers including their location and networking information), as well as managing MEC applications.

C.1 *Hybrid* combines cloud computing and edge computing so that they can cooperate. For instance, edge computing processes real-time data and makes real-time decisions, while cloud computing processes non-real-time data and makes non-real-time decisions. The hybrid infrastructure combines the advantages of both edge computing (i.e., real-time responses) and cloud computing (i.e., high computational and storage capabilities). Computation can be performed in different *layers*, particularly the cloud (or the upper) layer and the fog or the edge (or the bottom) layer. In general, real-time tasks are executed in the fog layer, and tasks requiring high computation are executed in the cloud layer. Compared to the traditional cloud, edge computing increases throughput and reduces latency, which are important to support delay-sensitive applications. Nevertheless, the hybrid platform is more complex compared to the separate cloud computing and edge computing platforms.

C. ATTRIBUTES

Edge computing has *three* main attributes as follows:

T.1 *Low latency and close proximity* enables edge computing to reduce the response delay (or round-trip time) suffered by UEs while accessing the traditional cloud. There are three main components in a response delay: a) communication delay that depends on data rate; b) computational delay that depends on computational time; and c) propagation delay that depends on propagation distance. In general, in cloud computing, the end-to-end delay is greater than 80ms (or 160ms for response delay) [25]. This is not suitable for delay-sensitive applications, such as remote surgery and VR, that require tactile speed with a response delay of at most 1ms [41]. In edge computing, UEs experience reduced overall end-to-end delay and response delay due to their close proximity to edge servers. The strategic location of edge cloud reduces the communication and propagation delays. For instance, the propagation distance is reduced to tens of meters via D2D communication and in small cells, and it is generally limited within a kilometer from the UEs [42].

T.2 *Location awareness* enables edge servers to collect and process data generated by UEs on the basis of the geographical location of UEs. This allows location-based and personalized service provisioning to UEs, whereby

edge servers can gather data generated by sources in its proximity without sending it to the cloud.

T.3 *Network context awareness* enables edge servers to acquire network context information. This is because edge servers tend to possess network context information, particularly the real-time network conditions (e.g., traffic load in a network cell, and radio access network information) and UEs' information (e.g., allocated bandwidth and user location). The information allows edge servers to adapt and respond to the varying network conditions and UEs, and subsequently to optimize network resource utilization. This helps edge servers to handle a massive amount of traffic in order to improve network performance. Fine-granular information (e.g., precise individual resource reservation information) can also be used to provide specific services to traffic flows in order to cater for individual user requirements.

D. USE OF 5G FUNCTIONS

Five major enablers of edge computing in 5G are as follows:

U.1 *Software-defined network (SDN)* defines a network architecture that separates a network into control and data planes to provide flexible and agile networks, which helps to simplify network management and deploy new services [43]. In general, the control plane handles policy on cloud, while the data plane forwards traffic according to decisions made by the control plane. Network functions (e.g., routing) that require real-time response can be handled by edge servers [44].

U.2 *Network function virtualization (NFV)* performs network functions (also known as virtual functions) in virtual machines on servers, which can handle a massive amount of data to provide flexible, automated, and scalable networks [45], [46]. Network demands can be processed either at the cloud or at the edge, which prevent all data and information from being sent to the cloud.

U.3 *Massive MIMO* deploys multiple antenna elements to increase an antenna array at transmitter and receiver. This is in accordance to the Shannon theorem [47] in which the signal-to-noise ratio increases without the need to increase the transmission power, leading to increased network capacity and energy efficiency. Using massive MIMO, multiple UEs can offload tasks to an edge server simultaneously to reduce latency and energy consumption [48].

U.4 *Dynamic access to radio access technologies* provide access to conventional access technologies, such as Wi-Fi, and new radio access technology (RAT) in 5G, such as NR [49]. 5G NR is a new standard that provides connection to a diverse range of devices for achieving low latency and scalable networks, which allow future extension to existing networks.

U.5 *D2D communication* enables direct communication between neighboring UEs using ad-hoc links without passing through BSs, which improves system throughput, energy efficiency, and spectrum utilization [50], [51]. UEs can offload tasks and computations to edge servers in order to empower UEs with computational capabilities. This feature of edge computing ensure successful D2D communication [52], [53].

E. ROLES OF EDGE COMPUTING IN 5G

There are *six* main roles of edge computing to support real-time and interactive applications and services as follows:

R.1 *Local storage.* Edge computing offloads a massive amount of data from UEs to edge clouds. While edge servers offer distributed local storage for a significant amount of data, yet their storage is much lower than that in the cloud, which has virtually unlimited storage capacity. Examples of data being stored are computing strategies (e.g., computation offloading strategy [54]), metadata (e.g., timestamps and geographical locations), and monitoring data. The edge server provides different types of storage strategies to support different kinds of data. For instance, ephemeral storage provides temporary data storage to a set of interconnected mobile devices [55], [56].

R.2 *Local computation.* Edge computing offloads computation and process from less complex (e.g., smart phone) and highly complex (e.g., surgical tools and smart factories) UEs to edge clouds. While traditional cache and access technologies (e.g., IEEE 802.11) provide simple computation, the edge cloud is an intelligent computing system that provides local computation and data processing capabilities close to UEs in an independent and autonomous manner [57]. The outcomes of the computations and processes can be valuable inputs to other UEs, such as those in a smart factory. The advantage is that edge clouds perform small tasks and provide real-time responses locally, which help to reduce the cost and delay incurred to send the required data to the cloud.

R.3 *Local data analysis.* Edge computing processes and performs critical and real-time data analysis on a massive amount of raw data gathered from different applications in close proximity to generate valuable information [58]. The capability to make data analysis locally reduces the latency required to send data to, as well as to wait for responses from, the cloud. Subsequently, the outcomes of the local data analysis are used for decision making [59].

R.4 *Local decision making.* Edge computing helps entities to make real-time decisions and corresponding actions in an automated manner based on well-processed data [60]. The capability to make decisions locally reduces involvement from more components and data or information exchange, leading to: a) improved system availability, particularly the cloud; and b) improved bandwidth availability. As an example, edge computing

TABLE 2. summary of objectives, challenges, metrics, characteristics, and performance measures of clustering schemes for 5G networks.

