Service Migration in Fog Computing Enabled Cellular Networks to Support Real-Time Vehicular Communications

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 18, 2018, accepted January 13, 2019, date of publication January 21, 2019, date of current version February 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2893571

Service Migration in Fog Computing Enabled Cellular Networks to Support Real-Time Vehicular Communications

JUN LI[1], XIAOMAN SHEN[1,2], LEI CHEN[3], DUNG PHAM VAN[4], JIANNAN OU[5], LENA WOSINSKA[6], (Senior Member, IEEE), AND JIAJIA CHEN[1], (Senior Member, IEEE)

[1]Optical Networks Laboratory, KTH Royal Institute of Technology, 16440 Stockholm, Sweden
[2]Centre for Optical and Electromagnetic Research, Zhejiang University, Hangzhou 310007, China
[3]RISE Viktoria, 41756 Gothenburg, Sweden
[4]Ericsson, 187 66 Täby, Sweden
[5]MOE International Laboratory for Optical Information Technologies, South China Academy of Advanced Optoelectronics, South China Normal University, Guangzhou 511400, China
[6]Department of Electrical Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden

Corresponding author: Jiajia Chen (jiajiac@kth.se)

This work was supported in part by the Göran Gustafssons Stiftelse, in part by the Natural Science Foundation of Guangdong Province under Grant 508206351021, and in part by the National Natural Science Foundation of China under Grant 61550110240 and Grant 61671212.

ABSTRACT Driven by the increasing number of connected vehicles and related services, powerful communication and computation capabilities are needed for vehicular communications, especially for real-time and safety-related applications. A cellular network consists of radio access technologies, including the current long-term evolution (LTE), the LTE advanced, and the forthcoming 5th generation mobile communication systems. It covers large areas and has the ability to provide high data rate and low latency communication services to mobile users. It is considered the most promising access technology to support real-time vehicular communications. Meanwhile, fog is an emerging architecture for computing, storage, and networking, in which fog nodes can be deployed at base stations to deliver cloud services close to vehicular users. In fog computing-enabled cellular networks, mobility is one of the most critical challenges for vehicular communications to maintain the service continuity and to satisfy the stringent service requirements, especially when the computing and storage resources are limited at the fog nodes. Service migration, relocating services from one fog server to another in a dynamic manner, has been proposed as an effective solution to the mobility problem. To support service migration, both computation and communication techniques need to be considered. Given the importance of protocol design to support the mobility of the vehicles and maintain high network performance, in this paper, we investigate the service migration in the fog computing-enabled cellular networks. We propose a quality-of-service aware scheme based on the existing handover procedures to support the real-time vehicular services. A case study based on a realistic vehicle mobility pattern for Luxembourg scenario is carried out, where the proposed scheme, as well as the benchmarks, are compared by analyzing latency and reliability as well as migration cost.

INDEX TERMS Connected vehicles, fog computing, service migration.

I. INTRODUCTION

Wireless connectivity is becoming an important feature of modern vehicles to enhance situational awareness, providing an information-rich travel environment. It extends the vision of vehicle perception systems beyond the line-of-sight and can potentially overcome many difficulties faced by traditional sensors based on radar and camera. It was recently reported that a tragic accident happened when a driver of Tesla was using autopilot mode [1], due to a misdetection occurred in the local sensing system. Such a fatal accident could have been avoided, if timely communications with other road users would be enabled in the vehicles, assisting the vehicle sensors for accurate road and traffic information. Thanks to ubiquitous availability as well as capability to offer high data rate and low latency communication, cellular networks are promising to support the vehicle connectivity.

In view of this, Vehicle-to-Everything (V2X), including Vehicle-to-Vehicle (V2V), Vehicle-to-Pedestrian (V2P), and Vehicle-to-Infrastructure/Network (V2I/N), has been standardized by the 3rd Generation Partnership Project (3GPP). An initial solution is based on the current Long Term Evolution (LTE) networks and included in 3GPP Rel-14 [2]. Such a LTE-V2X solution allows vehicles and infrastructure to exchange information of local sensors and cameras, thus enabling a global environmental perception for vehicles. This helps the traffic system to achieve higher safety, efficiency, and comfort. The cellular networks are evolving, so are the V2X solutions. The 5th Generation (5G) V2X has been under investigation for the enhanced vehicular services. In 3GPP Rel-15, the LTE platform has been extended to meet the evolving requirements of the automotive industry. These enhancements are driven by many new use cases identified for the advanced V2X services, such as vehicles platooning, extended sensors, advanced autonomous driving, and remote driving, with ultra-high reliability and low latency requirements [3]. In 3GPP Rel-16, potential architecture enhancements of 5G systems have been identified and evaluated to support the above advanced V2X services [4].

Meanwhile, many new data-driven applications and technologies related to traffic safety and efficiency (such as augmented reality techniques, intelligent transport) have been developed. The amount of data that is generated and needs to be analyzed by a vehicle is increasing sharply, reaching one gigabyte per second and even more for some traffic safety applications with video recording [5]. Due to the limited computational resources, data processing capability in the vehicles may not always satisfy the stringent delay requirement of real-time services. One of the existing solutions is to transfer the computation tasks from user equipment to a cloud, called computation offloading. The purpose is to make a full use of powerful computational capability in remote datacenters [6], [7]. However, in such cases, the computation resources are centralized and typically located in a large-scale datacenter (for backup or other purpose, one or more additional datacenters may be needed), which are not always close to the end users. Therefore, transmitting data from the vehicles to a centralized cloud suffers disadvantages in terms of communication latency. It is thus not suitable for the safety-related services, such as critical-event warning that requires response time of less than 10 milliseconds [2], [3]. Fog computing, which was initialized by Cisco as a new computing paradigm [8], is envisioned to support the real-time services. In contrast to the centralized cloud computing, the core idea of the fog computing is to distribute computational resources as close as possible to end users, allowing data to be processed in close proximity to the vehicles and roadside sensors/units. It is worth mentioning that, in addition to the fog computing, the idea of moving cloud servers to the network edge is also referred to as cloudlet and Mobile-Edge Computing (MEC) [9]. Such a distributed cloud paradigm enables low latency and enhances service availability and reachability. Therefore, deploying fog servers at base stations such as LTE's evolved Nodes B (eNBs), henceforth referred to as fog-enabled cellular networks, can be a promising solution to enhance the real-time services for connected vehicles [10].

One significant challenge in the fog-enabled cellular networks is mobility as the vehicles traverse different cells with high speeds. Therefore, in addition to the conventional cellular handover transferring users' connectivity between cells [11], service migration is needed to maintain service continuity. It means that the mobility must be properly handled to guarantee both low latency and high reliability, particularly for the real-time services. To minimize the negative impact on Quality of Service (QoS), both computation (e.g., computing architecture and virtualization techniques) and communication (not only the one between the vehicles and access points but also the one among the access points) have to be taken into account. In the context of the MEC or fog computing, active service applications are encapsulated in Virtual Machines (VMs) or containers. There have been several studies that deal with the mobility problem in the MEC or fog computing [12]–[17]. In [12], general architectural components supporting VM migration and interactions among such components are defined and discussed. In [13] a general layered framework is proposed, which allows the migrated applications to be decomposed into multiple layers. In such a framework, only the layers missing at the destination need to be transferred, thus reducing a big amount of data to be handled during the service migration. A VM handoff mechanism for the service migration is proposed in [14], in which the migration files are compressed before being migrated to adaptively reduce the total migration time. Huang *et al.* [15] and Wu *et al.* [16] focused on the mobility pattern of edge computation devices and developed a cost model for the service migration using a Markov decision process based approach. In [17], a time window based service migration is proposed to search the optimal service placement sequence. However, most of the existing studies, e.g. [15]–[17], are based on abstract models, and do not reflect the real situation, where many parameters need to be optimized. Furthermore, since in the service migration data needs to be transferred via the communication infrastructure, communication protocols (such as the ones for handover) and strategies to handle the service migration have to be considered.

