Understanding the Impact of Cellular RAN-induced Delay on Video Conferencing

Fan Yi
Princeton University
Princeton, USA
fanyi@princeton.edu

Oliver Michel
Princeton University
Princeton, USA
omichel@princeton.edu

Haoran Wan
Princeton University
Princeton, USA
hw8161@princeton.edu

Kyle Jamieson
Princeton University
Princeton, USA
kylej@princeton.edu

ABSTRACT

Congestion-control algorithms for video-conferencing applications work well in wired networks but are fragile in cellular networks due to high delay variations and variable capacity in these networks. This paper investigates the causes of delay variations in cellular networks using a cross-layer approach. By measuring a WebRTC application over LTE at both the physical and network layers, we identify the effects of such delay inflation caused by physical-layer resource scheduling and link-layer retransmissions.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Mobile networks**.

KEYWORDS

Cellular Network, Video Conferencing, Measurement

ACM Reference Format:
Fan Yi, Oliver Michel, Haoran Wan, and Kyle Jamieson. 2024. Understanding the Impact of Cellular RAN-induced Delay on Video Conferencing. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3636534.3697445

1 INTRODUCTION

Video-conferencing applications (VCAs), including Zoom [4], have gained popularity over the past several years and have become essential in businesses and educatione. VCAs rely on congestion control to dynamically adapt bitrate and adjust their sending rate to varying network conditions. These algorithms are designed based on the bottleneck-link model, which is effective in wired networks but falls short in the context of cellular networks. The inherent variability in capacity and latency (e.g., due to scheduling logic or link-layer retransmissions) of cellular networks poses substantial challenges for VCAs, hindering the accurate estimation of network conditions. This undermines the ability of VCAs to determine the appropriate bitrate for video and audio streams, causing applications to oversaturate the link or encode video at a lower than possible quality. Both problems negatively affect the meeting quality experienced by users. While some existing research [2, 3] leverages machine learning to learn congestion control logic for heterogenous networks, there remains a significant gap in understanding the underpinnings of delay variations, particularly in cellular networks.

This paper aims to bridge this gap by investigating the cause of delay variations in cellular networks. Specifically, we employ a cross-layer approach to measure WebRTC [1] over LTE networks at both the physical layer, by decoding the LTE control channel, and at the network layer. Through these measurements, we uncover the mechanisms behind delay variations caused by physical-layer resource scheduling and link-layer retransmissions. Such variations mislead rate-adaptation and congestion-control algorithms in VCAs as they are not directly related to link overuse (i.e., congestion) or underuse. By understanding these factors, we can better address the unique challenges posed by cellular networks and improve the design of congestion control in VCAs.

2 BACKGROUND

Delay-based Congestion Control. Video-conferencing applications (VCAs) require low latency and are highly sensitive

to network delay, necessitating the use of delay-based congestion control mechanisms to maintain optimal performance. WebRTC uses GCC [1], which operates by computing the one-way delay gradient to monitor network utilization. It applies a trendline filter to the raw delay gradient to smooth out fluctuations, and compares the smoothed delay gradient to a threshold to detect network overuse or underuse.

LTE TDD Frame Structure. In this work, we focus on LTE networks in TDD mode. Fig. 1 illustrates an example frame structure of a LTE TDD cell. It has a UL-DL switching periodicity of five subframes, with four downlink subframes and one uplink subframe in each period. LTE base station assigns transport blocks (TB) to user equipment (UE) to manage data transmission. The Transport Block Size (TBS) refers to the size of the amount of data that can be sent in a single subframe. Previous studies [8] have demonstrated telemetry tools capable of decoding the LTE control channel to obtain physical-layer TBS at the millisecond level. In this work, we utilize such tools to obtain the physical layer resource-scheduling information relevant to our target UE at the millisecond level.

3 MEASUREMENT SETUP

To investigate the LTE RAN-induced delay variance in VCAs, we set up an operational private LTE network using a Sercomm 4G cell [6] and the Aether OnRamp 4G core [5], as illustrated in Fig. 2. In our setup, two clients run a customized WebRTC application. We send video and audio from a sender to a receiver where the sender is connected to the private 4G small cell while the receiver is wired to the 4G core subnet. All hosts are time-synchronized using the Network Time Protocol (NTP). We conduct a 10-minute WebRTC video call, capturing packets and WebRTC internal statistics at both the sender and receiver sides. Concurrently, we employ NGScope [8] to collect TBS information.

