Chapter

Low-Latency Strategies for Service Migration in Fog Computing Enabled Cellular Networks

Jun Li, Xiaoman Shen, Lei Chen and Jiajia Chen

Abstract

This chapter presents a fog computing enabled cellular network (FeCN), in which the high user-mobility feature brings critical challenges for service continuity under stringent service requirements. Service migration is promising to fulfill the service continuity during mobility. However, service migration cannot be completed immediately and may lead to situations where the user-experience degrades. For this, a quality-of-service aware service migration strategy is proposed. The method is based on existing handover procedures with newly introduced distributed fog computing resource management scheme to minimize the potential negative effects induced by service migration. The performance of the proposed schemes is evaluated by a case study, where realistic vehicular mobility pattern in the metropolitan network of Luxembourg is used. Results show that low end-to-end latency for vehicular communication can be achieved. During service migration, both the traffic generated by migration and the other traffic (e.g., control information, video) are transmitted via mobile backhaul networks. To balance the performance of the two kinds of traffic, a delay-aware bandwidth slicing scheme is proposed. Simulation results show that, with the proposed method, migration data can be transmitted successfully within a required time threshold, while the latency and jitter for nonmigration traffic with different priorities can be reduced significantly.

Keywords: fog computing, mobile backhaul, low-latency service migration, bandwidth slicing, distributed fog resource management, vehicular communications

1. Introduction

The future cellular network is envisioned to support a variety of emerging mission-critical services such as industrial automation, cloud robotics, and safety-critical vehicular communications [1]. These mission-critical services usually have stringent requirements on latency, jitter, and reliability. In general, the required end-to-end latency is in the order of millisecond, while the probability that this requirement is met is expected to be as high as 99.999%. For example, the communication latency between sensors and control nodes for industrial automation has to be lower than 0.5 milliseconds, while that for virtual and augmented reality has to be lower than 5 milliseconds [1]. As an integral part of the cellular network, the transport network, referred to as the segment in charge of the backhaul of radio base stations and/or the fronthaul of remote radio unit, plays an especially important role to meet such a stringent requirement on latency.

The latency in transport networks can be reduced by moving the computing, storage, control, and network functions to the edge of the network, referred to as fog computing or edge computing, instead of performing all the functions in remote data centers. Fog computing is a new paradigm that can be integrated with the existing cellular networks (e.g., aggression points, base stations) to provide ultra-low-latency communication for time-critical services [2]. Thus, end users can access the applications (e.g., remote driving) hosted in fog nodes with low transport latency.

A fog node can be a terminal or a stand-alone node, which can be co-located with the existing cellular network infrastructure, such as router, gateway, aggregation points, and base stations (BS) [3]. Among them, BS (e.g., LTE evolved Nodes B) is a promising segment that can be integrated with fog nodes, which forms BS-Fog, giving rise to a new concept of fog enabled cellular networks (FeCNs). Such FeCN can be a promising candidate to support real-time services (e.g., real-time vehicular services) due to the ubiquitous access to radio access network (RAN) infrastructure as well as low communication delay enabled by fog computing. **Figure 1** illustrates the overall FeCN architecture.

In the FeCN, the BSs are responsible for providing network functions (e.g., handovers), whereas computational resources (e.g., computing and storage capability) can be provided by the fog nodes locally. One BS-Fog can cooperate with other BS-Fogs or cloud to allocate tasks dynamically. We design the FeCN with minimal changes on the current network architecture and reuse the existing interfaces. The S1 interface, which has been defined as the interface between the BS and the evolved packet core in LTE networks, is considered to realize the communications between the BS-Fog and cloud for the FeCN, while the interface X2, which has been defined as the interface between two BSs, is considered to support the communications among the BS-Fogs.

In the FeCN, it is possible to provide computing and storage capability closer to end users to support time-critical services. However, there are several challenges remaining to be addressed. Firstly, a fog node may be overloaded due to limited computing resources. Secondly, for high-mobility end users, the limited coverage of

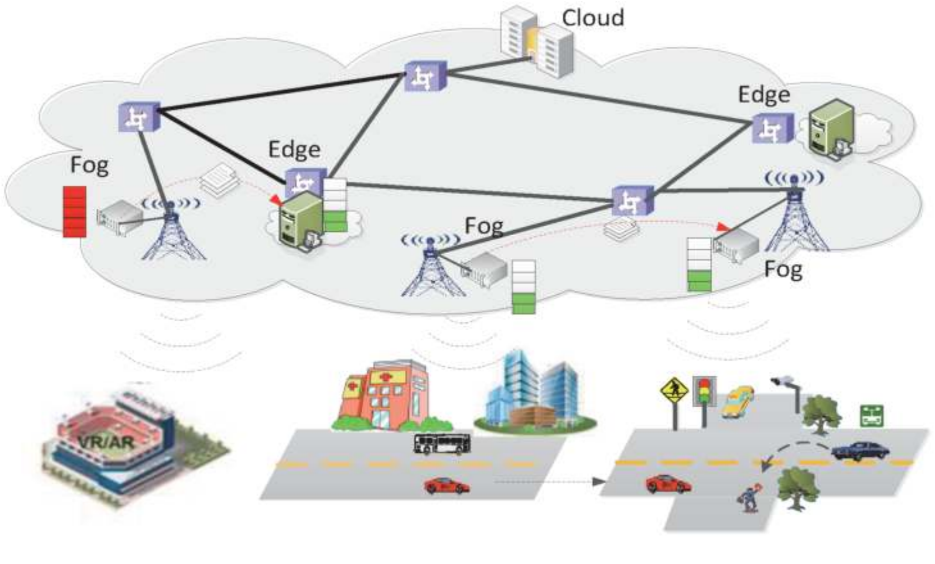

Figure 1.
Service migration in fog computing enabled cellular networks.

a single fog node may result in performance degradation in terms of latency when the user is moving far away from the serving node. In both cases, services need to be migrated to satisfy the quality-of-service (QoS) requirement. Service migration [4], which is referred to as relocating services from one fog node to another, has been proposed to deal with such challenges. As shown in **Figure 1**, once a single fog node is overloaded, service migration is triggered to offload this fog node to other fog nodes or edge servers. Besides, when users move from the area covered by one fog node to another, time-critical services are required to be migrated accordingly in order to follow the users' movement and maintain the service continuity with satisfying stringent latency requirements for these services. Since service migration may take time, which may result in service interruption, it needs to be handled carefully with consideration of the service requirements of differentiated traffic and the available network resources.

This chapter describes mechanisms and algorithms to deal with service migration in the FeCN. In Section 2, a QoS-aware service migration strategy is proposed. The strategy is based on the existing handover procedures, and the performance is studied with connected vehicle use cases. Following the proposed service migration strategy, in Section 3, a distributed fog computing resource management scheme is proposed to deal with limited computational resources at fog nodes. The scheme considers services with differentiated priority level, and in case of resource shortage, low-priority services may be migrated to other fog nodes to guarantee sufficient computation resources for the migrated high-priority services. The performance of the proposed schemes is evaluated by a case study, where realistic vehicle mobility pattern in the metropolitan network scenario of Luxembourg is used to reflect the real-world environment. Results show that low end-to-end latency (e.g., 10 ms) for vehicular communication can be achieved with typical vehicle mobility.

During service migration, both the traffic generated by migration (referred to as migration traffic) and other traffic (referred to as non-migration traffic, e.g., control information, video) are transmitted via mobile backhaul networks. To balance the performance of the two kinds of traffic, in Section 4, a delay-aware bandwidth slicing scheme is proposed in PON-based mobile backhaul networks. The proposed slicing scheme on one hand tries to guarantee the performance of the migration traffic, while on the other hand trying to minimize the negative impact on non-migration traffic. Simulation results show that, with the proposed method, migration data can be transmitted successfully within a required time threshold, while the latency and jitter for non-migration traffic with different priorities can be reduced significantly.

2. Service migration strategy in FeCN

Paper [5] presents a QoS-aware service migration strategy with connected vehicles as a use case to represent the high-mobility characteristic. In the context of FeCN, a vehicle is traveling while accessing a fog node. To maintain the service continuity, the vehicle needs to continue accessing the fog node and the provisioned running services through backhaul networks. When the vehicle travels away from the serving BS-Fog, the end-to-end (E2E) latency will increase, especially for the case that the vehicle does not have fixed routes. In order to keep the vehicle always accessing the fog services in one hop, the ongoing service can be migrated following the vehicle's trace. One straightforward strategy is to perform migration in combination with the handover procedure. This can guarantee one-hop access where a vehicle can directly access the services at its associated BS-Fog. However, since service migration cannot always be completed immediately, this may lead to a situation where users experience loss of service access. Therefore, frequent service migration is not always a good choice, and service migration needs to consider the QoS requirements of services.

