INVITED PAPER

SDN/NFV-Empowered Future IoV With Enhanced Communication, Computing, and Caching

By Weihua Zhuang, *Fellow IEEE*, Qiang Ye, *Member IEEE*, Feng Lyu, *Member IEEE*, Nan Cheng, *Member IEEE*, and Ju Ren, *Member IEEE*

ABSTRACT | Internet-of-Vehicles (IoV) connects vehicles, sensors, pedestrians, mobile devices, and the Internet with advanced communication and networking technologies, which can enhance road safety, improve road traffic management, and support immerse user experience. However, the increasing number of vehicles and other IoV devices, high vehicle mobility, and diverse service requirements render the operation and management of IoV intractable. Software-defined networking (SDN) and network function virtualization (NFV) technologies offer potential solutions to achieve flexible and automated network management, global network optimization, and efficient network resource orchestration with cost-effectiveness and are envisioned as a key enabler to future IoV. In this article, we provide an overview of SDN/NFV-enabled IoV, in which SDN/NFV technologies are leveraged to enhance the performance of IoV and enable diverse IoV scenarios and applications. In particular, the IoV and SDN/NFV technologies are first introduced. Then, the state-of-the-art research works are surveyed comprehensively, which is categorized into topics according to the role that the SDN/NFV technologies play in IoV, i.e., enhancing the performance of data communication, computing, and caching, respectively. Some open research issues are discussed for future directions.

KEYWORDS | Edge caching; fifth generation (5G); Internet of Vehicles (IoV); mobile edge computing (MEC); quality of service; resource slicing; software-defined networking (SDN)/network function virtualization (NFV); vehicle mobility.

I. INTRODUCTION

As fundamental technologies for the information dissemination and Internet access, communication networks have experienced a long-term evolution (LTE) from the first-generation (1G) systems to the fifth-generation (5G) systems for accommodating dramatically increased and diversified data communication demands. The evolving trend is getting faster and broader in terms of the unprecedented growth speed of mobile data traffic volume and the shifting demands from increasing the service data rate to multidimensional service quality enhancement (e.g., bandwidth, connectivity, latency, and energy efficiency). According to the IMT-2020 specifications [1], the 5G is envisaged to support a diversity of use cases in three broad service categories, namely enhanced mobile broadband (eMBB), ultrareliable and low latency communications (URLLCs), and massive

Manuscript received July 3, 2019; revised September 25, 2019; accepted October 24, 2019. This work was supported in part by a research grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada, in part by 111 Project under Grant B18059, and in part by the National Natural Science Foundation of China under Grant 61702562. *(Corresponding author: Qiang Ye.)*

W. Zhuang and **F. Lyu** are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: wzhuang@uwaterloo.ca; feng.lyu@uwaterloo.ca).

Q. Ye is with the Department of Electrical and Computer Engineering and Technology, Minnesota State University, Mankato, MN 56001 USA (e-mail: qiang.ye@mnsu.edu).

N. Cheng is with the School of Telecom Engineering and the State Key Lab of ISN, Xidian University, Xi'an 710071, China (e-mail: nancheng@xidian.edu.cn).

J. Ren is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: renju@csu.edu.cn).

Digital Object Identifier 10.1109/JPROC.2019.2951169

machine-type communications (mMTCs) [2]. The goal of eMBB is to achieve immerse experience for highly mobile users, such as virtual reality (VR) and augmented reality (AR), by providing higher and more stable data rates. The URLLC will support mission-critical applications, such as e-healthcare and industrial automation, while the mMTC provides seamless interconnection of a large number of Internet-of-Things (IoT) devices. These three 5G application categories capture key performance indicators (KPIs) for 5G and even beyond 5G (B5G) networks.

Internet of Vehicles (IoV) is one of the typical 5G networking scenarios to support different vehicle-related applications (e.g., safety and infotainment, and onboard sensing and processing services), which may require ultrahigh reliability and low-latency data transmissions, high throughput, or massive connectivity [3]. An advanced mobile communication system is required for IoV to enable timely information sharing and computing among vehicles and between vehicles and other "things" using vehicle-to-everything (V2X) communications, including vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-pedestrian (V2P), and vehicle-to-sensor (V2S) communications. During the past decade, many research and industrial standardization activities related to vehicular communications have significantly promoted the development of IoV and facilitated new vehicular and transportation services and applications, such as real-time traffic-aware navigation, collision warning, and infotainment. More importantly, the rapid development of IoV also brings autonomous driving into reality [3]. However, the diversified service demands of IoV applications not only require advanced vehicular communication techniques but also pose significant challenges in the management of communication, computing, and caching resources.

In order to meet various computing requirements from diverse services, offloading computation tasks to the cloud has been a dominant trend for IoV systems [4]. However, due to the increasing data traffic load in core networks, the long computing response latency and unstable connectivity in a vehicular environment make cloud computing unsuitable for supporting delay-sensitive and context-aware computing tasks. This consequently leads to the emergence of edge computing/caching that can exploit the computing and caching capabilities of "near-the-edge" infrastructures and devices [e.g., edge servers, base stations (BSs), and access points (APs) at roadside]. However, edge computing requires augmented capital and operational expenditures (CapEx and OpEx) for efficiently managing different types of system resources and flexibly provisioning different services in a complex and highly dynamic vehicular environment [5]. For managing vehicle access to edge servers, the increasing numbers of vehicles and distributed edge servers pose challenges on the allocation of communication and computation resources to meet various task requirements and balance the computation resource distribution. Meanwhile, due to high movement speed of vehicles causing a highly dynamic network topology, task offloading needs to cope with task migration among different edge servers, and caching strategies should be tailored to improve the prefetching accuracy for mobile vehicles [6].

Software-defined networking (SDN) is an emerging and promising solution to simplify network-layer data traffic forwarding and optimize network-level resource orchestration by decoupling the control plane from the data plane, which has been utilized along with mobile edge computing (MEC) for building advanced IoV systems [7]. It uses a centralized control module (called SDN controller) to collect network information, such as the vehicle density, vehicle location and mobility, supported service types and traffic load, and local resource allocation policies. With the gathered network information, the SDN controller can make network-level decisions for radio resource allocation, access control policies for end vehicles, and traffic routing path configuration among network elements, including BSs/APs, edge servers, and network switches in both access and core networks, to enhance service quality and improve the overall resource utilization. On the other hand, network function virtualization (NFV) offers a potential solution to use the virtualization technology for flexibly programming service functionalities, e.g., firewall, domain name system (DNS), network address translation (NAT), and video transcoding, as software instances referred to as virtual network functions (VNFs) at edge servers without adding significant costs [8]. The NFV can be combined with MEC to realize computation-oriented service provisioning at the network edge. By employing NFV, applications with different requirements of computation-intensive functionalities can be flexibly supported in a cost-effective way. The integration of SDN and NFV can potentially achieve network-level resource allocation and flexible service provisioning to enhance the performance of IoV systems, in terms of end-to-end quality-of-service (QoS) guarantee and service customization.

In leveraging the potential advantages, SDN/NFV-enabled IoV faces technical challenges that require further research for performance-enhancement techniques from the communication, computing, and caching perspectives. First, since the SDN controller needs to have both local and global control views, a challenge lies in how to apply SDN to IoV without significant signaling overhead in the presence of high network dynamics. Second, with SDN-/NFV-based MEC, a set of VNFs needs to be flexibly operated on edge servers at different locations near BSs to balance the computation burdens and enhance the computation efficiency. It brings challenging technical issues, such as how to determine the optimal amount of offloaded computation tasks between edge and cloud servers and how to optimize task migration among different edge servers, by taking vehicle mobility into consideration. Third, the existence of SDN/NFV can facilitate content caching at the network edge to reduce the service delay and alleviate the backhaul link congestion. However, further

research on caching strategies is required for optimizing cache deployment, improving cache hit ratio, and minimizing cache update cost.

This article provides a comprehensive survey of various techniques for SDN/NFV-enabled IoV. The remainder of this article includes an overview of IoV and the state-of-the-art SDN/NFV-enabling technologies for enhancing the performance of IoV services, ranging from reliable real-time services and data-intensive services to computing and caching oriented applications. We also discuss some open research issues for further studies and draw a conclusion at the end of this survey.

II. INTERNET OF VEHICLES: AN OVERVIEW

As an evolution of vehicular *ad hoc* networks (VANETs) in the era of 5G and IoT, the IoV integrates the emerging technologies, such as cloud/edge computing, SDN, and artificial intelligence (AI), with the conventional VANETs to further improve road safety, efficiency, and comfort. In IoV, vehicles act as multisensor smart objects, equipped with communication and computing capabilities. The IoV should be capable of sensing and processing in-car and surrounding information and communicating with nearby vehicles and other IoT devices and have the IP connectivity to the cloud servers and the Internet. The advanced communication and networking technologies in IoV facilitate new vehicular and transportation services and applications.

A. Architecture and Communication Paradigms

The IoV relies on efficient, fast, reliable, and cost-effective communication systems to guarantee ubiquitous network connectivity. V2X communications are the fundamental function of IoV, which allows vehicles to communicate with other vehicles, human, road sensors, transportation infrastructure, IoT devices, access networks, and the Internet [9]. V2X communications include different types of vehicular communications, such as V2V, V2I, V2P, and V2S, each of them employs various wireless and networking technologies to facilitate data transmission between vehicles and other network components.

The dedicated short-range communication (DSRC) and the IEEE Wireless Access in Vehicular Environments (WAVE) protocol stack are the key enabler of V2V and V2I communications. The IEEE 802.11p was released in 2010 [10], which specifies technical details of medium-access control (MAC) and physical layers of WAVE. In DSRC, one distinctive feature is the distributed and uncoordinated channel access, owing to the 802.11p carrier sensing multiple access with collision avoidance (CSMA/CA). It avoids the need of a centralized network coordinator for access control and resource allocation and, therefore, reduces the cost and improves the network scalability. However, low radio resource utilization, especially at a high vehicle density, is evident due to frequent channel access transmission collisions. Therefore, DSRC technologies are suitable for safety applications and lightweight proximate services, such as advertisement broadcasting, rather than data-craving applications.

