Supporting Low-Latency Service Migration in 5G Transport Networks

Jun Li and Jiajia Chen
Department of Electronic Engineering, Chalmers University of Technology, Göteborg, Sweden
Email: jiajiac@chalmers.se

Abstract: This paper concentrates on low-latency service migration in transport networks, where edge computing is employed for ultra-low end-to-end latency communications in 5G, and demonstrates that rapid service migration significantly reduces end-to-end packet delay.

1. Introduction

Future communication systems are envisioned to support diverse mission-critical services, where ultra-low end-to-end latency is highly required. The end-to-end latency is often defined as the packet delay from the instant of the beginning of transmission by the sender to the complete reception by the receiver (listener) [1], where delay components experienced in all network segments need to be included. The ultra-low end-to-end latency commonly refers to the very short packet delay, that can be on the order of a few milliseconds (ms) or even a few microseconds (μs). For instance, industrial applications for critical control have very tight delay bounds, where only a few μs are allowed [2], while services, like autonomous driving, augmented reality, and virtual reality request the end-to-end latency to be less than 10 ms [3]. In order to achieve such a stringent level of the end-to-end latency, edge computing is introduced, where computation and storage resources are deployed in close proximity to end users, greatly reducing communication delay in transport network segments. Recently, European telecommunications standards institute (ETSI) industry specification group has documented multi-access edge computing in 5G networks, to promote the adoption of edge computing for radio access networks [4].

Although edge computing is considered promising to address stringent latency requirements, it brings technical challenges, particularly for the transport networks that interconnect various edge nodes. One of the most crucial problems is how to handle user mobility considering limited coverage of a single edge node. Migrating the services from the source edge node to the destination edge node via the transport networks to follow the user mobility may result in interrupting ongoing services, and hence significantly affect the end-to-end latency performance [5]. To avoid/mitigate performance degradation caused by the service interruption, during the migration procedure the access to the source edge node needs to be kept as long as possible for the being-migrated services. Nevertheless, the end users may still suffer the extra latency caused by the transport network segment that interconnect the source and destination edge nodes [3]. Therefore, minimizing the service migration delay (also referred to as the time duration from the instant that the migration is initiated until the being-migrated service is successfully transferred to the destination node) would be very beneficial. However, the data generated for the service migration is usually huge (e.g., on the order of hundreds of Mbytes), which makes it difficult to realize low-latency service migration in the current transport networks.

In this regard, this paper attempts to address low-latency service migration for 5G transport networks. An envisioned framework for edge computing enabled 5G networks is presented, based on which three low-latency strategies for service migration are discussed. A use case of real-time vehicular communications employing mobility pattern and traffic of Luxembourg is carried out to show how the end-to-end latency can benefit from low-latency service migration strategies employed in 5G transport networks.

2. Low-latency strategies for service migration in transport networks

Figure 1 illustrates an envisioned framework for edge computing enabled 5G networks including both data

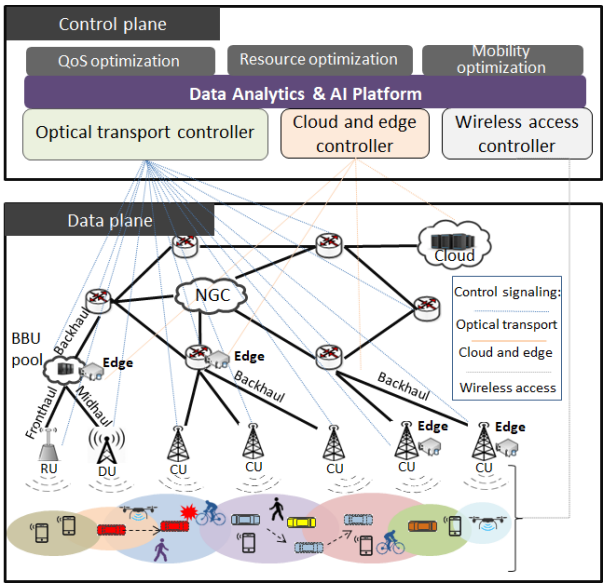


Fig. 1: Edge computing enabled 5G networks. AI: Artificial intelligence, RU: Radio unit, DU: Distributed unit, CU: Centralized unit, BBU: Baseband unit, NGC: Next generation core.

plane (Fig. 1 bottom part) and control plane (Fig. 1 upper part)). In the data plane, severs for edge computing can be located in the places after baseband processing function, e.g., baseband unit (BBU) pools, centralized units (CUs) and/or aggregation nodes of core networks, which are in line with recommendations from the ETSI [4]. According to the function splitting defined for 5G [5], distributed units (DUs) and radio units (RUs), which are lack of higher-layer functions, are not proper to co-locate edge computing facility. Therefore, mobile backhaul is a critical segment for service migration in 5G transport, where passive optical networks (PONs) are often considered promising thanks to its low cost and power consumption. In the control plane, the controllers for optical transport, wireless access, and cloud and edge computing are powered by the data analytics and artificial intelligence (AI) techniques for optimizing quality of service (QoS), resource allocation and mobility plan. Cross-layer approaches that take into account both data and control planes are essential. We identify three important means, namely connectivity enhancement, migration strategy and bandwidth slicing, for low-latency service migration, which are able to either improve latency experienced in the transport networks or optimize migration frequency.