In this regard, we investigate the service migration in the fog computing enabled cellular networks to support the real-time vehicular communication. More specifically, the main contributions of this paper include: 1) A framework of the fog-enabled cellular networks for the connected vehicles is introduced, 2) A QoS aware scheme enhancing the existing handover procedure and realizing information exchange for the service migration is proposed, and 3) A performance study of the proposed scheme as well as the benchmarks is carried out by simulation with a realistic traffic pattern in terms of end-to-end communication latency and reliability. Furthermore, migration cost is investigated in terms of migration frequency and migration time, providing a guideline on selecting the proper options in a given scenario.

TABLE 1. Real-time vehicular services supported by fog computing [2], [3], [19].

Applications	latency	Reliability	The amount of data for communications between vehicles and access points	Priority	Examples
Advanced driving	1ms ~20 ms	>99.999%	10Mbps	1	Autonomous driving; remote driving
Efficient driving	<100ms	90%~99.999%	1Mbps~25Mbps	2	Road sign notification; automated parking; real-time navigation
Infotainment	<100ms	Not concern	0.5Mbps~15Mbps	3	Augmented reality game, local advertisement

Note: Level of priority: 1 corresponds to the highest and 3 corresponds to the lowest priority.

The remainder of the article is organized as follows. First, Section II provides a high-level view of a fog-enabled V2X architecture and the envisioned applications that can be supported. In Section III three schemes to handle the service migration in the presented fog-enabled V2X architecture are introduced. These schemes extend the existing handover procedure to support information exchange for properly accessing the service when user mobility occurs. Then, in Section IV a performance assessment in terms of delay and reliability, as well as migration cost of all three investigated schemes is carried out. Finally, in Section V the conclusions are drawn and the relevant future research directions are identified.

II. FOG-ENABLED CELLULAR NETWORKS FOR CONNECTED VEHICLES

Fig. 1(a) presents a high-level view of the fog-enabled cellular networks for the connected vehicles. As illustrated, the decentralized fog and the centralized cloud co-exist and are complementary to each other to support different kinds of vehicular services. Fig. 1(b) categorizes various vehicular applications that fit either the centralized cloud in the remote data center or the distributed fog close to the users [2], [18]. The applications that are better hosted in the cloud are mainly used for the service management, which needs a global view of traffic information (e.g., traffic management), while the fog computing is responsible for the real-time vehicular services, whose characteristics are summarized in Table 1 [2], [3], [19]. For example, in an intelligent traffic system, optimal routes can be calculated by an application in the cloud, while collision avoidance at intersections can be supported by the services running in the fogs.

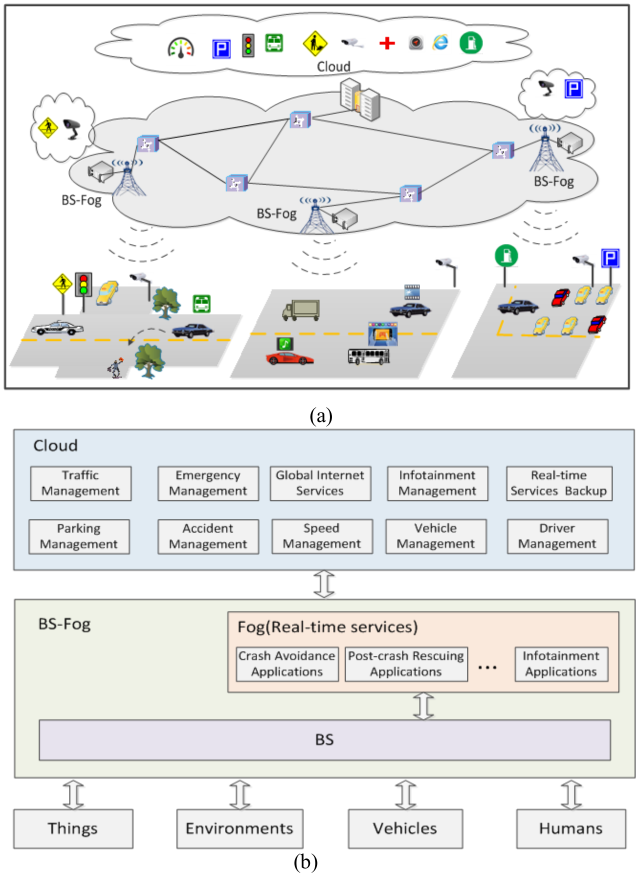


FIGURE 1. (a) A high-level view of the fog-enabled cellular network for the connected vehicles, (b) various vehicular applications that fit the cloud and the fog [2], [18].

For the real-time services, the fog servers that provide the required computing and storage resources should be deployed as close as possible to the mobile users. In the fog-enabled cellular network, base stations (For example, eNB in LTE) can be a good location for the fog servers, allowing only one-hop communication (between the user equipment and the base station) to access the services [8], [10]. Such a Base Station (BS) can be referred to as BS-Fog as shown in Fig. 1(a). The BS-Fog is an integrated entity, in which the BS is responsible for functions of the cellular networks, such as handovers, whereas the fog provides computation and storage capability locally. One BS-Fog can cooperate with other BS-Fogs or the cloud to allocate tasks dynamically. The scheduling can be performed by centralized controllers, where the global information about availability of resources (e.g., bandwidth, CPU, storage) is utilized and the applications can be migrated from the source BS-Fog node to the target BS-Fog node through mobile backhaul network with the guaranteed quality of service. In the 3GPP standards, two interfaces are defined at each BS, namely S1 and X2 [20]. S1 is assigned for the communications between BSs and the central aggregation switch in the mobile core network. This interface can be used for the communications between the BS-Fog and cloud in fog-enabled cellular based V2X solution. X2 is a logical interface for direct information exchanges between the BSs, which can be used for the communications between the BS-Fogs.

In the fog-enabled cellular-based V2X solution, when the vehicles move between the areas covered by two different BSs, apart from the handover, the service migration in many cases is needed to keep the ongoing services running at the closest BS-Fog in order to meet the QoS requirements in terms of latency. However, the service migration cannot always be completed immediately, which may lead to loss of service access or degraded QoS. Therefore, effective mechanisms are needed to alleviate such problems.

III. SERVICE MIGRATION SCHEMES

In this section, we study the various strategies to handle the service migration in the fog-enabled cellular networks for the connected vehicles and analyze the profile of end-to-end latency (D) of the vehicular traffic. Here, we consider the end-to-end latency in upstream (the end-to-end delay in downstream can be derived in a similar way), which is referred to as the total time experienced by a packet from the moment when it is sent from the User Equipment (UE) to the moment when it is finally received by the BS-Fog that hosts the offered service. Given that a vehicle is traveling while accessing a fog server, the end-to-end latency of a packet generated by the vehicle is composed of several components: wireless access delay (D_w), interruption time (D_h) during the handover, migration time (D_m), backhaul delay (D_b), and processing and queuing delays at the BS-Fogs (D_p), which are explained in Table 2. To simplify the delay analysis and without loss of generality, we assume that each fog server is associated with only one BS, and the computation and storage at the BS-Fogs are sufficient. Therefore, the processing and queuing delays are negligible with respect to other delay components. In addition, for the purpose of estimating the delay trend, the wireless access delay D_w and interruption time D_h are assumed constant since they are not affected by a migration strategy. Thus, migration time and backhaul delay are considered as the two major components affecting the final end-to-end latency of the migration schemes.

