Online Resource Management in Fog-enhanced Cellular Networks for Real-Time Vehicular Services

Jun Li [1], Carlos Natalino[2], Xiaoman Shen[3], Lei Chen[4], Jiannan Ou[5] , Lena Wosinska[2] and Jiajia Chen[1,2,*]

[1] Department of Communication Systems, KTH Royal Institute of technology, Sweden;
[2] Department of Electrical Engineering, Chalmers University of Technology, Sweden;
[3] Zhejiang University, Zhejiang, China;
[4] RISE Viktoria, Sweden;
[5] South China Normal University, Guangzhou, China
* Correspondence: jiajiac@kth.se;

Abstract: Fog computing is expected to be integrated with existing communication infrastructures, giving rise to the concept of fog-enhanced cellular networks (FeCNs) to support real-time services. In such FeCNs, service migration is necessary to maintain the service continuity and satisfy stringent latency requirements of real-time vehicular services, where the service is migrated from a source fog node to a target fog node following the vehicle's moving trace. Fog servers, however, need to have sufficient computational resources available to support such a migration. Also, provisioning resource for the migrated real-time services needs to be completed as soon as possible to minimize the service interruption. This paper proposes a distributed online resource management (ORM) scheme, in which resources for real-time vehicular services are provisioned with high priority. Once resources are scarce in one fog node, services with low priority can be migrated to neighboring fog nodes and their resources can be released in a distributed fashion. We propose two algorithms tailored to reduce the negative effects on the affected services. As a case study, the Luxembourg traffic volume model has been considered to verify the performance of the proposed scheme. Simulation results show that the performance of the proposed scheme is dependent on the backhaul capacity. Compared with other schemes, the one-hop access probability for real-time vehicular services implying low delay performance can be effectively improved, while the performance of other services can also be well maintained by providing sufficient backhaul capacity.

Keywords: Fog computing; online resource management; connected vehicle

---

1. Introduction

Fog computing brings the computation resources as close as possible to the end users and has a great potential to overcome the latency challenge and satisfy the delay requirement of real-time vehicular services [1, 2]. On the other hand, widely deployed cellular networks are regarded as one of the most promising access technologies for the connected vehicles due to their ability to provide high data rate and low latency to mobile users as well as their large coverage area [3, 4]. In this regard, fog servers (referred to as cloudlets) can be co-located with cellular base station (BS), e.g., long term evolution (LTE) evolved Node B (eNB), forming BS-Fog node and giving rise to the concept of fog-enhanced cellular networks (FeCNs), as shown in Fig.1 [5, 6]. Such FeCNs can be a promising candidate to support vehicle-to-everything (V2X) connectivity, i.e., to facilitate information exchange not only among vehicles, but also between the vehicles and the other entities, such as vulnerable road users (pedestrians, cyclists), road side sensors, cameras, thereby supporting various vehicular services and improving safety, efficiency, and convenience of transportation systems.

Although the concept of FeCN has many potential advantages in supporting connected vehicles, a number of challenges have to be tackled to make it become a reality. In FeCN-based V2X networks, when a vehicle moves from one area to another, a traditional handover is followed by a service migration to maintain the service continuity and high performance, which is particularly critical for safety-related vehicular services. In addition, vehicular users make decisions based mainly on the information from local environment (such as sensors, cameras and/or roadside units). Thus, service migration, referring to the process of relocating services from a source fog node to another (target) fog node, should follow the mobility of the respective vehicles and maintain a one-hop service access delay (e.g., below 5 milliseconds for remoted driving) [7, 8]. Such a one-hop access means that the service can be accessed directly at the BS with which the user equipment (UE) is associated, where only the last wireless segment between the UE and BS is involved. The one-hop access delay represents the minimum communication latency that the vehicle has to experience in the FeCN. Meanwhile, the target fog node should have enough resources for hosting the migrated services. Otherwise, the services have to stay at the source fog node, implying an increased access delay, which may be larger than the stringent latency requirement for real-time vehicular services.

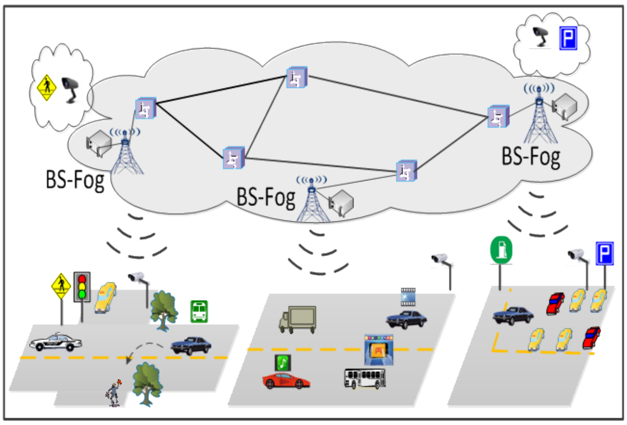

Fig. 1. A high-level view of FeCN for connected vehicles.