Reference	Objectives					Computing			Edge computing		Attributes			5G functions					Performance			Roles of Edge computing					
	O.1 Improving data management	O.2 Enhancing QoS	O.3 Predicting network demand	O.4 Managing location awareness	O.5 Managing resource	C.1 Cloud computing	C.2 Edge computing	C.3 Hybrid computing	E.1 MEC	E.2 Fog Computing	T.1 Low latency	T.2 Location awareness	T.3 Network context information	U.1 Software defined networks	U.2 Network function virtualization	U.3 Massive MIMO	U.4 Dynamic access to RAT	U.5 D2D communication	P.1 operational cost	P.2 QoS	P.3 Energy efficiency	R.1 Local storage	R.2 Local computation	R.3 Local data analysis	R.4 Local decision making	R.5 Local operation	R.6 Local security enhancement
Guo et al [65]		✓			✓			✓		✓	✓			✓						✓			✓				
Kitanov et al [66]		✓					✓			✓	✓						✓				✓		✓			✓	
Markakis et al [67]		✓			✓		✓			✓	✓		✓	✓	✓				✓	✓	✓	✓	✓				
Zhang et al [68]	✓	✓					✓		✓		✓						✓				✓		✓		✓		
Taleb et al [12]		✓	✓				✓		✓		✓	✓	✓					✓		✓		✓	✓		✓		
Chen et al [52]		✓			✓		✓		✓		✓		✓					✓			✓		✓			✓	
Bastug et al [70]	✓		✓				✓		✓		✓		✓					✓	✓			✓	✓				
Huang et al [71]	✓	✓					✓		✓		✓		✓				✓	✓	✓	✓			✓		✓	✓	
Hsieh et al [72]	✓	✓					✓		✓		✓		✓		✓				✓	✓			✓		✓		
Rimal et al [73]					✓		✓		✓		✓		✓				✓	✓		✓	✓		✓		✓	✓	
Huang et al [75]	✓	✓			✓		✓		✓		✓		✓	✓						✓			✓		✓	✓	
Solozabal et al [76]		✓					✓		✓		✓			✓	✓		✓			✓			✓		✓	✓	
Singh et al [77]	✓	✓						✓	✓	✓	✓							✓	✓	✓		✓	✓	✓			
Tran et al [78]					✓			✓	✓		✓		✓				✓		✓	✓			✓		✓	✓	
Ridhawi et al [17]	✓	✓						✓			✓		✓				✓		✓	✓			✓	✓	✓	✓	

facilitates local decision making by automated factories. Multiple entities can make decisions in a collaborative manner.

R.5 *Local operation.* Edge computing enables remote control and monitoring – particularly critical devices including those under unsafe environment – from a distance, or a more comfortable or safer place [61].

R.6 *Local security enhancement.* Edge computing serves as an additional layer between the cloud and connected devices in order to improve network security, including UEs with limited resources [62]. The edge clouds can serve as secured distributed platforms that provide security credentials management, malware detection, software patches distribution, and trustworthy communications, to detect, validate, and countermeasure attacks. The advantage is that, due to the close proximity of edge computing, malicious entities can be quickly detected and isolated, and real-time responses can be initiated to ameliorate the effects of the attacks. This helps to minimize service disruptions. In addition, the scalability and modularity nature, as well as the capabilities, of edge computing can facilitate the deployment of block chain [64] among UEs with limited capabilities.

F. PERFORMANCE MEASURES

There are *three* main performance measures as follows:

P.1 *Lower operational cost*. Edge computing reduces the operational cost by providing local functions (R.1)-(R.5), instead of offloading (or sending) tasks and data to the cloud. This reduces offloading overhead (or reduces network resource consumption), such as bandwidth.

P.2 *Higher QoS*. Edge computing improves QoS by providing local functions (R.1)-(R.5). This reduces the amount of tasks and data offloaded to the cloud, and so it increases network performance (e.g., higher throughput and lower latency), which are important to delay-sensitive applications (e.g, remote surgery and online gaming).

P.3 *Energy efficiency* reduces energy consumption by providing local functions (R.1)-(R.5). This reduces the amount of energy incurred to offload tasks and data to the cloud (i.e., the energy incurred in communication), and so it increases network lifetime.

IV. STATE OF THE ART

The state of the art of edge computing in 5G networks is presented according to the three categories, namely fog-based, MEC-based, and hybrid solutions. Table 2 presents a summary of qualitative comparison, which has been used in the literature [63], among the existing schemes, covering their objectives, computational platforms, attributes, performance measures, and roles.

A. FOG BASED SOLUTIONS

In [65], a cross-layer resource management scheme is presented between optical network and fog computing over fiber networks in order to incorporate delay requirements to edge servers. The proposed scheme achieves the objectives of improving QoS (O.2) and improving resource management (O.5) by providing local computation (R.2). The proposed scheme uses the hybrid computational platform (C.3), whereby the edge clouds perform real-time tasks

and the cloud servers perform highly computational and resource-intensive tasks. 5G function, including SDN (U.1), is used. The proposed scheme has the attribute of low latency and close proximity (T.1). The services for different applications are performed in three layers (i.e., the cloud, fog, and UE layers). Highly computational and resource-intensive services are executed at the centralized cloud layer and real-time services are offloaded to the fog layer; while the UE layer performs functions locally at the UEs (or end devices), which have less computational power and storage due to their limitations. The proposed scheme has shown to provide higher QoS (P.2) (i.e., lower end-to-end delay).

In [66], a comparison of energy efficiency between cloud computing and fog computing is made under different modulation schemes, including 64 quadrature amplitude modulation (QAM), 16 QAM, phase shift keying (PSK), and quadrature phase shift keying (QPSK), in 5G in order to propose an energy-efficient model to improve average throughput and energy consumption per user. The proposed scheme achieves the objectives of improving QoS (O.2) by providing local computation (R.2) and local operations (R.5). The proposed scheme uses edge computing (C.2), particularly fog computing, to analyze fog computing and its energy consumption as compared to cloud computing. 5G function, including dynamic access to RATs (U.4), is used. The proposed scheme has the attribute of low latency and close proximity (T.1). Different RATs (i.e., 3G, 4G, and 5G) serve a number of different UEs. An energy efficiency model is proposed based on throughput, energy consumption, and the energy consumption level under fog environment. The proposed scheme has shown to improve energy efficiency (P.3).