TABLE 2. Explanation of symbols.

Symbol	Explanation
D_w	Wireless access delay, which represents the time period starting from the instant when a packet is generated at UE until it is served by the associated BS, including the propagation delay in the air between the vehicle and its BS, transmission delay and queuing delay at UE.
D_h	Interruption time during the handover.
D_m	Migration time, which is referred to as the total time to transmit and resume the migration files in each migration procedure [13].
D_b	Backhaul delay, which represents the time period starting from the instant when a packet is sent from the currently associated BS until its arrival to the BS-Fog that hosts the offered service.
D_p	Processing and queuing delays at the BS-Fogs.

The migration time D_m is largely dependent on the processing time of migration files in the fog servers and transferring time in the network. Meanwhile, the backhaul delay D_b depends on the physical distance from the BS-Fog that hosts the offered service to the BS that the UE is associated with. In turn, the D_b relies on the chosen migration strategy, the fog computing capability, the backhaul link capacity, the actual network deployment, as well as the moving characteristics of the vehicle. There is a trade-off between the D_m and D_b related to the decision regarding the service migration. On the one hand, service migration helps to bring the services to the proximity of vehicles and thus maintaining a low value of the D_b once the migration is done. On the other hand, it requires a certain time to transfer service related files to the target BS-Fog. Particularly, when the migrated files are big, the D_m may become quite large. However, without the service migration, the packet needs to traverse a long distance to reach the BS-Fog that runs VMs for the service and thus incurring the large D_b when the vehicle travels away. This trade-off is related to the characteristics and requirements of the considered applications. In addition, when the QoS requirements are imposed, it is challenging to design a scheme that is able to properly handle the service migration while guaranteeing the required level of the QoS. In this article, we investigate the trade-off by studying two schemes, that is, no migration (Scheme 1) and always migration (Scheme 2), both of which have shown certain problems in supporting the QoS. In this regard, we propose a QoS aware scheme (Scheme 3) in which service migration decision is made based on whether the QoS metrics are satisfied or not.

A. SCHEME 1: NO SERVICE MIGRATION

Scheme 1 is considered as benchmark where no service migration is performed while the vehicle is moving, as illustrated in Fig. 2. To maintain the service continuity, the vehicle keeps accessing the fog server running the VM for service provisioning through the backhaul. We consider a scenario where the vehicles have a random route, e.g., private cars moving inside a city, as shown in Fig. 2(a). In such a scenario, a vehicle starts from Cell_1 to Cell_3 via Cell_2 and service subscribed by the vehicle is originally hosted by BS-Fog_1.

Fig. 2(c) shows the detailed handover procedure for the fog-enabled cellular network based on the conventional LTE handover protocol [21], in which the fog service location reporting is included. We refer to the original BS as the one that runs the VM for the services. In this procedure, the BS that the vehicle is currently associated with needs to forward the packets to the original BS, which carries the information required for computation offloading. A Handover Request message should contain information about the location of the fog services to make sure that the BS recently accessed by the vehicle is also aware of the VM location of the subscribed service. After the handover procedure, the UE can receive data from the core network via the newly associated BS-Fog,

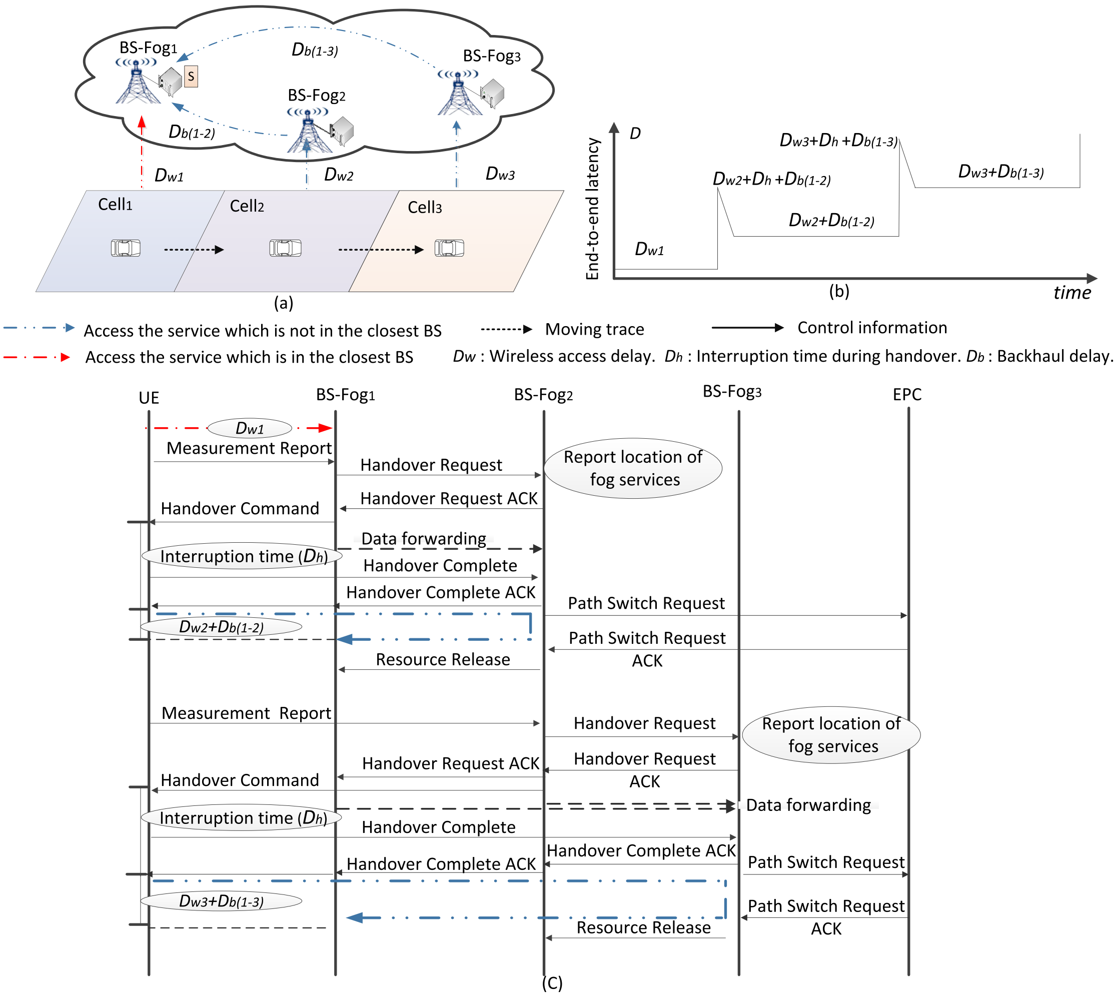


FIGURE 2. Scheme 1: (a) An example scenario: A private car with a random route, (b) the end-to-end latency profile, and (c) the corresponding communication protocol.

which sends A Resource Release message to the previously accessed BS-Fog to release the corresponding communication resources. Also, the newly associated BS-Fog should establish a connection with the BS-Fog where the service is provisioned. Then, the UE can keep accessing the services via such a newly established path. It should be noted that fog service locations are handled by the Mobility Management Entity (MME) in the Evolved Packet Core (EPC).