Furthermore, each BS-Fog usually has a limited amount of computation resources and its workload is highly burst, mainly due to the mobility of vehicles. It is therefore important to achieve high reliability, low latency, and scalability by employing efficient resource management mechanisms even in the scenarios with high load and fast mobility. In [6], two schemes are proposed, namely fog resource reservation and fog resource reallocation. In the first scheme some fog resources at each BS-Fog are reserved based on a predicted vehicular traffic load. The performance of this scheme depends on the traffic flow predication methods. Overestimating leads to low resource utilization, whereas underestimating significantly decreases the one-hop access probability for the high priority (HP) services, such as safety related vehicular services, including remote driving, pre-crash sensing warning and so on. The second scheme is based on a strategy to release part of fog resources used for low-priority (LP) services that are not critical, e.g., online game, real-time navigation, road sign notification, and reallocate it to HP services. However, in such a scheme, the performance of LP services may be affected when traffic load is high. In fact, not all the LP services, e.g. games, need to have one-hop latency requirement and local awareness. Therefore, such services can be placed in its neighboring fog nodes with low-load.

Recently, resource sharing among a cluster of neighboring fogs or cloudlet nodes has been studied to increase the fog resource utilization [9-11]. For example, in [9], R. Yu et al. proposed a resource sharing scheme, in which resource utilization is improved through the inter-cooperation between cloudlets. M. Jia et al. [10] proposed a load balancing algorithm by dynamically relocating the tasks that run in the heavy-loaded cloudlets to the light-loaded ones. Y. Xiao et al. [11] investigated a fog node cooperation strategy by optimizing workload allocation to maximize user's quality-of-experience under the given power efficiency. These schemes typically require complex computation or high communication overhead to achieve optimal solutions resulting in large processing time, especially in a large network, and therefore they are usually performed offline. Besides, scalability is also one of the challenges for the centralized offline resource sharing schemes, particularly for the vehicles that are normally traversing across a large area [2].

This paper proposes a distributed online resource management (ORM) scheme, in which once a fog node becomes overloaded and does not have sufficient resource for the incoming HP services, some LP services are selected to be migrated to proper neighbouring fog nodes and their resources are released for the HP services. The proposed ORM scheme is composed of two key algorithms: 1) LP service selection that determines proper LP services to be migrated in order to host coming HP services, and 2) neighbouring fog node selection that attempts to find best adjacent fog nodes as the targets to migrate selected LP services. The objective of the proposed LP service selection algorithm and neighbouring fog node selection algorithm is to minimize negative effects on the affected LP services. As a case study, the realistic mobility pattern for the city of Luxembourg is considered. Simulation results show that the performance of the proposed scheme mainly depends on the backhaul capacity. Compared with the existing schemes, the proposed scheme can effectively increase one-hop access probability for HP services, while service unavailability of LP services can also be reduced by sufficient backhaul capacity.

The rest of the paper is organized as followings. Section 2 introduces the service migration mechanism in fog-enhanced cellular networks for connected vehicles. In Section 3, the proposed online resource management algorithm is described. Section 4 conducts a case study and presents simulation results. A conclusion is drawn in Section 5.

2. Service Migration in Fog-enhanced Cellular Networks for Connected Vehicle

In the FeCN, fog servers are co-located with the base station, forming a BS-Fog node, which integrates the communication resource for radio access network (RAN) and computation resource for fog computing. The architecture of a BS-Fog is shown in Fig. 2(a), where the BS is responsible for network functions (e.g., handovers), whereas the fog servers provide computational resources (e.g., computing and storage capability) locally one BS-Fog can cooperate with the other BS-Fogs or cloud to allocate tasks dynamically. In the 3GPP standards, two interfaces are defined at each BS, namely S1 and X2 [5]. S1 is assigned for the communications between BSs and the central aggregation switch in the mobile core network, which 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 among the BS-Fogs.

To illustrate the service migration procedure in FeCNs, we consider a scenario where vehicles have a random route, e.g., private cars moving inside a city, as shown in Fig. 2(b). In such scenario, a vehicle starts from Cell1 to Cell3 via Cell2 and the service subscribed by the vehicle is originally hosted by BS-Fog1. When a vehicle moves from one BS-Fog's coverage area to another, a handover is triggered and the target BS-Fog is selected accordingly. At the same time, ongoing real-time services hosted at the source BS-Fog also need to be migrated following the vehicle. In the service migration, the technique of live virtual machine (VM) is deployed, where services are encapsulated

