IEICE Communications Express, Vol.13, No.7, 256–259

ComEX

LETTER

Proactive congestion control by fast optical switching within 1-ms delay at mobile backhaul utilizing traffic data volume

Yuka Okamoto[1, a)], **Hirotaka Ujikawa**[1], **Kota Asaka**[1], **Tatsuya Shimada**[1], **and Tomoaki Yoshida**[1]

Abstract The development and introduction of real-time remote control of robots using mobile NWs is expected. To realize this control, a network configuration in which the User Plane Function (UPF) and Mobile Edge Computing (MEC) are located near the Central Unit (CU) is considered. If traffic exceeds the processing capability of the UPF, there is the concern of congestion. In this paper, we propose and demonstrate a method for obtaining the uplink traffic volume from the CU and predicting the future traffic to switch the optical path and UPF before congestion occurs within the Mobile Backhaul (MBH) delay of 1ms.

Keywords: traffic steering, machine learning, low latency, congestion control.

Classification: Network

1. Introduction

As technology development and introduction of cyber-physical systems (CPS) are progressing, the introduction of real-time remote control of robots using mobile NWs is expected [1, 2]. According to the delay requirement for remote robot control summarized by 3GPP, the end-to-end (E2E) delay must be reduced to 5 ms [3]. To satisfy this requirement, with the assumptions of an optical path length of 600 km (maximum one-way transmission distance in the East or West Japan area), a transmission delay of 3 ms in the optical network, and a radio network delay of 1 ms [4], other delays such as those due to congestion should be less than 1 ms. For accurate remote control of robots, the use of a real-time streaming video signal, which is sent from the robot to a remote operator, is necessary. Furthermore, advanced remote operation is also expected to implement Augmented Reality (AR) technology over video signals to assist remote operators [5]. To realize video processing in real time, it is necessary to implement Mobile Edge Computing (MEC) at the edge of the network that performs processing using the computing resources of the network [6]. Since MEC performs video processing at the network edge near the robot instead of core NW, the amount of data to the core NW can be reduced, allowing for ultra-low delay communication. To realize MEC, mobile network endpoints, the User Plane Function (UPF) and MEC, are installed near the Central Unit (CU). At this time, the UPF must handle traffic that routes both video traffic handled by the MEC and background traffic transmitting to the core NW according to the traffic destination. Therefore, when the background traffic increases with the increase in the number of users, there are difficulties achieving a low delay, such as congestion delays due to the traffic exceeding the processing capability of the UPF. This leads to difficulty in real-time remote control. To avoid the effects of such congestion in the UPF, we propose a novel method for achieving low delay by predicting future traffic on the basis of the transmitted data volume from the CU and switching the optical path before congestion occurs. By using the proposed method, we successfully confirmed a congestion delay at the Mobile Backhaul (MBH) of 1 ms or less.

2. Related works

Related congestion control methods for avoiding congestion by switching optical paths and CUs or UPFs are mainly categorized into three types of traffic steering. The first method is (1) switching based on the amount of traffic arriving at the UPF. The second is (2) switching based on scheduled traffic information from the Distributed Unit (DU). The third is (3) switching based on the prediction of future traffic at the CU. In method (1), congestion delay occurs because switching starts after traffic congestion occurs [7]. An MBH delay of 1 ms or less may not be achievable. In method (2), it is possible to obtain the traffic allocation amount, such as the transport block size (TBS) from the cooperative transport interface (CTI) before traffic arrival [8]. However, the scheduling of the radio section is performed at the DU, so it is difficult to use it to control congestion at the MBH, where the distance between the DU and the UPF is wide. In method (3), the future traffic amount at the CU is predicted from the traffic allocation amount acquired from the CTI, and switching is performed on the basis of the prediction result [9]. In our previous study [10], the feasibility of traffic steering under a Mobile Midhaul (MMH) delay of 1 ms, suitable for the use case of remote control, was confirmed with a prototype. However, considering real-time remote control using video information, it is assumed that video processing is performed at the MEC/UPF close to the CU. Therefore, it is also assumed that the processing capability of the UPF is small. In addition, it is considered that congestion occurs not only at the CU but also at the UPF. At this time, in switching the UPF, method (3) is difficult to collect traffic from the existing CTI in advance for the same reason as method (2). In this

