5G Latency Analysis for Optimal TCP Cubic Congestion Control for 5G

Junha Park and Saewoong Bahk
Department of Electrical and Computer Engineering and Institute of New Media and Communications,
Seoul National University, Seoul, Korea
Email: jhpark@netlab.snu.ac.kr, sbahk@snu.ac.kr

***Abstract*—In this study, we investigate the latency characteristics of commercial 5G networks through empirical measurements and analyze their implications for TCP Cubic congestion control. Our measurement results reveal that 5G latency exhibits variable round-trip times and sudden delay spikes, creating irregular time intervals that may indirectly affect congestion control dynamics. While TCP CUBIC, the default algorithm in Linux, was originally optimized for stable high-bandwidth networks, the observed 5G latency behavior suggests potential challenges such as premature congestion window reductions and throughput inefficiency. Our study highlights how the temporal patterns of 5G latency can influence transport-layer performance, thereby motivating the need for congestion control mechanisms specifically tailored to the variability of 5G networks.**

***Index Terms*—5G, cellular latency, TCP, congestion control**

I. Introduction

The fifth generation (5G) of mobile communication systems promises ultra-low latency, high reliability, and enhanced data rates, enabling a wide range of applications such as real-time streaming, cloud gaming, autonomous driving, and industrial automation. While significant advances have been made in the physical and network layers of 5G, the end-to-end transport performance still largely relies on Transmission Control Protocol (TCP), which was originally designed for wired networks with relatively stable latency characteristics. Consequently, understanding how TCP congestion control interacts with 5G latency behavior is a crucial step toward achieving seamless performance in next-generation mobile environments.

TCP CUBIC, the default congestion control algorithm in Linux and one of the most widely deployed schemes in today's Internet, has been optimized for high-bandwidth, high-delay networks. [1]. Its cubic window growth function enables efficient bandwidth utilization in stable environments, yet its behavior in mobile networks—particularly in 5G with highly variable round-trip times (RTTs), scheduling delays, and transient latency spikes—remains insufficiently understood. Latency fluctuations in 5G can cause premature congestion window reduction, misinterpretation of network conditions, and suboptimal throughput, thereby limiting the potential of emerging 5G applications.

To address this gap, this paper conducts an empirical study of 5G latency characteristics and evaluates the performance of TCP CUBIC under real-world conditions. By combining latency measurement experiments with congestion control performance analysis, we aim to uncover the extent to which 5G latency patterns influence TCP dynamics. Furthermore, we discuss the limitations of current congestion control mechanisms and explore directions for developing 5G-tailored TCP algorithms that can better accommodate latency variability and ensure reliable, high-throughput communication.

The main contributions of this work are as follows: (1) Empirical analysis of 5G latency characteristics observed in commercial network environments, (2) Performance evaluation of TCP CUBIC under measured conditions, (3) Highlighting the sensitivity of congestion window dynamics to latency variations

By bridging the gap between 5G latency behavior and transport-layer protocol design, this work provides insights into optimizing TCP performance for next-generation mobile networks and paves the way for congestion control mechanisms that are more resilient to the dynamic nature of wireless latency.

II. background

A. Round-Trip-Time in 5G

Round-Trip Time (RTT) is a fundamental performance metric in transport-layer protocols, reflecting the elapsed time between sending a packet and receiving its acknowledgment. In TCP, RTT directly influences congestion window growth, retransmission timeout estimation, and overall throughput efficiency. Thus, understanding RTT dynamics is essential for evaluating the performance of congestion control algorithms.

With the advent of the fifth generation (5G) mobile communication system, RTT characteristics have undergone significant changes compared to previous generations [2]. 5G New Radio (NR) is designed to achieve ultra-reliable low-latency communication (URLLC), targeting end-to-end latencies below 1 ms in ideal conditions. This is made possible through shorter Transmission Time Intervals (TTIs), flexible numerologies, and advanced scheduling mechanisms at the radio access network (RAN). In practice, however, RTT in commercial 5G networks is affected by multiple factors including radio resource allocation, core network processing, mobility, and backhaul conditions [3]. As a result, while the average RTT may be lower than that of LTE or legacy systems, transient spikes and fluctuations are frequently observed.