In [67], an architecture that enables edge servers to provide caching, computing, and communications functions (also known as 3Cs) is proposed so that content and service providers can deploy their functions, services, and contents closed to UEs. The proposed architecture achieves the objectives of improving QoS (O.2) and improving resource management (O.5) by providing local storage (R.1) and local computation (R.2). The proposed architecture uses edge computing (C.2), particularly fog computing, to reduce processing delay. 5G functions, including SDN (U.1) and NFV (U.2), are used. The proposed architecture has the attributes of low latency and close proximity (T.1) and network context awareness (T.3) to acquire network information and traffic distribution. The architecture consists of: a) *virtual fog (vFog)* which is a framework that empowers UEs with 3Cs using NFV so that service provisioning becomes flexible; b) *hyper fog* which is a constellation of vFogs that allows data exchange and processing among the vFogs in order to provide resources from more than a single vFog; c) *regular extreme node*, which is a UE with processing and communication capabilities; and d) *super extreme node* is a UE with 3Cs that manage and manipulate the edge node of vFogs. In this architecture, a regular extreme node informs its corresponding super extreme node about its available resources, and then receive and execute networking tasks assigned by the super extreme node. The proposed architecture has shown to provide lower operational cost (P.1) and higher QoS (P.2) (i.e., lower decision making delay).

B. MEC BASED SOLUTIONS

In [68], an architecture is presented to perform energy-aware offloading, whereby each mobile UE decides whether to perform or offload computational tasks to MEC server, in order to reduce energy consumption of MEC. The UEs are heterogeneous in nature as they have different communication and computing capabilities, and the energy consumption of the computational tasks at the mobile UEs is higher than that in the MEC server [69]. The proposed architecture achieves the objectives of improving data management (O.1) and improving QoS (O.2) by providing local computation (R.2) and local decision making (R.4). The proposed architecture uses edge computing (C.2) to perform energy-aware offloading. 5G function, including dynamic access to RATs (U.4), is used. The proposed architecture has the attribute of low latency and close proximity (T.1). There are three main steps. *Firstly*, mobile UEs are classified according to their energy consumption in computation and file transmission, and the transmission delay between the mobile UEs and the MEC. There are three main categories: a) *type 1 UEs* use MEC server for computation; b) *type 2 UEs* perform computation themselves; and c) *type 3 UEs* can choose to perform computation either at MEC servers or by themselves. *Secondly*, priorities are given to the different UEs based on their energy consumption, as well as available channels and their channel quality. In general, type 1 UEs enjoy higher priorities due to their limited computational capabilities and the need to offload computational tasks to the MEC server in order to satisfy the delay constraint. *Thirdly*, channels are allocated for UEs with different priorities. Since UEs with higher priorities are offloaded, there are lower number of UEs competing for channels. The proposed architecture has shown to provide higher energy efficiency (P.3).

In [12], MEC services are autonomously created by the nearest edge server in order to provide mobile UEs with seamless QoE in video streaming. The proposed scheme achieves the objectives of improving QoS (O.2) and predicting network demand (O.3) by providing local storage (R.1), local computation (R.2), and local decision making (R.4). The proposed scheme uses edge computing (C.2) to perform uninterrupted video streaming. 5G function, including D2D communication (U.5), is used. The proposed scheme has the attributes of low latency and close proximity (T.1), location awareness (T.2), and network context awareness (T.3). The edge server receives all or part of a content (e.g., video) from the cloud, so that the content can be transmitted to UEs with reduced delay. Hence, the quality of the content is good as long as a UE is in the vicinity of the edge server. In general, the UEs receive contents from the edge server to reduce delay (and hence, higher quality streaming); however, if the contents are unavailable in the edge server, the UEs would receive contents from the cloud, which increases delay (and hence,

lower quality streaming). There are two main mechanisms to ensure seamless content transmission. Firstly, *migration* enables seamless content transmission when a UE moves from the vicinity of an edge server to another. Secondly, *handover* enables seamless content transmission when a UE handover from a network provider to another, which reduces delay (and hence, higher quality streaming). The proposed scheme has shown to provide higher QoS (P.2) (i.e., lower end-to-end delay).

In [52], a D2D architecture is proposed for a massive number of UEs to execute collaborative tasks in an energy-efficient manner. The proposed architecture achieves the objectives of improving QoS (O.2) and improving resource management (O.5) by providing local computation (R.2), local decision making (R.4), and local operation (R.5). The proposed architecture uses edge computing (C.2) to perform energy-efficient task offloading. 5G function, including D2D communication (U.5), is used. The proposed architecture has the attributes of low latency and close proximity (T.1), and network context awareness (T.3) to acquire network information and traffic distribution. The UEs are categorized based on their computational capacity and links (i.e., cellular and D2D links). The UEs use graph matching to determine whether to perform tasks locally or to offload them to the edge nodes via D2D in order to achieve energy efficiency. The graph matching algorithm, which represents nodes and links in a graph, has two main stages: a) prunes out nodes without tasks; and b) creates a sub-graph that consists of nodes with tasks (also called task nodes), replicates them, and connects them to edge nodes of the graph. A task node performs a task locally if it can be matched by its own replica, and it offloads the task to another node if it can be matched with the other node. The proposed architecture has shown to provide higher energy efficiency (P.3).

In [70], a predictive and proactive caching approach is introduced in order to reduce peak traffic demands. The proposed scheme achieves the objectives of improving data management (O.1) and predicting network demand (O.3) by providing local storage (R.1) and local computation (R.2). The proposed scheme uses edge computing (C.2) to perform proactive caching at the edge of the network or at UEs. 5G function, including D2D communication (U.5), is used. The proposed scheme has the attributes of low latency and close proximity (T.1), and network context awareness (T.3) to acquire network information and traffic distribution. Popular contents are cached in edge servers, BSs, or UEs during off peak times.WThe popularity of a content is based on the UEs' behavior and the frequency of the BS requesting for the content. When a BS requests for a particular content, there are two possibilities: a) the content is available at an influential UE, who had possessed or processed the content in the past, and so the content is delivered from the influential UE to the BS via D2D; and b) the content is unavailable at any influential UEs, and so the content is delivered from the core network to the BS. The proposed scheme has shown to provide lower operational cost (P.1).

In [71], an application-aware traffic redirection mechanism is proposed for MEC in order to reduce response time and bandwidth consumption. The proposed scheme achieves the objectives of improving data management (O.1) and improving QoS (O.2) by providing local computation (R.2), local decision making (R.4), and local operation (R.5). The proposed scheme uses edge computing (C.2). 5G functions, including dynamic access to RATs (U.4) and D2D communication (U.5), are used. The proposed scheme has the attributes of low latency and close proximity (T.1) and network context awareness (T.3) to acquire network information and traffic distribution (T.3). The MEC controller allows UEs to offload (or redirect) the traffic of an application to MEC at the edge of the network when the bandwidth requirement of the traffic exceeds a preset threshold. Subsequently, the UEs can access the application and its traffic. The proposed scheme has shown to provide lower operational cost (P.1) and higher QoS (P.2) (i.e., lower response time).