In such a scheme, when a car moves to Cell$_2$, the UE keeps accessing the service in BS-Fog$_1$. When the car is in Cell$_1$, the UE can access BS-Fog$_1$ directly, that is, $D = D_{w1}$, where D_{w1} is the wireless access delay for vehicles in Cell$_1$. At the start of the handover procedure, the UE has to experience a long waiting time D_h to access the services, which is a consequence of the interruption caused by the handover. Notice that handover is managed by the cellular network and the D_h may be reduced through soft handover schemes, which is beyond the scope of this paper. After the handover is completed, the UE can access the service hosted in BS-Fog$_1$ via the backhaul network. Thus, in addition to D_{w2}, the packet from the vehicle experiences $D_{b(1-2)}$ as well, which represents the delay caused by the backhaul between Cell$_1$ and Cell$_2$. Therefore, $D = D_{w2} + D_{b(1-2)}$. During the handover process, the D decreases from $D_{w2} + D_{b(1-2)} + D_h$ to $D_{w2} + D_{b(1-2)}$. It is because packets that are generated in the beginning of the handover only experience the remaining handover interruption time. Thus, for the car moving from Cell$_1$ to Cell$_2$, the end-to-end latency D increases from D_{w1} to $D_{w2} + D_{b(1-2)} + D_h$ and then decreases to $D_{w2} + D_{b(1-2)}$, as shown in Fig. 2(b).

When the car continues to move from Cell$_2$ to Cell$_3$, BS-Fog$_2$ needs to exchange information with both BS-Fog$_1$ and BS-Fog$_3$ to establish a new communication path between BS-Fog$_1$ and BS-Fog$_3$. As shown in the bottom part of Fig. 2(c), after receiving a Handover Request ACK message from BS-Fog$_3$, BS-Fog$_2$ forwards it to BS-Fog$_1$ to suspend the ongoing services. After receiving a Handover Complete message, the BS-Fog$_3$ needs to send a Handover Complete ACK message to BS-Fog$_2$. Then, BS-Fog$_2$ notifies BS-Fog$_1$ via another Handover Complete ACK message. After path switching, the UE can access the services hosted in BS-Fog$_1$. It is worth noting that in this scenario, the minimum number of the involved BS-Fogs is three since it needs at least one intermediate BS-Fog to facilitate the process. In the example shown in Fig. 2(b), BS-Fog$_2$ is such an intermediate BS-Fog. If the vehicle would travel to a new cell after Cell$_3$, then BS-Fog$_3$ would become the intermediate node.

Fig. 2(b) shows the profile of end-to-end latency for the case presented in Fig. 2(a), where BS-Fog$_1$ always hosts the services. It is obvious that the end-to-end latency becomes larger as the car moves away from BS-Fog$_1$. Note that in this work, the end-to-end latency profile is used to demonstrate that the end-to-end latency performance varies in time. It is plotted in form of a straight line due to the simplified assumptions (see Fig. 2(b)). In a realistic case, the actual delay profile may not be linear. For instance, in a cellular network, the D_w can fluctuate even in the same cell, as it depends on several factors, such as the traffic load, the wireless communication link condition, and so on. However, the end-to-end delay trend would be preserved even when the D_w is not constant.

B. SCHEME 2: SERVICE MIGRATION TRIGGERED BY HANDOVER

As discussed in Scheme 1, the end-to-end latency becomes larger when the vehicles travel away from the serving BS-Fog, particularly for the ones that do not have fixed routes. To reduce the delay, in Scheme 2 the migration is performed in combination with the handover in order to always provide one-hop access to the fog services for the UEs. As shown in Fig. 3(a), when the car moves from Cell$_1$ to Cell$_2$, the service is migrated from BS-Fog$_1$ to BS-Fog$_2$, accordingly. The corresponding protocol for the service migration triggered by handover is shown in Fig. 3(b). We consider that the service is migrated by using pre-copy technique, which is widely adopted for live VM migration [12], [13]. The migration can be divided into two phases. Firstly, the memory pages are transferred iteratively to the target BS-Fog without suspending VM. Secondly, once the sufficient memory pages are transferred, the VM is suspended at the source BS-Fog and the remaining memory pages are transferred to the target BS-Fog. The duration in which the VM is suspended is referred to as downtime (D_t), during which the services cannot be properly accessed. In Scheme 2, Handover Request messages need to be extended by containing the migration-related information including size of migrated application, categories of application, etc. The target BS-Fog makes a decision according to the request information and its available resources. The source BS-Fog executes migration after receiving a Handover Request ACK message. Otherwise, the UEs still access the source BS-Fog. After the migration is completed, the target BS-Fog sends a Resource Release message to BS-Fog$_1$ to release the corresponding computation and storage resources in the fog. Also, the target BS-Fog should update the MME with the location of the migrated services via a Path Switch Request message. Note that the target BS-Fog sends another Resource Release message to the source BS-Fog to release the corresponding communication resources after receiving a Path Switch Request ACK message.

In Scheme 2, the service migration is triggered by the handover. It means that before the service migration the UE first experiences the handover procedure. After the handover is completed but before the service is migrated, the UE still accesses the source BS-Fog. The UE has to wait during the downtime D_t before being able to access the services in the target BS-Fog. Fig. 3(c) shows the end-to-end latency profile for the example presented in Fig. 3(a) in the case where the migration time is shorter than the time the vehicle stays in the new cell (e.g., the UE can directly access the services in the closest BS except during the migration time D_m). It can be seen that after the service migration is completed, the end-to-end latency is decreased to D_{w2}. However, if the D_m is longer than the time that the UE stays in the new cell, the delay profile is different, as shown in Fig. 3(d). This happens if the size of VM is very large so that a long migration time is required or when the vehicle travels at a very high speed and hence the time that the vehicle stays in the new cell becomes very short. In such a case, the end-to-end latency cannot be minimized by the service migration (e.g., it hardly reaches D_{w2} as shown in Fig. 3(d)). This indicates that the frequent service migration is not always a good choice.