4 RAN-INDUCED DELAY VARIATION

In the Section, we delve into three main factors contributing to LTE uplink delay variations: resource scheduling, HARQ retransmissions and RLC-layer retransmissions.

4.1 Resource Scheduling

We present a time-series analysis (Fig. 3(a)) to illustrate the impact of resource allocation on VCA traffic latency. The upper portion of the diagram depicts packet transmission, with each horizontal line representing a single packet. The line's leftmost point indicates the sending time, while its rightmost point shows the reception time. Line length corresponds to the sender-to-receiver delay. Concurrently, the lower section displays the TBS during the same time interval, representing PHY-layer bit transmission per subframe. Vertical dashed lines connect both sections, highlighting which packets are conveyed by each Transport Block (TB).

In LTE networks, the UE transmits Buffer Status Reports (BSR) to communicate the amount of pending data it needs to send. The base station then assigns uplink grants based on these reports, determining the allocated TBS (depicted by green vertical bars in Fig. 3(a)). However, a lag exists between BSR transmission and grant utilization [7]. Our observations reveal that private 4G cells employ proactive grants (blue vertical bars in Fig. 3(a)). These pre-allocated uplink resources are given to active UEs, with the goal of mitigating BSR scheduling delays at the expense of bandwidth efficiency and computational resources.

In VCA applications, each video frame typically consists of multiple packets sent in a burst. When a packet is ready at the UE, an initial proactive TB can transmit one or two packets which is usually not enough for the entire frame. Given our cell's 5ms downlink-uplink switching pattern, subsequent proactive grants become available at 5ms intervals, enabling the UE to send additional packets. This cycle continues until the BSR-requested grant arrives, at which point the remaining buffered packets are transmitted via the BSR-requested TB. This scheduling mechanism induces a large delay spread between the first and last packets of a video frame, as indicated by the yellow double arrows in Fig. 3(a).

4.2 Link-layer Retransmissions

Cellular networks employ various mechanisms to handle transmission errors and data loss, particularly in challenging environments with high interference or mobility. These mechanisms include HARQ and RLC-layer retransmissions.

HARQ Retransmissions. We utilize another time-series example Fig. 3(b) to illustrate the impact of HARQ retransmissions, with failed and retransmitted TBs highlighted in red and purple bars, respectively. When a TB fails to transmit, the HARQ protocol initiates a retransmission. This process typically extends the packet delay by approximately 10 milliseconds, as depicted by the green double arrows in Fig. 3(b). In cases where retransmitted TBs encounter subsequent failures at the base station, the HARQ performs additional retransmission attempts. Each retry increments the packet delay by 10 milliseconds, which adds additional latency variation.

RLC-layer Retransmissions. In LTE networks, a maximum of three HARQ retransmissions are attempted for the same TB. If all three retransmissions fail, it triggers an RLC-layer retransmission. The delay for RLC-layer retransmissions is determined by several parameters set by the cell, including the RLC retransmission timer and the status report feedback mechanism. As presented in Fig. 3(c), when RLC-layer retransmissions occur, the packet delay are significantly increased, leading to an overall delay of more than 100 ms.

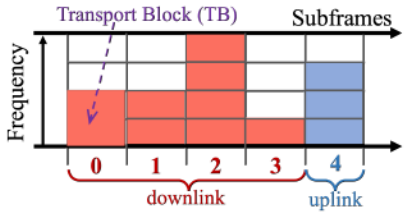


Fig. 1 — The LTE TDD frame structure of our private small cell.