2.1 QoS-aware service migration strategy

In view of the disadvantages of handover-based service migration strategy, we present a QoS-aware service migration strategy which is based on the existing handover procedure and considers the service QoS requirements. The key idea is to minimize the migration overhead while maintaining the E2E latency at an acceptable level under certain QoS requirements. E2E latency is a key QoS metric for real-time vehicular communication. When the E2E latency is at an unacceptable level, the performance of other metrics (e.g., reliability and packet drop) can also get worse. Therefore, in the proposed scheme, we focus on the metric of E2E latency to explain our proposed QoS-aware service migration strategy. The generalization of the proposed scheme to other QoS metrics is straightforward.

Figure 2 illustrates the service migration strategy, and **Figure 3** shows the communication protocol for the proposed QoS-aware service migration scheme. As illustrated, once the QoS requirements cannot be satisfied, the source BS-Fog node will trigger the service migration procedure and sends a Migration Request message that contains the information about the QoS requirements of the affected services to the target BS-Fog. After receiving the Migration Request message, the target BS-Fog will first make a decision whether to accept or not, and then sends back a Migration Request ACK to inform the source BS-Fog of its decision. If the request is agreed, the source Fog-BS will start implementing the migration.

In the proposed scheme, service migration can be achieved by pre-copy technique which is widely used for live virtual machine (VM) migration, as presented in [6]. The migration can be performed in two phases. In the first phase, the transfer of memory pages to the target BS-Fog is completed iteratively without suspending VM. In this phase, the UE still accesses the source BS-Fog (see blue dashed line in **Figure 3**). In the second phase when sufficient memory pages are transferred, the

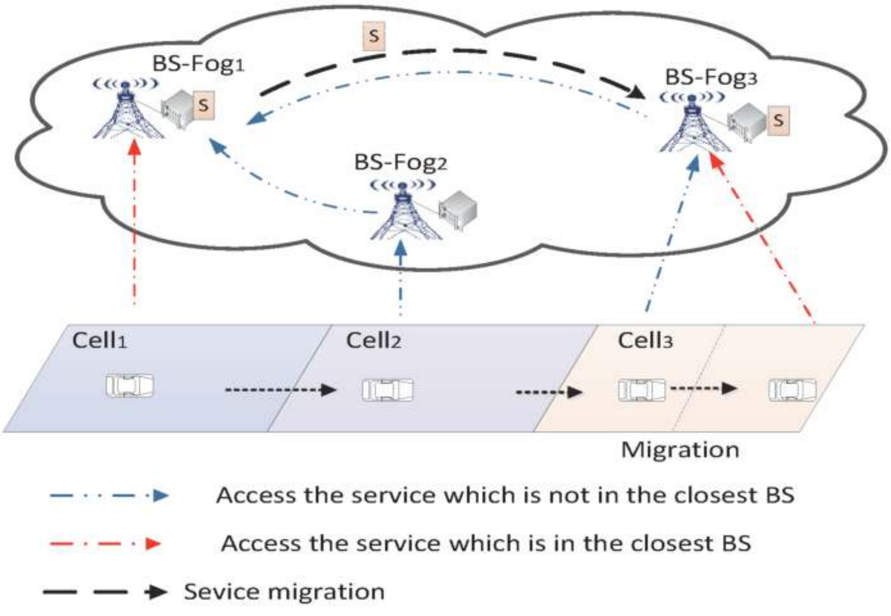
BS-Fog1
S
BS-Fog3
S
BS-Fog2
Cell1
Cell2
Cell3
Migration
Access the service which is not in the closest BS
Access the service which is in the closest BS
Sevice migration

Figure 2.
Illustration of a QoS-aware service migration [5].

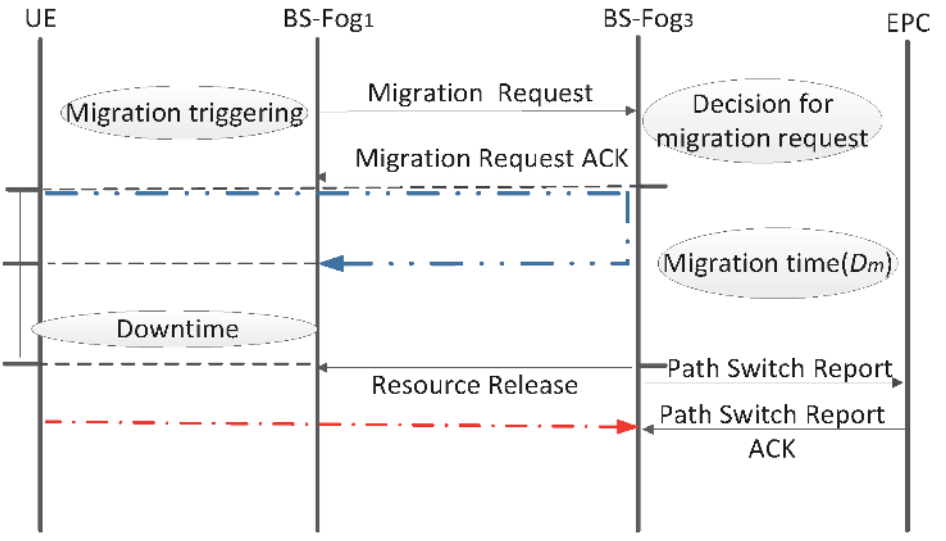

Figure 3.
Communication protocol to support QoS-aware service migration [5].

source BS-Fog will suspend the VM and finish transferring the remaining memory pages to the target BS-Fog. The services cannot be properly accessed during this period when the VM is suspended. Such duration is denoted as downtime. After the migration is completed, the UE can directly access the service in one hop (see red dashed line in **Figure 3**).

2.2 Performance evaluation

In this section, the performance of the proposed migration strategy is evaluated by using simulation. Here, the fog computing resources allocated for migrated services are assumed to be sufficient. This can be realized by designing efficient fog computing resource management schemes. We consider a case study for a small service area by using the realistic mobility pattern for the country of Luxembourg, and BS-Fog entities are evenly distributed over the city. The parameters used in this simulation are described in **Table 1**.

The performance of the proposed delay-aware migration (named Scheme 3) is evaluated in comparison to two benchmarks: no service migration (named Scheme 1) and always service migration (named Scheme 2). The details of Scheme 1 and Scheme 2 are introduced in paper [5]. The handover interruption time and wireless delay cannot be ignored and are not affected by migration strategies. Therefore, it is assumed that the uplink delay in the wireless segment is within 0.5 ms and the handover interruption time is a constant.

The average E2E latency for the three schemes is shown in **Figure 4**. Here, E2E latency consists of wireless access delay, interruption time during the handover, migration time, backhaul delay, and processing and queuing delays at the BS-Fogs. The transmission capacity is denoted as B and refers to the bandwidth allocated to the X2 interface between two Fog-BSs. It can be seen that the E2E latency in all three schemes decreases as B increases, especially for Scheme 1. This is because higher bitrate leads to shorter packet transmission time. Thus, the packet queueing delay can be reduced, resulting in smaller access latency. When B is high enough (e.g., B = 240 Mbps in **Figure 4**), the queueing delay is as minor as negligible. Meanwhile, in Scheme 2, the E2E latency is mainly affected by downtime (D_t), during which the ongoing services need to be suspended.

Parameter	Value
Coverage of the country of Luxembourg	155 km^2 [7]
Total number of vehicles	5500 [7]
Vehicle density	35.5 per km^2
Bitrate of traffic generated by the vehicles	(2 Kbps, 10 Mbps)
Size of the applications encapsulated in VMs	(10,100) Mbits
Link speed in upstream and downstream in PONs	10 Gbps
Repeated times of simulations	10
Simulation time	1000 s
Coverage of a single BS-Fog	1 km^2
Handover interruption time	20 ms [8]
Wireless delay	0.5 ms
Vehicle speed	(1, 45) m/s [7]
Processing time at active node	0.2 ms
Number of ONUs in each PON	16
Number of PONs	10
Confidence level	95%

Table 1.
Simulation parameters.