C. SCHEME 3: QOS AWARE SERVICE MIGRATION

Given that both Scheme 1 and Scheme 2 have their own drawbacks, it is desirable to design a new scheme (referred to as Scheme 3) that takes the advantages of Schemes 1 and 2 in an adaptive fashion based on the QoS requirements. The key idea of Scheme 3 is to flexibly combine the two strategies presented in the previous sections to minimize the migration overhead while maintaining the end-to-end performance at an acceptable level to satisfy the QoS requirements. For the vehicular applications with low latency and high reliability requirements, the end-to-end latency is the key QoS metric considered in the proposed migration scheme. When the end-to-end latency exceeds a maximum value, the performance of the other QoS metrics such as reliability and packet drop ratios cannot be guaranteed. In this regard, we extract the latency as the key metric to realize QoS-aware migration management. It should be noted that the threshold-based scheme does not limit its applicability to only the latency metric. The other QoS metrics can be considered by applying the similar design principle. In particular, Scheme 3 aims at avoiding the migration as long as the stable delay is acceptable. In Scheme 3, given a required end-to-end latency threshold, the scheme starts with no migration (similarly to Scheme 1) by keeping accessing the original fog server, which runs the VM for the subscribed service. Once the end-to-end latency is not tolerable anymore the scheme then performs the service migration to reduce the delay. Note here the threshold is for the stable delay, which is referred to as the delay excluding the surge during the handover interruption or downtime (see Fig. 4(c)). As shown in Fig. 4(a) and (c), when

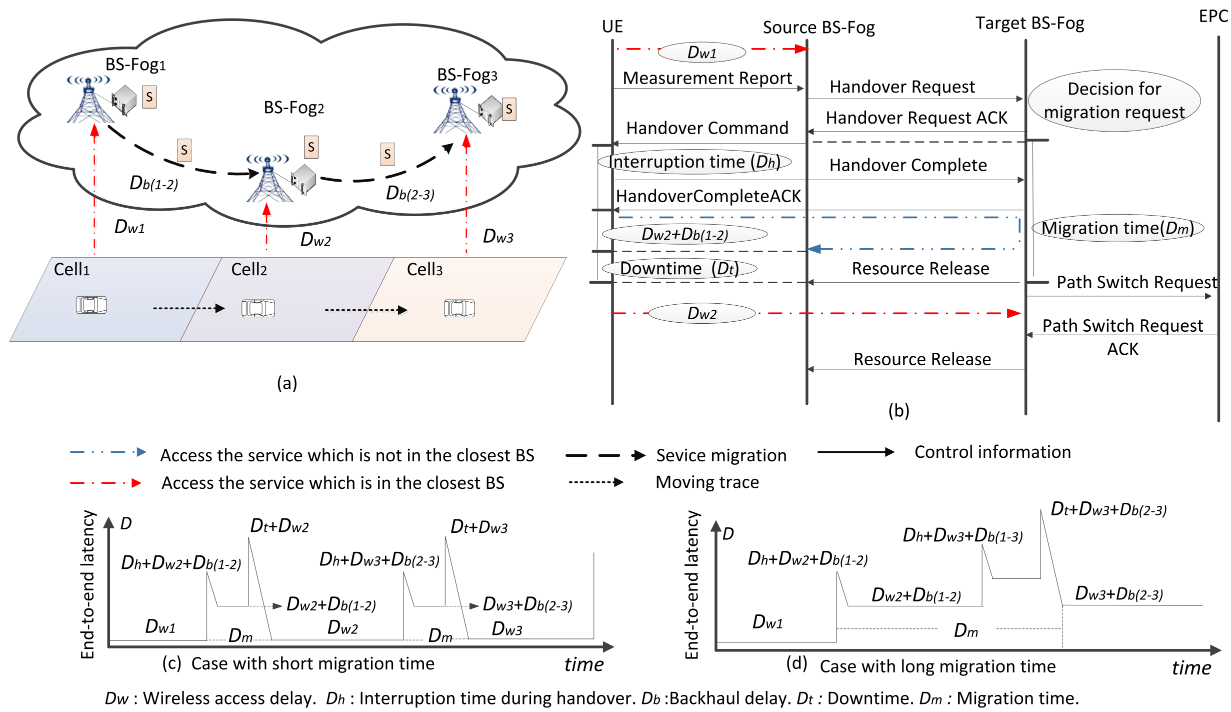


FIGURE 3. Scheme 2: (a) An example scenario: A private car with a random route, (b) the communication protocol to support the service migration triggered by the handover, (c) the end-to-end latency profile for the case with a relatively short migration time, and (d) the end-to-end latency profile for the case with a relatively long migration time.

the car is in Cell_1 or Cell_2, the UE can access the services hosted by BS-Fog_1 with the delay of D_{w1} and $D_{w2} + D_{b(1-2)}$, respectively, which are lower than the threshold of the end-to-end latency. When the vehicle moves from Cell_2 to Cell_3, the D increases to $D_{w3} + D_h + D_{b(1-3)}$ and then decreases to a stable value of $D_{w3} + D_{b(1-3)}$ after the handover is completed. The delay is however higher than the imposed delay threshold, so that the service migration to BS-Fog_3 is triggered. The service migration procedure is similar to the one considered in Scheme 2, as shown in Fig. 4(b). It should be noted that, since the migration may not be performed every time the handover takes place, extended control messages based on the handover-related messages need to be defined to embed the migration-related information.

IV. CASE STUDY AND DISCUSSION

In this section, we evaluate the performance of the three migration schemes by simulation using Urban Mobility (SUMO) [22] and Python. We use a realistic mobility pattern for the country of Luxembourg, which can be considered as a case study for a small service area [23]. As discussed in the previous section, the wireless delay and handover interruption time cannot be avoided and are not dependent on migration strategies. Therefore, we follow the requirements of Ultra-High Reliability Low Latency Communication (URLLC), where uplink delay in the wireless segment is assumed to be within 0.5 ms, and the handover interruption time is considered as a constant [24]. On the other hand, the backhaul delay and migration time are two important delay components that are distinct in different migration strategies. Passive Optical Network (PON) is widely adopted for mobile backhaul because of its energy efficiency and high capacity [25]. In the PON based mobile backhaul each BS-Fog can be associated with one Optical Network Unit (ONU), so that the traffic for both S1 and X2 is sent from the ONUs to the Optical Line Terminal (OLT) located at a central office. In order to obtain low latency, the X2 traffic can be directly handled at the central office without a need to involve the EPC [25]. In such a PON-based backhaul, transmission capacity becomes the main factor that affects delay performance, which has been pointed out in [26]. Here, the bandwidth allocation algorithm introduced in our previous work [26] is implemented for the delay analysis. The detailed parameters used in simulation are introduced in Table 3.

Fig. 5 shows the average end-to-end latency for the considered three migration schemes. All delay curves decrease as the transmission capacity in the backhaul increases, and then become saturated. That is because higher bitrate in the backhaul leads to shorter packet transmission time, which then can reduce the packet queuing delay and result in the smaller end-to-end access delay. Once the bitrate is sufficiently high (e.g., 240 Mbps in Fig. 5), the queuing time is minor and can be negligible. In such cases, the backhaul delay is mainly determined by the transmission delay and processing delay at active nodes (e.g., ONU and OLT). The end-to-end latency in Scheme 1 is more sensitive to the bitrate than in the other two