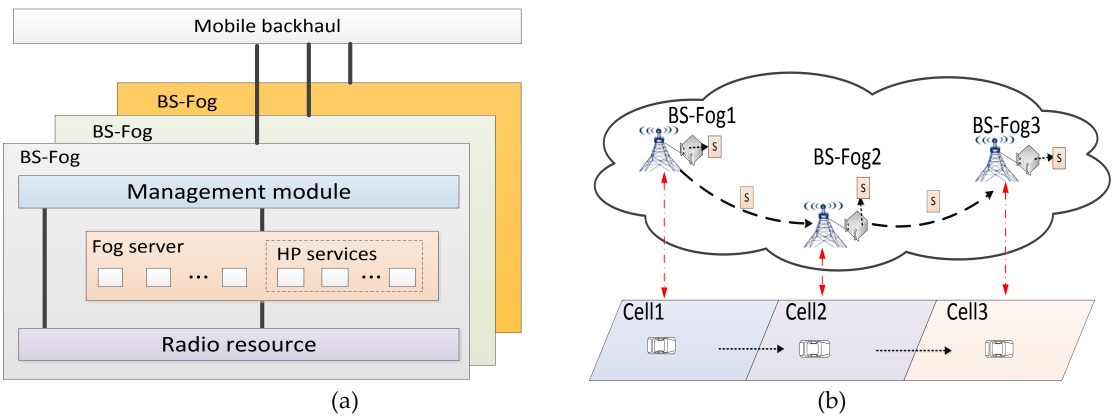

Fig. 2: (a) BS-Fog architecture; and (b) an example of service migration in FeCNs.

into VM and migrated from the source BS-Fog to the target BS-Fog [3]. In such a case, the vehicle can access the HP service with one hop. It should be noted that the proposed scheme is also suitable for other virtualization techniques (e.g., container). The detailed protocol is described in Fig. 3. Once the handover is triggered, the source BS-Fog checks the resources (e.g., CPU, memory) used for the ongoing services, and sends a Migration Request message which includes the information of the required resources to the target BS-Fog. After receiving the Migration Request message, the target BS-Fog makes a decision whether to accept the service migration request or not, according to the resource management strategy.

Two scenarios are defined in the resource management strategy. First, if the target BS-Fogs has sufficient computing resource, the request is approved and the corresponding resource is assigned. Then, the target BS-Fog sends a Migration Request ACK message to the source BS-Fog (see Scenario ① in red dotted box in Fig. 3). After receiving the Migration Request ACK message, the source BS-Fog starts to perform the service migration and service can still be accessed with one hop after the vehicle moves to the area covered by the target BS-Fog.

Otherwise, the target BS- Fog selects ongoing LP services to be migrated to its neighboring BS-Fogs, and releases their resources for HP services (see Scenario ② in blue dashed box in Fig. 3). After selecting the LP services for migration, the target BS-Fog broadcasts a Migration Request message to its neighboring BS-Fogs through X2 interface. Here, neighboring BS-Fogs refer to the BS-Fogs that can directly communicate with the target BS-Fog via X2 interface. Once the Migration Request message is received, the neighboring BS-Fogs check their available resources and send back Migration Request ACKs. The target BS-Fog makes a final decision on which neighboring BS-Fogs are selected among all the BS-Fogs that accept the Migration Request. When the selected LP services complete migration and release their resources, the target BS-Fog sends a Migration Request ACK to the source BS-Fog. If there are no available LP services or neighboring BS-Fogs, the to-be-migrated HP service still needs to run at the source BS-Fog and the vehicle has to access the service with more than one hop (i.e., the hop(s) between two neighboring BS-Fogs have to be counted for access), which may result in an obviously higher access delay. The selection strategies of LP services and neighboring BS-Fogs are discussed in the following section.

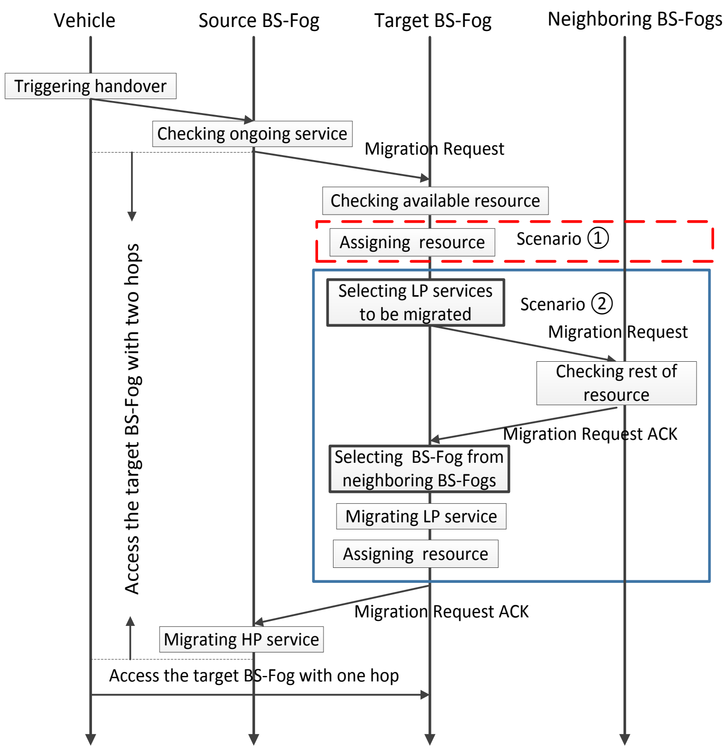

Fig. 3. Resource management in service migration procedure.