Cellular V2X (C-V2X) is a series of advanced technologies to provide IoV with coverage, guaranteed QoS, reduced infrastructure deployment cost, and enhanced network security. The recent LTE and 5G new radio (NR) cellular networks adopt sophisticated wireless transmission technologies to enhance the network capacity and can, thus, offer high data rates to IoV applications. In addition, the centralized resource allocation, such as orthogonal frequency-division multiple access (OFDMA), guarantees the QoS of each IoV node in terms of data rate, delay, and reliability, crucial in many IoV services. The International Standard Organization, Third Generation Partnership Project (3GPP), has been working on the standardization of C-V2X and defining its architecture, fundamental functions, and use cases [11]. The architecture of C-V2X is also specified and two fundamental operation modes are discussed, i.e., traditional uplink/downlink communication via Uu interface and direct V2X communication via PC5 interface [12]. There are many works investigating the theoretical and practical performance of C-V2X. The reliability of DSRC is shown to be better at shorter distances, while LTE-V2V is more reliable at longer distances [13]. The transmission performance measures of DSRC and C-V2X are characterized and compared through field tests in Ann Arbor and San Diego. The results show that C-V2X outperforms DSRC in terms of coverage, reliability, and interference [14].

Although C-V2X communication has high performance, it is often an expensive approach to support the exponentially increasing IoV data traffic. In addition, the limited radio resources in the cellular system render it vulnerable to network congestion. In order to reduce the data service costs and offload the cellular IoV data traffic, alternative communication spectrum bands and paradigms are proposed for IoV, which includes cognitive radio (CR), roadside wireless local area networks (WLANs), and vehicular device-to-device (D2D) communications. These vehicular data offloading paradigms can employ much richer spectrum resources to provide cost-effective data pipes. Due to the primary usage of radio spectrum bands, limited coverage, and vehicle mobility, these communication modes are typically available tempospatially, which are, therefore, referred to as opportunistic communication methods. For instance, TV white spaces (TVWSs) are considered as a potential solution to accommodate the heavy IoV data traffic by employing CR technologies. In roadside WLANs, the communication takes place only when the IoV terminals are within the network coverage area, which depends on the vehicle speed and the deployment of roadside WLANs. Consequently, an important research issue is how to efficiently harvest the available spectrum and communication opportunities in order to improve the network performance of IoV, which has been investigated

for CR-IoV [15], roadside WLAN traffic offloading [16], and vehicular D2D communications [17].

Space–air–ground integrated networking (SAGIN) is another promising solution to provide IoV with ubiquitous network coverage, pervasive communication and computing, scalable network deployment, and flexible network services and management [18]. The satellite communication network can provide IoV users with telecommunication services, especially on the highway or in rural areas where a terrestrial network is not available. With technical advances in high-throughput satellites (HTSs) and low-earth orbit (LEO) constellation, the satellite network can offload the heavy data traffic burden from terrestrial cellular networks. Another important category of components, such as communication balloons and unmanned aerial vehicles (UAVs), forms an airborne network that can be deployed on demand and provide flexible services to IoV. Compared to the solution of deploying ultra-dense terrestrial cellular networks [19], UAVs can be deployed according to service demand variations over time and space, serving as flying relays, BSs, or caching units for different services, which is more flexible and cost-effective [20].

With the development of intelligent transportation systems (ITSs), smart city, and autonomous vehicles, a wide range of computation-intensive IoV services has emerged, such as VR/AR, self-driving, and crowdsourcing. To support these services, a comprehensive cloud/edge computing system is required. Cloud computing technologies are mature, and many commercial cloud computing services are available, including Amazon Web Services (AWS), AliCloud, and Microsoft Arezu. IoV users can offload the computation-intensive tasks to a cloud computing platform through a variety of V2I communication techniques to reduce the latency and energy consumption [21]. Yet, the latency of cloud computing can be unacceptable for some delay-critical applications, such as autonomous driving. In this case, edge computing should be utilized, where the computing resources are deployed at the network edge, much closer to the IoV users than cloud computing. Recently, another novel IoV computing paradigm, referred to as vehicular cloud (VC), has attracted much attention. Since connected vehicles are equipped with increasingly powerful onboard computing capabilities, they can form a local cloud through reliable V2V communications to cooperatively execute their computing tasks [22]. To reduce the latency for content retrieval, edge caching can be applied by caching popular contents at edge servers to improve the network performance when data-intensive services are delivered. Fig. 1 shows a comprehensive networking architecture of IoV systems, where V2X communications are provided by the space, aerial, and terrestrial communication networks, which can cooperatively offer seamless, efficient, and secure wireless connections to IoV users. A three-tier (vehicle-edge-cloud) computing system with edge caching is enabled to realize a flexible and cost-effective IoV service provisioning.

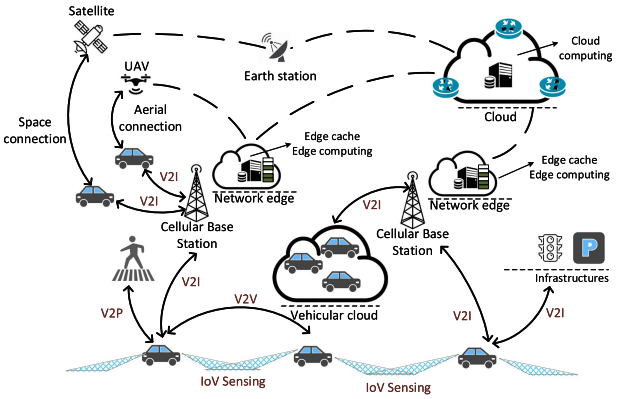

Fig. 1. ***Illustration of the IoV system.***

B. IoV Applications and Services

A wide range of IoV applications and services are at the core of IoV, the requirements of which should be fully satisfied by IoV sensing, communication, networking, and computing technologies. In the following, several typical IoV applications are described, and their requirements are listed and compared in Table 1.

1) Vehicle Collision Warning: Collision warning is an important road safety application, aiming at reducing the risk of accidents caused by collisions among vehicles, road structures, and pedestrians. Collision warning includes many different use cases, such as forward collision warning, emergency vehicle warning, signal violation warning, and precrash sensing warning [23]. Periodical and event-driven broadcasting is usually employed to warn the nearby vehicles with vehicle movement status, traffic condition, and potential danger. Such an application has strict requirements on message delivery latency and reliability, ranging between 50 and 100 ms for the delay and 10^{-5} for maximal message delivery failure rate. Short-range V2V and V2P communications are proper candidates for collision warning applications.

2) Real-Time Vehicle Traffic Management: It is to manage and optimize the traffic flow in a timely manner based on real-time vehicle traffic and road information [24]. In IoV, both vehicles and roads can be equipped with cameras, light detection and ranging (LiDAR) devices, and other sensors. These cameras and sensors can collect rich road and traffic information, which is then transmitted to a centralized ITS server via V2I communications. City-wide traffic optimization is executed in the ITS server, and navigation guidelines are then disseminated to vehicles considering the traveling time, safety, fuel, and comfort. To guarantee real-time traffic optimization and management, efficient IoV data collection, traffic information transmission, and traffic management optimization are required.

Table 1 Characteristics and Requirements of IoV Applications

Application	Communication support	Bandwidth Req	Delay Req	Computation Req
Vehicle collision warning	V2V, V2P	low	high	none
Real-time navigation	V2R, V2I	high	high	low
AR/VR	V2I, V2V	very high	high	low to high
Autonomous driving	V2V, V2I, V2S	high	very high	very high

3) Virtual Reality and Augmented Reality: One important category of IoV applications is infotainment, i.e., to provide users with a variety of information and entertainment contents and services to make road traveling more pleasant and comfortable. Among them, VR and AR have attracted much attention since they can facilitate the immersive experience, windshield road information display, and driving assistance. However, supporting IoV VR/AR services requires ultrahigh transmission rates and computing resources, which can pose severe challenges on the IoV infrastructure. Although cloud computing can provide ultrahigh computing capabilities, the latency is likely unacceptable for AR/VR applications. Therefore, MEC/caching technologies are the potential solution to jointly provide low latency and alleviate the data traffic burden of the backbone network [2].

4) Autonomous Driving: Autonomous vehicles undoubtedly represent the future of transportation systems. The Society of Automotive Engineers (SAE) International has defined six levels of car autonomy, from none to full automation [25]. The low levels of autonomy, i.e., driving assistance system, can be solely supported by the conventional V2V and V2I communications. The high levels of autonomy, such as level 3 and level 4, are currently in development based on individual vehicle perception, data fusion, machine learning, and maneuver control. However, the individual perception-based autonomy suffers from the limited perception capability and the lack of vehicle cooperation and, thus, cannot support the level 5 full autonomy. IoV and V2X communications are promising approaches to address these limitations [26].

III. SDN/NFV FOR IoV INFRASTRUCTURE

The conventional vehicular networking architecture employs V2V and V2I communications to interconnect different vehicles for supporting both road safety applications and infotainment services. With an increasing number of vehicles and the diversity of services, the overall network resource utilization needs to be enhanced substantially in order to accommodate differential QoS demands [27]. For example, in autonomous driving, different computation-intensive applications, including high-definition (HD) map downloading and onboard sensing, require ultralow processing and communication delays, whereas other data-hungry services demand high throughput. To achieve more efficient resource usage for diversified service accommodation, it is necessary to have a multitier of network coverage deployment (i.e., a macrocell deployment underlaid by different tiers of small cells) to exploit the spatial multiplexing gain of radio resources. In addition, a variety of wireless access technologies (e.g., LTE-V and DSRC) will coexist to offer a wide range of radio spectrum opportunities. However, the current networking architecture, i.e., an integration of heterogeneous wireless networking technologies with multiple layers of BS deployment, requires distributed coordination among communication infrastructures (e.g., BSs) for resource allocation, leading to increased signaling overhead when the number of BSs becomes large [28]. In the core network, traffic flows of different services aggregated from the wireless access networks need to traverse different sets of functionalities to fulfill E2E service delivery. For instance, traffic from a video streaming service may need to go through a firewall function and a transcoding function before reaching end clients for a secured E2E video downloading. These functionalities are often provided and operated in different function-specific servers. When supporting more diversified services, an increasing number of hardware servers are added into the network, forming a cloud computing platform with augmented capital and operational expenditures.