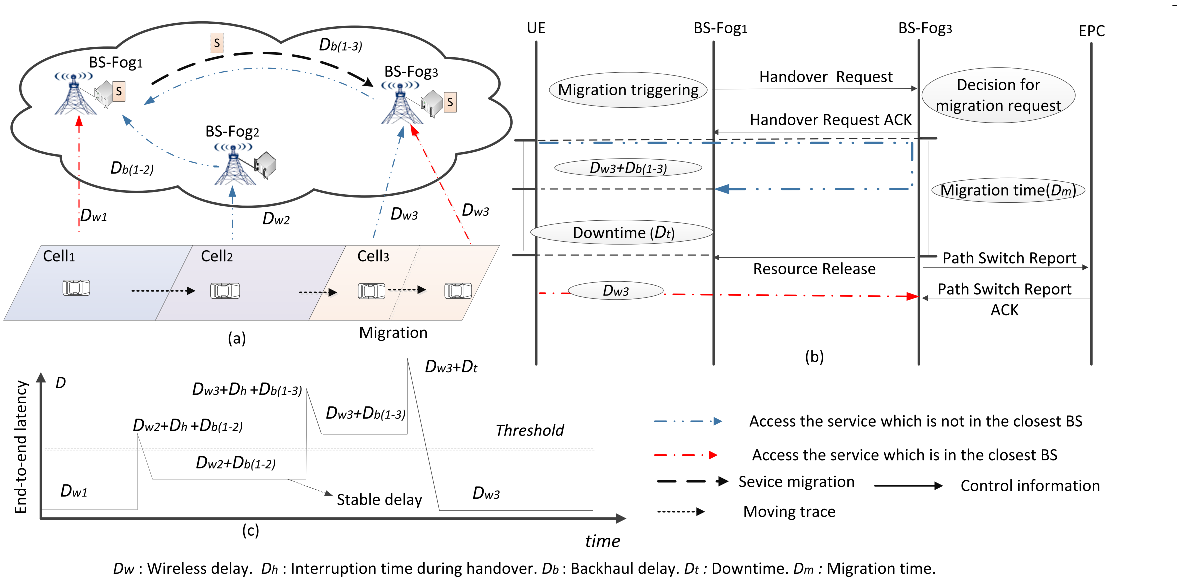


FIGURE 4. Scheme 3: (a) An example scenario: A car with a random route, (b) the communication protocol to support QoS aware service migration, and (c) the end-to-end latency profile.

TABLE 3. Parameters in simulation.

Parameter	Value
Coverage for the country of Luxembourg	155 Km^2 [14]
Total number of vehicles	5500 [14]
Bitrate of the traffic generated by the vehicles	[2Kbps, 10Mbps] [2,3,4]
Size of the applications encapsulated in VMs	[10,100] Mbits [17]
Link speed for upstream and downstream in PONs	10Gbps [17]
Repeated times of simulations	10
Simulation time	1000 s
Coverage for a single BS-Fog	1 Km^2
Handover interruption time	20 ms [12]
Speed of vehicles	[1, 45] m/s [14]
Processing time at active nodes	0.2 ms [16]
Number of ONUs in each PON	16
Number of PONs	10
Confidence level	95%

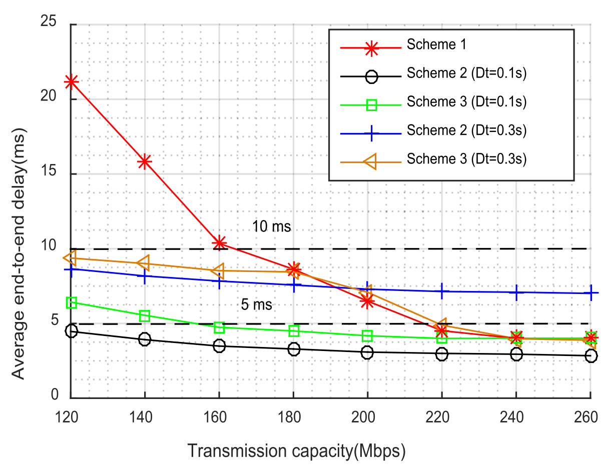


FIGURE 5. Transmission capacity in the backhaul versus average end-to-end latency.

schemes. Meanwhile, in Scheme 2, changes in the downtime (D_t) may result in a large variance of the end-to-end latency, during which the ongoing services need to be suspended. When D_t increases from 0.1 s to 0.3 s, the end-to-end latency is almost doubled regardless of the bitrate in the backhaul. In Scheme 3, the service migration is triggered when the delay exceeds 5 ms. When the bitrate in the backhaul is low, Scheme 3 performs similarly to Scheme 2, because frequent migrations are performed in both of them. As the bitrate increases, less and less migrations are triggered in Scheme 3, thus the effects on the end-to-end latency caused by the downtime are reduced. That is why Scheme 3 performs similarly to Scheme 1 when the bitrate is high, and has better performance than Scheme 2 when downtime is large (e.g., 0.3 s). It is noted that due to the small service area considered in our simulations the average end-to-end latency for Scheme 1 can be lower than that in Scheme 2 when the bitrate and downtime are large (e.g., the downtime is 0.3 s and the bitrate is higher than 200 Mbps). It is expected that in an area with large coverage, the performance of Scheme 1 would become worse, as discussed in the previous section.

Reliability is another important performance metric of vehicular communications. Here, it is defined as the probability that the end-to-end latency does not exceed a maximum allowable latency level [27]. Here, the Cumulative Distribution Function (CDF) of the end-to-end latency is used to

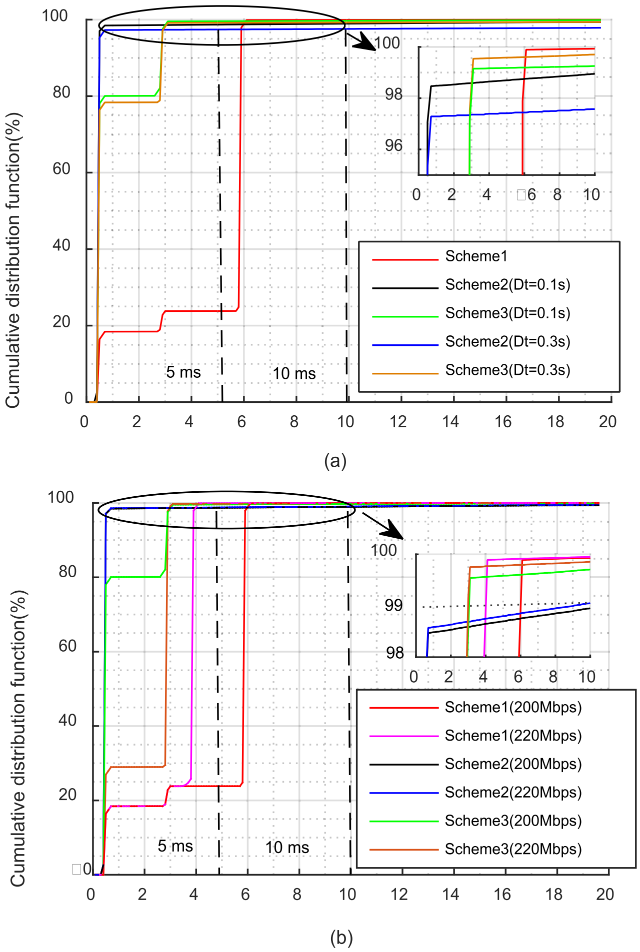


FIGURE 6. **(a) Cumulative distribution function versus end-to-end latency with transmission capacity of 200 Mbps, (b) cumulative distribution function versus end-to-end latency with Dt = 0.1 s.**