Such variability in RTT poses unique challenges for TCP congestion control. Conventional algorithms, such as TCP CUBIC, often assume relatively stable RTTs and interpret

delay variations as signals of congestion. In 5G environm however, sudden RTT increases may not necessarily cate congestion but rather reflect scheduling delays, hanc events, or temporary link quality degradation. This discrep can cause TCP to misreact, leading to premature conge window reductions and degraded throughput performanc

Therefore, characterizing 5G RTT behavior and unders ing its interaction with TCP congestion control mechar is critical. A comprehensive analysis can provide insights the limitations of existing algorithms and guide the d of congestion control schemes better aligned with the lat dynamics of 5G networks.

B. TCP Cubic congestion control

Transmission Control Protocol (TCP) has long served as the dominant transport-layer protocol for reliable communication over the Internet [4]. To efficiently utilize network resources, TCP employs congestion control mechanisms that adapt the sending rate according to perceived network conditions [5]. Among the various algorithms developed, TCP CUBIC has emerged as one of the most widely deployed schemes and is the default congestion control algorithm in Linux operating systems.

TCP CUBIC was introduced as an enhancement over traditional loss-based algorithms such as TCP Reno [1]. Its key innovation lies in its cubic window growth function, which replaces the linear growth model of Reno with a cubic function of elapsed time since the last congestion event. This design enables faster probing of available bandwidth in high-bandwidth, high-latency networks while maintaining stability in steady-state conditions. Specifically, CUBIC grows the congestion window (cwnd) more aggressively when it is far from the previous maximum, and more conservatively when it approaches that maximum, thereby achieving a balance between throughput efficiency and fairness.

Another important feature of CUBIC is its reduced dependence on Round-Trip Time (RTT). Unlike Reno, where cwnd growth is linearly tied to RTT, CUBIC's growth function is primarily time-based, making it more scalable across heterogeneous network environments. This RTT-fairness property has contributed to its adoption as the de facto standard in modern operating systems.

However, despite its advantages, TCP CUBIC was designed primarily with relatively stable wired networks in mind. In highly dynamic wireless environments such as 5G, the latency characteristics differ significantly due to scheduling delays, variable link quality, and mobility-induced events. In such cases, CUBIC may misinterpret transient RTT fluctuations as congestion signals, triggering premature window reductions and degrading throughput performance. These limitations underscore the importance of re-examining TCP CUBIC's behavior under 5G conditions and exploring potential enhancements or alternative algorithms tailored to next-generation mobile networks.

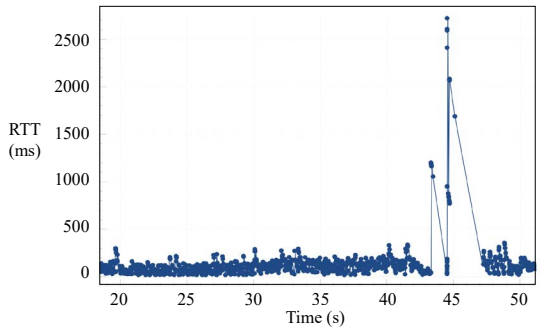

Fig. 1: Round-Trip-Time of 5G

III. Motivation

Although 5G networks are designed to provide ultra-low latency and high reliability, practical measurements show that latency behavior in 5G is far from stable. Factors such as dynamic scheduling, radio resource allocation, and mobility lead to highly variable round-trip times and occasional latency spikes. These irregularities directly affect transport-layer protocols, especially TCP, which interprets delay variations as potential congestion signals.

TCP CUBIC, the default congestion control algorithm in most operating systems, is optimized for high-bandwidth and relatively stable latency environments. However, in 5G networks, its cubic window growth mechanism may misinterpret transient delay increases as congestion events, resulting in unnecessary congestion window reductions and throughput degradation. This mismatch between 5G latency characteristics and the assumptions of conventional congestion control highlights a critical gap: current TCP algorithms are not fully adapted to the dynamics of next-generation mobile networks.

Understanding the relationship between 5G latency and TCP congestion control is therefore essential. By empirically analyzing how TCP CUBIC behaves under real commercial networks, we can identify its limitations and derive insights for designing new congestion control mechanisms that are resilient to latency variability. Ultimately, such improvements are crucial to unlock the full potential of 5G applications, from interactive multimedia services to ultra-reliable low-latency communications (URLLC).

IV. Measurement Methodology

To analyze the relationship between 5G latency behavior and TCP congestion control performance, we designed an empirical measurement study consisting of two components: (i) latency measurements in commercial 5G networks and (ii) congestion window dynamics of TCP CUBIC under the same network conditions. The overall goal of this methodology is to obtain complementary datasets that enable us to explore how 5G latency patterns can indirectly influence TCP behavior.