By enabling a flexible and programmable network configuration, SDN is a promising networking technology to simplify network-layer data forwarding and optimize network-level resource allocation [29]. With SDN, the network control plane is decoupled from the data plane and is migrated to a centralized controller placed in a network server. The SDN controller is physically connected to underlying network elements, including BSs in wireless networks and network switches/servers in core networks, through programmable links operating OpenFlow protocol [30]. In the core networks, the control function for routing configuration is migrated from each forwarding device to the SDN controller that determines and optimizes an edge-to-edge routing path for each service flow based on global network information. In this way, the underlying data forwarding structure of each device is greatly simplified, and the SDN control module customizes routing configurations for different service flows to meet different QoS requirements. In the wireless network domain, an SDN controller can be physically installed at an edge server connecting BSs through programmable control links. As shown in Fig. 2, with the information collected from underlying physical networks (e.g., the number of vehicles, vehicle location and mobility, and data traffic load) through the

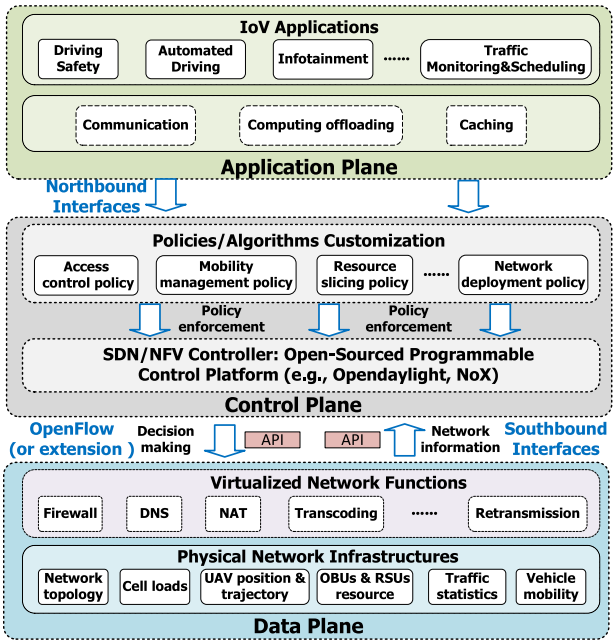

Fig. 2. *Layered SDN/NFV architecture supporting IoV.*

southbound application–program interface (API), the controller can make network-level decisions for radio resource slicing and access control to improve the overall service performance and resource utilization [31]. On the other hand, to reduce CapEx and OpEx for service provisioning, NFV is to decouple network (or service) functions from underlying hardware servers and to represent the functions as programmable software instances (i.e., VNFs) using virtualization technology [32]. This virtualization makes it possible to use generic servers instead of the conventional function-specific servers, and the VNFs managed by a centralized NFV control module can be flexibly placed at different network locations with optimal computing resource allocation to achieve cost-effectiveness. As a matter of fact, NFV provides a way of flexibly programming an increasing number of VNFs at appropriate network servers with the minimal hardware installation cost. This flexibility in function provisioning supports service customization, especially in the 5G era, when newly emerging applications require different sets of computation-intensive functionalities to be installed and operated in the network. Therefore, SDN and NFV are complementary technologies that can be integrated, as shown in Fig. 2, to achieve cost-effective VNF placement and service-oriented traffic routing, and network-level resource slicing with proper control of channel access from vehicles under different IoV scenarios.

With the new SDN/NFV integrated networking architecture, the overall network performance can be enhanced in terms of satisfying E2E QoS requirements (e.g., high throughput, low latency, and high reliability) and realizing a sequence of computation-intensive functionalities (e.g., data classification and video transcoding) with low response delay for different services. First, a new network-level resource allocation framework with proper BS-vehicle access control is necessary, in which resource combinations are created and customized for different services to satisfy their unique QoS requirements. This process is called resource slicing. The SDN controller has direct programmability on resource allocation for heterogeneous BSs, and the process of resource slicing aims at improving the overall resource utilization. Second, with SDN/NFV control, a set of VNFs can be flexibly executed in network servers at different network locations to enhance the computation efficiency. With the emergence of edge/fog computing technology, computation tasks can be offloaded from cloud servers to the edge servers connecting to BSs or even to end vehicles to reduce the two-way response delay. How to optimize computation offloading with the consideration of high vehicle mobility needs investigation. Finally, using SDN/NFV integrated control for content caching at the network edge becomes promising in terms of optimizing the caching deployment and minimizing the cost of caching content updates. In the following, we provide a detailed discussion on how SDN/NFV can enhance communication, computing, and caching for IoV applications.

IV. IoV COMMUNICATION SERVICES

To enhance road safety, transportation efficiency, and infotainment services to users on roads, many location-based and Internet services are essential, which rely on real-time, reliable, and data-intensive information exchanges among vehicles as well as other roadside equipment [33]. With the advanced SDN techniques [34], the communication paradigm is geared for better performance in the following aspects.

A. Latency

Ensuring low-latency communication is of paramount importance for delivering road-hazard messages, which are exchanged among nearby vehicles and roadside units (RSUs). The 3GPP standard (release 15) has specified latency requirements to support V2X services. Specifically, for vehicular cooperation applications in terms of platooning information exchange, sensor information sharing, and information sharing for automated driving, the end-to-end latency levels are required to be lower than 20, 50, and 100 ms, respectively [35], which is challenging to achieve in highly dynamic vehicular environments. The delivery latency mainly consists of the over-the-air transmission delay and channel access delay. As the transmission delay is deterministic when the data rate is given and negligible due to the high light traveling speed, the existing studies in the VANET literature focus primarily on guaranteeing channel access delay [36].

In the literature, the low-latency communications have been studied in terms of balancing the tradeoff between throughput and delay, such as the delay-constrained link capacity and the use of effective network capacity [37], [38]. These works aim at minimizing the long-term

time-averaged latency while falling short of guaranteeing the tail and worst case latency, which may fail to deliver critical messages for danger warning. To essentially guarantee the access delay for delivering road-hazard messages by broadcasting, efficient MAC protocol design is requisite [39]. Even though the IEEE 802.11p-based MAC protocol has been standardized, it suffers from long access delay and unfairness in support of periodical broadcast due to the contention-based distributed random access [10]. To overcome the limitations, time-division multiple access (TDMA)-based MAC protocols are proposed to support periodical broadcasting [40], [41]. The predefined frame and slot channel structure naturally support delay guarantee for urgent messages. Based on the time-slotted medium structure, other centralized and distributed TDMA-based MAC protocols [40], [42] are proposed for efficient time slot assignment and transmission CA.

Advanced SDN techniques potentially can help to reduce the random access delay when a vehicle fails to obtain a share of medium resources after contending or negotiating. First, by applying the network softwarization and slicing, a virtual dedicated network can be created and customized for the periodical broadcast service, whereby other bandwidth-intensive services are completely isolated out. Hence, the delay performance of broadcasting emergency messages is independent of the bursty heavy loads from other bandwidth-hungry services. Even though the resource utilization in the dedicated network may degrade without multiplexing with other services in resource usage, it is worthwhile as fast channel access has high priority in road safety applications. To enable isolated V2X services, Campolo *et al.* [43] justify the role of network slicing and provide suggestions to design dedicated V2X slices. Ye *et al.* [31] present a dynamic radio resource slicing framework for a two-tier heterogeneous wireless network to determine the optimal bandwidth slicing ratios. In a vehicular scenario, the challenge is to guarantee the stable performance for each network slice as vehicles move dynamically, while the allocation of physical radio resources is relatively stationary in interference management with frequency reuse, leading to the mismatch between supply and demand. Therefore, to deal with the mismatch, a dynamic and observant algorithm is in need to update the bandwidth slicing ratios frequently, in order to achieve both QoS satisfaction and efficient resource management. Given the available spectrum resources in the dedicated virtual network, the SDN controller can be utilized to control the medium access from vehicles and dynamically update the length of time frames and the time slot assignment parameters in accordance with vehicle density dynamics in both time and space domains.

B. Reliability

Communication reliability is another hurdle to overcome for IoV service satisfaction, especially for safety and online vehicle-traffic management applications. The wireless channel fading and network topology vary dramatically in a vehicular environment. For safety message delivery, the reliability is generally measured in the probability that a packet is successfully received by the target vehicle within a time limit [44]. As specified in the 3GPP standard [35], the reliability for most V2X applications is required to be above 99.99%, which call for developing new engineering solutions for V2X transmission and networking in presence of rapid-changing network topology and fast-varying wireless channels. In vehicular scenarios, two main factors can significantly degrade the reliability: 1) channel access collisions by concurrent transmissions [45] and 2) multipath fading due to both mobile (high-speed vehicles) and stationary (buildings) scatterers, as well as shadowing due to in-channel obstacles, such as trees, slopes, trucks, and buses. The channel impairments can significantly degrade transmission performance at the physical and link layers.

To reduce channel access collisions among nearby vehicles, various TDMA-based MAC protocols are proposed to assign disjoint time slots to vehicles in a neighborhood [40], [42]. In addition to transmission collisions caused by vehicle mobility, link-layer transmission performance deteriorates in a high vehicle density scenario. Transmission power control [46] and transmission rate control [47] are some approaches to deal with transmission congestion. In particular, cooperative communication at the link layer can enhance transmission reliability. For instance, Bharati and Zhuang [48] propose a cooperative MAC, named CAH-MAC, to mitigate the channel impairments and improve the network throughput. In cooperative communication, how to select the helper vehicle for retransmission is critical, and there are receiver- [49] and sender-oriented [23] approaches to determine the helper vehicle.