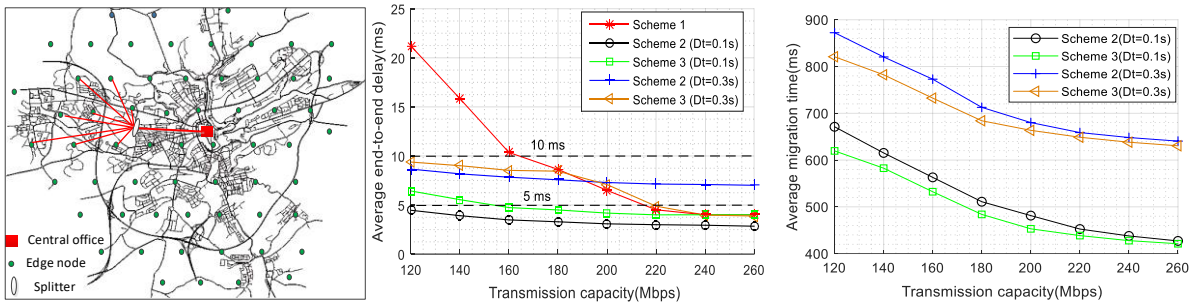
Connectivity enhancement: A straightforward way to reduce service migration latency is to improve connectivity in the mobile backhaul, allowing directly connected adjacent edge nodes as much as possible. Some research works have been carried out to enhance connectivity among various optical network units (ONUs) often co-located with base stations, which can also be the places for servers used in edge computing, referred to as edge nodes. [6] introduced a splitter-box containing several passive combiners and diplexers employed at the remote node of the PON, allowing for direct interconnections among the edge nodes within the same PON. [7] considered wavelength division multiplexing (WDM) PON, which passively interconnects the ONUs by using a multi-port-in and multi-port-out arrayed waveguide grating at the remote node. In [8], a loopback mechanism was introduced at the remote node, in which the upstream data transmitted by each ONU can be directly sent back to all other ONUs belonging to the same PON through a passive coupler. However, for the edge nodes belonging to different PONs, the traffic still has to be sent to the central office and even farther to the core networks, which is not sufficient to guarantee low-latency for the users characterized by fast mobility that can easily move to the outsides of the area covered by the edge nodes associated to one PON. Furthermore, the enhanced connectivity in [6-8], on the other hand, potentially increases the risk of traffic conflicts and hence raises the issue of the control plane design. The standardized media access control (MAC) protocol for PONs, i.e., multipoint control protocol [9], is not sufficient. In [10], a cross-layer design was presented for efficient PON-based mobile backhaul that can significantly enhance the connectivity between any adjacent edge nodes by adding extra fiber links among different remote nodes belonging to different PONs, while a tailored MAC protocol and dynamic bandwidth allocation algorithm were introduced. This scheme can support ultra-low latency (i.e., less than 1 ms packet delay) for communications among adjacent edge nodes, which could be low enough for services, such as autonomous driving, augmented reality, and virtual reality, demanding the end-to-end latency less than 10 ms [3].

Migration strategy: Service migration is often needed for two reasons: 1) Resource at the current edge node is not enough to carry out all the services, some of which need to be migrated to offload the computing tasks; and 2) users who subscribe the services at the current edge node move away, causing some QoS requirements unable to be satisfied anymore. For the second reason, assuming the computing resource at the edge node is sufficient, the mechanism that determines how and when the service migration is triggered is of key importance to mitigate QoS degradation, particularly for the latency. [11] investigated different strategies, including no migration (Scheme 1), migration always together with handover (Scheme 2), and QoS aware migration (Scheme 3) in which decision-making is based on whether the QoS metrics are satisfied or not. Apparently, the end-to-end latency cannot be kept sustainable in Scheme 1 when the vehicles travel far away from the serving edge node. Scheme 2 attempts to provide one-hop access to the edge node, which is able to minimize the delay caused in the transport networks. However, if the migration time is longer than the time that the user stays in the area covered by the new edge node, the service migration becomes inefficient and the migration overhead may significantly affect the end-to-end latency. This indicates that the frequent service migration is not always a good choice. The key idea of Scheme 3 is to flexibly combine Scheme 1 and 2 to minimize the migration overhead while maintaining the end-to-end performance at an acceptable level to satisfy the QoS requirements.