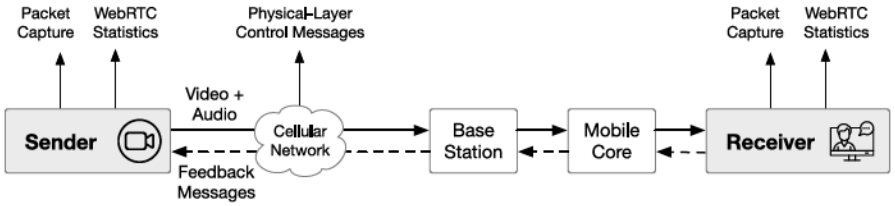
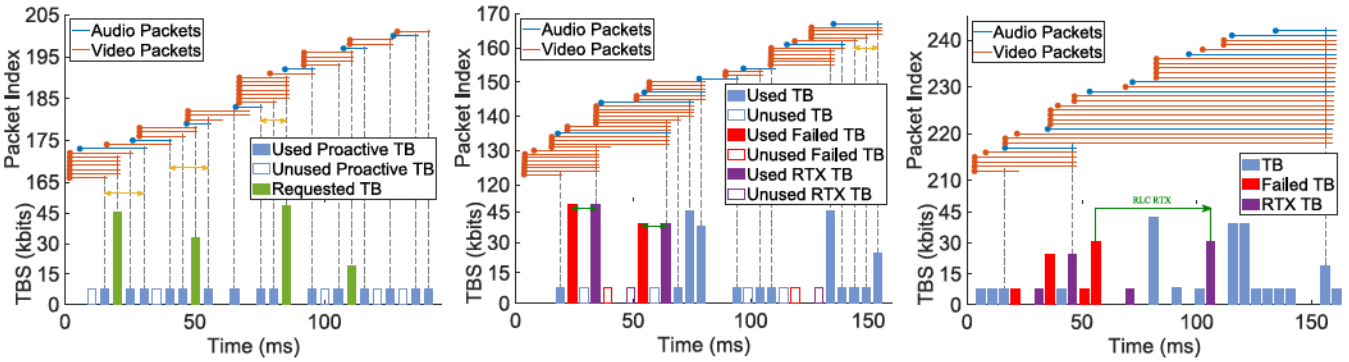


Fig. 2 — The measurement setup consists of two clients running WebRTC applications with the sender accessing the network via LTE.



(a) Resource scheduling induces frame-Level delay spreads in 5 ms increments, denoted as yellow double arrows.

(b) The HARQ retransmissions increases the packet delay by multiples of 10 ms, denoted as green double arrows.

(c) The RLC-layer retransmissions increase the packet delay up to 100 ms level, denoted as green double arrows.

Fig. 3 — Time series analysis of collected traffic traces, integrating transport layer packet transmissions and physical layer TB allocations. Dashed lines connect each packet to its corresponding TB.

Additionally, since the RLC layer guarantees in-order delivery of packets, an RLC-retransmitted packet will cause all subsequent packets in the burst to be delayed.

5 CONCLUSION

To sum up, our study highlights the significant impact of delay variations caused by resource scheduling and link-layer retransmissions in cellular networks on VCAs. By employing a cross-layer measurement approach, we provide valuable insights that can improve the design of VCA congestion control algorithms over cellular networks.

6 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. CNS-2223556 and DARPA grant HR001120C0107.

WORKS CITED

[1] Gaetano Carlucci, Luca De Cicco, Stefan Holmer, and Saverio Mascolo. 2016. Analysis and design of the google congestion control for web real-time communication (WebRTC). In *MMSys '16*. 1–12.

[2] Mo Dong, Qingxi Li, Doron Zarchy, P Brighten Godfrey, and Michael Schapira. 2015. PCC: Re-architecting congestion control for consistent high performance. In *NSDI '15*. 395–408.

[3] Mo Dong, Tong Meng, Doron Zarchy, Engin Arslan, Yossi Gilad, Brighten Godfrey, and Michael Schapira. 2018. PCC vivace: Online-Learning congestion control. In *NSDI '18*. 343–356.

[4] Oliver Michel, Satadal Sengupta, Hyojoon Kim, Ravi Netravali, and Jennifer Rexford. 2022. Enabling Passive Measurement of Zoom Performance in Production Networks. In *IMC '22*. 244–260.

[5] Open Networking Foundation. 2024. Aether: An ONF Project. [org].

[6] Sercomm 2024. Sercomm 4G Cell. [url].

[7] Zhaowei Tan, Jinghao Zhao, Yuanjie Li, Yifei Xu, and Songwu Lu. 2021. Device-Based LTE latency reduction at the application layer. In *NSDI '21*. 471–486.

[8] Yaxiong Xie and Kyle Jamieson. 2022. Ng-scope: Fine-grained telemetry for NextG cellular networks. *ACM SIGMETRICS* 6, 1 (2022), 1–26.