Proactive congestion control within 1-ms delay at mobile midhaul utilizing parallel traffic prediction and fast switchover of CU and optical path

Yuka Okamoto, Hirotaka Ujikawa, Kota Asaka, Tatsuya Shimada, and Tomoaki Yoshida
NTT Access Network Service Systems Laboratories, NTT Corporation, 1-1 Hikarinooka, Yokosuka, Kanagawa 239-0847, Japan
Author e-mail address: yuka.okamoto@ntt.com

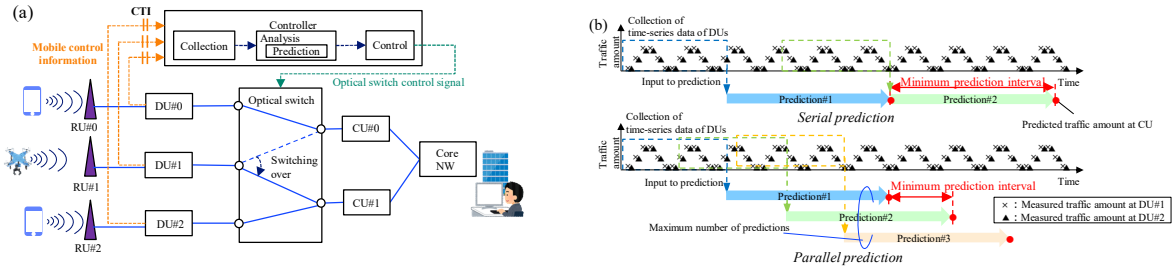
Abstract: We propose a proactive congestion control method that utilizes parallel traffic prediction and fast switchover of the CU and optical path. Our prototype controller can perform these tasks within a 1-ms delay at the MMH. © 2024 The Author(s)

1. Introduction

Remote inspection for outdoor infrastructures using a remotely controlled robot through mobile networks has recently attracted interest because it can address the reduced workforce challenge currently facing advanced countries. For a remote inspection service, it is necessary to transmit high-definition video from a robot to a remote operator in real-time with low delay. According to the delay requirements of remote robot control summarized by 3GPP, the end-to-end (E2E) delay must be reduced to 5 ms [1]. To satisfy this requirement, under the assumptions of a 600-km optical path length (maximum one-way transmission distance in the East or West Japan areas), 3-ms transmission delay in the optical network, and 1-ms delay in the radio network [2], other delays (e.g., those due to congestion) need to be less than 1 ms.

For the implementation of a remote inspection service using the centralized radio access network (C-RAN) architecture for 5G and beyond, the congestion delay at the central unit (CU) that aggregates uplink signals from multiple distributed units (DUs) should be within 1 ms. At present, the related congestion control technique is to switch the uplink traffic from a DU of the congested CU to one of the other available CUs by means of an optical path (DU-CU) switching [3]. In this case, the delay at the mobile midhaul (MMH), which is sum of the congestion delay and the switching delay, should be within 1 ms. There are two main techniques, a reactive technique and proactive technique. In the reactive technique, the controller collects traffic information, and if it detects congestions, then it controls the optical path [4]. However, this causes a congestion delay corresponding to the elapsed time from when the congestion is detected to when the switching is completed. In the proactive method, the controller collects mobile control information from a DU via the cooperative transport interface (CTI), which is the future uplink traffic amount (e.g., grant information) arriving at the CU [5]. Switching the CU and optical path by using mobile control information before the arrival of future traffic at the CU enables congestion to be avoided. This approach can receive the information in advance, but how many milliseconds ahead traffic can be obtained from mobile control information depends on the scheduling interval, which is how often the DU outputs mobile control information. When the scheduling interval is a short value like in 5G (e.g., 0.125 ms), it cannot switch the CU and the optical path before traffic congestion happens.

As an alternative congestion control approach to satisfy the MMH delay requirement of 1ms in 5G and beyond, we previously proposed a proactive optical path switching method based on machine learning prediction and mobile control information via CTI [3]. Although our simulation using offline processing demonstrated the possibility of achieving proactive congestion control, the feasibility of real-time and online processing utilizing a prototype controller has yet to be confirmed. In this paper, we investigate our method by utilizing a prototype controller equipped with the proactive congestion control software, an emulated MMH system, and a newly proposed parallel prediction technique with an optimized prediction period. Our findings showed that we can achieve the switchover



(a) Configuration of proposed technique. (b) Prediction interval for serial and parallel predictions.

Fig. 1 Schematic views of proposed technique.

of the CU and the corresponding optical path without congestion within a 1-ms delay at the MMH.