In [72], a virtualized multi-access edge computing framework is proposed to increase available bandwidth and reduce end-to-end delay in an intelligent manner in Internet of things. The proposed framework achieves the objectives of improving data management (O.1) and improving QoS (O.2) by providing local computation (R.2), and local decision making (R.4). The proposed framework uses edge computing (C.2). 5G function, including NFV (U.2), is used. The proposed framework has the attributes of low latency and close proximity (T.1), and network context awareness (T.3) to acquire network information and traffic distribution. The proposed framework uses MEC to perform virtualized multi-access computing at the edge of the network. Hardware devices are disaggregated and virtualized into layers that provide different control functions (e.g., traffic offloading), services (e.g., computational and storage capabilities), and resources (e.g., computing and storage resources) using NFV. In addition, traffic offloading provides network traffic information, such as the number of packets, as well as the priority level and type of traffic, based on data flow. Traffic is prioritized and segregated into three categories, namely high-, medium-, and low-priority traffic, based on the packet flow rate and type of traffic, as well as the number of packets in the queue. Low-priority packets are dropped when signal strength is low and congestion occurs. The proposed framework has shown to provide lower operational cost (P.1) and higher QoS (P.2) (i.e., lower end-to-end delay).

In [73], a fiber wireless (FiWi) access architecture is introduced to improve MEC services (e.g., traffic and network performance monitoring). The proposed architecture achieves the objectives of improving resource management (O.5) by providing local computation (R.2), local decision making (R.4), and local operation (R.5). The proposed scheme uses edge computing (C.2). 5G functions, including dynamic access to RATs (U.4) and D2D communication (U.5), are used. The proposed architecture has the attributes of low latency and close proximity (T.1) and network context awareness (T.3) to acquire network information and

traffic distribution. In the proposed architecture, BSs serve as rational service centers that provide updated information (i.e., traffic demand and RAT) to backhaul in order to provide intelligent and energy-efficient schemes. MEC is operating over FiWi [74], and ethernet is used to transfer traffic from RAN. The FiWi, along with ethernet, provides a framework for back-haul and broadband access. The proposed architecture has shown to provide higher QoS (P.2) (i.e., lower queuing delay in the data buffer) and lower energy consumption (P.3).

In [75], a group of vehicular neighboring nodes (or VNG) is dynamically managed using SDN to improve control over network and its resources in vehicular networks. The proposed scheme achieves the objectives of improving data management (O.1), improving QoS (O.2), and improving resource management (O.5) by providing local computation (R.2), local decision making (R.4), and local operation (R.5). The proposed scheme uses edge computing (C.2). 5G function, including SDN (U.1), is used. The proposed architecture has the attributes of low latency and close proximity (T.1), and network context information and traffic distribution (T.3). The proposed scheme integrates SDN to MEC in order to strengthen network control (e.g., a unified network control of heterogeneous networks) at the edge of the network for achieving a flexible network control and management. Real-time instructions (e.g., safety messages) is passed from road side units to vehicles in order to monitor network states (i.e., the available resources of vehicles) in order to make effective decisions (i.e., road blocks and route changes). Using SDN, the edge of the network is segregated into three layers: a) the *control plane* enables the MEC to obtain the global knowledge of network states for making optimal decisions (i.e., network-level decisions for efficient networking and fault diagnosis) with lower response time; b) the *social plane*, which is abstracted for communication among VNGs, enables the SDN switch to separate and forward sociality flows, each of which consists of data packets that indicate the key features of a VNG (e.g., the strength of a relationship, contact time, contact frequency, and the contact method) among vehicles so that suitable vehicles can be selected to form strong and weak ties. As an example, two workmates from the same office leaving a parking area on a daily basis can form a strong tie. As another example, random vehicles on the road can form temporary weak ties; and c) the *data plane* provides data transmission. The proposed scheme has shown to provide higher QoS (P.2) (i.e., lower end-to-end delay).

In [76], a non-standalone (i.e., disconnected from the Internet) MEC-based architecture is presented for mission-critical public safety services in order to achieve the delay requirement (i.e., less than 1 ms (ideal) or 10 ms (maximum) of round trip time) of 5G. The proposed architecture achieves the objective of improving QoS (O.2) by providing local computation (R.2), local decision making (R.4), and local operation (R.5). The proposed architecture uses edge computing (C.2). 5G functions, including SDN (U.1), NFV (U.2), and dynamic access to RATs (U.4), are used. The proposed architecture has the attribute of low latency and close proximity (T.1). MEC is used to provide a flexible architecture, whereby the user plane, which is the bottom layer, consists of UEs that can be grouped into virtual groups (or clusters) based on their ownership, as well as co-location and co-service relationships that define the relative location between a cluster and a service requested by the UEs. MEC is deployed close to the UEs, and the flexibility of the architecture allows the location and structure of the MEC to be customized and redefined as time goes by. The computational resources (e.g., servers, processors, and cloud), as well as radio interfaces and schemes (e.g., modulation schemes and TDMA) are distributed in different slices in network slicing, which enables virtualization by running multiple logical networks on a shared physical network infrastructure. The key benefit of network slicing is that it provides an end-to-end virtual network encompassing networking, computation, and storage functions. Urgent services (e.g., mission critical services) are executed in higher priority slices (e.g., real-time services such as life-saving services in e-health). Hence, additional resources are allocated to higher priority slices to serve the urgent services. The proposed architecture has shown to provide higher QoS (P.2) (i.e., lower end-to-end delay).

C. HYBRID SOLUTIONS

In [77], a D2D-based mobile edge and fog computing architecture is introduced to enable collaborative computing, which performs tasks in more than a single computing platforms or paradigms, in order to enhance MEC. The proposed architecture achieves the objectives of improving data management (O.1) and improving QoS (O.2) by providing local storage (R.1), local computation (R.2), local data analysis (R.3), and local decision making (R.4). The proposed architecture uses hybrid computing (C.3) to exploit D2D communication in collaborative environment. 5G function, including D2D communication (U.5), is used. The proposed architecture has the attribute of low latency and close proximity (T.1). Each UE initiates a service request and send it to the nearest relay gateway, which has connection to the core network (or cloud). The service handler of a relay gateway, which has information about the available services, decides whether the requested service should be performed locally or forwarded to another relay gateway that can perform the service. The decision is based on the availability of the service (e.g., the processing, computational, and storage capabilities, as well as delay requirements) at the relay gateway and its neighboring gateways. The proposed architecture has shown to provide lower operational cost (P.1) and higher QoS (P.2) (i.e., lower end-to-end and round-trip delays).