In Scheme 3, the service migration is triggered once the latency exceeds the threshold (e.g., 10 ms in **Figure 4**). Scheme 3 is considered as a tradeoff between Scheme 1 and Scheme 2. When *B* is low, Scheme 3 performs similar to Scheme 2, that is, frequent migrations are performed in both schemes (see **Figure 5**). As B increases, there are fewer and fewer migrations triggered in Scheme 3. Thus, the E2E latency is less and less affected by downtime. Therefore, Scheme 3 has similar

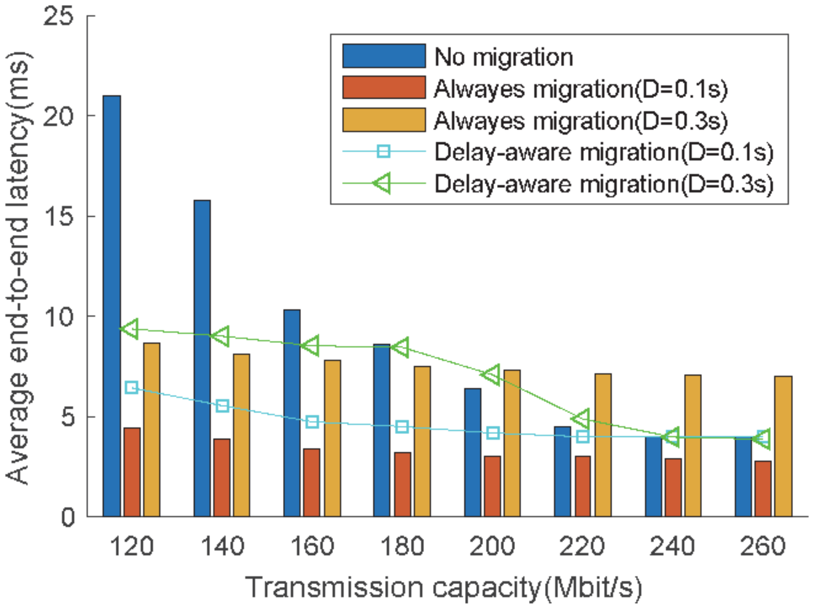

Figure 4.
Transmission capacity (B) in the backhaul versus average end-to-end latency.

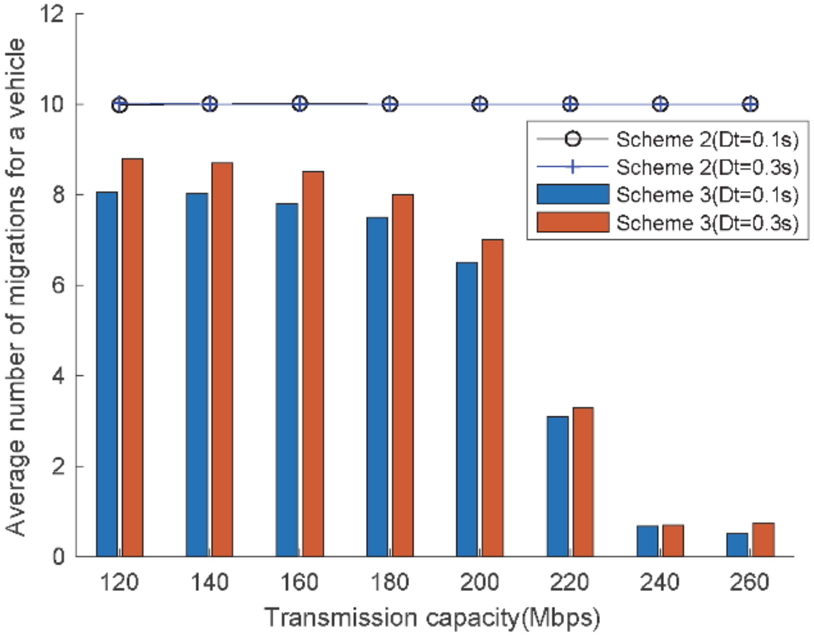

Figure 5.
The average number of migrations for a vehicle as a function of transmission capacity.

performance with Scheme 1 when *B* is high and performs better than Scheme 2 when downtime is large (e.g., 0.3 s), as shown in **Figure 5**.

3. Distributed fog computing resource management

During the service migration procedure, sufficient computation resource is needed to host the migrated services. Also, provisioning resource for the migrated real-time services needs to be completed as soon as possible to minimize the service interruption. Otherwise, the services have to stay at the source fog node, which may increase the access delay. On the other hand, one single fog node only has limited amount of computation resource, and its load is highly burst due to the mobility of vehicles. The service migration strategy discussed in the previous section assumes sufficient resources for the migrated services, which may not always hold in practice. Thus, it is important to employ efficient resource management scheme, especially in the scenarios with fast mobility and high load.

To guarantee sufficient computation resource for the migrated vehicular services and thus reducing the service migration blocking caused by the lack of computing resources, a distributed fog computing resource management scheme is proposed [9]. Two distributed fog computing resource management schemes, namely, fog resource reservation (FRR) and fog resource reallocation (FRL), have been considered. In FRR scheme, a certain amount of computation resources for vehicular services in each fog node are reserved based on the predicted vehicular traffic load. The performance of this scheme depends on the traffic flow prediction methods. Overestimating leads to low resource utilization, while underestimating significantly decreases one-hop access probability for high-priority (HP) vehicular services (e.g., remote driving, pre-crash sensing warning). For FRL, the key idea is to release part of fog resources used for low-priority (LP) services (e.g., online game, navigation, sign notification) by suspending those services and reallocate them to HP services. However, in such a scheme, the one-hop access probability for

LP services may be affected, especially when traffic load is high. In fact, not all the LP services (e.g., online game) need to have one-hop latency requirement and local awareness. Therefore, such services can be placed in its neighboring fog nodes with low load.

3.1 Distributed computing resource management

As introduced above, in both FRR and FRL, each BS-Fog node manages its resource independently without cooperating each other. Once it is overload, the one-hop access probability for LP services may be affected. Considering that some LP services also have one-hop latency requirement, we proposed an online resource management (ORM) scheme, in which fog nodes can cooperate to allocate resources for both HP and LP services [10]. In comparison to FRL, LP services are not suspended, but instead, they are migrated to other neighboring BS-Fogs. This can guarantee the resource requirement of the to-be-migrated HP services as well as LP service continuity. The details are as follows.

When a vehicle moves from one BS-Fog's coverage area to another, both handover and migration of ongoing services need to be handled. As shown in **Figure 6**, once the handover is triggered, the source Fog-BS sends a Migration Request message to the target BS-Fog, which includes the information of the requested resources. After receiving this message, the target Fog-BS will make a decision