[1] NTT Access Network Service Systems Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan

[a)] yuka.okamoto@ntt.com

DOI: 10.23919/comex.2024XBL0055
Received March 18, 2024
Accepted April 8, 2024
Publicized May 9, 2024
Copyedited July 1, 2024

paper, the proposed method using prediction and switching method of (3) without CTI is demonstrated for the first time in an MBH system using a prototype controller that implements proactive traffic steering software. Moreover, thanks to the collected CU uplink (UL) transmitted data volume, we successfully achieve switchover of the UPF and the correspondent optical path without congestion within a 1-ms delay at the MBH.

3. Proposed method

Figure 1 shows a schematic configuration of the proactive traffic steering method that we propose. The controller is composed of three parts, 1) collection, 2) analysis, and 3) control. 1) The collection part collects the UL transmitted data volume from the CUs. This part also converts the information into time series data for prediction. 2) The analysis part predicts the UL UPF traffic amount, which is the sum of the future UL traffic amounts of each CU in different threads predicted by machine learning using time series data. It calculates the congestion amount of traffic from the UL traffic of the UPF and the maximum bandwidth at the MBH. When UPF congestion is predicted, the analysis part decides where/when to switch the UPF and sends the switching instruction including the optical path of the MBH to the control part. 3) Upon receiving the instruction, the control part instructs the optical switch to change ports at the designated timing. In the analysis part, it is necessary to predict the future traffic ahead of the processing time. The processing time is defined as the elapsed time at the analysis and control parts including the time needed for physical optical switching. These three parts are periodically repeated in order from 1) to 2) to 3). The period at that time is defined as the prediction period.

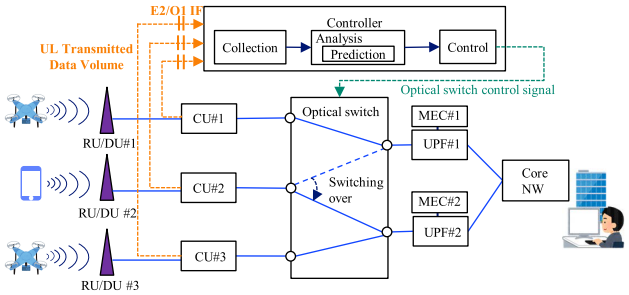

Fig. 1 Configuration of proposed method.

4. Experimental evaluations

4.1 Experimental setup

Figure 2 shows the experimental setup used to verify our proposed proactive traffic steering within a delay at the MBH of less than 1 ms. The setup is composed of a traffic generator, which transmits UL traffic corresponding to the CUs, a traffic collection system, which calculates the traffic amount from the collected traffic and transmits it from the CU to the controller corresponding to the UL transmitted data volume, a traffic receiver corresponding to the UPFs, and a traffic shaper emulating UPF bandwidth control. In this experiment, two UPFs were connected by using optical fibers to three CUs via a PLZT optical switch, and each UPF aggregated UL traffic from the CUs. The traffic collection system detected the traffic at the CU and sent the calculated traffic amount, which correspond to the UL transmitted data volume, to the controller using a simple network management protocol management information base (SNMP-MIB). The controller, composed of collection, analysis, and control parts as software components, was implemented on a GPU server (LGA-4189 3rd Gen Intel® Xeon® Scalable processors), and the measured processing time of analysis part was 10 ms [10]. We introduce a parallel prediction scheme using Transformer [11], which is one of the major time series prediction methods. On the basis of the collected UL transmitted volume every SNMP period (5, 7, 10, 15, 20, 30, and 50 ms), the controller converted it into time series data every prediction period the same as the SNMP period. In addition, the controller predicted future traffic on each thread in parallel with the time series data for each CU. We assumed that CU#1 and CU#3 transmitted only one video traffic, of which the frame rates were 10, 30, and 60 fps, the max bit rate was 180 Mbps, and the average bit rate was 36 Mbps. The MBH link rate of 300 Mbps was configured by the traffic shaper, CU#2 accommodated 24 UEs, and the UE data traffic had an average bit rate of 8 Mbps. In this experiment, the delay at the MBH was measured from the reception time difference of each packet, which was captured at the CU output and UPF input.