2. Newly proposed prediction method

Figure 1(a) shows a schematic configuration of the proactive congestion control technique that we previously proposed [3]. The controller is composed of three parts: 1) collection, 2) analysis, and 3) control. 1) The collection part collects mobile control information from the DUs through the CTI every scheduling interval. This part also converts the information into time-series data for prediction. 2) The analysis part predicts the uplink CU traffic amount, which is the sum of predicted future uplink traffic amounts at each DU, by machine learning using time-series data. It calculates the congestion amount of traffic from the uplink traffic at the CU and the maximum bandwidth at the mobile back haul. When CU congestion is predicted, the analysis part decides where/when to switch the CU and sends the switching instructions including the optical path of the MMH, to the control part. The analysis part predicts the future traffic at each DU in different threads and calculates the future traffic at the CU. As shown in Figure 1(b), the prediction interval of the analysis part includes serial prediction (top) and parallel prediction (bottom). The serial prediction starts the prediction processing (DU traffic prediction and calculation of future CU traffic) once the previous processing completes. Therefore, the prediction period is equal to the prediction time. In contrast, the parallel prediction starts the processing before the previous processing completes, so the prediction period is shorter than the prediction time. 3) Upon receiving the instructions, the control part instructs the optical switch to change the ports at the designated timing. During the analysis part, it needs to predict the future traffic ahead of the processing time every short period (roughly a few milliseconds). The processing time is defined as the elapsed time at the analysis and control parts, including the time needed for physical optical switching. Our previous study using serial prediction estimated that the processing time would be more than 10 ms [3]. Therefore, it is difficult to control congestion using serial prediction every scheduling interval.

As a contribution of this paper, we newly propose utilizing parallel prediction to predict the future traffic ahead of the processing time every short period. To achieve the parallel traffic prediction, it is important to optimize the prediction period, as there is a tradeoff between the calculation amount and the prediction accuracy. Figure 2 shows the calculation amount (pink line, left vertical axis) and prediction accuracy (light blue line, right vertical axis) as a function of the prediction period. As indicated by the pink line, the calculation amount at the analysis part increases inversely proportional to the prediction period because of the large number of predictions. As the prediction period gets shorter, the calculation amount exceeds the processing capacity of the CPU (red dotted line). Therefore, the prediction sometimes fails due to the increased CPU loads when the period is too short. If the prediction period is longer than a burst period of video traffic, microburst congestion cannot be detected, and the prediction accuracy (light blue line) becomes worse. Therefore, it is necessary to set the prediction period optimally. Since the processing capacity depends on the specifications of the implemented CPU, we first need to experimentally clarify the optimal prediction periodic zone, then we define the smallest period in this zone as the optimal prediction period.

3. Experimental setup and results

Figure 3 shows the setup of the experimental we conducted to determine whether our proposed proactive congestion control can be performed within a delay at the MMH of less than 1 ms. The setup is composed of a traffic generator, which transmits mobile control information and DU uplink traffic corresponding to the DUs, a traffic analyzer, which corresponds to the CUs, and a traffic shaper, which emulates the CU bandwidth control. In this experiment, two CUs are connected by optical fibers to three DUs via a PLZT optical switch and the shaper, and each CU aggregates uplink traffic from the DUs. Mobile control information is a transport block size (TBS) corresponding to the traffic amount 1ms ahead, and it is sent from the DU to the controller through the CTI every 1ms. The controller consisting of the collection, analysis and control parts is implemented on the server (Intel(R) Xeon(R) Silver 4310), and the measured prediction time is 10 ms. We introduce our parallel prediction scheme using Transformer [6], which is one of the major time-series prediction methods. Every prediction period, the controller converts the TBS

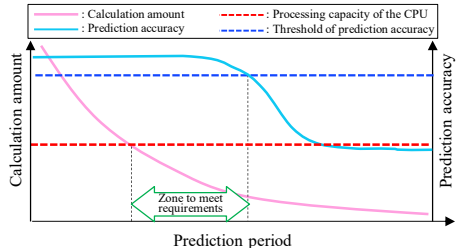


Fig.2. Relationship among prediction period, calculation amount, and prediction accuracy.

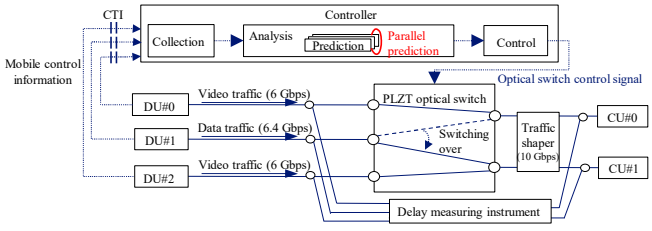
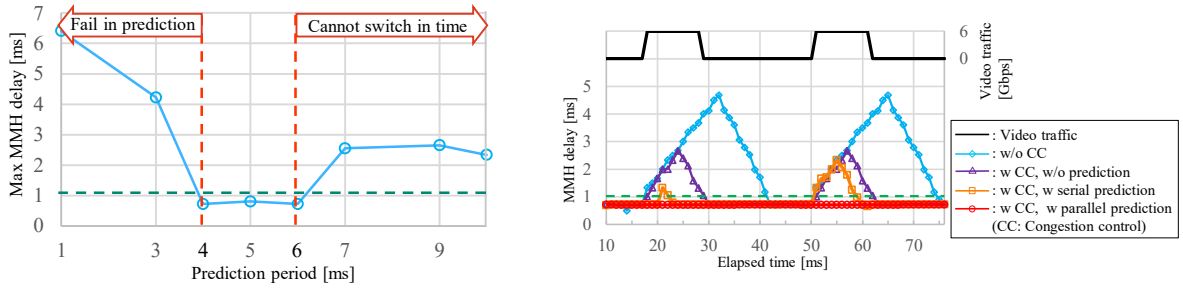