In [78], a context-aware, real-time collaborative architecture is proposed to manage heterogeneous resources (e.g., different storage and computational capabilities in different computational platforms/ layers) at the edge of the network. The proposed architecture achieves the objective of improving resource management (O.5) by providing local

computation (R.2), local decision making (R.4), and local operation (R.5). The proposed architecture uses hybrid computing (C.3) to optimally distribute tasks among cloud, MEC, and mobile UEs. 5G function, including dynamic access to RATs (U.4), is used. The proposed architecture has the attributes of low latency and close proximity (T.1) and network context awareness (T.3). Tasks are split and offloaded among cloud, MEC, and mobile UEs based on the task requirements: a) UEs process tasks that require less processing and computational capabilities; b) MEC server processes delay-sensitive tasks; and c) cloud processes non-delay-sensitive tasks. Both MEC and cloud process tasks that require higher processing and computational capabilities. The proposed architecture has shown to provide lower operational cost (P.1) and higher QoS (P.2).

In [17], a real-time, context-aware, service-composition, and collaborative architecture is proposed to deliver fast composite service, which is the consolidation of multiple services supported by the collaboration of different hardware (e.g., UEs, edge clouds, and cloud) and software with different capabilities. The proposed architecture achieves the objectives of improving data management (O.1) and improving QoS (O.2) by providing local computation (R.2), local data analysis (R.3), local decision making (R.4), and local operation (R.5). The proposed architecture uses hybrid computing (C.3) that enables collaboration among cloud, MEC, and UEs. 5G function, including dynamic access to RATs (U.4), is used. The proposed architecture has the attributes of low latency and close proximity (T.1) and network context awareness (T.3). Frequently accessed blocks, which are small units decomposed from a file, are stored (or cached) in MEC servers. Blocks requested by more than a single server are replicated and cached in other MEC servers based on file types and contents. This helps to reduce the end-to-end delay incurred to access the cloud. The proposed architecture has shown to provide lower operational cost (P.1) and higher QoS (P.2).

V. OPEN RESEARCH ISSUES

This section highlights the open research issues for a successful deployment of edge cloud in the 5G environment.

A. SERVICE ENHANCEMENT: QOE

QoE is a measure of the overall customer satisfaction level with a service provider. QoE is related to, but differs from, QoS, which embodies the notion that the hardware characteristics (e.g., the storage capacity and the number of processors in the servers [79]) and software characteristics (e.g., the interface development) can be measured, improved, and guaranteed. The challenge is to achieve a balanced trade-off between: a) higher availability or seamless connectivity of an application, which can be provided by the cloud when a UE is out of the vicinity of the edge server; and b) higher QoE of the application, which can be provided by the edge cloud when the UEs are in the vicinity of the edge server, in order to reduce delay and jitter. Hence, collaborative computational approaches, such as hybrid computing (C.3), can be used. Edge computing can be used to maintain network or service states (e.g. the availability and cost of the links, as well as the way a switch forwards the traffic) for evolving applications (e.g., 4K video streaming [80]) and offer proxying functionality on behalf of UEs. By maintaining the network states, the trade-off between the availability and QoE performance can be achieved with reduced signaling overhead incurred by network processes (e.g., handover). The signaling messages can also be aggregated to reduce signaling overhead. This leads to reduced network congestion, hence improving network scalability and network performance (e.g., higher throughput [81]). Addressing this open issue can provide improvement in QoS (P.2).

B. STANDARDIZATION OF PROTOCOLS

Standardization of protocols requires standardizing bodies or organizations to provide a set of universally acceptable rules for edge computing in 5G environment. There are *two* main challenges. *Firstly*, it is difficult to agree upon a standard (e.g., the location and capabilities of the edge cloud) due to its flexibility and diversified customization by different vendors. *Secondly*, a large number of heterogeneous UEs use different interfaces to communicate with the edge cloud. Standardization effort, such as the initiative from the European Telecommunications Standards Institute (ETSI) [82], has been put in place so that heterogeneous UEs can communicate with edge servers, and different layers and computation paradigms can collaborate among themselves, in a multi-vendor environment.

C. ADDRESSING HETEROGENEITY

Heterogeneity in communication (e.g., transmission range and data rate) and computing (e.g., hardware architecture and operating systems) technologies in edge computing for 5G has resulted in difficulties in developing a solution that is portable across different environment. Software-based (or programming-based) schemes may develop a programming-model for edge nodes to facilitate the execution of workloads simultaneously at multiple hardware levels [2]. However, a comprehensive distributed computing system must allow the different schemes to operate in a collaborative manner. Data and task-level parallelism splits workload into independent and smaller tasks that can be executed in parallel across different hardware and layers in edge clouds [12]. The proposed solutions enable heterogeneous UEs to communicate with edge servers.

D. SECURITY AND PRIVACY

While security and privacy is enhanced in edge computing as data do not travel across a network, there are two main problems that can increase network vulnerability at the edge of network. *Firstly*, the dynamic environment causes the data and network requirements of different network entities to vary rapidly. *Secondly*, the increasing number of devices communicating with each other must require a

scalable solution. Hence, trust and security management must address the aforementioned problems in order to address network vulnerability; however, this may incur high complexity and cost. Enhancing security and privacy is significant due to the importance of the data (e.g., health information). There are two potential solutions. *Firstly*, applications running on edge cloud must be blind/ unaware to the raw information (or unprocessed data). So, the raw information (e.g., personal data including healthcare information) must be encrypted or processed. *Secondly*, raw information can be removed prior to reaching the edge cloud to ensure privacy [83], [84].