Under the SDN architecture, with the knowledge of real-time network status (e.g., the number of vehicles, vehicle mobility, QoS requirements, and available medium resources) via OpenFlow (or extension), the logically centralized control plane is able to update the channel access and transmission policy in a timely manner. The programmability of protocol-agnostic digital transceivers in the data plane makes the underlying hardware adaptable to all channel access and transmission policies by enabling/disabling transmissions, varying the transmission power and transmission rate [50]. Under this working model, the transmission collisions caused by mobility and data traffic congestion can be reduced via efficient control implemented at a local SDN controller. For instance, without affecting other applications, the transmission policy can require a transceiver to lower the data rate and enlarge the transmission power for reliability enhancement. In addition, in cooperative communication, the regional view of local controller benefits the helper vehicle selection since the rich information about

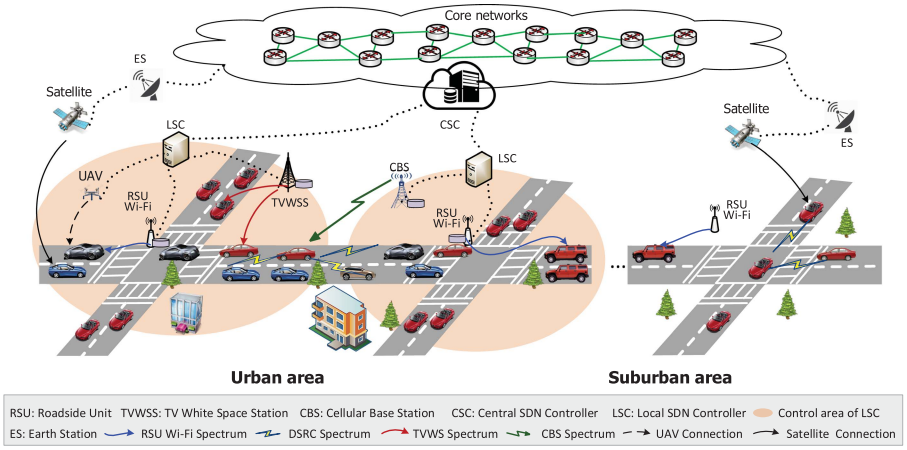

Fig. 3. ***SDN-based IoV communication system illustration.***

candidate helpers (e.g., position, historical transmitting/receiving status, and velocity) is available. However, to achieve the centralized control, the overhead of collecting the network status from vehicles is considerable as all vehicles under the control have to update the status with a high frequency, which limits the coverage area of each controller. In addition, as the controller deployment is costly, it is impossible to deploy local controllers to cover all road segments. Fig. 3 illustrates an SDN-based enhanced IoV communication architecture for both urban and suburban areas. In urban areas with busy and complicated vehicle traffic, centralized control can potentially play an important role.

1) As the radio spectrum resources are likely insufficient in such a high vehicle density scenario, the controller with a regional view is to enhance the overall efficiency of resource management. Taking account of the signaling overhead, multiple local SDN controllers should be placed to work cooperatively for an area of interest.
2) As the vehicle traffic condition is more complicated and changes more dramatically in the scenario (in comparison with that in suburban and rural areas), a robust policy is needed to work in all conditions for service performance guarantee.

On the other hand, in suburban areas where the vehicle traffic density is low and traffic condition is relatively simple, distributed network operation and control (via direct interactions among nearby vehicles) are potentially applicable, which avoids deployment of local SDN controllers.

C. Ubiquitous and Efficient Connections

In addition to the preceding lightweight V2V communications, data-intensive communication is essential to accommodate the high-throughput Internet services for vehicle users and for big data informatics, where vehicles equipped with various sensors are information source nodes. As the DSRC radio spectrum is insufficient, the radio channels are likely congested when supporting data-craving applications [47]. To this end, the heterogeneous vehicular network architecture is required. As shown in Fig. 3, in addition to DSRC and cellular BS, roadside Wi-Fi and TVWS should be leveraged to provide IoV connections [51]. However, terrestrial networks suffer from two inherent limitations in supporting IoV applications. First, the vehicle density varies dramatically with time, while the terrestrial network resources are more or less tied to geographical areas, leading to both resource wastage and shortage in different times and locations [18]. Even with an optimized deployment strategy, it is difficult to achieve both efficient resource utilization and service quality satisfaction in the long run. Second, for the terrestrial network, coverage holes exist especially in rural and remote areas, where radio signals are not sufficiently strong. The coverage holes restrain the ubiquitous connections for vehicles. Note that a full connection to and from a vehicle anywhere and anytime is required for future autonomous driving.

To address the two limitations, aerial and space networks are advocated to provide connections for vehicles, where the on-demand deployment of aerial networks (i.e., UAVs) can potentially make the network adaptive to a diverse environment, and a space satellite network can provide the ubiquitous wireless coverage [52]. To support IoV applications, how to determine in real time which network for vehicles to connect to [referred to as radio access technology (RAT) selection] is a critical yet challenging task, as various access technologies have unique advantages in terms of delay, throughput, and coverage range. For instance, roadside Wi-Fi can provide a high data rate when a vehicle is nearby, but the data rate is unstable as the vehicle moves away and the connection duration is short. Cellular networks demonstrate stable performance, but its high service cost may not be acceptable to users. Although a satellite network can provide the ubiquitous coverage, the transmission delay from the ground to space is long,

making it unsuitable for delay-sensitive applications [18]. In addition, complex algorithms are required to dispatch on-demand UAVs and schedule their trajectories [53].

SDN can potentially enhance the IoV connection performance by determining the access mode of each vehicle in real time. The SDN architecture decouples the underlying data packet forwarding from network control and can utilize protocol-agnostic hardware to support dynamically customized control policies. At the application plane, the precise application-specific requirements can be predetermined and delivered to the control plane via northbound API. Based on the network status information, including available resources in network segments (i.e., the terrestrial, aerial, and space network segments), vehicle density, QoS requirements, and vehicle mobility statistics, the control plane can customize the access policy and parameters for different vehicles toward differentiated applications [54]. However, to simultaneously manage all resources in the terrestrial, aerial, and space networks is difficult as the networks have disparate interference ranges, while the coverage of each controller is limited due to the signaling overhead. Therefore, a hierarchical control architecture is requisite, with a number of local controllers, each in charge of V2X communications in its coverage area (e.g., via a terrestrial network), and a central controller in charge of the local controllers and to coordinate service provisioning in a large area (including managing resources in the satellite network), as shown in Fig. 3. Under the hierarchical architecture, the local controllers can be deployed at network edge to provide quick responses to service demands from vehicles, while the central controller can be deployed at the core/backhaul network for large time and geographical scale (global) network management and operation.

V. IoV COMPUTING SERVICES

Data traffic flows from an increasing number of computation-intensive IoV applications (e.g., onboard video streaming, environment sensing, and HD map downloading) are required to traverse different sets of network/service functionalities (e.g., firewall, transcoding, data classification, and data fusion) for service customization, which is different from conventional communication-oriented vehicular networking for road safety applications. For instance, a secure HD-video streaming service requires video traffic traverse a sequence of firewall function and transcoding function before being displayed on end vehicles supporting differentiated video file formats. In a vehicular networking scenario, access and mobility management function (AMF) is another important network-level functionality in the 5G core network, which deals with user location registration, connection management, and vehicle-to-BS or vehicle-to-new-network (vertical) handover management [55]. By using the NFV technology, these network-level and service-level functionalities are decoupled from hardware servers and are virtualized as VNFs that can be flexibly placed by an NFV control module on generic servers in the network core, reducing the function execution and operational cost. For example, the AMF can be instantiated near the edge of the core network to improve the efficiency of connection and mobility management for vehicles moving among different BS coverage areas. In addition, with SDN, virtual network topology configured for both routing and VNF placement can be optimized and customized for different types of services. However, executing VNFs only at cloud servers incurs a two-way communication delay. For some real-time applications, a low response delay is required for task computation, such as object identification in environment sensing for autonomous driving [56], HD map data downloading and processing [57], and VR gaming. Therefore, MEC emerges to offload parts of the computing tasks from cloud servers to edge servers, in order to reduce the task transmission delay on backhaul links [4]. As shown in Fig. 4, a three-tier vehicular computing architecture is proposed to accommodate various IoV services, where nearby vehicles provide opportunistic computing, edge nodes (e.g., MEC servers and cloudlets) support fast computing and processing with limited resources, and the remote cloud guarantees sufficient resources for sustainable computing. Many existing studies present different integrated architectures for IoV by combining MEC/cloud computing with SDN/NFV [50], [58], to enable the flexibility of VNF orchestration with computing task offloading near the network edge and to improve the QoS performance for different computation-intensive applications. Several main research issues under this integrated architecture are discussed in the following.

A. Computation Offloading With Edge-Cloud Interplay

Combining NFV with MEC for vehicular networking is a promising approach to extend flexible VNF placement with computation offloading to the network edge for fulfilling the communication delay requirements for computing task execution. The system architecture has already been standardized by ETSI in 2018 [59]. Although the computation offloading reduces communication resource consumption over backhaul links and shortens the response latency for delay-sensitive computing task execution, it increases architectural complexity (i.e., more MEC servers need to be deployed) and computation redundancy at the network edge. Moreover, some delay-tolerant computing tasks from certain IoV applications require more powerful computation capability and should be placed at cloud servers for execution. Hence, there exists an edge-cloud interplay in terms of computation offloading. A key issue is how to balance the tradeoff between reducing communication delay for task processing and achieving desired computation performance with low cost. The SDN architecture provides a potential solution to enhance the NFV- and MEC-integrated vehicular networking architectures while