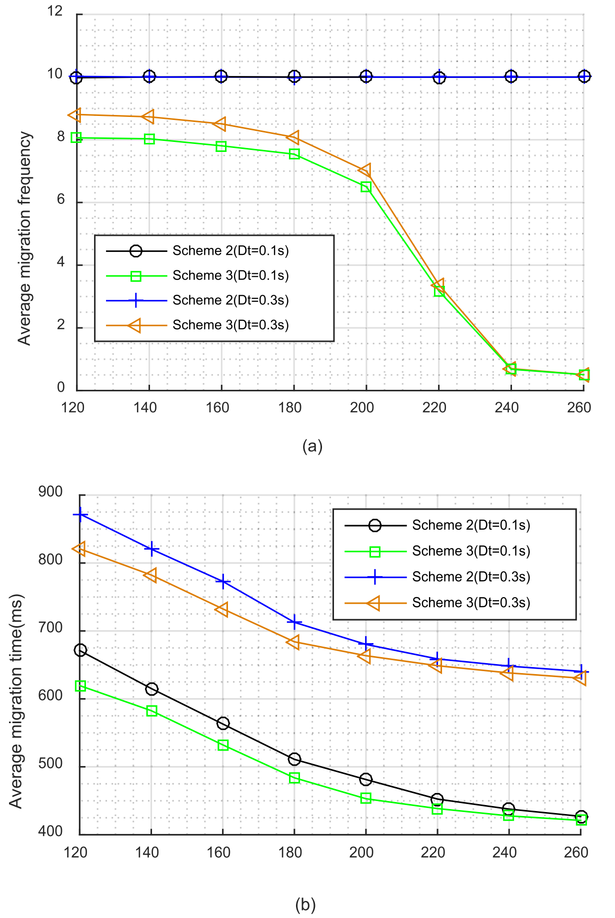


FIGURE 7. **(a) The average number of migration for a vehicle versus transmission capacity, (b) the average migration time versus transmission capacity.**

derive the reliability. The packet with its end-to-end delay higher than the predefined maximum limit is dropped. Hence, to some extent the reliability can be reflected by the packet drop ratio. Fig. 6(a) shows the CDF of the end-to-end latency when the backhaul bitrate is 200 Mbps. Without loss of generality, we set the maximum allowable end-to-end latency to 5 ms and 10 ms to satisfy the requirements of use cases in 5G (e.g., remote driving) [3], [19]. Clearly, a higher CDF value that can be achieved at the maximum delay implies higher reliability. It can be seen that Scheme 3 achieves the highest CDF among the three schemes when the maximum delay limit is less than 5 ms, which is about 99.6% and 99.1% for $D_t = 0.1$ s and 0.3 s, respectively. That is because the service migration will be triggered once the end-to-end latency is larger than 5 ms. In such a case, the downtime is the major factor affecting the reliability. Compared with Scheme 3, service migration is more frequent in Scheme 2, and thus the corresponding CDF is much lower (97.5% at $D_t = 0.1$ s and 98.5% at $D_t = 0.3$ s). When the maximum delay limit is relaxed to more than 10 ms, Scheme 1 has the best performance in terms of CDF in a small service area, which can be up to 99.9%. In such a scheme, the reliability is mainly affected by the handover interruption time, which is usually shorter than the downtime. Furthermore, we find that providing a higher bitrate is an effective way to increase the CDF in Scheme 1. For example, under the end-to-end latency requirement of 5 ms, the CDF increases from 25% to more than 99.9% when transmission capacity increases from 200 Mbps to 220 Mbps, as shown in Fig. 6(b). On the other hand, in Scheme 2 the increment of the backhaul bitrate has a little effect on the CDF. Fig. 6(b) also shows that the CDF profile for Scheme 3 becomes more similar to Scheme 1, when the bitrate increases and fewer migrations are triggered.

As discussed above, the end-to-end latency for vehicular communications can be reduced with an efficient service migration strategy. On the other hand, the frequent migrations and long migration time may have negative impacts on QoS. Thus, we further investigate average migration frequency and average migration time (defined in Table 2) in Scheme 2 and Scheme 3. Here, the migration frequency is referred to as the number of migrations that a vehicle experiences during its journey recorded in the simulation. Fig. 7(a) shows the average migration frequency versus transmission capacity in the backhaul. In Scheme 2, service migrations are triggered by the handovers, so the number of migrations is not correlated with the downtime or transmission capacity. In contrast, in Scheme 3, the higher transmission capacity leads to the smaller migration frequency, verifying the results shown in Fig. 5 and Fig. 6. Furthermore, a shorter downtime also results in a lower frequency of service migrations. Regarding the average migration time, it decreases with increasing transmission capacity and downtime for both Scheme 2 and Scheme 3, as shown in Fig. 7(b). Compared to Scheme 2, the average migration time for Scheme 3 is smaller for both $D_t = 0.1$ s and 0.3 s. That is because obviously fewer migrations are triggered in Scheme 3 when the bitrate increases, and the traffic generated by service migration decreases correspondingly.

As a general summary, when the backhaul bitrate is sufficient, in small service areas Scheme 1 can be a good choice in terms of reliability. However, for services that are time critical or location aware, and can only be satisfied by one-hop access, Scheme 2 may be necessary. In such a case, downtime is the main factor that can affect the performance in term of latency and reliability. Therefore, downtime should be minimized. Scheme 3 is a tradeoff between Scheme 1 and Scheme 2, which is suitable for the backhaul with a variable bitrate. It should be noted that the choice of migration strategies might be different in other backhaul architectures. For example, in the backhaul with a mesh topology the number of active nodes traversed by the packets is a very important factor in addition to the backhaul bitrate.

V. CONCLUSIONS AND FUTURE WORK

In this article, we have proposed a fog-enabled cellular based V2X solution supporting vehicular services. In order to improve the service continuity and QoS, we have investigated various service migration schemes that handle the mobility of vehicles. The performance of these schemes has been evaluated in terms of end-to-end delay, reliability, migration time and frequency by simulation with realistic traffic pattern in a small-size European country. It can be concluded that the choice of service migration schemes should adapt to the QoS requirements and the backhaul capability.

The results in this paper can be served as a general guideline, whereas we have realized future work is needed for enhancement. To realize seamless and fast service migration, the connectivity between BS-fog nodes should be enhanced via high-capacity links, such as based on fiber or millimeter wave. In this context, novel network architectures and protocols for high-capacity backhaul are necessary to be further developed. In addition, to improve backhaul network performance, virtualization techniques are required to reduce the size of the migration-related data and release the negative impacts on backhaul networks. The assumption of the wireless access delay and the interruption time is quite simple in this paper, and the impact of realistic values is important to be investigated. Finally, in order to guarantee sufficient resources for the service migration, particularly for the time-critical services, effective resource management is essential to reduce the blocking of service migration requests caused by the lack of computing/storage resources at the fog nodes. Considering that in practice resource availability status is complex and varies in time, artificial intelligence may be a promising tool for making smart decisions related to the service migration.

REFERENCES

[1] (Mar. 2018). *An Update on Last Week's Accident*. [Online]. Available: https:// www.tesla.com/blog/tragic-loss

[2] *Study on LTE Support for V2X Services*, document TR 22.885, v14.0.0, 3GPP, 2015. [Online]. Available: http://www.3gpp.org/DynaReport/22885.htm

[3] *Study on Enhancement of 3GPP Support for 5G V2X Services*, document TR 22.886, v15.1.0, 3GPP, 2016. [Online]. Available: http://www.3gpp.org/DynaReport/22886.htm

[4] *Study on Architecture Enhancement for EPS and 5G System to Support for Advanced V2X Services*, document TR 23.786, v0.6.0, 3GPP, 2018. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3244

[5] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[6] A. Ashok, P. Steenkiste, and F. Bai, "Adaptive cloud offloading for vehicular applications," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Dec. 2016, pp. 1–8.

[7] M. Azizian, S. Cherkaoui, and A. S. Hafid, "Vehicle software updates distribution with SDN and cloud computing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 74–79, Aug. 2017.

[8] Y.-J. Ku *et al.*, "5G radio access network design with the fog paradigm: Confluence of communications and computing," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 46–52, Apr. 2017.