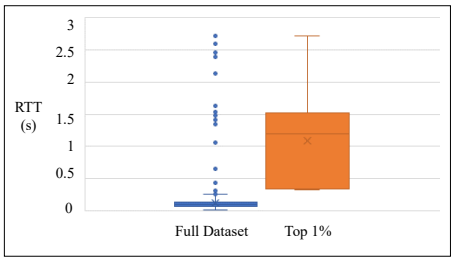

Fig. 2: Round-Trip-Time Dataset Analysis

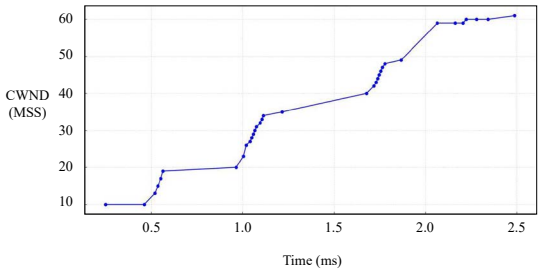

Fig. 3: TCP Cubic Measurement

A. Round-Trip-Time in 5G Measurement

We first measured the latency characteristics of a commercial 5G network, as shown in Fig. 1. Latency was captured using iperf3 to generate controlled TCP flows and Wireshark to record packet traces. Round-Trip Time (RTT) values were extracted from the captured traces by analyzing TCP acknowledgments and timestamp information. This setup allowed us to observe the time-varying behavior of 5G latency, including average RTT, variability, and occasional spikes.

B. TCP Cubic congestion control measurement

In parallel, we measured the dynamics of the TCP CUBIC congestion control algorithm, with results presented in Fig. 3. Unlike the latency measurements, this experiment was not conducted in a commercial 5G environment. Instead, we used iperf3 in TCP mode with CUBIC enabled as the congestion control algorithm (default in Linux) to observe its behavior in a controlled network setting. Packet traces were collected using Wireshark, and the congestion window (cwnd) values were extracted over time. While these measurements were obtained outside of 5G, they provide a baseline for understanding CUBIC's inherent behavior and serve as a reference point for analyzing how 5G latency patterns could potentially influence congestion control dynamics.

C. Correlation Analysis between Latency and Congestion Control

Although the two measurement datasets—latency traces and congestion window dynamics—were collected separately, our objective was to examine the temporal correspondence between them. By comparing the intervals of sudden latency increases with the cwnd evolution of TCP CUBIC, we aim to identify whether 5G latency spikes can indirectly trigger congestion responses in CUBIC. This methodology provides the foundation for understanding the interaction between 5G latency variability and transport-layer performance.

V. Results and Analysis

Fig. 1 illustrates the latency behavior of a commercial 5G network over time. The measurement results reveal that, although the average latency remains relatively low, there are distinct intervals where latency exhibits sudden spikes. These abrupt increases indicate the presence of transient delay events, likely caused by scheduling dynamics, radio resource variations, or mobility-related factors. Such characteristics highlight the non-negligible variability of 5G latency, which can significantly influence transport-layer performance and congestion control behavior.

Fig. 2 presents a boxplot comparison between the full 5G latency dataset and the top 1 percent latency values. The results clearly demonstrate a substantial disparity: while the majority of latency measurements remain within a relatively stable range, the extreme top 1 percent values are significantly higher, forming heavy-tailed outliers. This observation indicates that although 5G generally achieves low-latency performance, rare but severe delay events can disproportionately affect overall latency distribution and transport-layer behavior.

Fig. 3 shows the congestion window (cwnd) dynamics of the conventional TCP CUBIC congestion control algorithm measured over time. The results illustrate how cwnd evolves under real-world network conditions, reflecting both the cubic growth pattern and the sensitivity of window adjustments to latency fluctuations. This measurement provides an empirical basis for analyzing the interaction between 5G latency variability and TCP CUBIC's congestion control behavior.

A. Measured 5G RTT Characteristics

As shown in Fig. 1, the measured RTT in the 5G environment remained on the order of several tens of milliseconds on average, but with noticeable spikes and variability. Such RTT fluctuations not only reflect lower-layer transmission delays but also directly affect the performance of the transport-layer congestion control. In particular, when RTT becomes large or unstable, the acknowledgment (ACK) feedback interval lengthens, delaying the update cycle of the congestion window (cwnd).