3. Online Resource Management

In this section, we propose an online resource management (ORM) scheme which consists of migrating LP service selection algorithm and neighboring BS-Fog selection algorithm. In the ORM scheme, the ongoing LP services can be migrated from one BS-Fog to its neighbors and the corresponding resources are released for HP services.

3.1. Migating LP Service Selection Algorithm

The ongoing LP services can be migrated from the target BS-Fog to its neighboring BS-Fogs to release resources for HP services. Once LP services are selected, they need to be migrated to the neighboring BS-Fogs and will be accessed via backhaul network. During this procedure, a certain amount of backhaul bit rate is needed for VM migration and for running the selected LP services, which should be minimized. According to this principle, a migrating LP services selection algorithm is proposed. The total needed amount of data (Mbits) for i^{th} LP service (α_i) is calculated by

$$\alpha_i = C_i \cdot T_i + V_i \tag{1}$$

where, C_i is the needed bit rate (Mbps). T_i is the remaining time for the i^{th} LP service to be completed (seconds), and V_i is the amount of data needed for migrating the VM of the i^{th} LP service (Mbits), which is equal to the size of the VM . The service with the lowest α_i is selected to be migrated. If the released resources of the i^{th} LP service cannot satisfy the required resources of HP service (Δ), more than one LP service will be selected for migration.

Algorithm1: Migrating Low-Priority Service Selection

```
Input: n_i , l_i
Output: LP_m
1. Set LP_n ← {n_i, i = 1,2 ... K}; LP_r ← {l_i, i = 1,2 ... K}
2.  for  i ←1:1: length (LP_n) do
3.       α_i ← C_i · T_i + V_i
4.       V_c ← [V_c,  α_i]
5.       if  LP_r[i] ≥ Δ  then
6.           V_p ← [V_p,  α_i]
7.       end if
8.  end for
9.  if  V_p ≠ ∅  then
10.      Find LP service (LP_n[i]) with the lowest  α_i ∈ V_p
11.      LP_m ← [LP_m, LP_n[i]]
12. else
13.      Sort (V_c) according to ascending of α_i
14.      for j←1:1: length (V_c) do
15.          if  Δ> 0  then
16.              Δ← Δ − LP_r[j]
17.          else
18.              LP_m ← [LP_m, LP_n[j]; Break
19.          end if
20.      end for
21. end if
```

The proposed migrating LP service selection algorithm is described in Algorithm 1, whereas the symbols used in the pseudocode are explained in Table 1. Firstly, the amount of the required bandwidth for each LP service is calculated (see line 3-4 in Algorithm 1). Then, the LP services with available computing resources larger than the requested amount are selected (see Lines 5-7 in Algorithm 1). Among these services, the one with the lowest communication cost is further selected (see Lines 9-11 in Algorithm 1). If there is no service that satisfies the requirement, more than one service can be selected according to the total needed amount of data for the i^{th} LP service (α_i) in ascending order (see Lines 13-20 in Algorithm 1). In order to avoid a ping-pong effect on LP service

Table 1. Explanation of symbols.

Symbol	Description
n_i	Series number for the i^{th} LP service, belonging to set LP_n
l_i	Released resource for the i^{th} g LP service, belonging to set LP_r
V_p	Set of cost parameter for the i^{th} LP service ($l_i > \Delta$)
V_c	Set of cost parameter for the i^{th} LP services
r_i	Available resource for the i^{th} neighbor BS-Fog, belonging to set R
s_i	Series number for the i^{th} neighbor BS-Fog, belonging to set S
g_i	Backhaul delay between the target BS-Fog and the i^{th} neighbor BS-Fog, belonging to set G
d_j	Budget of backhaul delay for the j^{th} migrated LP service, belongs to set D
x_j	Required resource for the j^{th}LP service, belonging to set X
y_j	Serial numbers for the j^{th} LP service, belonging to set Y
LP_m	Set of selected LP services
F	Set of selected neighbor BS-Fogs

migration, LP services are only allowed to be migrated once. Here, the ping-pong effect is referred to the procedure, in which one LP service is migrated among the target BS-Fog and the neighbouring BS-Fogs backwards and forwards. It may occur when the load in fog node is highly bursty due to the frequent mobility of vehicles. The complexity of Algorithm 1 is $O(M)$, where M is the number of LP services for the target BS-Fog.