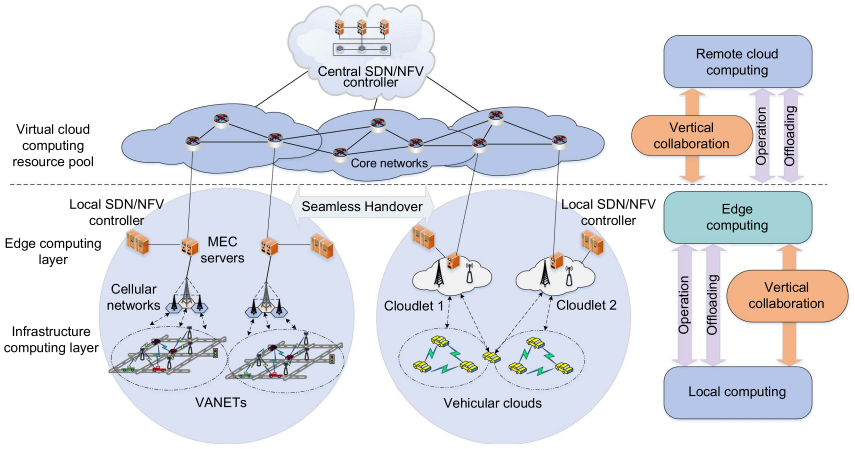

Fig. 4. ***Three-tier vehicular computing architecture.***

dealing with the edge-cloud interplay. The local SDN control can be implemented on a virtualized MEC server as a control system to handle the computing task scheduling among the edge server and end vehicles, or a central SDN controller can be implemented in a cloud data center to execute the computation task migration among different edge servers. By decoupling and migrating the control functionalities for data delivery from vehicles to edge servers, the underlying vehicles equipped with radio transceivers are abstracted as SDN switches [50]. Then, each local SDN controller can collect the vehicular network state information, the communication, and computing requirements of IoV applications through either DSRC or cellular networks operating the OpenFlow protocol in the application layer [4]. Based on the collected information, an optimized computing task scheduling policy can be executed as an application-plane program over the local SDN controller through the northbound API. On the other hand, the central SDN controller in the core network manages computing task offloading and migration policies among different edge servers based on the collected data traffic information and the prediction on vehicle mobility for service satisfaction and service continuity. Many existing works investigate the edge-cloud interplay problem in an SDN-/NFV-based MEC architecture. For example, a task scheduling framework is proposed for IoV based on an SDN/NFV-enabled MEC architecture, where challenges of the edge-cloud interplay problem are identified in terms of resource and energy limitation on edge servers and mobility management [60]. With the consideration of resource, energy, and delay constraints, developing a comprehensive computation offloading scheme to satisfy differentiated service requirements falls into the category of multiobjective optimization problems. Specifically, to obtain optimal decisions for task offloading, the edge-cloud interplay problem can be formulated to optimize the tradeoff between energy efficiency and backhaul bandwidth resource consumption under the latency constraints for task execution. For tractability, multiobjective evolutionary algorithms can be used to solve the problem based on decomposition methods.

B. Mobility-Aware Computing Task Migration and Data Offloading

Under the NFV- and MEC-integrated architectures, another important research problem is how to conduct computing task migration among nearby edge servers to maximize service completion probability within a certain time limit when a vehicle moves across different BS coverage areas. Each MEC server is physically connected to a set of BSs and is responsible for computing task offloading (from the vehicles) and execution under the BS coverages, as shown in Fig. 4. The SDN can be incorporated into the architecture to obtain computing resource usage information on edge servers and facilitate task migration over a properly configured path. There exist proposals for SDN-based wireless connection handover and inter-MEC server computing task migration in SDN-/NFV-based vehicular network architectures [61], [62], in which a local SDN controller manages the wireless connection reassociation, when vehicles move out of the previous BS coverage, and determines computing task migration among MEC servers. Based on the collection of vehicular network context information, how to optimize the edge server deployment and computing resource allocation based on vehicle mobility models is essential to maximize the success probability of task completion. In an IoV scenario, the SDN can help to establish an end-to-end routing path between two vehicles at different network areas through V2V and V2I communication modes for data offloading. Because of the increasing number of vehicles and densified network deployment, how to route data traffic with low communication overhead and to accommodate high source/destination node mobility are the key issues. In an SDN-based edge computing architecture, the SDN controller can assist to optimize the traffic routing paths to achieve load balancing and adapt the path selection to customized services with

differentiated delay requirements [63]. V2V-based traffic routing for data offloading from cellular networks is a cost-effective solution in an SDN-based IoV scenario due to its simplicity, where a local SDN controller can calculate an optimal V2V routing path across different RSU coverages based on the collected vehicular network information [64] and vehicle mobility prediction [65].

C. Vehicular Edge Cloud

Because of the deployment cost of MEC servers, another way is to explore computation opportunities among nearby vehicles for computing task offloading, which can reduce the response delay for task execution to satisfy strict latency requirements and relieve the computation burden on edge servers. Therefore, by employing vehicular networking technologies, some computation powerful vehicles can form a vehicular edge cloud, referred to as VC, to cooperatively execute various computing tasks [4], [22], [56]. The concept of VC has become relevant, especially in the area of autonomous driving where vehicles need to consistently collect a large amount of sensing data from the environment for prompt processing and proper actions. Forming an effective VC for cooperative computing can significantly reduce the two-way response delay for task execution on MEC servers and improve the computing performance by efficient task dissemination. Due to vehicle dynamics in joining and leaving a VC, the computing resource availability for a VC can be intermittent. Based on the computation requirements from different services, how to select vehicles to form an effective VC and how to schedule tasks among the selected vehicles to ensure the task completion within the vehicle contact duration need to be customized, which is technically challenging. Combining SDN with MEC is a promising architecture to facilitate VC establishment and make locally centralized decisions for computing task execution and data traffic routing among vehicles within a VC. In the literature, how to leverage SDN to dynamically form a VC and distribute tasks among vehicles for collaborative computing has been studied based on the information of resource availability, where the discrepancy of resource availability and task interdependence are considered for minimizing the average response time for task completion. The local SDN controller deployed at each MEC server helps to group nearby vehicles to form a VC, and the central SDN controller deployed at a cloud server coordinates distributed MEC servers for centralized management of VC groups. How to design the hierarchical controller architecture (in terms of determining the placement locations, the necessary number of controllers, and the networking topology among controllers) remains a key and challenging research issue, as it depends on the size, stability, resource availability, and task computation load of each VC. The controller operational cost also needs to be specified by considering the information exchange between each VC and the controllers, including the types, amount, and interaction frequency of network state information. Therefore, the objective is to determine an appropriate SDN controller architecture to improve the VC formation and computing task allocation performance under certain controller deployment and operational cost.

VI. IoV CACHING SERVICES

With more and more onboard mobile devices, IoV is envisioned to support ever-increasing infotainment applications for drivers and passengers to improve their road experience. However, these infotainment applications, such as HD-video streaming, online gaming, and future AR/VR, can consume a huge amount of spectrum resources and have rigorous real-time service delivery requirements, which poses challenges on vehicular networking. In addition, to enable autonomous driving, delivering real-time and high-volume HD map to vehicles is essential, which helps to localize exact vehicle positions (with the precision of 10–20 cm) on the road with the consumption of large amounts of computing and communication resources. Mobile edge caching plays an important role in providing satisfactory vehicular service provisioning, upon which contents can be cached at edge nodes close to end vehicles. It can significantly reduce the service latency and alleviate the backhaul pressure by allowing users to fetch contents within one transmission hop without traversing multihop core/backhaul networks to reach the Internet or cloud servers for contents. However, to create an effective caching system is not easy, as the problems of caching deployment, caching content update, and content dissemination are significant yet unsolved. The caching deployment problem deals with where to deploy caching functions and how much cache space to allocate, which is essential to increase the probability of successful downloading of cached data but is also labor-intensive with high capital and operational expenditure [66]. The performance of a caching system can be measured by three metrics: 1) caching gain—the amount of successfully downloaded data over the total amount of requested data; 2) content hit ratio—the number of content hits over the number of requests; and 3) caching diversity—the types of cached contents at different BSs. Therefore, an effective caching deployment aims at maximizing the caching gain under constraints of deployment cost (including installation cost and hardware maintenance cost) [66]. Considering vehicle mobility and differentiated spatiotemporal content request patterns from vehicles, how to update the cached contents to achieve a high content hit ratio is critical to improve the caching performance [67]. When a requested content from a vehicle is hit, how to disseminate the content to the end vehicle also affects service quality. Caching duplicated contents at multiple locations improves the transmission efficiency by exploiting high channel qualities at BSs close to vehicles, which, however, lowers the caching diversity. Achieving efficient content dissemination is to balance the tradeoff between transmission efficiency and

caching diversity. SDN/NFV technologies can potentially be leveraged to deal with the aforementioned problems with caching performance enhancement.

A. Caching Deployment

The edge caching deployment is a nontrivial problem as the content demands from vehicles are unpredictable and vehicle locations are time-varying. In a typical vehicular scenario, moving vehicles, stationary RSUs, and BSs are the possible executors for caching functions. Vehicles have limited capability for computing and data storage, and their mobility leads to difficulties to fully utilize their caching functions, especially with unstable communication links. RSUs are lightweight, with potential large-scale deployment. To achieve satisfactory caching performance, the deployment strategy for caching locations among all candidate RSUs aims to balance the tradeoff between deployment cost and caching gain. For cellular BSs, the cache sizes should be properly determined. Insufficient cache sizes can fail to fully unleash the caching potentials, while an excessive cache space leads to resource wastage.

In order to maximize the long-term caching gain, a statistical analysis on user association and data traffic should be carried out for developing a strategy in a large-scale Wi-Fi system [66]. The spatiotemporal vehicle traffic variations can be learned by mining the existing large-scale GPS traces [68] for caching deployment at RSUs and BSs. Likewise, to deploy caching functions at moving vehicles, the wireless interconnection time between a pair of vehicles [69] and their social behaviors (e.g., vehicle traveling routes) [70] can be studied based on the GPS trace to improve the caching performance. However, the inherent one-shot deployment approach cannot accommodate all vehicle traffic conditions with maximized caching gain. NFV technology can help to facilitate dynamic caching deployment. With NFV, computing and storage resources on vehicles are virtualized to support different functionalities as programmable software instances. In this way, caching functions can be flexibly programmed on different vehicles to exploit their resource availability, especially at rural or remote areas where both RSU and BS coverages are insufficient [71]. In the areas where vehicle traffic is light, some virtual resources used for sensing and computing can be further released to support more caching function instances to serve nearby vehicles with on-demand content requests [72]. In urban areas, superfluous caching function instances at RSUs and BSs can be deactivated during off-peak hours, and the released virtual resources can be utilized to support other network functions, such as firewall and retransmission. Therefore, resource virtualization technology can improve caching deployment performance by accommodating more caching function instances in different spatiotemporal scenarios [73]. Furthermore, for dynamic caching deployment, the SDN can be leveraged to make decisions on when and where to activate/deactivate caching function instances and how many virtualized storages to allocate to support caching function instances, by collecting real-time network state information and customizing dynamic caching policies [74].