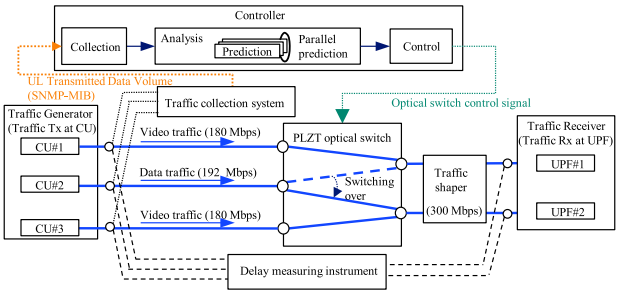

Fig. 2 Experimental setup.

4.2 Discussion on optimal prediction period

First, we experimentally identified the optimal prediction period to satisfy the MBH delay of 1 ms or less when the frame rate of the video traffic was changed. This is because there are three problems, and the first two problems occur when the prediction period is too short, while the third problem occurs when the period is too long. The first problem is that the UL transmitted data volume cannot be acquired in time through SNMP-MIB. This is because when the prediction period, which is the same period as the SNMP-MIB, becomes too short, the SNMP request frequency from the controller to the traffic collection system drastically increases. As a result, the system cannot respond to requests at such a high frequency and fails in acquiring the UL transmitted data volume. The second problem is that the congestion control is delayed because the prediction calculation amount exceeds the processable capacity of the GPU server. Under this problem, as the computation time increases because of the too

short prediction period, the prediction cannot be completed before congestion occurs, which causes congestion delays since the traffic steering is not conducted in time. The third problem is that the controller is not able to predict the congestion because smooth traffic is predicted due to prediction period being too long which results in loss of the burstiness of the traffic. To accurately predict burst traffic when the frame rate is large, the prediction period of the traffic should be short compared with the burst interval. That is, the allowable prediction period differs depending on the frame rate.

Therefore, it is considered that the optimal prediction period differs depending on the frame rate of the traffic. For frame rates of typical traffic, that is, 10, 30, and 60 fps, the prediction period is changed, and we measure the delay at the MBH. The prediction period that satisfies the MBH delay of less than 1 ms is defined as the optimal prediction period.

4.3 Experiment results

Figure 3 shows the max delay at the MBH when the prediction period was changed for each frame rates (10, 30, and 60 fps). To achieve remote control, we experimentally identified the optimal prediction period for satisfying the MBH delay of 1 ms or less. When the prediction period is less than the prediction time of 10 ms, the prediction cannot be made in time regardless of the video period, and the MBH delay exceeds 1 ms. Therefore, the prediction period should be 10 ms or more in our experimental setup (Note that the optimal prediction time depends on the capability of the processing unit of controller). In the case of the 60-fps frame rate (green line), when the prediction time was 10 ms, the MBH delay was 0.99 ms which met the MBH delay requirement of 1 ms or less. The MBH delay increased to over 1 ms for the prediction periods of 15 ms and 20 ms and then stabilized at a constant value of 1.26 ms beyond 20 ms. The long prediction period beyond 20 ms smoothed the traffic and the prediction results, so congestion could not be detected and switching did not occur. In the case of the case of the 30-fps frame rate (orange line), when the prediction period was 10 ms, the MBH delay was 0.88 ms. The MBH delay increased to over 1 ms for the prediction periods of 15, 20, and 30 ms and then stabilized at a constant value of 1.9 ms beyond 30 ms. The reason for the constant MBH delay is the same as the case of the 60-fps frame rate. In the case of the 10-fps frame rate (blue line), when the prediction period was 10 ms, the MBH delay was 0.69 ms. The MBH delay increased to over 1 ms for the prediction periods of 15, 20, and 30 ms and then stabilized at a constant value of 3.1 ms beyond 30 ms. The constant MBH delay of 3.1 ms was caused for the same reason as in the cases of the 60-fps and 30-fps frame rate. Therefore, the optimal prediction period is 10 ms for the frame rate of the 60-fps and 30-fps, the optimal period is between 10-ms and 15-ms for the frame rate of 10-fps.