Bandwidth slicing: For infrastructure sharing, the migration traffic is sent together with other traffic (e.g., voice, surfing), referred to as non-migration traffic. If no specific scheduling algorithm is applied, transmission windows could be occupied by the non-migration traffic for a long time in the case with a high load of the non-migration traffic. It leads to migration time increases, and consequently the end-to-end latency requirements may not be satisfied. In this regard, bandwidth slicing, where the cycle time can be dynamically divided into several slices for different kinds of services, is a powerful mechanism to schedule traffic fairly and effectively [12]. To address low-latency service migration, delay-aware bandwidth slicing mechanism is beneficial, where the large-size migration

traffic is partitioned into small pieces and allowed to be transmitted within a certain time constraint, thus minimizing the impact on the latency of the non-migration traffic while assuring latency requirements for the service migration.

3. Case study of real-time vehicular communications



(a) (b) (c)

Figure 2: (a) Luxembourg map, in which edge nodes are evenly distributed and co-located with ONUs connected by PON based backhaul networks, (b) average end-to-end delay and (c) average migration time as a function of mobile backhaul transmission capacity different migration strategies previously reviewed.

Based on the real vehicular traffic profile and mobility pattern of Luxembourg (see Fig. 2a), a use case for service migration in the scenario of real-time vehicular communications is presented. We take migration strategy as an example to explain the relationship between the service migration time and the end-to-end latency. Performance evaluation of the three migration schemes reviewed previously is carried out by simulation using Urban Mobility (SUMO) and Python. The wireless delay and handover interruption time are not dependent on migration strategies. The requirements of ultra-high reliability low latency communication (URLLC) for radio access networks are followed, where the uplink delay in the wireless segment is assumed to be within 0.5 ms and the handover interruption time is considered as a constant. The service downtime (D_t), during which the services cannot be properly accessed, varies in Schemes 2 and 3. As shown in Figs. 2b and 2c, if no migration (i.e., Scheme 1), the end-to-end delay can be extremely high when the transmission capacity in the mobile backhaul is limited. Thus, service migration is obviously necessary to reach ultra-low end-to-end packet delay. Meanwhile, the end-to-end delay demonstrates a similar trend as the migration time for both Scheme 2 and 3. It implies lowering the migration delay is an efficient mean to reduce the end-to-end latency in edge computing enabled 5G networks. The service downtime significantly affects both types of delays, which should be minimized during the migration procedure.

4. Summary

To realize low end-to-end latency in edge computing enabled 5G networks, fast service migration is essential. Connectivity enhancement, migration strategy and bandwidth slicing are three possible means to speed up service migration. We believe two aspects that are worth exploring more in future: 1) Advanced strategies that can combine two or all of the low-latency means for service migration to improve the latency, and 2) AI techniques that can be introduced in any of low-latency strategies for service migration in transport networks to optimize QoS metrics.

Acknowledgement: 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.

References

[1] A. Nasrallah et al., "Ultra-Low Latency (ULL) Networks: The IEEE TSN and IETF DetNet Standards and Related 5G ULL Research," IEEE Communications Surveys & Tutorials, vol. 21, pp. 88-145, Firstquarter 2019.
[2] B. Rudy et al., “IEC/IEEE TSN Profile for Industrial Automation”, document IEC/IEEE 60802 V0.61, Piscataway, NJ, USA, Apr. 2018.
[3] ”5G for Mission critical communication” Nokia white paper, 2016.
[4] S. Kekki, et al., “MEC in 5G networks,” ETIS, June, 2018.
[5] ITU-T GSTR-TN5G Transport network support of IMT-2020/5G, Feb. 2018
[6] T. Pfeiffer, “Converged heterogeneous optical metro-access networks,” in European Conf. on Optical Communication (ECOC), Sept. 2010.
[7] C. Choi, et al., “Mobile WDM backhaul access network with physical inter-base-station links for coordinated multipoint transmission/reception system,” in IEEE Global Telecommunications Conf. (GLOBECOM), Dec. 2011.
[8] C. Ranaweera, et al., “Next generation optical-wireless converged network architectures,” IEEE Netw., vol. 26, pp. 22–27, Mar. 2012.
[9] G. Kramer, “Ethernet Passive Optical Networks” New York: McGraw-Hill, 2005.
[10] J. Li and J. Chen, “Passive Optical Network Based Mobile Backhaul Enabling Ultra-Low Latency for Communications among Base Stations”, IEEE/OSA JOCN, vol. 9, pp. 855-863, Oct. 2017.
[11] J. Li, et al., "Service Migration in Fog Computing Enabled Cellular Networks to Support Real-Time Vehicular Communications," *IEEE Access*, vol. 7, pp. 13704-13714, 2019.
[12] J. Li, et al., "Delay-Aware Bandwidth Slicing for Service Migration in Mobile Backhaul Networks," IEEE/OSA JOCN, vol. 11, pp. B1-B9, April. 2019.