B. Cached Content Update

In addition to HD maps that are required to be cached with priority, each user may have a unique preferable file list for other Internet contents. Given certain cache resources, what contents to cache and when to update become an important research problem, the goal of which is to maximize the cached content hit ratio. There are two conventional caching strategies in the literature: 1) least recently used (LRU): when the cache is full, the least recently requested contents are replaced and 2) least frequently used (LFU): when the cache is full, the least frequently requested contents are replaced. Both strategies are general approaches, but their granularity is not small enough to meet user request demands with spatiotemporal patterns. For instance, by learning the real-world data set that stores 0.3 million unique video viewing records by two million users, Ma *et al.* [67] investigate the user request patterns and confirm their spatiotemporal variation features. Some works aim at predicting content popularity in order to enhance the content hit ratio, based on a noisy linear model [75] or human-centric information (e.g., visited locations and requested contents) [76].

To adapt to the spatiotemporal popularity variations of Internet contents, the big data [3] and emerging AI techniques [77] are essential methodology. The centralized SDN architecture provides a suitable platform for implementation. Particularly, with the OpenFlow (or its extension) protocol, SDN controllers can facilitate the collection of content requests with user IDs, times, and locations and generate a statistically sufficient data set. Given a location, the time series of file requests can be learned and modeled offline, based on which the location-aware online popularity prediction is enabled [78]. Such information is useful for a dynamic content placement policy to enhance the overall content hit ratio. With offline prelearning for the dynamic network status as well as other features, such as location and time, the convergence speed for making real-time learning decisions can be improved to keep the pace of the fast-changing vehicular traffic conditions. Even though the information of historical file requests can be obtained, due to the SDN controller, the random nature of content demands reduces the caching performance. For instance, the requested content distribution for gaming can change significantly, which deviates from the prediction results. Likewise, if a traffic accident happens, bursty and unpredictable file requests may be generated, thus reducing the accuracy of AI models. Therefore, a backup scheme should be in place when the trained model loses efficacy. It should react to unexpected events in time and run as an alternative to compensate for the degraded performance from AI decisions. In addition, the popular

file list normally changes rapidly with time [67], which requires the SDN platform to continuously collect the up-to-date user request records, in order to update the training model. The large-scale data collection and computationally costly model training can put a high burden on the platform.

C. Cached Content Dissemination

The main objective of SDN-/NFV-based cached content dissemination in IoV is to minimize the content retrieval time [73]. However, it is technically challenging to achieve this goal as the transmission efficiency and the caching diversity are conflicting objectives. Caching duplicated contents at multiple BSs improves the transmission efficiency by successfully fetching the content over better channels but degrades the content caching diversity with underutilized cache spaces; on the other hand, caching different contents at multiple locations, referred to as cooperative caching, increases the content caching diversity but degrades the transmission efficiency for the cached contents when the channel quality between a BS and a vehicle is low. The essence of cached content dissemination is to investigate a multidimensional resource (e.g., caching resources and radio resources) management problem to balance the tradeoff between caching diversity and transmission efficiency, with the consideration of available caching locations, cache spaces and radio resources on BSs, and vehicle mobility. Mathematically, a multidimensional resource allocation problem can be formulated to jointly optimize two caching performance metrics: 1) content hit ratio and 2) content delivery time. Due to the problem complexity, it is difficult to develop a simple and accurate algorithm for making real-time decisions in a highly dynamic vehicular network. Alternatively, model-free reinforcement learning (RL) methods remain a tractable online-learning approach to solve the problem with fast convergence and high accuracy [18]. Specifically, the multidimensional resource allocation problem in a dynamic vehicular scenario can be described as a Markov decision process (MDP), and the RL method can be applied to solve the MDP by making real-time decisions on content placements and user associations, in order to maximize the time-averaged reward. A reward function can be expressed as a weighted function of content diversity and content delivery time. Upon interactions with the network environment, the RL-based approach tries to balance the exploration (randomly trying an action) and exploitation (choosing the most promising action based on current knowledge) processes to achieve time-averaged reward maximization. Due to vehicle mobility and limited wireless communication range, the connections between vehicles and caching places are intermittent, leading to the possible failure of downloading a complete cached content. Therefore, with the consideration of mobility prediction, how to cooperatively prefetch portions of contents at multiple locations to maximize the content hit ratio is important to enhance the content dissemination.

VII. OPEN RESEARCH ISSUES

There are many open technical issues for IoV in terms of improving the network performance with the consideration of high vehicle mobility, increased data traffic load, and diverse service requirements. In this section, some potential research issues are discussed from the aspects of enhancing IoV performance by applying SDN/NFV technologies. From the communication perspective, we identify the importance and effectiveness of leveraging the SDN/NFV technologies in IoV scenarios for conducting network-level resource slicing in a multitier heterogeneous communication infrastructure (e.g., macrocell BSs, small-cell BSs, and RSUs/APs) for service customization. From the computing perspective, a VNF chaining and placement problem at the edge of vehicular networks is discussed. Another potential research issue is how to jointly allocate communication, computing and caching resources in SDN/NFV-enabled vehicular networks. We also discuss hierarchical SDN/NFV controller deployment for multidimensional resource allocation.

A. Resource Slicing and Vehicle Access Control for IoV

SDN and NFV are the key technologies to enhance the integration of network-level communication and computing resource orchestration and end-device association. An SDN/NFV integrated control module is often placed at the edge of the core network and is physically connected to a set of BSs/APs through backhaul links [79]. Through programmable control links, the integrated control module collects important network state information from underlying vehicular networks, such as resource reservation information on BSs, vehicle density and mobility, data traffic load, statistics of different service demands, and supported service types with QoS requirements. Accordingly, the control module makes network-level decisions to update vehicle access control policies and reconfigure the overall network resources among BSs/APs to create resource "slices" for different services. Due to the interdependence of radio resource allocation and BS-user association, how to jointly deal with the resource slicing and vehicle access control to maximize the network-level resource utilization is a key research issue. The 5G system standards by 3GPP, release 15 [80] and release 16 [81], specify the split network functionalities between 5G radio access networks and core networks, such as AMF, network slice selection function (NSSF), to support network slicing, and introduce new interfaces (e.g., X_n) to facilitate inter-gNodeB resource sharing. A general description of slice characteristics is provided in terms of QoS requirements (e.g., bit rate and latency) for service differentiation. Although the network slicing concept has been proposed in 3GPP standards, and the radio resource slicing problem has been studied for both 5G wireless and core networks [31], [82], designing an optimal resource slicing framework to support diversified IoV applications is still in its infancy, as the

problem becomes much more complex in an IoV scenario with the consideration of mobility patterns of vehicles and small moving cells (e.g., covered by UAVs), resource heterogeneity, and differentiated IoV service requirements in terms of latency, reliability, and data rate. SDN-/NFV-based resource slicing frameworks are proposed for IoV in existing works [43]. Joint communication and computing resource slicing problem is studied to minimize the connection outage probability of vehicles for edge-assisted vehicular networks [83]. In [84], a resource slicing problem with content pushing and caching is studied to support different vehicular network applications in air-assisted IoV, with a focus on how much computing tasks and popular contents that can be offloaded onto ground vehicles for differentiated QoS guarantee. Further studies are needed to determine proper resource slices at heterogeneous BSs to maximize the overall resource utilization via BS-vehicle association.

B. VNF Chaining and Placement for Computation Offloading

Under the SDN-/NFV-based MEC architecture, as shown in Fig. 4, some computing tasks can be executed on network edge servers near BSs or even on individual vehicles with sufficient computing resources. As available computing resources are often unevenly distributed among network elements, offloading computation-intensive missions to the network edge for task processing improve the computation efficiency, especially for applications requiring ultralow latency for task execution. For many IoV applications, their computing tasks need to go through a sequence of VNFs for E2E traffic delivery. For instance, HD map data traffic needs to traverse data classification and data fusion functionalities before being analyzed on vehicles. Therefore, computing offloading also comes with VNF migration and corresponding computing resource allocation. For a VC network, as shown in Fig. 4, a local SDN/NFV integrated control module deals with how VNFs are chained and placed at different vehicles, with the consideration of vehicular mobility, to ensure service continuity and differentiated QoS satisfaction. The coordination between a central SDN/NFV controller and local SDN/NFV controllers also needs to be considered for VNF migration when vehicles move among different BS coverages. For IoV scenarios, computation offloading with VNF chain orchestration is a challenging research issue, where the VNF chaining and placement and resource allocation on the edge servers and vehicles depend on dynamic resource availability, function provisioning cost, and vehicle mobility. Preliminary studies on VNF placement and resource allocation include a resource-oriented VNF scheduling based on traffic classification [85] and VNF chain mapping and resource provisioning to maximize the overall service acceptance ratio [86]. A key issue in jointly determining computation offloading and VNF placement is how to maximize overall resource utilization, under the constraints of VNF offloading and service provisioning costs.

C. Joint Communication, Computing, and Caching Resource Allocation

With the emergence of edge computing and edge caching, future vehicular networks are expected to integrate communication, computing, and caching resources to support a variety of IoV applications. The SDN/NFV framework integrated with MEC and edge caching provides centralized network control over different types of resources, which can improve the efficiency of multiresource orchestration. However, diverse demands for the multidimensional resources from end devices pose technical challenges over conventional model-based network optimization approaches [87], [88]. First, the model-based methods require a large amount of network state information obtained from the underlying physical vehicular networks in the 5G (and B5G) era, which increases the complexity of resource allocation and computation overhead at SDN/NFV control modules. Second, the heterogeneity and interdependence of resource demands make it difficult to determine the tradeoff among multidimensional resources. For example, a computing task request sent from a vehicle to an edge server consumes a combination of radio bandwidth resources for wireless transmission and computing resources on the edge server for task execution. Requesting a cached video content in a BS needs transcoding before being fetched by an end vehicle, which consumes a sequence of caching, computing, and communication resources. Modeling the relation among multiple types of resource consumption is essential, leading to complex resource allocation problems. To overcome the complexity, an alternative approach is RL-based methods to learn an efficient multidimensional resource allocation policy by including the present action's impact on future network states, to achieve long-term accumulated maximal rewards [89]. The features of available network state information (e.g., user demands, vehicle mobility, and network service load) can be extracted and predicted to some extent, which can facilitate the RL module design to improve the learning convergence and accuracy.