Figure 4 shows the measured delay the MBH for the 30-fps frame rate, which is the most popular period for video streaming signals, as a function of the elapsed time for our proposed method and several other methods for comparison. Here, we investigated the delay behavior at the MBH in the case 1) without traffic steering (blue circle), 2) with traffic steering and without prediction (orange diamond) as related methods, and 3) with traffic steering based on the optimal prediction period of 10 ms (green cross) as a our proposed method. In this evaluation, the congestion delay is defined as when the MBH delay exceeds 1 ms.

1) Without traffic steering, congestion delay occurred during an elapsed time of 7–17 ms and 40–50 ms at UPF#1 when the video burst traffic of CU#1 overlapped with the background data traffic of CU#2. The measured congestion delay increased to 1.6 ms, which significantly exceeded the delay requirement of 1 ms. After the end of the burst traffic over the course of 17–20 ms and 50–52 ms, the delay decreased to less than 1 ms because the traffic buffered at the UPF#1 (shaper) was gradually transmitted. 2) With traffic steering and without prediction, the maximum congestion delay decreased to 1.4 ms at elapsed times of 14 ms and 37 ms thanks to the switchover of the CU and optical path after congestion happened. However, it still exceeded the requirement. The reason for the congestion delay between 27–37ms, is that the UPF#2 to which the UPF was switched also experienced congestion, so the video burst traffic of CU#3 overlapped with the background data traffic of CU#2, and a delay corresponding to the rise of the traffic of CU#3 occurred. 3) With traffic steering based on predic-

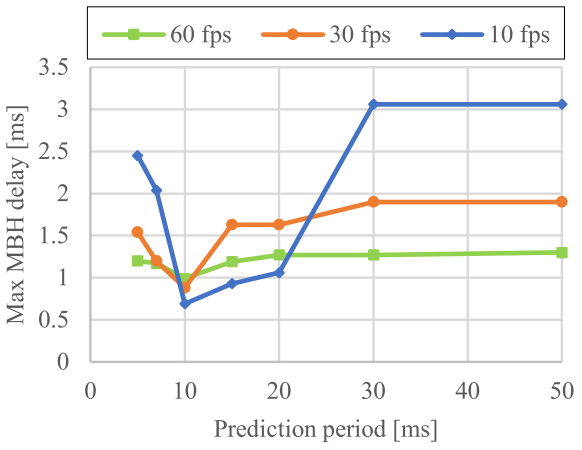

Fig. 3 Relationship between prediction period and MBH max delay

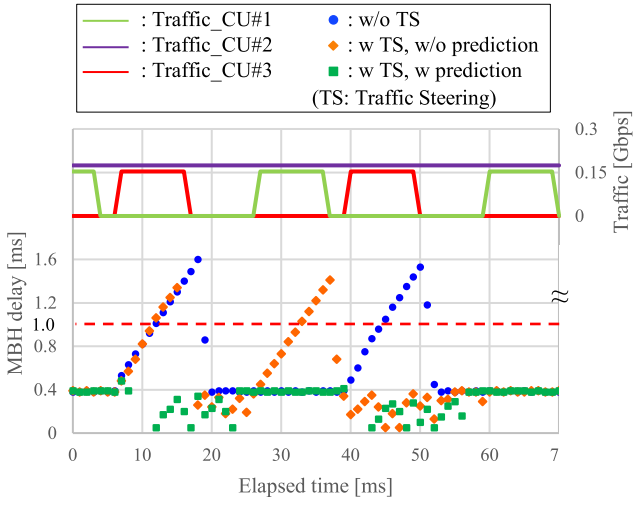

Fig. 4 MBH delay as a function of elapsed time.