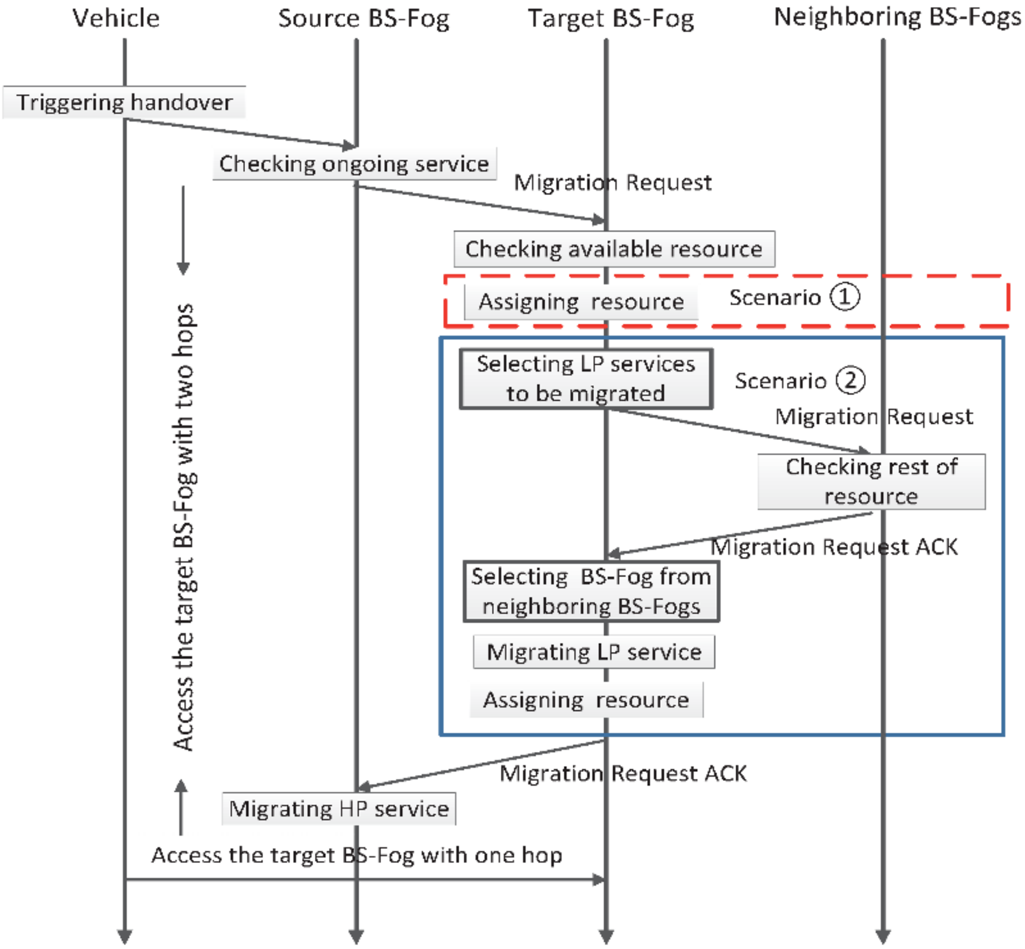

Figure 6.
Illustration of resource management in service migration procedure [10].

whether to accept the request or not according to the resource management strategy.

In this part, we consider two scenarios used for the resource management strategy. In the first scenario, there is sufficient available resource in the target BS-Fog, the request is approved, and the corresponding resource is assigned to the migrated HP services (see Scenario ① in red dotted box in **Figure 6**).

In the second scenario, the target BS-Fog does not have enough available resource. In such a case, the target BS-Fog selects ongoing LP services and migrates them to its neighboring BS-Fogs. After that, the resources of the selected LP services are released for HP services. As shown by Scenario ② in blue dashed box in **Figure 6**, the procedure is divided into two phases. In the first phase, LP services to be migrated are selected according to service selection strategies, while in the second phase, the target BS-Fog selects one of its neighboring BS-Fogs to host the migrated LP services. For such a purpose, the target BS-Fog firstly broadcasts a Migration Request message to its neighboring BS-Fogs that have direct communication with the target BS-Fog via X2 interface. Once the neighboring BS-Fogs receive the Migration Request messages, they will check their available resource and then send Migration Request ACK to the target BS-Fog. A final decision on the selection of neighboring BS-Fog is made by the target BS-Fog immediately. When the selected LP services complete the migration and release their resources, the target BS-Fog sends a Migration Request ACK to the source BS-Fog for HP service migration. If there are no available LP services or neighboring BS-Fogs, the to-be-migrated HP service has 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 obviously result in a higher access delay. The selection strategies of LP services and neighboring BS-Fogs are discussed in the following section.

3.1.1 Low-priority service selection algorithm

In the proposed scheme, the selected LP services are migrated from the target BS-Fog to the selected neighboring BS-Fog, which will be accessed via backhaul networks. In such a procedure, a certain amount of backhaul bandwidth is needed for service migration and running the selected LP services. To minimize the required bandwidth, we propose a LP service selection algorithm to minimize the communication cost. The amount of the required bandwidth resources for each LP service is firstly counted. Then, the LP services whose available computing resources are larger than the requested amount are selected. Among these services, the one with the lowest communication cost is finally selected. If there is no service that satisfies the requirement alone, more than one service can be selected. In order to avoid ping-pong effect in LP service migration, LP services are only allowed to be migrated once.

3.1.2 Neighboring fog-BS selection algorithm

Once the LP services to be migrated are decided, neighboring BS-Fogs that will host the migrating services need to be decided. The decision needs to consider the QoS of those services since after migration, the services will be 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 transmission delay between the target BS-Fog and its neighboring BS-Fog should thus be smaller than the budget of backhaul delay. The key idea of the proposed algorithm is to select the neighboring BS-Fog with the most available resources under the acceptable backhaul delay of the to-be-migrated LP service.

3.2 Performance evaluation

In this section, the performance of the proposed scheme is investigated through simulation. The realistic mobility pattern for the city of Luxembourg is used. **Figure 7** shows the vehicular traffic profile in Luxembourg, which varies in time over a day. Also, the vehicular traffic is spatially diverse. For example, the inserted chart (a) in **Figure 7** shows the numbers of vehicles in each coverage area of BS-Fogs at 8:00 am, while the inserted chart (b) in **Figure 7** 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.

Each vehicle is assumed to only require 1 HP service (i.e., safety-related service). The data traffic distribution is proportional to the vehicular traffic. 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). 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 **Figure 7(a)**), while the LP service arrival rate is distributed among BS-Fogs evenly. Both HP and LP services arrive according to Poisson Procedure. The parameters are shown in **Table 2**.

The performance of the proposed scheme is investigated in terms of access probability for HP services and service unavailability for LP services. Here, one-hop access probability is defined as the ratio of the one-hop 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 discussed, 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. Transmission capacity B becomes the main factor that affects the delay performance.

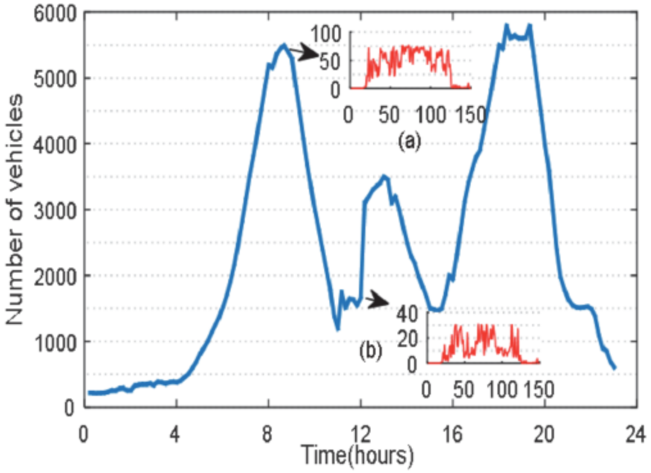

Figure 7.
Vehicular traffic profile in Luxembourg [10].

Parameter	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 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)
Data rate generated by end users (bps)	(2 K, 10 M) [11]
Amount of data encapsulated in application VMs (Mbits)	(10, 100)
Downtime in live VM migration (millisecond)	20
Confidence level	95%

Table 2.
Simulation parameters [10].

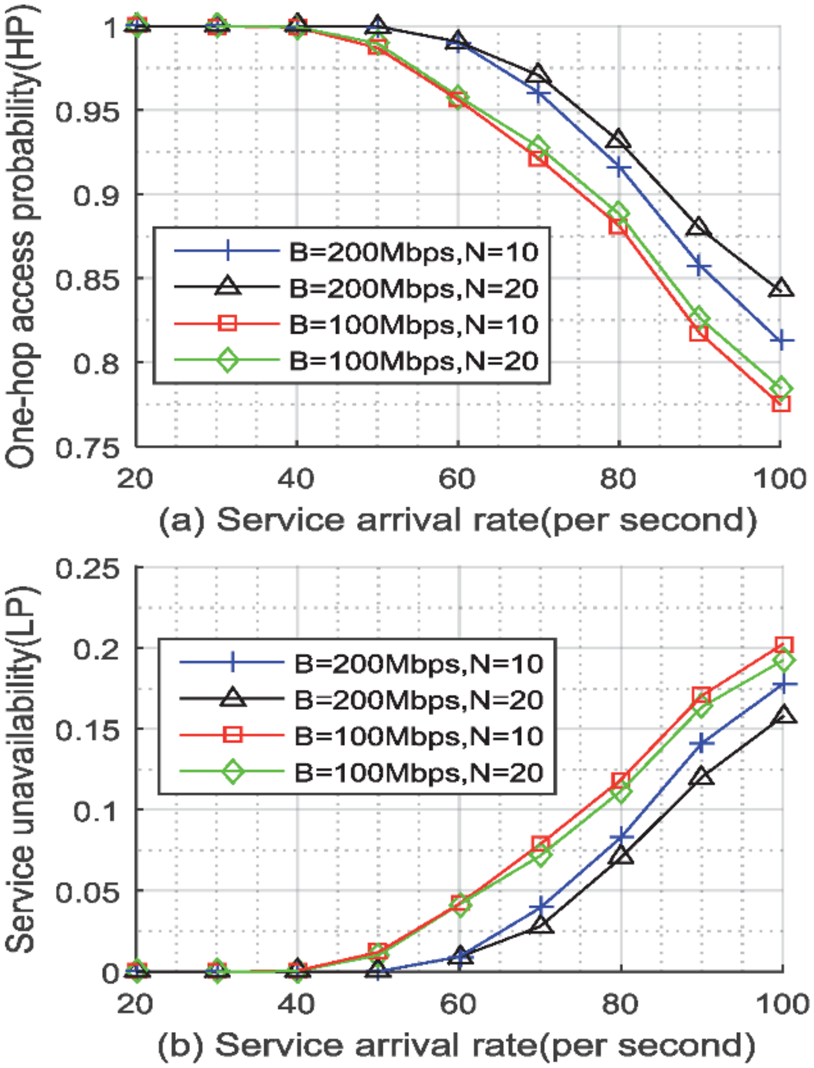

Figure 8.
One-hop access probability and service unavailability versus service arrival rate.