D. Hierarchical SDN/NFV Controller Deployment

The placement of SDN/NFV control modules is of paramount importance to improve the communication, computing, and caching performance in terms of optimizing the multiresource allocation with minimal controller deployment cost and control message exchange overhead. One way to deal with scalability issue of a centralized controller is to have a hierarchical SDN/NFV controller architecture. Local controllers are placed at each MEC server connecting to a set of BSs, and MEC servers are managed by a higher level central controller placed at a cloud server or a data center node through backhaul and core network connections, as shown in Fig. 4. This controller architecture is promising in distributing the overall controlling tasks at different levels of control modules to reduce the implementation complexity with increased

efficiency and accuracy and, at the same time, to reduce signaling delays to/from vehicles. A local controller is in charge of radio resource slicing among its connected BSs, vehicle access control and traffic routing among vehicles, and computation offloading with VNF chain orchestration among end vehicles. From the coordination with local controllers, a central controller determines the policy updates for computing task migration among neighboring edge servers, VNF placement with edge-cloud computation offloading, and content caching placement. Many existing studies present different conceptual approaches for the controller architectural design in different IoV networking scenarios (e.g., MEC-assisted hierarchical control for VC grouping [4]). Several important research aspects need further investigation for the controller deployment.

1) Given the large scale of a physical vehicular network, we need to determine the minimum number of local/central controllers, their location placement and networking topology, and association patterns between each local controller and its connected BSs to achieve the optimal allocation of multiple types of resources. The solution depends on the types, amount, and update frequency of collected network state information by each controller, and the cost of bidirectional information exchange between each control module and underlying network elements.
2) An efficient networking protocol is required to interconnect the control modules for control-plane message exchanges, which affects the E2E QoS performance of supported IoV applications.

To justify the proposed SDN/NFV-enabled IoV architecture, it is necessary to build a real test bed to conduct experiments in terms of performance evaluation of different resource allocation algorithms/schemes for IoV. MEC servers with both computing and caching capabilities can be deployed to physically connect to cellular, Wi-Fi, or TVWS BSs, and the SDN/NFV control modules can be realized in edge and cloud servers to provide centralized hierarchical control over different types of resources (for communication, computing, and caching). Some high-performance deep-learning servers are expected to connect to BSs, for data traffic trace training and RL algorithm implementation to make learning decisions in terms of multidimensional resource allocation. The interworking among the deployed network elements for synchronization, algorithm/scheme implementation, and hierarchical control needs to be properly specified to achieve system-level stability.

VIII. CONCLUSION

SDN/NFV technologies are a key enabler of future IoV systems in terms of multidimensional resource allocation and service provisioning. This article introduces the emerging trend of enhancing IoV with SDN/NFV technologies. According to various QoS requirements of IoV applications and services, three important roles of the SDN/NFV networking framework in IoV are described and surveyed, respectively, for enhancing the communication, computing, and caching capabilities. Network softwarization, resource slicing, and VNF orchestration can improve service delivery, reliability, and enhance ubiquitous and efficient connectivity. On the other hand, MEC and NFV technologies can be utilized to offload IoV computing tasks to edge and cloud servers, and SDN can be integrated to configure routing paths for data offloading and computing task migration. Furthermore, the SDN-/NFV-based framework can facilitate adaptive caching deployment, cached content update, and dissemination. To achieve the SDN/NFV/MEC potentials for IoV applications, there are many open research issues, including joint resource slicing and access control, VNF chaining and placement for computation offloading, joint multidimensional resource orchestration, and hierarchical SDN/NFV controller deployment. ■

REFERENCES

[1] *ITU Towards IMT for 2020 and Beyond*. Accessed: May 31, 2019. [Online]. Available: https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx

[2] *The Top 5G Use Cases*. Accessed: May 31, 2019. [Online]. Available: https://www.sdxcentral.com/5g/definitions/top-5g-use-cases/

[3] N. Cheng *et al.*, "Big data driven vehicular networks," *IEEE Netw.*, vol. 32, no. 6, pp. 160–167, Nov./Dec. 2018.

[4] 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.

[5] K. Wang, H. Yin, W. Quan, and G. Min, "Enabling collaborative edge computing for software defined vehicular networks," *IEEE Netw.*, vol. 32, no. 5, pp. 112–117, Sep./Oct. 2018.

[6] X. Wang, Z. Ning, and L. Wang, "Offloading in Internet of vehicles: A fog-enabled real-time traffic management system," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4568–4578, Oct. 2018.

[7] X. Ge, Z. Li, and S. Li, "5G software defined vehicular networks," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 87–93, Jul. 2017.

[8] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May 2015.

[9] H. T. Cheng, H. Shan, and W. Zhuang, "Infotainment and road safety service support in vehicular networking: From a communication perspective," *Mech. Syst. Signal Process.*, vol. 25, no. 6, pp. 2020–2038, Aug. 2011.

[10] *IEEE Standard for Information Technology—Telecommunications and Information Exchange Between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band*, IEEE Standard 802.11-2016, 2016.

[11] *Study on LTE Support for Vehicle to Everything (V2X) Services*, document TS 22.885, 3GPP, 2016.

[12] *Architecture Enhancements for V2X Services*, document TS 23.285, 3GPP, 2017.

[13] A. Bazzi, B. M. Masini, A. Zanella, and I. Thibault, "On the performance of IEEE 802.11p and LTE-V2V for the cooperative awareness of connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10419–10432, Nov. 2017.

[14] 5GAA. (2018). *V2X Technology Benchmark Testing*. [Online]. Available: https://www.fcc.gov/ecfs/filing/109271050222769

[15] M. Li, L. Zhao, and H. Liang, "An SMDP-based prioritized channel allocation scheme in cognitive enabled vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7925–7933, Sep. 2017.

[16] G. S. Aujla, R. Chaudhary, N. Kumar, J. J. Rodrigues, and A. Vinel, "Data offloading in 5G-enabled software-defined vehicular networks: A stackelberg-game-based approach," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 100–108, 2017.

[17] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850–3860, Jun. 2018.

[18] N. Cheng *et al.*, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.

[19] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[20] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, and L. Wang, "Power control in UAV-supported ultra dense networks: Communications, caching, and energy transfer," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 28–34, Jun. 2018.

[21] I. Jang, S. Choo, M. Kim, S. Pack, and G. Dan, "The software-defined vehicular cloud: A new level of sharing the road," *IEEE Veh. Technol. Mag.*, vol. 12, no. 2, pp. 78–88, Jun. 2017.

[22] F. Sun *et al.*, "Cooperative task scheduling for computation offloading in vehicular cloud," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11049–11061, Nov. 2018.

[23] F. Lyu *et al.*, "Characterizing urban vehicle-to-vehicle communications for reliable safety applications," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2019.2920813.

[24] Y. Wang, Z. Su, Q. Xu, T. Yang, and N. Zhang, "A novel charging scheme for electric vehicles with smart communities in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8487–8501, Sep. 2019, doi: 10.1109/TVT.2019.2923851.

[25] SAE International. (Jun. 2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. [Online]. Available: https://www.sae.org/standards/content/j3016_201806/

[26] L. Hobert, A. Festag, I. Llatser, L. Altomare, F. Visintainer, and A. Kovacs, "Enhancements of V2X communication in support of cooperative autonomous driving," *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 64–70, Dec. 2015.

[27] H. Jiang, W. Zhuang, X. Shen, and Q. Bi, "Quality-of-service provisioning and efficient resource utilization in CDMA cellular communications," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 1, pp. 4–15, Jan. 2006.

[28] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, Mar. 2015.

[29] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1617–1634, 3rd Quart., 2014.

[30] N. McKeown *et al.*, "OpenFlow: Enabling innovation in campus networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Apr. 2008.

[31] Q. Ye, W. Zhuang, S. Zhang, A-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[32] F. Bari, S. R. Chowdhury, R. Ahmed, R. Boutaba, and O. C. M. B. Duarte, "Orchestrating virtualized network functions," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 4, pp. 725–739, Dec. 2016.

[33] H. A. Omar, N. Lu, and W. Zhuang, "Wireless access technologies for vehicular network safety applications," *IEEE Netw.*, vol. 30, no. 4, pp. 22–26, Jul./Aug. 2016.

[34] D. Kreutz, F. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[35] *Service Requirements for Enhanced V2X Scenarios, Release 15*, document TS 22.186 V15.3.0, 3GPP, Jul. 2018, pp. 1–17.

[36] J. He, L. Cai, P. Cheng, and J. Pan, "Delay minimization for data dissemination in large-scale VANETs with buses and taxis," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 1939–1950, Aug. 2016.

[37] G. Zhang *et al.*, "Multicast capacity for vanets with directional antenna and delay constraint," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 818–833, Apr. 2012.

[38] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1937, Jun. 2005.

[39] P. Wang and W. Zhuang, "A token-based scheduling scheme for WLANs supporting voice/data traffic and its performance analysis," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1708–1718, May 2008.

[40] H. A. Omar, W. Zhuang, and L. Li, "VeMAC: A TDMA-based MAC protocol for reliable broadcast in VANETs," *IEEE Trans. Mobile Comput.*, vol. 12, no. 9, pp. 1724–1736, Sep. 2013.

[41] M. Hadded, P. Muhlethaler, A. Laouiti, R. Zagrouba, and L. A. Saidane, "TDMA-based MAC protocols for vehicular ad hoc networks: A survey, qualitative analysis, and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2461–2492, 4th Quart., 2015.

[42] X. Jiang and D. H. Du, "PTMAC: A prediction-based TDMA MAC protocol for reducing packet collisions in VANET," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9209–9223, Nov. 2016.

[43] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 38–45, Dec. 2017.

[44] F. Bai, D. D. Stancil, and H. Krishnan, "Toward understanding characteristics of dedicated short range communications (DSRC) from a perspective of vehicular network engineers," in *Proc. ACM MobiCOM*, Sep. 2010, pp. 329–340.

[45] W. Alasmary and W. Zhuang, "Mobility impact in IEEE 802.11 p infrastructureless vehicular networks," *Ad Hoc Netw.*, vol. 10, no. 2, pp. 222–230, 2012.