VI. CONCLUSION

In this paper, we present a review of the state-of-the art development in edge computing, including fog-based, MEC-based, and hybrid solutions, in 5G networks. A taxonomy is established in which the edge computing approaches are classified according to different characteristics (e.g, objectives, computational platforms, and attributes) and the features of edge computing are presented. The key requirements of edge computing are to provide real-time interaction, local processing, high data rate, and high availability. Edge computing improves network performance to support and deploy different scenarios, such as remote surgery. Open issues for the successful deployment of edge computing in 5G are identified, including service enhancement, standardization, as well as addressing heterogeneity and security vulnerabilities. A qualitative comparison among the existing schemes in the literature is presented, and it shows the research gaps in this topic whereby the missing ticks represent potential open issues that can be further investigated. Although the deployment of edge computing in 5G provides numerous benefits, the convergence of both edge computing and 5G brings about new issues that should be resolved in the near future.

REFERENCES

[1] W. Z. Khan, E. Ahmed, S. Hakak, I. Yaqoob, and A. Ahmed, "Edge computing: A survey," *Future Gener. Comput. Syst.*, vol. 97, pp. 219–235, Aug. 2019.

[2] N. Hassan, S. Gillani, E. Ahmed, I. Ibrar, and M. Imran, "The role of edge computing in Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 110–115, Nov. 2018.

[3] E. Ahmed, A. Ahmed, I. Yaqoob, J. Shuja, A. Gani, M. Imran, and M. Shoaib, "Bringing computation closer toward the user network: Is edge computing the solution?" *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 138–144, Nov. 2017.

[4] J. Shuja, S. Mustafa, R. W. Ahmad, S. A. Madani, A. Gani, and M. K. Khan, "Analysis of vector code offloading framework in heterogeneous cloud and edge architectures," *IEEE Access*, vol. 5, pp. 24542–24554, 2017.

[5] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[6] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. INFOCOM*, Apr. 2018, pp. 1970–1978.

[7] J. M. Hamamreh, E. Basar, and H. Arslan, "OFDM-subcarrier index selection for enhancing security and reliability of 5G URLLC services," *IEEE Access*, vol. 5, pp. 25863–25875, 2017.

[8] P. Gallo, K. Kosek-Szott, S. Szott, and I. Tinnirello, "CADWAN: A control architecture for dense WiFi access networks," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 194–201, Jan. 2018.

[9] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. 10th Int. Conf. Intell. Syst. Control (ISCO)*, Jan. 2016, pp. 1–9.

[10] Q. V. Pham , F. Fang, V. N. Ha, M. Le, Z. Ding, L. B. Le, and W. J. Hwang, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," 2019, *arXiv:1906.08452*. [Online]. Available: https://arxiv.org/abs/1906.08452

[11] K. Antonakoglou, X. Xu, E. Steinbach, T. Mahmoodi, and M. Dohler, "Toward haptic communications over the 5G tactile Internet," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3034–3059, 4th Quart., 2018.

[12] T. Taleb, S. Dutta, A. Ksentini, M. Iqbal, and H. Flinck, "Mobile edge computing potential in making cities smarter," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 38–43, Mar. 2017.

[13] Y. Yu, "Mobile edge computing towards 5G: Vision, recent progress, and open challenges," *China Commun.*, vol. 13, no. Supplement2, pp. 89–99, 2016.

[14] S. Raza, S. Wang, M. Ahmed, and M. R. Anwar, "A survey on vehicular edge computing: Architecture, applications, technical issues, and future directions," *Wireless Commun. Mobile Comput.*, vol. 2019, Feb. 2019, Art. no. 3159762.

[15] S. York and R. Poynter, "Global mobile market research in 2017," in *Mobile Research*. New York, NY, USA: Springer, Aug. 2017, pp. 1–14.

[16] Z. Cui and J. Wang, "Enhanced software-defined network controller to support ad-hoc radio access networks," U.S. Patent 15 054 180, Oct. 23, 2018.

[17] I. A. Ridhawi, M. Aloqaily, Y. Kotb, Y. A. Ridhawi, and Y. Jararweh, "A collaborative mobile edge computing and user solution for service composition in 5G systems," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 11, Jun. 2018, Art. no. e3446.

[18] B. Yang, Z. Yu, J. Lan, R. Zhang, J. Zhou, and W. Hong, "Digital beamforming-based massive MIMO transceiver for 5G millimeter-wave communications," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 7, pp. 3403–3418, Jul. 2018.

[19] N. C. Luong, P. Wang, D. Niyato, Y.-C. Liang, Z. Han, and F. Hou, "Applications of economic and pricing models for resource management in 5G wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, to be published.

[20] Z. Pi, J. Choi, and R. W. Heath, Jr., "Millimeter-wave gigabit broadband evolution toward 5G: Fixed access and backhaul," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 138–144, Apr. 2016.

[21] A. Ateya, A. Muthanna, I. Gudkova, A. Abuarqoub, A. Vybornova, and A. Koucheryavy, "Development of intelligent core network for tactile Internet and future smart systems," *J. Sens. Actuator Netw.*, vol. 7, p. 1, Jan. 2018.

[22] Z. Ning, X. Kong, F. Xia, W. Hou, and X. Wang, "Green and sustainable cloud of things: Enabling collaborative edge computing," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 72–78, Jan. 2018.

[23] K. Yeow, A. Gani, R. W. Ahmad, J. Rodrigues, and K. Ko, "Decentralized consensus for edge-centric Internet of Things: A review, taxonomy, and research issues," *IEEE Access*, vol. 6, pp. 1513–1524, 2017.

[24] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[25] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "The brewing storm in cloud gaming: A measurement study on cloud to end-user latency," in *Proc. 11th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nov. 2012, pp. 1–6.

[26] Z. Zhao, K. Hwang, and J. Villeta, "Game cloud design with virtualized CPU/GPU servers and initial performance results," in *Proc. 3rd Workshop Sci. Cloud Comput. Date ScienceCloud*, 2012, pp. 23–30.

[27] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, 2016.

[28] W. Hu, Y. Gao, K. Ha, J. Wang, B. Amos, Z. Chen, and P. Pillai, "Quantifying the impact of edge computing on mobile applications," in *Proc. 7th ACM SIGOPS Asia–Pacific Workshop Syst.*, 2016, pp. 1–5.

[29] M. Fahad, F. Aadil, S. Khan, P. A. Shah, K. Muhammad, J. Lloret, H. Wang, J. W. Lee, and I. Mehmood, "Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks," *Comput. Elect. Eng.*, vol. 70, pp. 853–870, Aug. 2018.