Fig.3. Experimental setup.



(a) Max MMH delay when prediction period is changed. (b) MMH delay as a function of elapsed time.

Fig. 4. Experimental results.

collected every 1ms into time series data. In addition, time-series data for each DU and each prediction period are utilized to forecast future traffic on each thread in parallel. As a preliminary study, we assume that DU#0 and DU#2 transmit only one video traffic with a frame rate of 30 fps (burst period and length are 33 ms and 11 ms, respectively), max bit rate of 6 Gbps, and average bit rate of 2 Gbps. The MMH link rate of 10 Gbps is configured by the traffic shaper. DU#1 accommodates 800 UEs, and UE data traffic has an average bit rate of 8 Mbps. As discussed in Section 2, we experimentally identified the optimal prediction period zone before the validation of proactive congestion control. Figure 4(a) shows the maximum delay at the MMH when the prediction period is changed. As shown in Figure 4(a), we successfully identified the optimal prediction period zone as the periods between 4-ms and 6-ms, which meets the MMH delay requirement of less than 1 ms (green dotted line in Figure 4(a)). We set 4 ms as the optimal prediction period because the shorter period has an advantage for microburst traffic. In our experiment, we set the maximum number of parallel predictions to three, as derived from the optimal prediction period of 4 ms and the prediction time of 10 ms.

Figure 4(b) shows the measured delay at the MMH as a function of elapsed time for our proposed method and several other methods for comparison. Here, we investigated the delay behavior in the MMH when 1) without congestion control, 2) with congestion control and without prediction, 3) with congestion control and serial prediction and 4) with congestion control and parallel predictions. The main results are as follows. 1) Without congestion control, the MMH delay occurred at the elapsed times of 18-32 ms and 51-65 ms when the video burst traffic overlapped with the background data traffic. The measured delay increased to 4.6 ms, which significantly exceeds the MMH delay requirement of 1 ms. At 32-42 ms, the delay decreased to less than 1 ms because the traffic buffered at the CU (shaper) was gradually transmitted after the end of the burst traffic. 2) With congestion control and without prediction, thanks to the switchover of the CU and optical path after congestion happened, the maximum MMH delay was decreased to 2.7 ms (at 24 ms and 57 ms). However, it still exceeded the requirement. 3) With congestion control and serial prediction, the maximum MMH delay was further decreased to 2.3 ms (at 55 ms) thanks to using the traffic prediction. However, it does not meet the requirement due to the insufficient prediction accuracy caused by the long prediction period of 10 ms. 4) With congestion control and parallel prediction, thanks to the improved prediction accuracy brought about by the short prediction period of 4 ms in the parallel prediction, we successfully avoided the congestion. As a result, the MMH delay dramatically decreased to 0.72ms, which meets the requirement of less than 1 ms.

4. Conclusion

We have proposed a proactive congestion control method that performs parallel traffic prediction and fast switchover of the CU and optical path in the MMH. Thanks to the parallel prediction with the experimentally optimized prediction period of 4 ms, our prototype controller successfully achieved a delay at the MMH as low as 1 ms or less without congestion. These results demonstrate that our proposed congestion control technique is promising for emerging services that utilize remotely controlled robots via mobile networks.

References

[1] 3GPP TS22.261 v18.6.1, “Service requirements for the 5G system,” (2022).
[2] 3GPP TS38.913 v17.0.0, “Study on scenarios and requirements for next generation access technologies,” (2022).
[3] Y. Okamoto, et al., “Short-term traffic prediction based on mobile control information for proactive optical switching to lower congestion delay,” in IEICE Proceeding Series 72 O6-2, (IEICE 2022).
[4] A. Sahara, et al., “Congestion-Controlled Optical Burst Switching Network With Connection Guarantee: Design and Demonstration,” IEEE J Lightwave Technol **26**, 2075 - 2086 (2008).
[5] ITU-T G. Supplement 66, Revision 2, “5G Wireless Fronthaul Requirements in a PON Context,” (2019).
[6] Z. Liu, et al., “Forecast Methods for Time Series Data: A Survey,” IEEE Access **9**, 91896-91912 (2021).