[46] M. Torrent-Moreno, J. Mittag, P. Santi, and H. Hartenstein, "Vehicle-to-vehicle communication: Fair transmit power control for safety-critical information," *IEEE Trans. Veh. Technol.*, vol. 58, no. 7, pp. 3684–3703, Sep. 2009.

[47] F. Lyu *et al.*, "DBCC: Leveraging link perception for distributed beacon congestion control in VANETs," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4237–4249, Dec. 2018.

[48] S. Bharati and W. Zhuang, "CAH-MAC: Cooperative ADHOC MAC for vehicular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 470–479, Sep. 2013.

[49] N. Wisitpongphan, O. K. Tonguz, J. S. Parikh, P. Mudalige, F. Bai, and V. Sadekar, "Broadcast storm mitigation techniques in vehicular ad hoc networks," *IEEE Wireless Commun.*, vol. 14, no. 6, pp. 84–94, Dec. 2007.

[50] J. Liu, J. Wan, B. Zeng, Q. Wang, H. Song, and M. Qiu, "A scalable and quick-response software defined vehicular network assisted by mobile edge computing," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 94–100, Jul. 2017.

[51] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of DSRC and cellular network technologies for V2X communications: A survey," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9457–9470, Dec. 2016.

[52] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[53] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[54] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101–109, Jul. 2017.

[55] *System Architecture for the 5G System (Release 15)*, document TS 23.501, 3rd Generation Partnership Project, Sophia Antipolis, France, 2017.

[56] Z. Su, Y. Hui, and T. H. Luan, "Distributed task allocation to enable collaborative autonomous driving with network softwarization," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2175–2189, Oct. 2018.

[57] H. Peng, Q. Ye, and X. Shen, "Spectrum management for multi-access edge computing in autonomous vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2019.2922656.

[58] R. Bruschi, F. Davoli, P. Lago, and J. F. Pajo, "A multi-clustering approach to scale distributed tenant networks for mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 499–514, Mar. 2019.

[59] *Multi-Access Edge Computing (MEC); Study on MEC Support for V2X Use Cases*, Standard ETSI GR MEC 022 V2.1.1, European Telecommunications Standards Institute, 2018. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/MEC/001_099/022/02.01.01_60/gr_MEC022v020101p.pdf

[60] K. Kaur, S. Garg, G. S. Aujla, N. Kumar, J. J. P. C. Rodrigues, and M. Guizani, "Edge computing in the industrial Internet of Things environment: Software-defined-networks-based edge-cloud interplay," *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 44–51, Feb. 2018.

[61] Y. Kim, N. An, J. Park, and H. Lim, "Mobility support for vehicular cloud radio-access-networks with edge computing," in *Proc. IEEE CloudNet*, Oct. 2018, pp. 1–4.

[62] S. Choo, J. Kim, and S. Pack, "Optimal task offloading and resource allocation in software-defined vehicular edge computing," in *Proc. IEEE ICTC*, Oct. 2018, pp. 251–256.

[63] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, "Adaptive transmission optimization in SDN-based industrial Internet of things with edge computing," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1351–1360, Jun. 2018.

[64] C. Huang, M. Chiang, D. Dao, W. Su, S. Xu, and H. Zhou, "V2V data offloading for cellular network based on the software defined network (SDN) inside mobile edge computing (MEC) architecture," *IEEE Access*, vol. 6, pp. 17741–17755, 2018.

[65] 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.

[66] F. Lyu *et al.*, "Demystifying traffic statistics for edge cache deployment in large-scale WiFi system," in *Proc. IEEE ICDCS*, Jul. 2019, pp. 965–975.

[67] G. Ma, Z. Wang, M. Zhang, J. Ye, M. Chen, and W. Zhu, "Understanding performance of edge content caching for mobile video streaming," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1076–1089, May 2017.

[68] H. Y. Huang *et al.*, "Performance evaluation of SUVnet with real-time traffic data," *IEEE Trans. Veh. Technol.*, vol. 56, no. 6, pp. 3381–3396, Nov. 2007.

[69] Y. Zhang, C. Li, T. H. Luan, Y. Fu, W. Shi, and L. Zhu, "A mobility-aware vehicular caching scheme in content centric networks: Model and optimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3100–3112, Apr. 2019.

[70] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. M. Ni, "Recognizing exponential inter-contact time in VANETs," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–5.

[71] A. Abdrabou and W. Zhuang, "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 129–139, Jan. 2011.

[72] P. Hao, L. Hu, J. Jiang, J. Hu, and X. Che, "Mobile edge provision with flexible deployment," *IEEE Trans. Serv. Comput.*, vol. 12, no. 5, pp. 750–761, Sep./Oct. 2019, doi: 10.1109/TSC.2018.2842227.

[73] C. Liang, Y. He, F. R. Yu, and N. Zhao, "Enhancing QoE-aware wireless edge caching with software-defined wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6912–6925, Oct. 2017.

[74] J. Chen *et al.*, "An SDN-based transmission protocol with in-path packet caching and retransmission," in *Proc. IEEE ICC*, May 2019,

pp. 1–6.
[75] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. S. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.
[76] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.
[77] Z. Chang, L. Lei, Z. Zhou, S. Mao, and T. Ristaniemi, "Learn to cache: Machine learning for network edge caching in the big data era," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 28–35, Jun. 2018.
[78] T. Trzciński and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2561–2570, Nov. 2017.
[79] 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.
[80] *Technical Specification Group Services and System Aspects; Release 15 Description; Summary of Rel-15 Work Items (Release 15)*, document TR 21.915 V2.0.0, 3rd Generation Partnership Project, 3GPP, Sep. 2019, pp. 1–118.
[81] *Technical Specification Group Services and System Aspects; Release 16 Description; Summary of Rel-16 Work Items (Release 16)*, document TR 21.916 V0.1.0, 3rd Generation Partnership Project, Sep. 2019, pp. 1–23.
[82] Q. Ye, J. Li, K. Qu, W. Zhuang, X. S. Shen, and X. Li, "End-to-end quality of service in 5G networks: Examining the effectiveness of a network slicing framework," *IEEE Trans. Veh. Technol.*, vol. 13, no. 2, pp. 65–74, Jun. 2018.
[83] K. Xiong, S. Leng, J. Hu, X. Chen, and K. Yang, "Smart network slicing for vehicular fog-RANs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3075–3085, Apr. 2019.
[84] S. Zhang, W. Quan, J. Li, W. Shi, P. Yang, and X. Shen, "Air-ground integrated vehicular network slicing with content pushing and caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2114–2127, Sep. 2018.
[85] J. Wang, B. He, J. Wang, and T. Li, "Intelligent VNFs selection based on traffic identification in vehicular cloud networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4140–4147, May 2019.
[86] J. Wang, Q. Qi, S. Qing, and J. Liao, "Elastic vehicular resource providing based on service function-group resource mapping of smart identify network," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1897–1908, Jun. 2018.
[87] C. Wang, Y. He, F. Yu, Q. Chen, and L. Tang, "Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 7–38, 1st Quart., 2017.
[88] Q. Chen, F. R. Yu, T. Huang, R. Xie, J. Liu, and Y. Liu, "Joint resource allocation for software-defined networking, caching, and computing," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 274–287, Feb. 2018.
[89] T. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

ABOUT THE AUTHORS

Weihua Zhuang (Fellow, IEEE) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, since 1993, where she is currently a Professor and a Tier I Canada Research Chair in Wireless Communication Networks.

Dr. Zhuang is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. She is also an elected member of the Board of Governors and VP Publications of the IEEE Vehicular Technology Society. She was a recipient of the 2017 Technical Recognition Award from the IEEE Communications Society Ad Hoc & Sensor Networks Technical Committee and a co-recipient of several best paper awards from IEEE conferences. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2013, the Technical Program Chair/Co-Chair of the IEEE Vehicular Technology Conference (VTC) Fall 2017 and Fall 2016, and the Technical Program Symposia Chair of the IEEE Global Communications Conference (Globecom) 2011.

Feng Lyu (Member, IEEE) received the B.S. degree in software engineering from Central South University, Changsha, China, in 2013, and the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018.

Since September 2018, he has been a Postdoctoral Fellow with the BBCR Group, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His current research interests include vehicular *ad hoc* networks, space–air–ground integrated networks, cloud/edge computing, and big-data-driven application design.

Dr. Lyu is a member of the IEEE Computer Society and the IEEE Communications Society.

Qiang Ye (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2016.

He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, from December 2016 to November 2018, where he was a Research Associate from December 2018 to September 2019. He has been an Assistant Professor with the Department of Electrical and Computer Engineering and Technology, Minnesota State University, Mankato, MN, USA, since September 2019. His current research interests include 5G networks, software-defined networking and network function virtualization, network slicing, artificial intelligence and machine learning for future networking, protocol design, and end-to-end performance analysis for the Internet of Things.

Nan Cheng (Member, IEEE) received the B.E. and M.S. degrees from the Department of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, in 2016.

He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, from 2017 to 2018. He is currently a Professor with the School of Telecommunication Engineering, Xidian University, Shaanxi, China. His current research focuses on space–air–ground integrated systems, big data in vehicular networks, and self-driving systems. His current research interests include performance analysis, medium-access control, opportunistic communication, the application of artificial intelligence for vehicular networks, and space–air–ground integrated networks.

Ju Ren (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2009, 2012, and 2016, respectively.

From 2013 to 2015, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently a Professor with the School of Computer Science and Engineering, Central South University. His current research interests include the Internet of Things, wireless communication, network computing, and cloud computing.

Dr. Ren is a member of the Association for Computing Machinery (ACM). He has been a TPC Member of many international conferences, including the IEEE International Conference on Computer Communications (INFOCOM) 2018/2019 and Globecom 2017. He was a co-recipient of the Best Paper Award of the IEEE Internet of People (IoP) 2018 and the Most Popular Paper Award of the *Chinese Journal of Electronics* from 2015 to 2018. He has been an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and *Peer-to-Peer Networking and Applications*. He has served as the TPC Chair of the IEEE International Conference on Big Data Science and Engineering (BigDataSE) 2019, the Poster Co-Chair of the IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS) 2018, the Track Co-Chair of IEEE VTC 2017 Fall, and an active reviewer for over 20 international journals.