Figure 8(a) shows that one-hop access probability for HP services versus service arrival rate. As expected, due to the fact that the migration delay for both HP and LP services can be reduced by enlarging backhaul capacity, larger transmission capacity (*B*) leads to higher one-hop access probability, which benefits the reduction of migration time. As also shown, when B is large (e.g., *B* = 200Mbps), a higher number of neighbors (*N)* lead to a better one-hop access probability, while when *B* is small, the increase of *N* has little impact on the one-hop access probability for HP services. This is because with a smaller *B*, the backhaul delay between the target and the neighboring BS-Fogs is higher, and the number of neighboring BS-fogs that satisfy the latency requirement decreases, even when N is high. **Figure 8(b)** shows LP service unavailability as a function of service arrival rate. Similarly, increasing backhaul capacity *B* leads to a lower migration delay and thus a reduced service unavailability.

We further compare the performance of the proposed ORM scheme with two benchmarks. The first benchmark is based on the principle of first come first served (FCFS), in which HP and LP services are treated equally. The second benchmark is FRR where a certain amount of resource is reserved for HP. **Figure 9(a)** shows that, in comparison to FCFS and FRR, the one-hop access probability of HP services for ORM is higher when *B* is large (e.g., *B* = 200 Mbps). When *B* decreases to 100 Mps,

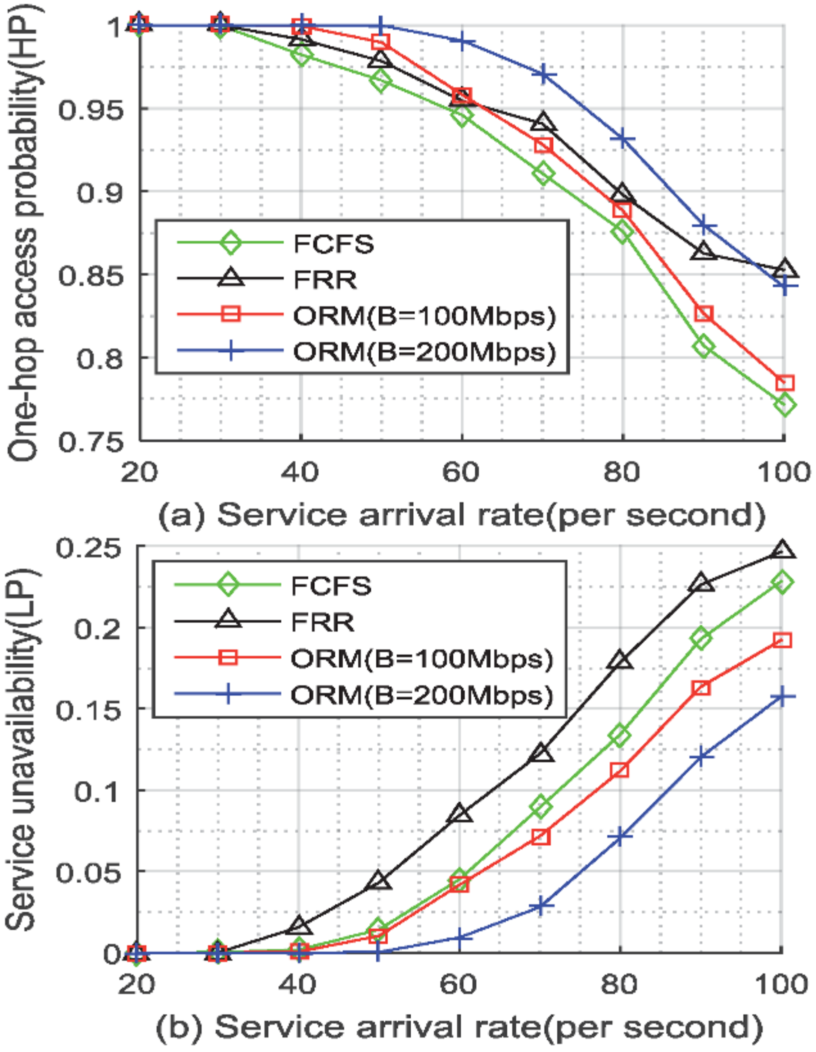

Figure 9.
One-hop access probability and service unavailability versus service arrival rate.

the one-hop access probability for ORM shows different results based on the service arrival rate. When the service arrival rate is below 60 arrivals per second, it is higher than that for both FCFS and FRR, while when the service arrival rate increases above 60, the one-hop access probability remains higher than that for FCFS but lower than that for FRR. The reason is that more LP services need to be migrated to the neighboring BS-Fogs when the service traffic arrival rate increases, which results in high migration traffic, thus increasing the migration delay.

Figure 9(b) shows that the service unavailability for LP services increases with service arrival rate for all methods, and that of ORM is shown to be the lowest in comparison with the two benchmarks. As also can be seen, transmission capacity B has impacts on the service unavailability for LP services. With a higher B = 200 Mbps, ORM demonstrates significant lower LP service unavailability at all service arrival rates, that is because a larger B results in a shorter time used for migrating LP services; thus service unavailability can be reduced. When B is decreased to be at 100 Mbps, service arrival rate is shown to affect the performance of ORM. With a service rate under 60 arrivals per second, ORM has very similar performance regarding LP service unavailability in comparison with that for FCFS. When the service rate is above 60, ORM demonstrates the advantages over FCFS with a lower LP service unavailability. The reason is that, in the case of FCFS, the migration delay is mainly introduced by the waiting time for the available resources. Since a larger arrival rate leads to a longer waiting time, the performance of FCFS degrades. As another general observation, though ORM demonstrates better performance, the migration delay is shown as an important factor that affects the performance in terms of one-hop access probability and service unavailability, indicating that migration delay needs to be properly handled.

4. Bandwidth slicing in mobile backhaul networks

In previous sections, service migration strategy and fog computing resource management scheme have been investigated to support real-time vehicular services. As indicated, mobile backhaul capacity is a main factor that affects the performance of the service migration schemes. Regarding this matter, passive optical network (PON) based mobile backhaul network can be considered to support FeCN due to its high capacity.

In PON-based mobile backhaul network supporting the FeCN, the BS-Fogs are integrated with optical network units (ONUs) through high-speed Ethernet interface, shown in **Figure 10**. The traffic generated by service migration, named migration traffic, is transmitted together with the non-migration traffic. On the one hand, the size of the data generated by service migration can be up to hundreds of MBytes [12]; thus, such migration traffic can be fragmented into Ethernet frames at ONUs and should be carefully handled by the optical line terminal (OLT) that is located at the central office. On the other hand, service migration is usually deadline-driven and has to be handled within a certain time limit. We define migration delay as the time duration from the moment when a migration is initiated until the affected service is successfully transferred to the target fog node. In order to minimize the service interruption, the migration delay should be lower than a pre-defined time threshold, which is usually specified in the QoS requirements and in a magnitude of seconds [13]. The non-migration traffic includes data generated by multi-type applications, which usually have different QoS requirements but less stringent in terms of latency, packet loss ratio, etc. These different types of the non-migration traffic can be queued independently and scheduled with different priorities according to medium access control (MAC) protocol in Ethernet PON [14].