3.2. Neighboring BS-Fog Selection Algorithm

Once the LP services to be migrated are decided, neighboring BS-Fogs that will host the migrating services need to be selected, and should have sufficient available resources to satisfy the resource requirement of the to-be-migrated LP services. Besides, after LP service is migrated, it is hosted at the selected neighboring BS-Fog and accessed with two hops, which results in extra backhaul delay. As LP services are only allowed to be migrated once, the access delay for the migrated LP services consists of radio access delay and backhaul delay. According to the delay requirement of the selected LP service, the budget of backhaul delay can be calculated. The key idea of the proposed algorithm is to select the neighboring BS-Fog with the most available resources under the backhaul delay requirement of the to-be-migrated LP service.

The proposed neighboring BS-Fog selection algorithm is shown in Algorithm 2, whereas the symbols used in the pseudocode are explained in Table 1. The proposed neighbouring BS-Fog selection algorithm is shown in Algorithm 2, whereas the symbols used in the pseudocode are explained in Table 1. Firstly, the neighbouring BS-Fogs are sorted according to their available resources (see Lines 8-9 in Algorithm 2). Then, the neighbouring BS-Fog with the largest amount of available resources is selected if the transmission delay between the target BS-Fog and the neighbouring BS-Fog is smaller than the backhaul delay (see Lines 11-13 in Algorithm 2). After that the selected neighbouring BS-Fog allocates the resources to the migrated LP services, the neighbouring BS-Fogs are sorted according to their available resources again (see Lines 14-16 in Algorithm 2). The complexity of Algorithm 2 is $O(N \times M)$, where N is the number of neighbouring BS-Fogs and M is the number of LP services for the target BS-Fog.

Algorithm2: Neighboring Base station-Fog Selection

```
Input: r_i , s_i , d_j , g_i , x_j , y_j;
Output: F
1. Set R ← {r_i, i = 1,2 ... N}
2. Set G ← {g_i, i = 1,2 ... N}
3. Set S ← {s_i, i = 1,2 ... N}
4. Set X ← {x_j, j = 1,2 ... M}
5. Set D ← {d_j, j = 1,2 ... M
6. Set Y ← {y_j, j = 1,2 ... M}
7. Sort (R) according to the descending order of r_i
8. Update ( S , G)
9.    for  j ← 1: 1: M  do
10.     for i ← 1: 1: N  do
11.           if  R[i] ≥ X[j]   & G[i] < D[j]  then
12.                F ← {S[i], Y[j]}
13.                β ← S[i]
14.                r_β ← r_β − X[j]
15.                Soft (R) according to the descending of r_i
16.                 Update (S, G)
17.                 Break
18.             end if
19.         end for
20.    end for
```

4. Performance evaluation

In this section, the performance of the proposed scheme is evaluated using Urban Mobility simulation tools Sumo together with Matlab [12].

4.1. Simulation Setup

We use a realistic mobility pattern for the city of Luxembourg [12], which can be considered as a case study for a small service area. The scenario topology is shown in Fig. 4, in which BS-Fog entities are evenly distributed in the city and a single entity has the coverage of 1 km² [13, 14]. Fig.5 shows the vehicular traffic profile in Luxembourg SUMO traffic which varies in time during a day [12]. Also, the vehicular traffic is spatially diverse. For example, the insert chart (a) in Fig. 5 shows the numbers of vehicles at each coverage area of BS-Fogs at 8:00 am, while the insert chart (b) in Fig. 5 shows the numbers of vehicles at 12:00 pm. The Y-axis shows the number of vehicles, while the X-axis is the series number of BS-Fog. We assume that each vehicle only requires one HP service (safety-related). The data traffic distribution is proportional to the vehicular traffic.

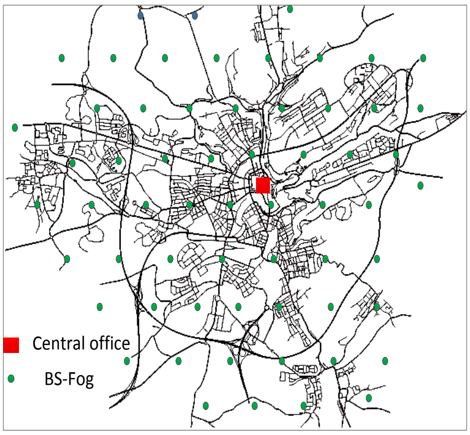

Fig. 4. Luxembourg scenario topology.

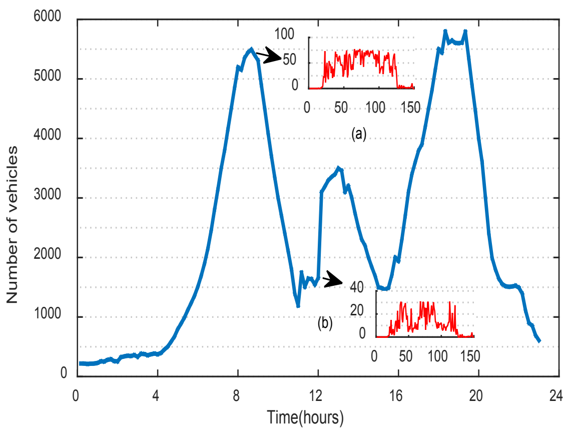

Fig.5 . vehicular traffic profile over a day in LuST.