B. TCP CUBIC cwnd Growth under 5G RTT

TCP CUBIC employs a cubic function to drive cwnd growth during the congestion avoidance phase. However, because cwnd is updated at the granularity of RTT, a larger RTT inevitably slows down the practical cwnd growth. As illustrated in Fig. 3, under identical CUBIC parameters, cwnd evolution in the 5G environment lags behind that in lower-latency networks. This results in slower convergence of the congestion window and a transmission rate that falls short of

the expected link capacity. In other words, the inherently high Bandwidth-Delay Product (BDP) of 5G networks exposes a structural inefficiency in the standard CUBIC algorithm.

C. RTT Impact on Congestion Control Efficiency

The measured results reveal that large RTTs in 5G environments significantly reduce the efficiency of TCP CUBIC. Longer acknowledgment intervals slow down the pace of cwnd growth, creating a persistent gap between the available link capacity and the achieved throughput. After a packet loss event, the cubic recovery curve is stretched across longer RTTs, which increases the time required for cwnd to return to its previous value. Consequently, the high bandwidth offered by 5G networks is underutilized, leading to degraded throughput and increased latency at the application layer.

D. Necessity of Optimized CUBIC for 5G

These findings demonstrate that the standard CUBIC algorithm is not fully adequate for high-bandwidth, high-latency environments such as 5G. To address this limitation, an optimized version of CUBIC tailored for 5G should be developed. Such an adaptation could include RTT-aware cwnd scaling that accelerates growth in long-delay networks, hybrid mechanisms that consider both RTT trends and packet loss when adjusting congestion states, and refined recovery tuning that enables faster cwnd restoration after loss events. Without such enhancements, TCP CUBIC cannot fully exploit the performance potential of 5G, leaving a substantial portion of available bandwidth unused.

VI. Conclusion and Future Work

In this work, we examined the impact of 5G latency characteristics on TCP congestion control through empirical measurements and real-world performance evaluation of TCP CUBIC. Our study confirmed that the latency dynamics of 5G—marked by variability, sudden spikes, and scheduling-induced fluctuations—significantly influence the behavior of TCP CUBIC, often leading to premature window reductions and throughput degradation. These findings highlight the limitations of conventional congestion control algorithms, which were primarily designed for wired or more stable network environments, when applied directly to next-generation mobile systems.

By analyzing both latency patterns and transport-layer responses, this research provides new insights into the interplay between 5G network behavior and TCP dynamics. The results suggest that future congestion control schemes must explicitly account for the unique latency characteristics of 5G, moving beyond loss- and delay-based heuristics toward adaptive, context-aware mechanisms. Such designs are essential for achieving robust end-to-end performance and ensuring that 5G networks can fully support latency-sensitive and bandwidth-intensive applications.

In summary, our work underscores the need for rethinking TCP congestion control in the 5G era. Future research should focus on developing algorithms that integrate cross-layer awareness, predictive latency modeling, and mobility adaptation, thereby enabling TCP to operate efficiently in highly dynamic 5G environments and paving the way for next-generation Internet transport protocols.

References

[1] S. Ha, I. Rhee, and L. Xu, "Cubic: a new tcp-friendly high-speed tcp variant," *ACM SIGOPS Operating Systems Review*, vol. 42, no. 5, pp. 64–74, 2008.

[2] "System architecture for the 5G system (5gs); stage 2," 3GPP, Tech. Rep. TS 23.501 V16.6.0, Oct. 2020. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/16.06.00_60/ts_123501v160600p.pdf

[3] D. Xu, A. Zhou, X. Zhang, G. Wang, X. Liu, C. An, Y. Shi, L. Liu, and H. Ma, "Understanding operational 5G: A first measurement study on its coverage, performance and energy consumption," in *Proceedings of the ACM SIGCOMM 2020 Conference*. ACM, 2020, pp. 479–494. [Online]. Available: https://dl.acm.org/doi/10.1145/3387514.3405882

[4] M. Allman, V. Paxson, and E. Blanton, "TCP congestion control," IETF, RFC 5681, Sep. 2009. [Online]. Available: https://www.rfc-editor.org/rfc/rfc5681.html

[5] I. Rhee, L. Xu, S. Ha, A. Zimmermann, L. Eggert, and R. Scheffenegger, "CUBIC for fast long-distance networks," RFC 8312, Feb. 2018. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8312.html