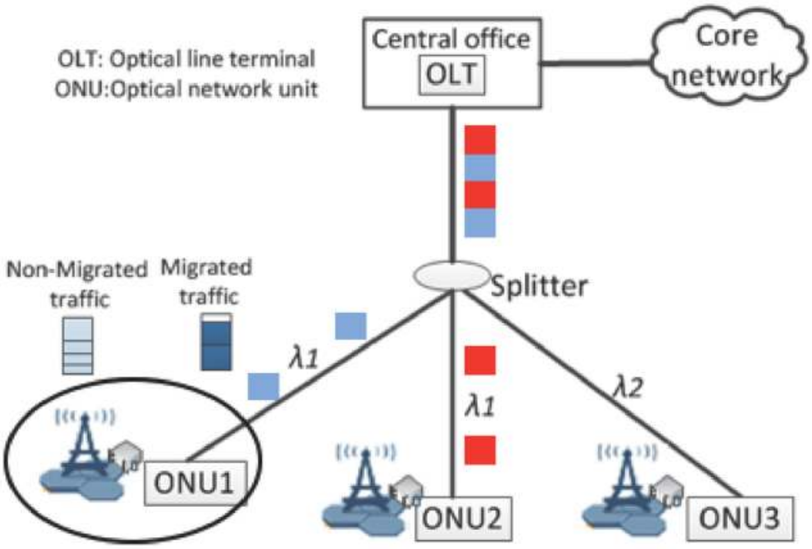

Figure 10.
PON-based mobile backhaul for the FeCN.

Likewise, the migration traffic can be also queued based on their deadline requirements.

Time and wavelength division multiplexing (TWDM-PON) has been regarded as a promising candidate for next-generation PON 2 (NG-PON 2), where dynamic bandwidth allocation (DBA) mechanisms are performed on each wavelength for efficient channel sharing [15]. In a classical DBA algorithm, migration traffic and non-migration traffic are scheduled with no distinction. Each ONU reports to the OLT the amount of data that needs to be transmitted in the next cycle and then receives a grant message. According to the information contained in the grant message, including the allocated time slots and polling cycle for transmission, each ONU transmits data based on the principle of FCFS without considering the traffic priorities. Once a service migration occurs, a large volume of migration data arrives, and more than one polling cycle may be needed for the transmission. In such a case, the non-migration data that arrives after the migration data has to wait before being transmitted, thus experiencing a long queuing delay, leading to high latency and jitter. One way to deal with this is to assign higher priority to non-migration traffic that arrives after the migration traffic based upon the existing QoS management mechanism in Ethernet PON. In such a case, the delay for the non-migration traffic can be reduced significantly. However, short delay for transmitting the migration traffic that comes after may not be guaranteed, particularly when the load of non-migration traffic is high.

For balancing the transmission of migration traffic and non-migration traffic, we propose a dynamic bandwidth slicing (DBS) scheme with a bandwidth slicing mechanism and a tailored delay-aware resource allocation algorithm. We present the DBS in the following part together with simulation results.

4.1 Bandwidth slicing mechanism

A bandwidth slicing scheme for service migration in PON-based mobile backhaul networks is proposed [16]. In the scheme, the cycle time can be cut into several slices dynamically, which are provisioned to different kinds of traffic (i.e., migration traffic and non-migration traffic with different priorities). Such a mechanism is based on the report-grant mechanism, as introduced in the previous section. In each polling cycle, request messages are first sent from each ONU to the OLT containing the information about their data size and delay requirements. According to such information, the polling tables (see **Figure 10**) for both non-migration data and migration data are then updated. Once service migration occurs, the lengths of the slices for both migration and non-migration traffic can be calculated by the resource management controller located at the OLT with the bandwidth allocation algorithms and the information contained in the polling table.

Figure 11 illustrates in more detail the proposed bandwidth slicing mechanism. Following similar principles of the considered DBA algorithms, in each polling cycle, the time slots in Slice 1 and Slice 2 are allocated to the non-migration traffic and the migration data, respectively. As mentioned, the lengths of the slices are decided dynamically based on the traffic and resource allocation algorithms. In the case where there is no migration data to be transmitted, the proposed mechanism performs in the same way as the classical DBA mechanism with a FCFS fashion. Note that the same principle as the proposed mechanism applies to both the upstream and downstream.

4.2 Delay-aware resource allocation algorithm

Following the proposed bandwidth slicing mechanism, a tailored delay-aware resource allocation algorithm is proposed with the aim to transmit migration traffic within the required deadline by cutting the large-size migration traffic into small pieces and transmitting them at each polling cycle. Such an algorithm is implemented at the end of Slice 1 in each polling cycle. First, the OLT acquires the information of the amount of the non-migration traffic that ONUs need to send in the next polling cycle and their priorities contained in the Report messages. If service migration occurs, such messages also contain the information of the sizes and deadline of the migration data that ONUs need to send. Then, the length of the slice for the migration traffic (S^m) can be calculated by OLT, and the migration mechanism will be triggered. In such a case, as illustrated in **Figure 11** the bandwidth slicing mechanism, the polling cycle will be divided into two slices for non-migration traffic (Slice 1) and migration traffic (Slice 2), respectively. The time slots in Slice 1 are allocated to ONUs for non-migration traffic according to their priority level, while the time slots in Slice 2 are allocated to the migration traffic according to the ascending order of the deadline for finishing migration. Here time slot allocation is purely based on incoming traffic, and with high migration traffic load, it can be expected that the time slots are monopolized. To avoid such situation,

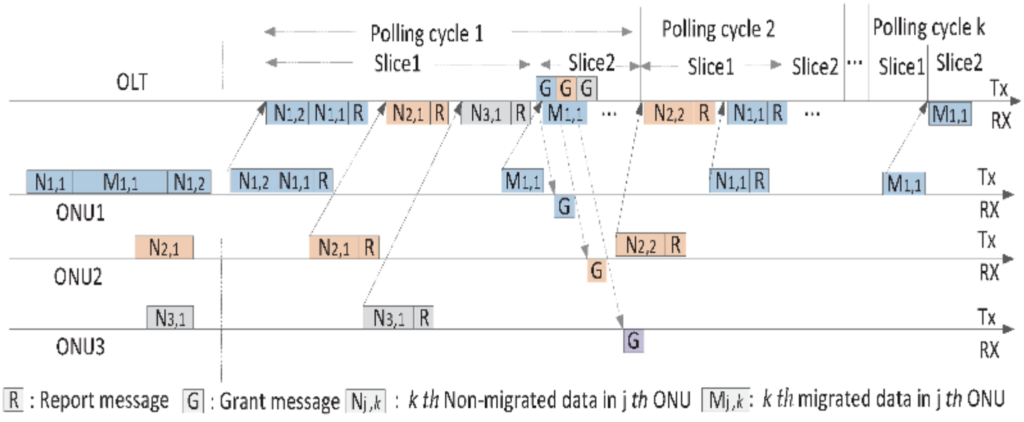

Figure 11.
Illustration of the proposed bandwidth slicing mechanism.

Symbol	Explanation
$R_i^{j,k}$	Required time slot for the remaining of the k^{th} migration traffic from the ONUj in the i^{th} polling cycle
$D^{j,k}$	Deadline for the k^{th} migration traffic sent by ONUj
$T_i^{j,k}$	Remaining time for transmitting the k^{th} migration traffic sent by ONUj, which starts from the beginning of the i^{th} polling cycle to its deadline
K_j	Number of the migration tasks in the ONUj
B^R	Transmission time for sending each report and grant massage
B^G	Guard time with a fixed value for the slice of the migration traffic in the i^{th} polling cycle
$B_i^{j,l}$	Length of the requested time slots for the non-migration traffic with the l^{th} priority at the in the i^{th} polling cycle
H_j	Number of priority levels for the non-migration traffic at the ONUj

Table 3.
Explanation of symbols.

a hard threshold θ is introduced as the percentage of the total time slots that can be allocated to migration traffic within each polling cycle.