Without loss of generality, we assume the total service request arrival rate for the BS-Fog network is in the range from 20 to 100 (per second) [10]. The arriving requests consist of 30% of HP and 70% of LP services. The HP service arrival rate is distributed among the BS-Fogs according to the traffic profile at 8:00 am (see Fig. 5(a)), while the LP service arrival rate is distributed among BS-Fogs evenly. We assume the Poisson arrivals of vehicles [15, 16]. Consequently, the arrivals of HP service requests are also according to the Poisson process. The time that vehicles are served by one BS-Fog is dependent on several factors such as the route, speed of the vehicles, and coverage area of one BS-Fog, which is explored in [6, 12]. Besides, we also assume that the LP service requests arrive according to the Poisson process, whereas their average service time is exponentially distributed [17]. The remaining parameters are described in Table 2 [6]. The simulation time is set to be 1000 seconds and all the results presented in the next section are the average of ten different simulations. The confidence interval of these results is less than 3.5%, and it has been calculated with a confidence level of 95%.

Table 2. Explanation of symbols.

Term	Value
Total number of computing units in one fog	400
Number of BS-Fogs in the network	100
Number of computing units for each HP service	3
Number of computing units for each LP service	[2,6]
Average serving time in one fog for HP service (second)	90
Standard deviation of the serving time in one fog for HP service (second)	10
Average serving time in one fog for LP service (second)	120
Budget of backhaul delay for LP services(millisecond)	[5,10]
Downtime in live VM migration (millisecond)	20

4.2 Results

In this subsection, the performance of the proposed scheme is evaluated in terms of one-hop access probability for HP services and service unavailability for LP services. Here, one-hop access probability is defined as the ratio of the one-hope service access duration to the total holding time. Similarly, service unavailability is defined as the ratio of the time when service is not available to the total holding time. Here, the services may be unavailable due to the lack of resources in the current BS-Fog and its neighbors, as well as due to the interruption during service migration. As shown in Fig. 3, to increase one-hop access probability while reducing service unavailability, migration time for both HP and LP services should be minimized, which is related to the transmission time in the mobile backhaul network. Passive optical network (PON) is widely adopted for mobile backhaul because of its energy efficiency and high capacity [18]. In such a PON-based backhaul, transmission capacity becomes the main factor that affects delay performance, which has been pointed out in [19]. Here, the backhaul capacity (B) is referred to as the bandwidth allocated to the X2 interface between two Fog-BSs. The bandwidth allocation algorithm introduced in our previous work [19] is implemented for the migration bandwidth provisioning, in which a large size of migration data can be divided into small pieces and transmitted under a required time threshold. Meanwhile, the non-migration traffic can be transmitted according to their priorities. We assume that the data rate of the traffic generated by the users is uniformly distributed between 2 Kbps and 10 Mbps, while the amount of data encapsulated in application VMs is between 10 Mbits and 100 Mbits [6,13,17]. The size of neighboring BS-Fog nodes (N) is set by considering that one BS may need connectivity up to around 20 neighboring cells [20].

Fig. 6(a) shows that one-hop access probability for HP services decreases with the service arrival rate. It can be seen that the one-hop access probability for B=200Mbps is larger than that for

B=100Mbps. That is because larger backhaul capacity can reduce migration time for both HP and LP services, and increase one-hop access probability of HP services. When B=200Mbps, a higher number of neighbors are beneficial to the one-hop access probability. Especially for the high service arrival rate, one-hop access probability increases about 4% when N increases from 10 to 20. However, when B=100Mbps, the increase of N has little effect on one-hop access probability. That is because reduced backhaul capacity results in high backhaul delay between the target and the neighboring BS-Fogs. The number of neighboring BS-fogs that satisfy the latency requirement decreases, even when the number of neighboring BS-Fog is higher. Fig. 6(b) shows LP service unavailability as a function of service arrival rate. Similarly, increasing backhaul capacity from 100Mbps to 200Mbps can also reduce service unavailability as there is a decrease in the migration delay