tion, thanks to the optimized prediction period of 10 ms, we successfully avoided congestion at any of the elapsed times. As a result, the max delay of MBH at an elapsed time of 7 ms dramatically decreased to 0.56 ms, which meets the requirement of less than 1 ms. The reason for the decrease in MBH delay is that the controller can successfully predict congestion before burst traffic occurs. The detailed behavior of MBH delay with the proposed method is as follows. Since the destination of CU#2 was switched from UPF#1 to UPF#2 thanks to the proposed method for an elapsed time of 7 ms, the MBH delay was maintained at less than 0.4 ms for an elapsed time between 8 ms and 27 ms when the burst traffic of CU#3 did not occur. At an elapsed time of 27 ms when the burst traffic of CU#3 started, the proposed traffic steering method predicted future congestion at UPF#2 and switched the destination of CU#2 from UPF#2 to UPF#1. Therefore, an MBH delay of 0.4 ms or less was successfully achieved at an elapsed time between 27 ms and 39 ms. Beyond the time of 39 ms, the controller repeatedly conducted the aforementioned procedures, which led to a low delay of 0.4 ms or less at the MBH.

5. Conclusion

Our previous work revealed that our congestion control method using the prediction of future traffic and switching of the optical and CU effectively worked for the MMH section. In this paper, we proposed a novel method for achieving low delay at the MBH section by predicting future traffic on the basis of the traffic volume information from a CU and switching the optical path and UPF before congestion occurs. Thanks to the experimentally optimized prediction period of 10 ms, our prototype controller successfully achieved a delay at the MBH as low as 1 ms or less without congestion. This study is promising for emerging services that utilize remotely controlled robots via mobile networks.

References

[1] H. Ou, K. Asaka, T. Shimada, and T. Yoshida, "SLA aware real time control technology across optical and mobile networks," OFC, San Diego, USA, M3A.4, March 2022. DOI: 10.1364/OFC.2022.M3A.4

[2] K. Asaka, H. Ou, T. Shimada, and T. Yoshida, "SLA-aware real-time control technology for all photonics network and beyond," OECC, Shanghai, China, pp. 1–4, OECC2023-0220-2, July 2023. DOI: 10.1109/OECC56963.2023.10209857

[3] 3GPP TS22.261 v18.6.1, "Service requirements for the 5G system," 2022.

[4] 3GPP TS38.913 v17.0.0, "Study on scenarios and requirements for next generation access technologies," 2022.

[5] K. Brunnström, E. Dima, T. Qureshi, M. Johanson, M. Andersson, and M. Sjöström, "Latency impact on quality of experience in a virtual reality simulator for remote control of machines," *Signal Processing: Image Communication*, vol. 89, 116005, Nov. 2020. DOI: 10.1016/J.IMAGE.2020.116005

[6] L. Lin, X. Liao, H. Jin, and P. Li, "Computation offloading toward edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1584–1607, Aug. 2019. DOI: 10.1109/JPROC.2019.2922285

[7] A. Sahara, T. Ono, J. Yamawaku, A. Takada, S. Aisawa, and M. Koga, "Congestion-controlled optical burst switching network with connection guarantee: design and demonstration," *IEEE J. Lightw. Technol.*, vol. 26, no. 14, pp. 2075–2086, July 2008. DOI: 10.1109/JLT.2008.922307

[8] ITU-T G. Supplement 66, Revision 2, "5G wireless fronthaul requirements in a PON context," 2019.

[9] Y. Okamoto, H. Ujikawa, Y. Sakai, T. Shimada, T. Yoshida, "Short-term traffic prediction based on mobile control information for proactive optical switching to lower congestion delay," IEICE Proceeding Series 72, IEICE, Tokyo, Japan, no. O6-2, Nov. 2022. DOI: 10.34385/proc.72.o6-2

[10] Y. Okamoto, et al., "Proactive congestion control within 1-ms delay at mobile midhaul utilizing parallel traffic prediction and fast switchover of CU and optical path," OFC, San Diego, USA, M4D.3, March 2024, in print.

[11] Z. Liu, Z. Zhu, J. Gao, and C. Xu, "Forecast methods for time series data: A survey," *IEEE Access*, vol. 9, pp. 91896–91912, June 2021. DOI: 10.1109/ACCESS.2021.3091162