In the proposed algorithm, the calculation of the lengths of slices in each polling cycle plays a very important role and is described in more detail in the following part. The symbols used are explained in **Table 3**. In each polling cycle (e.g., i^{th} polling cycle), the required time slots ($G_i^{j,k}$) for the k^{th} migration traffic from the ONUj can be calculated by

$$G_i^{j,k} = R_i^{j,k} / \left\lceil T_i^{j,k} / W_{max} \right\rceil \tag{1}$$

In the proposed resource allocation algorithm, the length of the polling cycle (W) varies dynamically with the traffic load. Thus, when calculating the required time slots in the current polling cycle, the maximum polling cycle (W_{max}) is used to guarantee that the transmission of the whole migration traffic can be finished before the deadline. Here, the time unit (μs) is used to represent the length of the time slots and polling cycles. Then, the total length of the time slots granted for the migration traffic ($TG_i^{j,k}$) can be calculated by

$$TG_i^{j,k} = \sum_{j=1}^{N} \sum_{k=1}^{K_j} G_i^{j,k} \tag{2}$$

To guarantee the fairness between the migration and non-migration traffic, the length of the granted time slots cannot exceed the maximum allowed length of the time slots in this polling cycle, which can be calculated by.

$$R_i^m = \left(W_{max} - N \times \left(B^R + B^G\right)\right) \times \theta \tag{3}$$

The maximum length of the allocated time slots for the migration traffic is set by the threshold (θ) for the slice of the migration traffic ($\theta \in [0, 1]$). Thus, the time slots granted for the migration traffic in the i^{th} polling cycle can be calculated by

$$TG_i^{j,k} = \begin{cases} TG_i^{j,k}, & TG_i^{j,k} < R_i^m \\ R_i^m, & TG_i^{j,k} \geq R_i^m \end{cases} \tag{4}$$

In the i^{th} polling cycle, the length of the slice for the migration traffic (S_i^m) equals to the total length of the granted time slots ($TG_i^{j,k}$). Then, for the non-migration traffic (C_i^t) with different priorities, the total length of the granted time slots can be calculated by

$$C_i^t = \sum_{j=1}^{N} \sum_{l=1}^{H_j} B_i^{j,l} \tag{5}$$

For the non-migration traffic, the maximum available time slot (C_i^a) can be calculated by

$$C_i^a = W_{max} - S_i^m - N \times \left(B^R + B^G\right) \tag{6}$$

Then, the granted time slots can be calculated by

$$C_i^t = \begin{cases} C_i^t, & C_i^t < C_i^a \\ C_i^a, & C_i^t \geq C_i^a \end{cases} \tag{7}$$

Similarly, in the i^{th} polling cycle, the length of the slice for the non-migration traffic (S_i^n) equals the total length of the granted time slots (C_i^t).

4.3 Performance evaluation

The performance of the proposed algorithm has been investigated through simulation and is also further compared with two benchmarks that are based on the conventional DBA algorithms [15]. In Benchmark1, the migration traffic and non-migration traffic follow FCFS, while in Benchmark2 a higher priority is given to the non-migration traffic. Besides, in both benchmarks, the non-migration traffic is assumed with two priority levels (e.g., low and high). **Table 4** summarizes the main parameters.

As mentioned, a threshold θ is introduced to regulate the allocation of time slots within each polling cycle. It has been shown that with θ set to 1, 85% of the time slots

Parameter	Value
Number of ONUs in a PON	8
Propagation delay in the optical links	5 μs/km
Packet size of Ethernet frame (bytes)	(64, 1518)
Guard time between two consecutive time slots	1 μs
Buffer size (Mbytes)	100
Amount of data encapsulated in application VMs (Mbits)	(10, 50)
Deadline for the migration data (second)	(1, 5)
Confidence level	95%

Table 4.
Simulation parameters [16].

in the overall polling cycle are allocated to the migration traffic at load = 0.9. In the following simulation, different values have been chosen to illustrate the impacts.

Figure 12 illustrates the migration success probability (MSP) versus traffic load. Here, MSP is defined as the ratio of the amount of services migrated before the required deadline over the total amount of services that are migrated. As shown, MSP decreases with increasing traffic load in Benchmark2 and DBS with different thresholds. At a lower traffic load (e.g., less than 0.4), all three schemes achieve high MSP, while when the traffic load is above 0.4, MSP starts to decrease. For Benchmark1, MSP shows very minor changes when traffic load increases with almost 1 when traffic load is 0.9. This is because according to the principle of FCFS, large-size migration traffic can be fully transmitted in several cycles once the migration starts. On the other hand, in Benchmark2, the MSP for the migration traffic decreases sharply due to the fact that non-migration traffic is prioritized. As shown, MSP is as low as 0.1 when the traffic load is 0.9. For DBS, all the migration tasks can be performed within the time constraints when the traffic load is under 0.5. When the traffic load is higher than 0.6, the MSP performance is mainly affected by the threshold of the allowable time slots that can be used for the migration traffic and increases as the threshold increases. For example, with the threshold set to 0.5, MSP can be up to 0.98 at load of 0.7.

The average E2E latency for the non-migration traffic with high priority is shown in **Figure 13(a)** as a function of load. It can be seen that the proposed scheme and two benchmarks have a similar trend, that is, the average latency increases with traffic load. Among the three schemes, Benchmark1 always has the highest average E2E latency which can be up to 100 ms when traffic load is 0.9. Such large latency may not be accepted for time-critical services (e.g., interactive voice). Compared with Benchmark1, the average E2E latency in Benchmark2 increases more slowly even when the traffic load is high. The reason is that the non-migration traffic in Benchmark2 has high priority to be transmitted. Compared with Benchmark1 and Benchmark2, the average E2E latency for DBS is much lower, which is less than 1 ms when the traffic load is lower than 0.5 and remains to be less than 10 ms even at high traffic load. This is due to the fact that the migration traffic can be transmitted in multiple cycles by partitioning those with large size into multiple smaller pieces; thereby the non-migration traffic that

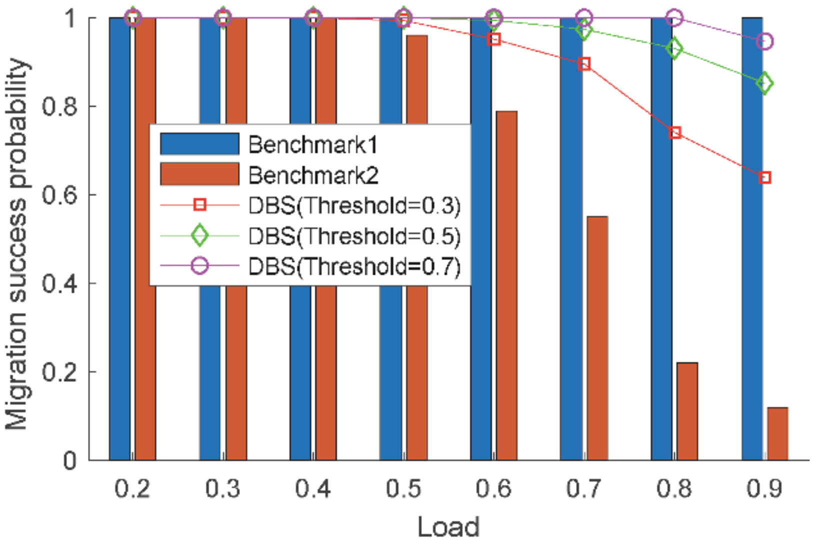

Figure 12.
The migration success probability versus traffic load.

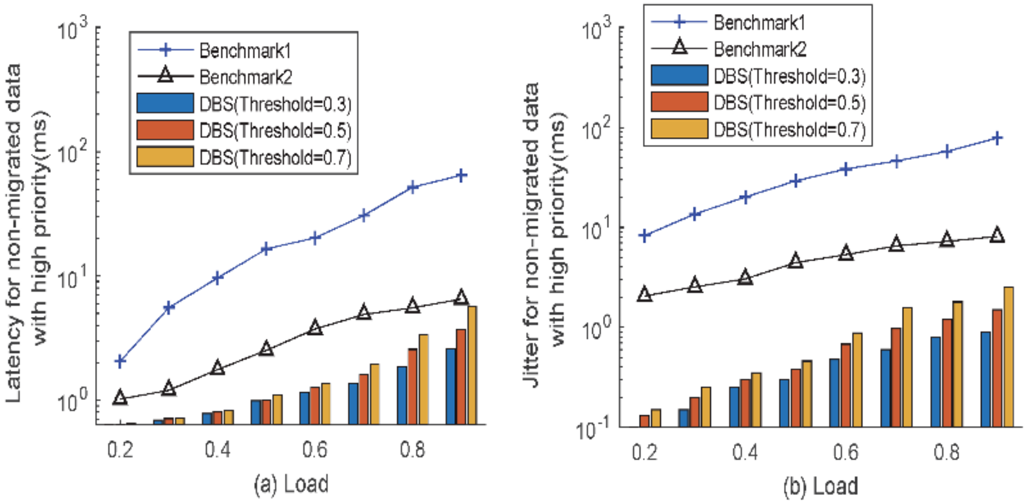

Figure 13.
(a) The average latency and (b) jitter for the non-migration data with high priority.

arrives after or during the transmission of migration traffic does not need to wait too long for transmission. Furthermore, when the threshold increases, the allocated time slots for transmitting the non-migration traffic decreases; thus the average E2E latency increases. Regarding the jitter for the non-migration data with high priority, a similar trend as the average E2E latency can be found, as shown in **Figure 13(b)**.