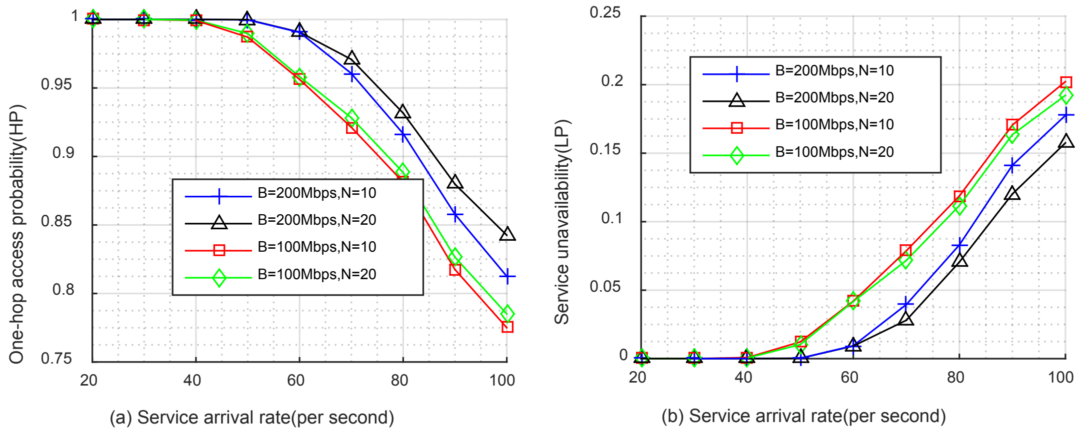

Fig. 6. One-hop access probability and service unavailability versus service arrival rate.

We compare our results with other two benchmark schemes. The first benchmark is based on first come first serve (FCFS) scheme which doesn't apply any specific resource management, i.e., without any differentiation for treating the HP and LP services. The second benchmark is based on the fog resource reservation (FRR), in which a certain amount of computing resources is reserved for HP services [6]. Without loss of generality, in the proposed ORM, the values of backhaul capacity B are selected as 200Mbps and 100Mbps, while N is set to 20. Fig. 7(a) shows the one-hop access probability for the HP services as a function of service arrival rate. The one-hop access probability for ORM with B=200Mbps is higher than that for FCFS and FRR when service arrival rate increases from 20 to 90 arrivals per second. However, when service arrival rate keeps increasing to 100 arrivals per second, the one-hop access probability is lower than that for FRR. That is because the migrated LP services also take up resources of neighboring BS-Fogs, and the one-hop access probability of HP services is reduced in the affected neighboring BS-Fogs. It also can be seen that the one-hop access probability for ORM with B=100Mbs is higher than for FCFS and FRR when service arrival rate is below 60 arrivals per second, but lower than FRR before service arrival rate increases to 60 arrivals per second. That is because more LP services need to be migrated to the neighboring BS-Fogs with the service traffic arrival rate increasing, which results in high migration traffic and thus increases migration delay.

Fig. 7(b) shows the service unavailability as a function of service arrival rate. The service unavailability for ORM with B=200Mbps is the lowest among all schemes with service arrival rate increasing due to the shorter time used for migrating LP services. When B=100Mbps, the service unavailability for ORM is similar to FCFS for service arrival rate lower than 60 arrivals per second, and becomes lower with increase of the service arrival rate. That is because the migration delay in ORM is similar to the time used for waiting for the available resources in the case of FCFS when the service arrival rate is low. A larger arrival rate results in a longer waiting time. It can be seen that migration delay is an important factor that affects the performance of the proposed ORM in terms of one-hop access probability and service unavailability.

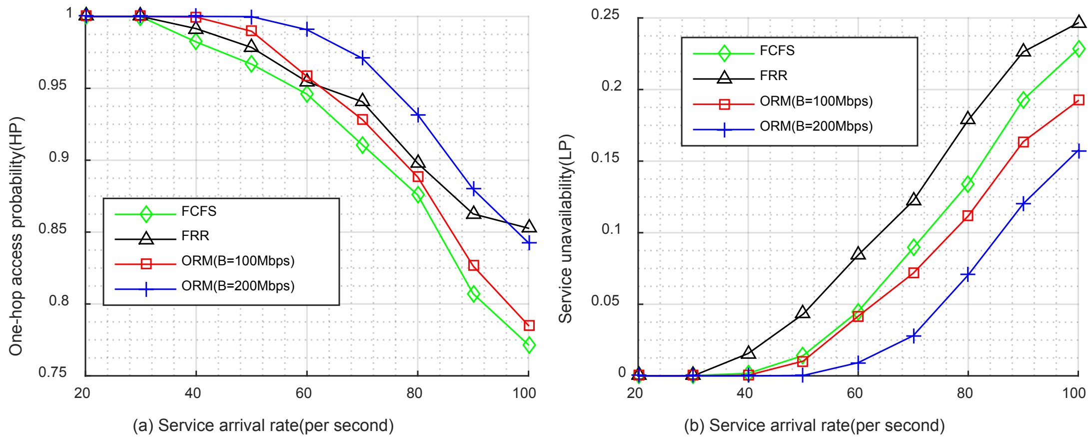

Fig. 7. One-hop access probability and service unavailability versus service arrival rate.

