Research Reviews on Network-based Congestion Control

Yeunwoong Kyung
Div. of Info. & Comm. Eng.
Kongju National University
Cheonan, Korea
ywkyung@kongju.ac.kr

Jaewook Lee
Dep. of Info. and Comm. Engineering
Pukyong National University
Busan, Korea
jlee0315@pknu.ac.kr

Seokwon Jang
Mobile Core Network Research Section
Electronics and Telecomm. Research Institute
Daejeon, Korea
seokwon.jang@etri.re.kr

Haneul Ko
Dep. of Electronic Eng.
Kyung Hee University
Yongin-si, Korea
heko@khu.ac.kr

Seungmin Oh, Jaehwan Lee
Dep. of Computer Science and Eng.
Kongju National University
Cheonan, Korea
smoh@kongju.ac.kr, jhnlee@kongju.ac.kr

Youngjun Kim
School of Computer Science and Eng.
Kyungnam University
Changwon, Korea
youngjun@kyungnam.ac.kr

***Abstract*—Congestion control has been considered as a principal technology for efficient end-to-end traffic transmission. Previous studies on congestion control are mainly focused on the control of the transmission rate or window size at the host. However, such host-based congestion controls have difficulties on short time-scale variable network conditions and increased host complexity. To address this problem, network-based congestion controls that utilize network information have been considered which determine and control the congestion using the network information. In this paper, we review the network-based congestion control methods.**

***Index Terms*—Congestion control, 5G network architecture, Network-based control**

I. Introduction

Currently, lots of applications use the Transmission Control Protocol (TCP) method as a transport layer protocol. TCP supports congestion and flow controls, which aim for efficient end-to-end traffic transmission. Congestion control is a traditional research topic that is still receiving a lot of attention for managing and controlling congestion where traffic flow in the network exceeds the capacity of the network, resulting in queueing delays or packet loss. Congestion control methods are needed to prevent these issues because congestion situations can directly affect the performance of services. Based on the congestion control, hosts that perform end-to-end traffic transmission can prevent excessive data transmission and bandwidth occupancy.

Previous studies on congestion control are mainly focused on the control of the transmission rate or window size at the host (user terminal, server, etc.) in charge of end-to-end transmission [1].

Recently, several studies have presented that such host-based congestion controls have difficulties on short time-scale variable network conditions such as micro-busts traffic and increased host complexity such as maintenance of software or protocol stacks [2], [3].

To address this problem, network-based congestion controls that utilize network information have been considered. In addition, there have been recent works to introduce congestion controls considering 5G network architecture. In this paper, Chapter 2 presents congestion control methods utilizing network information and Chapter 3 introduces congestion control methods considering 5G network architecture.

II. Congestion Control utilizing Network Information

Congestion control methods utilizing network information can be divided into two approaches: 1) centralized approach [4]–[6] and 2) distributed (or hybrid) approach [2], [3], [7].

Ruffy *et al.* [4] introduced an emulator that can apply centralized and distributed reinforcement learning algorithms in the data center network. In particular, the data center network is expected to be able to determine the optimal congestion control policy through information on these network equipment because operators can fully control all hosts and network equipment. Lei *et al.* [5] proposed a congestion control method based on multi-task DRL (Deep Reinforcement Learning) in the SDN architecture. Since the SDN architecture was assumed, SDN controller can collect information on the data plane, and the multi-task means congestion control and load distribution. For performance analysis, simulations were performed based on Mininet and Python RLib, and it was confirmed that the round trip time (RTT) was reduced compared to the existing method. Li *et al.* [6] presented a data-driven traffic engineering structure. In this structure, network state information can be obtained based on detailed granularity by assuming INT (In-band Network Telemetry). In this study, DRL was also used to solve the traffic engineering problem and the control plane controls the network through a flow

Identify applicable funding agency here. If none, delete this.

rule for optimal link utilization. For performance analysis, bmv2 and customized Python controllers were used, and it was confirmed that this study can reduce the max-utilization of the network compared to conventional works.

Mai *et al.* [2] proposed an in-network control structure based on a hybrid approach. In this study, a programmable data plane is assumed where local measurements such as network events are performed based on the telemetry probe. In addition, learning is performed based on the information, and data plane control is performed through P4 when the control method is determined. Based on global knowledge, the centralized management plane determines the coordination of data plane switches and control methods to improve overall network performance. The specific methodology used in this study is presented in [7] where the load distribution problem is solved by means of multi-agent factor-critical RL. Pan *et al.* [3] introduced a distributed congestion control method based on INT in a space information network composed of host, ground-based station, and space-based networks such as satellites. Each node in the network can collect congestion information on all neighboring links through the probe. Then, it can inform the ground-based station of feedback on congestion information. In addition, the ground-based station can provide the information to the host to reduce the sending rate, which can guarantee a lower RTT than conventional end-to-end congestion control methods. Mininet and bmv2 were used for performance analysis and it was confirmed that this study can have a lower queue depth than conventional methods.