The average E2E latency of the low-priority non-migration data with different traffic loads is shown in **Figure 14(a)**. Similar to high-priority non-migration traffic, a general trend is that E2E latency increases with the traffic load for all schemes. When the traffic load is low, all kinds of traffic can be assigned with sufficient time slots, while when the traffic load increases, the average E2E latency for the low-priority non-migration traffic increases sharply because of its large queueing delay. More specifically, Benchmark1 has the highest average E2E latency among the three schemes. Compared with Benchmark1, the average E2E latency in Benchmark2 is much lower. And when traffic load is low (e.g., less than 0.6), the average latency of the low-priority non-migration traffic in DBS is the lowest, which is smaller than 2 ms. However, the latency of DBS increases quickly with larger thresholds

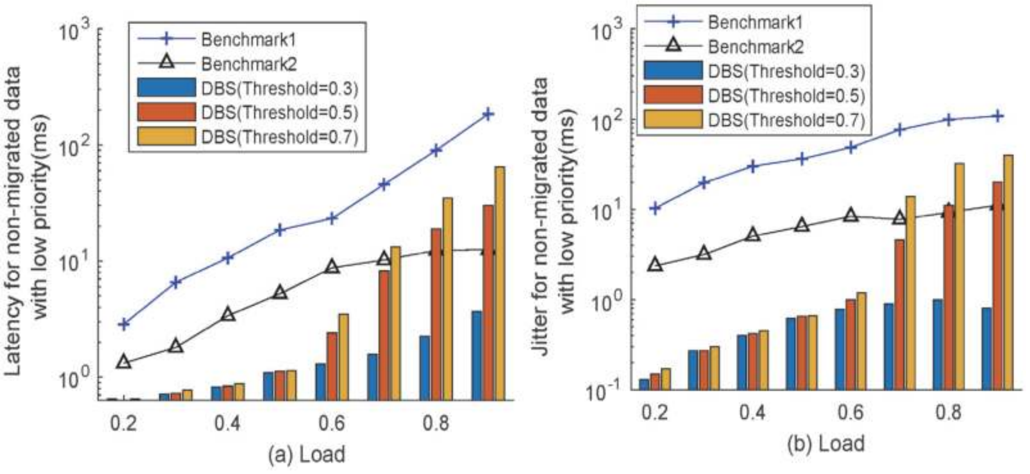

Figure 14.
The average (a) latency and (b) jitter for the non-migration data with low priority.

(e.g., larger than 0.5) and exceeds the level observed for Benchmark2 when traffic load is high (e.g., higher than 0.7). The reason is that the time slots are prioritized for the non-migration traffic with high priority and the migration traffic; thus the low-priority non-migration has to wait. The jitter shows similar trend as the E2E latency, as shown in **Figure 14(b)**.

5. Conclusions

This chapter presents a concept of fog enabled cellular networks (FeCN), where computing, storage, and network functions are provisioned closer to end users in order to improve the service QoS. In addition, to guarantee service continuity and QoS, service migration is introduced to ensure that services always follow the end users through migration from the current fog server to the target one. A QoS-aware service migration strategy based on the existing handover procedures is firstly proposed to balance the benefits and costs of migration. A case study using a realistic vehicle mobility pattern for Luxembourg scenario is carried out through simulation to evaluate the performance of the proposed schemes. Results show that low end-to-end latency (e.g., 10 ms) for vehicular communication can be achieved, while the total number of migrations for each user in the whole journey can be decreased significantly.

To deal with the situation that the target fog node does not have enough resources to support the migrated services, a distributed fog computing resource management scheme is introduced. The scheme purposely selects low-priority (LP) services and migrates those services to carefully selected neighboring fog nodes so that QoS for high-priority (HP) migration services can be served at the target fog node. LP service selection algorithm is proposed to minimize the migration costs for those services, and neighboring fog node selection algorithm is proposed for selecting a fog node that provides enough resources for LP services with also satisfied QoS. Simulation results show that the one-hop access probability for HP services increases significantly, while the service unavailability for LP services can also be well reduced.

During service migration, both the traffic generated by migration and other traffic (e.g., control information, video) are transmitted via mobile backhaul networks. To balance the performance of the two kinds of traffic, we propose a delay-aware bandwidth slicing mechanism in PON-based mobile backhaul networks. The method tries to guarantee the transmission of migration traffic within the deadline, while at the same time minimizing the negative impact on non-migration traffic. Simulation results show that migration data can be transmitted successfully in a required time threshold, while the requirements of latency and jitter for non-migration traffic with different priorities can be well satisfied.

Acknowledgements

This work is supported in part by SJTU State Key Laboratory of Advanced Optical Communication System and Networks Open Project 2018GZKF03001, Swedish Research Council (VR) project 2016-04489 “Go-iData,” Swedish Foundation for Strategic Research (SSF), and Chalmers ICT-seed grant.

Author details

Jun Li[1], Xiaoman Shen[2], Lei Chen[3]* and Jiajia Chen[1]*

1 Chalmers University of Technology, Göteborg, Sweden

2 Zhejiang University, Hangzhou, China

3 RISE Research Institutes of Sweden, Göteborg, Sweden

*Address all correspondence to: lei.chen@rise.se and jiajiac@chalmers.se

IntechOpen

References

[1] 3GPP TS 122261, Service requirements for next generation new services and markets; V 15.5.0, 2019

[2] Chiang M, Zhang T. Fog and IoT: An overview of research opportunities. IEEE Internet of Things Journal. 2016; **3**(6):854-864

[3] Ku Y. 5G radio access network design with the fog paradigm: Confluence of communications and computing. IEEE Communications Magazine. 2017;**55**(4): 46-52

[4] Wang S, Xu J, Zhang N, Liu Y. A survey on service migration in Mobile edge computing. IEEE Access. 2018;**6**: 23511-23528

[5] Li J, Shen X, Chen L, Pham D, Ou J, Wosinska L, et al. Service migration in fog computing enabled cellular networks to support real-time vehicular communications. IEEE Access. 2019;**7**: 13704-13714

[6] Machen A, Wang S, Leung KK, Ko BJ, Salonidis T. Live Service Migration in Mobile Edge Clouds. IEEE Wireless Communications. 2018;**25**(1): 140-147

[7] Codeca L, Frank R, Engel T. Luxembourg SUMO Traffic (LuST) Scenario: 24 Hours of Mobility for Vehicular Networking Research. Kyoto: IEEE Vehicular Networking Conference (VNC); 2015. pp. 1-8

[8] Han D, Shin S, Cho H, Chung J, Ok D, Hwang I. Measurement and stochastic modeling of handover delay and interruption time of smartphone real-time applications on LTE networks. IEEE Communications Magazine. 2015; **53**(3):173-181

[9] Li J, Natalino C, Van D.V, Wosinska L, Chen J, Resource Management in Fog Enhanced Radio Access Network to Support Real-Time Vehicular Services. Madrid: IEEE Information Conference on Fog and Edge computing; May 2017

[10] Li J. Ultra-Low Latency Communication for 5G Transport Networks. Universitetsservice US AB; 2019. ISBN: 978–91–7873-243-2

[11] Morabito R, Cozzolino V, Ding AY, Beijar N, Ott J. Consolidate IoT edge computing with lightweight virtualization. IEEE Network. 2018; **32**(1):102-111

[12] Zhang H, Chen K, Bai W, Han D, Tian C, Wang H, et al. Guaranteeing deadlines for inter-data Center transfers. IEEE/ACM Transactions on Networking. 2017;**25**(1):579-595

[13] Kramer G. Ethernet Passive Optical Networks. New York: McGraw-Hill; 2005

[14] Dixit A, Lannoo B, Colle D, Pickavet M, Demeester P. Energy Efficient DBA Algorithms for TWDM-PONs. In: IEEE International Conference on Transparent Optical Networks (ICTON). Budapest: IEEE; 2015. pp. 1-5

[15] Kramer G, Mukherjee B, Pesavento G. IPACT: A dynamic protocol for an Ethernet PON (EPON). IEEE Communications Magazine. Feb., 2002;**40**(2):74-80

[16] Li J, Shen X, Chen L, Ou J, Wosinska L, Chen J. Delay-aware bandwidth slicing for service migration in mobile backhaul networks. IEEE/OSA Journal of Optical Communications and Networking. 2019;**11**(4):B1-B9