[9] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 54–61, Apr. 2017.

[10] Y.-Y. Shih, W.-H. Chung, A.-C. Pang, T.-C. Chiu, and H.-Y. Wei, "Enabling low-latency applications in fog-radio access networks," *IEEE Netw.*, vol. 31, no. 1, pp. 52–58, Jan./Feb. 2017.

[11] S. Wang, J. Xu, N. Zhang, and Y. Liu, "A survey on service migration in mobile edge computing," *IEEE Access*, vol. 6, pp. 23511–23528, 2018.

[12] L. F. Bittencourt, M. M. Lopes, I. Petri, and O. F. Rana, "Towards virtual machine migration in fog computing," in *Proc. 10th Int. Conf. P2P, Parallel, Grid, Cloud Internet Comput. (3PGCIC)*, Krakow, Poland, Nov. 2015, pp. 1–8.

[13] A. Machen, S. Wang, K. K. Leung, B. J. Ko, and T. Salonidis, "Live service migration in mobile edge clouds," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 140–147, Feb. 2018.

[14] K. Ha *et al.*, "Adaptive VM handoff across cloudlets," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS15-113, 2015.

[15] T. Taleb *et al.*, "Follow-me cloud: When cloud services follow mobile users," *IEEE Trans. Cloud Comput.*, Feb. 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7399400

[16] S. Wang *et al.*, "Dynamic service migration and workload scheduling in edge-clouds," in *Proc. IFIP Perform.*, Oct. 2015, pp. 1–9.

[17] S. Wang, R. Urgaonkar, T. He, K. Chan, M. Zafer, and K. K. Leung, "Dynamic service placement for mobile micro-clouds with predicted future costs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 4, pp. 1002–1016, Apr. 2017.

[18] M. Whaiduzzaman, M. Sookhak, A. Gani, and R. Buyya, "A survey on vehicular cloud computing," *J. Netw. Comput. Appl.*, vol. 40, pp. 325–344, Apr. 2014.

[19] 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.

[20] *Feasibility Study for Evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN)*, document TR 25.912, V13.0.0, 3GPP, 2015. [Online]. Available: http://www.3gpp.org/DynaReport/25912.htm

[21] D. Han, S. Shin, H. Cho, J.-M. Chung, D. Ok, and I. Hwang, "Measurement and stochastic modeling of handover delay and interruption time of smartphone real-time applications on LTE networks," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 173–181, Mar. 2015.

[22] *Sumo-Simulation of Urban Mobility*. [Online]. Available: https://www.dlr.de/ts/en/desktopdefault.aspx/tabid-9883/16931_read-41000/

[23] L. Codeca, R. Frank, and T. Engel, "Luxembourg SUMO traffic (LuST) scenario: 24 hours of mobility for vehicular networking research," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Kyoto, Japan, Dec. 2015, pp. 1–8.

[24] *Ultra Reliability and Low Latency (URLLC) User-Plane Latency Analysis*, document R1-165028, 3GPP, May 2016.

[25] T. Orphanoudakis, E. Kosmatos, J. Angelopoulos, and A. Stavdas, "Exploiting PONs for mobile backhaul," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. S27–S34, Feb. 2013.

[26] J. Li, X. Shen, L. Chen, J. Ou, L. Wosinska, and J. Chen, "Delay-aware bandwidth slicing for service migration in mobile backhaul networks," *J. Opt. Commun. Netw.*, vol. 11, no. 4, pp. B1–B9, Dec. 2018.

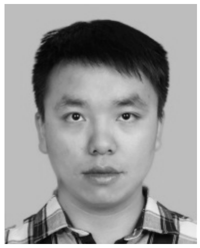
[27] P. Popovski *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

JUN LI received the M.E. degree from Shanghai Jiao Tong University, China, in 2015. He is currently pursuing the Ph.D. degree with the Optical Networks Laboratory, KTH Royal Institute of Technology, Sweden. His research interests include fog/edge computing, vehicular communication, and optical networks.

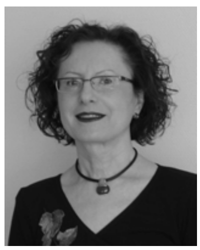
XIAOMAN SHEN received the B.E. degree from Zhejiang University, in 2014, where she is currently pursuing the Ph.D. degree. She was also a joint Ph.D. student with the Optical Networks Laboratory, KTH Royal Institute of Technology, from 2017 to 2018. Her research interests include vehicular communication and optical network for datacenter.

LEI CHEN received the B.E. and M.E. degrees in communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in infra-informatics from Linköping University, Sweden, in 2013. He is currently a Senior Researcher with Cooperative Systems, Viktoria Swedish ICT. His research interest includes cooperative intelligent transport systems, where he investigates communication technologies and cloud-based services for cooperative and automated transport systems to improve traffic safety, efficiency, and sustainability.

DUNG PHAM VAN received the B.Sc. degree in information technology from Hong Duc University, Thanh Hóa, Vietnam, in 2003, the M.Sc. degree in ICT from Waseda University, Tokyo, Japan, in 2009, and the Ph.D. degree *(cum laude)* in telecommunications from the Scuola Superiore Sant'Anna, Pisa, Italy, in 2014. He was a Postdoctoral Researcher with the KTH Royal Institute of Technology, Sweden, and INRS, Montreal, QC, Canada. He is currently an Experienced Researcher with Ericsson Research. He has authored more than 40 papers in international journals and conference proceedings.



JIANNAN OU received the B.Sc. degree from the Guandong University of Technology, in 2016. He is currently pursuing the master's degree with South China Normal University, China. His research interests include PON-based network resource management, the Internet of Things, and mobile edge cloud computing.



LENA WOSINSKA (SM'11) is currently a Professor with the Chalmers University of Technology. She was the Founder and Leader of the Optical Networks Laboratory, KTH Royal Institute of Technology. She has been involved in several EU projects and coordinating a number of national and international research projects. Her research interests include fiber access and 5G transport networks, energy-efficient optical networks, photonics in switching, optical network control, reliability and survivability, and optical datacenter networks. She has been involved in many professional activities, including the Guest Editorship of the IEEE, OSA, Elsevier, and Springer journals, serving as the General Chair and a Co-Chair of several IEEE, OSA, and SPIE conferences and workshops, serving in TPC of many conferences, and being a Reviewer of scientific journals and project proposals. She is currently serving on the Editorial Board of *Photonic Network Communications* journal (Springer) and the *Transactions on Emerging Telecommunications Technologies* (Wiley). She has been an Associate Editor of the *OSA Journal of Optical Networking* and the IEEE/OSA JOURNAL OF OPTICAL COMMUNICATIONS AND NETWORKING.

JIAJIA CHEN received the B.S. degree from Zhejiang University, China, in 2004, and the Ph.D. degree from the KTH Royal Institute of Technology, Sweden, in 2009, where she is currently an Associate Professor with the Optical Networks Laboratory. She has co-authored over 100 publications in international journals and conferences in the area of optical networking. Her main research interests include optical transport and interconnect technology supporting future 5G and cloud environments. She has been involved in various European research projects, such as European FP7 projects IP-OASE and IP-DISCUS, and EIT-ICT projects. Moreover, she is a Principal Investigator/Co-Principal Investigator of several national research projects funded by the Swedish Foundation of Strategic Research (SSF) and the Swedish Research Council (VR).