[30] F. Aadil, A. Raza, M. F. Khan, M. Maqsood, I. Mehmood, and S. Rho, "Energy aware cluster-based routing in flying ad-hoc networks," *Sensors*, vol. 18, no. 5, p. 1413, 2018

[31] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on multi-access edge computing for Internet of Things realization," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2961–2991, 4th Quart., 2018.

[32] M. Shahid, A. Bo, A. Dulaimi, and K. Tsang, "Guest editorial 5G tactile Internet: An application for industrial automation," *IEEE Trans. Ind. Inform.*, vol. 15, no. 5, pp. 2992–2994, May 2019.

[33] M. Aazam, K. A. Harras, and S. Zeadally, "Fog computing for 5G tactile industrial Internet of Things: QoE-aware resource allocation model," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3085–3092, May 2019.

[34] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis," *IEEE Netw.*, vol. 32, no. 2, pp. 32–40, Mar. 2018.

[35] R. Miller. (2017). *Data Center First: Intels Vision for a Data-Driven World*. Accessed: Apr. 24, 2019. [Online]. Available: https://datacenterfrontier.com/data-center-first-intels-vision-for-a-data-driven-world/

[36] X. Sun and N. Ansari, "Green cloudlet network: A distributed green mobile cloud network," *IEEE Netw.*, vol. 31, no. 1, pp. 64–70, Feb. 2017.

[37] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," *ETSI White Paper*, vol. 11, no. 11, pp. 1–16, Sep. 2015.

[38] Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. Braun, and S. Sargento, "Mobility prediction-assisted over-the-top edge prefetching for hierarchical VANETs," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1786–1801, Aug. 2018.

[39] E. Chirivella-Perez J. Gutiérrez-Aguado, J. M. Alcaraz-Calero, and Q. Wang, "Nfvmon: enabling multioperator flow monitoring in 5G mobile edge computing," *Wireless Commun. Mobile Comput.*, vol. 2018, Feb. 2018, Art. no. 2860452.

[40] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st Ed. MCC Workshop Mobile Cloud Comput.*, Aug. 2012, pp. 13–16.

[41] General Intelligence, "Understanding 5G: Perspectives on future technological advancements in mobile," Walbrook Building, London, U.K., White Paper 26, 2014.

[42] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[43] Y. Li and M. Chen, "Software-defined network function virtualization: A survey," *IEEE Access*, vol. 3, pp. 2542–2553, 2015.

[44] Y. Jararweh, C. Mavromoustakis, D. B. Rawat, and M. H. Rehmani, "SDN, NFV, and mobile edge computing with QoE support for 5G," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 11, p. 3536, Nov. 2018.

[45] A. J. Gonzalez, G. Nencioni, A. Kamisi ski, B. E. Helvik, and P. E. Heegaard, "Dependability of the NFV orchestrator: State of the art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3307–3329, 4th Quart., 2018.

[46] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos, and C. Verikoukis, "Application and network VNF migration in a MEC-enabled 5G architecture," in *Proc. 23rd IEEE Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Nov. 2018, pp. 1–6.

[47] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[48] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[49] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[50] Y. He, F. R. Yu, N. Zhao, and H. Yin, "Secure social networks in 5G systems with mobile edge computing, caching, and device-to-device communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 103–109, Jun. 2018.

[51] H. Wang, J. Wang, G. Ding, L. Wang, T. A. Tsiftsis, and P. K. Sharma, "Resource allocation for energy harvesting-powered D2D communication underlaying UAV-assisted networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 14–24, Mar. 2018.

[52] X. Chen, L. Pu, L. Gao, W. Wu, and D. Wu, "Exploiting massive D2D collaboration for energy-efficient mobile edge computing," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 64–71, Aug. 2017.

[53] Y. He, J. Ren, G. Yu, and Y. Cai, "D2D communications meet mobile edge computing for enhanced computation capacity in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1750–1763, Feb. 2019.

[54] J. Zhang, W. Xia, F. Yan, and L. Shen, "Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing," *IEEE Access*, vol. 6, pp. 19324–19337, 2018.

[55] N. Bessis and C. Dobre, *Big Data and Internet of Things: A Roadmap for Smart Environments*, vol. 546. New York, NY, USA: Springer, 2014.

[56] J. A. Silva, R. Monteiro, H. Paulino, and J. M. Lourenco, "Ephemeral data storage for networks of hand-held devices,"in Proc. IEEE Trustcom/BigDataSE/ISPA, Aug. 2016, pp. 1106–1113.

[57] A. Ndikumana, N. H. Tran, T. M. Ho, Z. Han, W. Saad, D. Niyato, and C. S. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Trans. Mobile Comput.*, to be published.

[58] C. Li, Y. Xue, J. Wang, W. Zhang, and T. Li, "Edge-oriented computing paradigms," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–34, Apr. 2018.

[59] E. Ahmed, I. Yaqoob, I. A. Hashem, I. Khan, A. I. Ahmed, M. Imran, and A. V. Vasilakos, "The role of big data analytics in Internet of Things," *Comput. Netw.*, vol. 129, pp. 459–471, Dec. 2017.

[60] M. M. Hussain, M. S. Alam, and M. M. S. Beg, "Feasibility of fog computing in smart grid architectures," in *Proc. 2nd Int. Conf. Commun., Comput. Netw.* Singapore: Springer, Sep. 2018, pp. 999–1010.

[61] C. Xu, K. Wang, P. Li, S. Guo, J. Luo, B. Ye, and M. Guo, "Making big data open in edges: A resource-efficient blockchain-based approach," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 4, pp. 870–882, Apr. 2019.

[62] M. Ali, H. S. Bilal, M. A. Razzaq, J. Khan, S. Lee, M. Idris, M. Aazam, T. Choi, S. C. Han, and B. H. Kang, "IoTFLiP: IoT-based flipped learning platform for medical education," *Digital Commun. Netw.*, vol. 3, no. 3, pp. 188–194, Aug. 2017.

[63] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, Sep. 2017.

[64] R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang, "Integrated blockchain and edge computing systems: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1508–1532, Feb. 2019.

[65] S. Guo, S. Shao, Y. Wang, and H. Yang, "Cross stratum resources protection in fog-computing-based radio over fiber networks for 5G services," *Opt. Fiber Technol.*, vol. 37, pp. 61–68, Sep. 2017.

[66] S. Kitanov and T. Janevski, "Energy efficiency of fog computing and networking services in 5G networks,"in Proc. EUROCON, Jul. 2017, pp. 491–494.