5. Conclusions

This proposed a distributed online resource management scheme in the FeCN to increase the one-hop access probability by prioritizing HP services for using computing resources. Meanwhile, LP services can be migrated to neighboring BS-Fogs with minor degradation by implementing the proposed migrating LP service selection algorithm and neighboring BS-Fog selection algorithm. To verify the performance of the proposed ORM scheme, the realistic mobility pattern for the city of Luxembourg has been investiagted as a case study. Simulation results show that the performance highly depends on backhaul capacity. With sufficient backhaul capacity, the proposed online resource management scheme can effectively increase one-hop access probability for HP services, while reducing service unavailability of LP services compared with existing schemes. On the other hand, the larger the backhaul capacity is, the better performance can be achieved by increaing the number of neigboring BS-Fog nodes for selection. In this regrad, the capacity of bakchaul networks that connect BS-Fogs is one of the key factors when deploying fog nodes with the existing cellular networks.

Funding: This work was supported by the Göran Gustafssons Stiftelse, Natural Science Foundation of Guangdong Province (Grant No. 508206351021) and National Natural Science Foundation of China (Grant No. 61550110240, 61671212).

Conflicts of Interest: The authors declare no conflicts of interest

References

1. C. Huang, R. Lu, and K.R. Choo, “Vehicular fog computing: architecture, use case, and security and forensic challenges,” *IEEE Communication Magazine*, vol. 55, no. 11, pp. 105-111, Nov. 2017.
2. W. Zhang, Z. Zhang and H. Chao, “Cooperative fog computing for dealing big data in the internet of vehicles: architecture and hierarchical resource management,” *IEEE Communication Magazine*, vol. 55, no. 12, pp.60-67, Dec., 2017.
3. J. Li, X. Shen, L. Chen, D. Van, J. Ou, W. Lena and J: Chen “Service migration in fog computing enabled cellular network to support real-time vehicular communication” *IEEE Access*, vol. 7, pp. 13704 – 13714, Feb. 2019.
4. 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 Network*, vol. 31, no. 1, pp. 52-58, Feb. 2017.
5. 3GPP, TR 25.912, V13.0.0, “Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN)”.
6. J. Li, C. Natalino, D. Van, L. Wosinska, and J. Chen, “Resource management in fog enabled radio access network to suppot real-time vehicular services,” *IEEE/ACM International Conference on Fog and Edge Computing*, May, 2017.
7. 3GPP, TR 22.886, v15.1.0, “Study on Enhancement of 3GPP Support for 5G V2X Services”.

8. T. Taleb, A. Ksentini, and P. A. Frasngoudis, “ Follow- me cloud: when cloud services follow mobile users,” *IEEE Transaction on Cloud Computing*, vol. pp, no. 99, Feb., 2016.
9. R. Yu, J. Ding, X. Huang, M. zhou, S. Gjessing, and Y. Zhang, “Optimal resource sharing in 5G-enabled vehicular network: A matrix game approach, ” *IEEE Transaction on Vehicular Technology*, vol. 65, no. 10, pp. 7844-7856, Oct., 2016.
10. M. Jia, W. Liang, Z. Xu, and M. Huang, “Cloudlet load balance in wireless metropolitan area Network,” *IEEE INFOCOM*, Apr., 2016.
11. Y. Xiao, and M. Krunz, “ QoS and power efficiency tradeoff for fog computing networks with fog node cooperation,” *IEEE INFOCOM*, Apr., 2017.
12. L. Codeca, R. Frank, and T. Engel, “Luxembourg SUMO Traffic(LuST) Scenario: 24 hours of mobility for vehicular networking research,” *IEEE Vehicular Networking Conference* , 2015.
13. 3GPP, TR 36.885 v14.0.0, “Study on LTE support for V2X services”.
14. W. Viriyasitat, F. Bai, and O.K. Tonguz, “Dynamic of network connectivity in urban vehicular networks,” *IEEE Journal of Selected Area in Communication*, vol. 29, no. 3, Mar., 2011.
15. H. Liang, L. Cai, D. Huang, X. Shen, and D. Peng, “An SMDP-based service model for inter domain resource allocation in mobile cloud networks, ” *IEEE Transactions on Vehicular Technology*, vol.61, no.5, Jun., 2012.
16. R. Yu,Y. Zhang, S. Gjessing, W. Xia, and K. Yang,“Toward cloud-based vehicular networks with efficient resource management,” *IEEE Network*, vol. 27, no. 5, pp. 48-55, Oct., 2013.
17. Y. Yu, T. Chiu, A. Pang, M. Chen, and J. Liu, “Virtual machine placement for backhaul traffic minimization in fog radio access networks,” IEEE International Conference on Communication, May, 2017.
18. T. Orphanoudakis, E. Kosmatos, J. Angelopoulos, and A. Stavdas, “Exploiting PONs for Mobile Backhaul,” *IEEE Communication Magazine*, vol. 51, no. 2, pp. 27-34, Feb., 2013.
19. J. Li, X. Shen, L. Chen, J. Ou, L. Wosinska, J. Chen, "Delay-aware bandwidth slicing for service migration in mobile backhaul networks", *IEEE/OSA Journal of optical network and communication*., vol. 11, no. 4, pp. B1-B9, Dec. 2018.
20. P. Unell, “Right sizing RAN transport requirements,” Ericsson.