III. Congestion Control considering 5G Network Architecture

Congestion control studies considering the 5G network architecture can be classified into 1) UPF (User Plane Function)-relocation [8], [9], 2) determining TCP window size in 5G cores [10], 3) API exposure [11], and 4) the feedback-based approach [11]–[13].

First, studies that relocate UPF mainly handle the problem of minimizing self-defined costs based on ILP (Integer Linear Programming) or heuristic algorithms [8], [9]. For example, in Pupo's study [8], UPF deployment and relocation costs considering the frequency of handover were used for overall cost. Then, the problem of minimizing cost was formulated as ILP. In addition, since the problem is NP-hard, it solves the problem by means of a near-optimal rearrangement algorithm.

Kanakaraj *et al.* [10] introduced a study to determine the TCP window size in 5G cores. In this method, SMF (Session Management Function) monitors the load and traffic of UPF to determine the maxRWND (receive window) of TCP based on a ML model. In addition, if the current RWND is greater than maxRWND by checking it in UPF, the corresponding RWND is replaced by maxRWND and transmitted. Since the host sets the corresponding maxRWND to the window size, congestion can be prevented.

Huang *et al.* [11] suggested three methods for providing congestion information. The first method is to detect congestion in RAN, mark ECN (Explicit Congestion Notification) on the IP layer, and send it to UE and UPF. The second method is to detect congestion in the RAN, inform it to UPF, and deliver it with ECN marking in UPF. The third method is to extend the function of the current API to request the average window size for a specific service flow from the application to the 5G system, or to notify the application of the congestion information or the supported transmission rate through the API from the 5G system.

Wu *et al.* [12] determines whether to increase or decrease the transmission rate based on the buffer status of the RLC (Radio Link Control) layer. Then, UPF marks the information in the packet header and transmits it to the host. Accordingly, the host can prevent congestion by changing the transmission rate. Similarly, in Keirkhah's study [13], an agent in PDCP (Packet Data Convergence Protocol) layer provides feedback to hosts based on state information of the RLC and MAC layers. This allows the host to change the transmission rate.

IV. Conclusion

This paper reviews studies that utilize network information or perform congestion control directly at the networks. Then, this paper introduces studies that suggest congestion control methods considering 5G network architecture. Recently, a number of studies have been published that use ML to perform congestion control. In 5G or beyond 5G, it is expected that NWDAF (Network Data Analysis Function) will be utilized to determine or predict congestion information based on various information of 5G systems.

Acknowledgment

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

References

[1] R. A-Saadi, G. Armitage, J. But, and P. Branch, "A Survey of Delay-Based and Hybrid TCP Congestion Control Algorithms," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, Fourth Quarter 2019.

[2] T. Mai, S. Garg, H. Yao, J. Nie, G. Kaddoum, and Z. Xiong, "In-Network Intelligence Control: Toward a Self-Driving Networking Architecture," *IEEE Network*, vol. 35, no. 2, pp. 53–59, March/April 2021.

[3] H. Pan, H. Yao, T. Mai, N. Zhang, and Y. Liu, "Scalable Traffic Control Using Programmable Data Planes in Space Information Network," *IEEE Network*, vol. 35, no. 4, pp. 35–41, Aug. 2019.

[4] F. Ruffy, M. Przystupa, and I. Beschastnikh, "Iroko: A Framework to Prototype Reinforcement Learning for Data Center Traffic Control," *in Proc.* NPIS, 2018.

[5] K. Lei, Y. Liang, and W. Li, "Congestion Control in SDN-Based Networks via Multi-Task Deep Reinforcement Learning," *IEEE Network*, vol. 34, no. 4, pp. 28–34, July/August 2020.

[6] Q. Li, J. Zhang, T. Pan, T. Huang, and Y. Liu, "Data-driven Routing Optimization based on Programmable Data Plane," *in Proc.* ICCCN, Aug. 2020.

[7] T. Mai, H. Yao, Z. Xiong, S. Guo, and D. T. Niyato, "Multi-Agent Actor-Critic Reinforcement Learning based In-network Load Balance," *in Proc.* IEEE Globecom, Dec. 2020.

[8] I. L-Pupo *et al.*, "A Framework for the Joint Placement of Edge Service Infrastructure and User Plane Functions for 5G," *Sensors*, 2019.

[9] I. L-Pupo *et al.*, "Dynamic UPF placement and chaining reconfiguration in 5G networks," *Computer Networks*, 2022.

[10] V. Kanakaraj *et al.*, "ML Assisted Feedback Mechanism for TCP Congestion Control in Next Generation Wireless Networks," *in Proc.* IEEE ANTS, 2022.
[11] Z. Huang *et al.*, "Standard Evolution of 5G-Advanced and Future Mobile Network for Extended Reality and Metaverse," *IEEE IoT-J*, 2023.
[12] Wu *et al.*, "AQM-based Buffer Delay Guarantee for Congestion Control in 5G Networks," *in Proc.* IEEE WCNC, 2023.
[13] M. Kheirkhah *et al.*, "XRC: An Explicit Rate Control for Future Cellular Networks," *in Proc.* IEEE ICC, 2022.