[67] E. K. Markakis, K. Karras, A. Sideris, G. Alexiou, and E. Pallis, "Computing, caching, and communication at the edge: The cornerstone for building a versatile 5G ecosystem," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 152–157, Nov. 2017.

[68] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.

[69] Y. Ai, M. Peng, and K. Zhang, "Edge computing technologies for Internet of Things: A primer," *Digit. Commun. Netw.*, vol. 4, no. 2, pp. 77–86, 2018.

[70] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[71] S.-C. Huang, Y.-C. Luo, B.-L. Chen, Y.-C. Chung, and J. Chou, "Application-aware traffic redirection: A mobile edge computing implementation toward future 5G networks," in *Proc. IEEE 7th Int. Symp. Cloud Service Comput. (SC2)*, Nov. 2017, pp. 17–23.

[72] H.-C. Hsieh, J.-L. Chen, and A. Benslimane, "5G virtualized multi-access edge computing platform for IoT applications," *J. Netw. Comput. Appl.*, vol. 115, pp. 94–102, Aug. 2018.

[73] B. P. Rimal, D. P. Van, and M. Maier, "Mobile edge computing empowered fiber-wireless access networks in the 5G era," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 192–200, Feb. 2017.

[74] J. Liu, H. Guo, H. Nishiyama, H. Ujikawa, K. Suzuki, and N. Kato, "New perspectives on future smart FiWi networks: Scalability, reliability, and energy efficiency," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1045–1072, 2nd Quart., 2016.

[75] X. Huang, R. Yu, J. Kang, Y. He, and Y. Zhang, "Exploring mobile edge computing for 5G-enabled software defined vehicular networks," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 55–63, Dec. 2017.

[76] R. Solozabal, A. Sanchoyerto, E. Atxutegi, B. Blanco, J. O. Fajardo, and F. and Liberal, "Exploitation of mobile edge computing in 5G distributed mission-critical push-to-talk service deployment," *IEEE Access*, vol. 6, pp. 37665–37675, 2018.
[77] S. Singh, Y.-C. Chiu, Y.-H. Tsai, and J.-S. Yang, "Mobile edge fog computing in 5G era: Architecture and implementation," in *Proc. Int. Comput. Symp. (ICS)*, Dec. 2016, pp. 731–735.
[78] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 54–61, Apr. 2017.
[79] M. Breternitz, K. A. Lowery, P. Kaminski, and A. Chernoff, "System and method for allocating a cluster of nodes for a cloud computing system based on hardware characteristics," U.S. Patent 13 568 395, Feb. 13, 2014.
[80] K. Kanai, K. Imagane, and J. Katto, "[Invited paper] overview of multimedia mobile edge computing," *ITE Trans. Media Technol. Appl.*, vol. 6, no. 1, pp. 46–52, 2018.
[81] Z. Guan, Y. Zhang, L. Wu, J. Wu, J. Li, Y. Ma, and J. Hu, "APPA: An anonymous and privacy preserving data aggregation scheme for fog-enhanced IoT," *J. Netw. Comput. Appl.*, vol. 125, pp. 82–92, Jan. 2019.
[82] K. Jain and S. Mohapatra, "Taxonomy of edge computing: Challenges, opportunities, and data reduction methods," in *Edge Computing*. New York, NY, USA: Springer, Nov. 2018, pp. 51–69.
[83] E. Ahmed and M. H. Rehmani, "Mobile edge computing: Opportunities, solutions, and challenges," *Future Generat. Comput. Syst.*, vol. 70, pp. 59–63, May 2017.
[84] K. Kaur, S. Garg, G. Kaddoum, M. Guizani, and D. Jayakody, "A lightweight and privacy-preserving authentication protocol for mobile edge computing," 2019, *arXiv:1907.08896*. [Online]. Available: https://arxiv.org/abs/1907.08896

NAJMUL HASSAN received the M.S. degree in computer sciences from Mohammad Ali Jinnah University, Islamabad, Pakistan, in 2010. He is currently pursuing the Ph.D. degree with the Department of Computing and Information Systems, Sunway University. He was a Visiting Faculty with the Department of Computer Science, COMSATS University Islamabad, Attock and Abbottabad Campus. His research interests include the Internet of Things, edge/cloud communication, wireless body area networks, the airborne Internet, and wireless sensor networks.

KOK-LIM ALVIN YAU (M'08–SM'18) received the B.Eng. degree (Hons.) in electrical and electronics engineering from Universiti Teknologi PETRONAS, Malaysia, in 2005, the M.Sc. degree in electrical engineering from the National University of Singapore, in 2007, and the Ph.D. degree in network engineering from the Victoria University of Wellington, New Zealand, in 2010. He is currently an Associate Professor with the Department of Computing and Information Systems, Sunway University. He is also a Researcher, a Lecturer, and a Consultant in cognitive radio, wireless networks, applied artificial intelligence, applied deep learning, and reinforcement learning. He serves as a TPC Member and a Reviewer for major international conferences, including ICC, VTC, LCN, GLOBECOM, and AINA. He was a recipient of the 2007 Professional Engineer Board of Singapore Gold Medal for being the best graduate of the M.Sc. degree, in 2006 and 2007. He also served as the Vice General Co-Chair for ICOIN'18, the Co-Chair for IET ICFCNA'14, and the Co-Chair (Organizing Committee) for IET ICWCA'12. He serves as an Editor for the *KSII Transactions on Internet and Information Systems*, an Associate Editor for IEEE Access, a Guest Editor for the Special Issues of IEEE Access, *IET Networks*, the *IEEE Computational Intelligence Magazine*, Springer *Journal of Ambient Intelligence and Humanized Computing*, and a Regular Reviewer for more than 20 journals, including the IEEE journals and magazines, the *Ad Hoc Networks*, the *IET Communications*, and others.

CELIMUGE WU received the M.E. degree from the Beijing Institute of Technology, China, in 2006, and the Ph.D. degree from The University of Electro-Communications, Japan, in 2010, where he is currently an Associate Professor. His current research interests include vehicular networks, sensor networks, intelligent transport systems, the IoT, and edge computing. He has been a TPC Co-Chair of Wireless Days 2019, ICT-DM 2019, ICT-DM 2018, a Track Co-Chair of many international conferences including IEEE VTC 2020-Spring, ICCCN 2019, and IEEE PIMRC 2016. He serves as an Associate Editor for IEEE Access, *IEICE Transactions on Communications*, the *International Journal of Distributed Sensor Networks*, and *MDPI Sensors*.