Athena: Seeing and Mitigating Wireless Impact on Video Conferencing and Beyond

Fan Yi
Princeton University
fanyi@princeton.edu

Haoran Wan
Princeton University
haoran.w@princeton.edu

Kyle Jamieson
Princeton University
kylej@princeton.edu

Jennifer Rexford
Princeton University
jrex@princeton.edu

Yaxiong Xie
University at Buffalo
yaxiongx@buffalo.edu

Oliver Michel
Princeton University
omichel@princeton.edu

ABSTRACT

Rapid delay variations in today's access networks impair the QoE of low-latency, interactive applications, such as video conferencing. To tackle this problem, we propose Athena, a framework that correlates high-resolution measurements from Layer 1 to Layer 7 to remove the fog from the window through which today's video-conferencing congestion-control algorithms see the network. This cross-layer view of the network empowers the networking community to revisit and re-evaluate their network designs and application scheduling and rate-adaptation algorithms in light of the complex, heterogeneous networks that are in use today, paving the way for network-aware applications and application-aware networks.

CCS CONCEPTS

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

KEYWORDS

Video Conferencing, Network Measurement, 5G Networks

1 INTRODUCTION

Interactive Video-Conferencing Applications (VCAs) such as Google Meet [15] and Zoom [46] are ubiquitous [13], yet unreliable [11, 30]. The vagaries of today's heterogeneous wireless access networks (4G, 5G, Wi-Fi, and low-earth orbit satellite)—in particular their capacity and latency variations—challenges VCAs' estimation of these variables, frustrating their task of encoding video and audio media streams that match this capacity [3, 8, 9, 24, 28] to maximize interactive video quality. Wireless access technologies are complex and necessarily employ sophisticated methods to enable multiple access to a shared medium and increase the reliability of data transmission at the link layer. Yet these same methods introduce various artifacts in the datagram stream higher layers see, such as rapidly changing packet delays and link capacities. Today, congestion control and VCA bit-rate adaptation algorithms are largely oblivious to such artifacts and instead operate on the assumption of the generic bottleneck link model, which has been used to design congestion-control algorithms for decades [19]. While some proposals [12, 22, 42] leverage machine learning-based approaches to deal with these hard-to-predict artifacts, we show here that they still largely see a clouded view of packet arrivals, filtered through a wireless network that introduces a number of pathological-seeming—yet in fact explainable—jitter patterns.

While the physical and link layers of the wireless network know exactly their network state and can provide the necessary millisecond-level telemetry information [14, 17, 23, 40, 43], today, this layer-specific information remains siloed away from higher layers. If higher-layer algorithms (e.g., for rate adaptation) had access to this information, they could track and match physical capacity more accurately, resulting in higher application performance. Conversely, higher layers know best about their demands such that the physical layer does not need to attempt to infer and predict future application requirements. Consequently, in this paper, we argue that (while functionality should remain within the respective layer) we need APIs to open up layer-specific information to adjacent layers to enable more efficient operation of

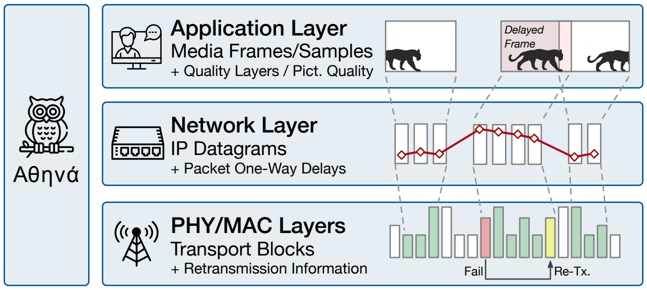


Fig. 1 — The Athena Measurement Framework: Athena synchronizes fine-grained measurements of video conferencing across all layers of the network stack, revealing new performance insights for application and transport protocol designers to improve their end-point adaptation algorithms.

today's heterogeneous networks.

To enable this vision, we propose **Athena**[1], a cross-layer measurement framework that correlates information across the physical, link, network, and application layers. In the specific context of 5G networks and VCAs, we demonstrate Athena's capability to remove the fog that hinders today's network applications when estimating the quality of the underlying network, suggesting the potential of our approach to do the same for myriad other types of access networks such as satellite, cable modem, 5G mmWave New Radio, and others.

Athena correlates measurements across layers (as depicted in Fig. 1), revealing the root causes of individual QoE impairments, such as video stalls, low resolution, or long mouth-to-ear delay. Specifically:

(1) We extract fine-grained 5G control channel telemetry of physical-layer data units (*transport blocks*), retransmissions, and scheduling decisions [40].

(2) We precisely time-synchronize this data with packet captures at the network layer and correlate physical transport blocks with network datagrams.

(3) We further correlate network datagrams with application-layer semantics, such as frames, different Scalable Video Coding (SVC) layers indicating the relative importance of a frame, and audio samples whose quality we also measure from the application side [28].

This broad, new perspective offers deep insights into the operation of the 5G Radio Access Network (RAN) and other access networks, and their immediate impacts on application QoE. We identify various causes of delay variation and delay spread, along with significant scheduling inefficiencies within the 5G RAN. Armed with this understanding, we propose a comprehensive agenda for future work, outlining concrete steps to mitigate these issues using mechanisms at the physical, network, and application layers. We specifically explore how application-layer information can be leveraged to inform the RAN scheduler, significantly reducing uplink delay. Additionally, we propose an approach where physical-layer information is fed to the application layer, enhancing delay-based congestion control mechanisms.

A Call to Cross-Layer Interactive Video Research

As users' QoE demands and the use of video conferencing, cloud gaming, and AR/VR in new wireless access networks increase, we urgently need research that can provide deep insights into the operation of cutting-edge access networks (L1, L2) and their impact on QoE (L7). To this end — and using Athena with Zoom as a starting point for this arc of research — we demonstrate and explain the intricacies of 5G networks that incur significant delay variations, leading to poor QoE, such as low frame rates and video stalls.

2 5G TELECONFERENCING PITFALLS

Today, video-conferencing applications (VCAs) generally deliver media signals in a similar way [2, 3, 27, 28, 32]: the sender captures media information, encodes it using a codec such as H.264 [34] or Opus [39], and transmits it over the network using the Real-Time Transport Protocol (RTP) [36] or similar transport. A congestion-control algorithm estimates network capacity by observing delay and loss, so the encoder may adjust media quality, resolution, and frame rate to match this capacity. A *jitter buffer* at the receiver smooths delay variations before it decodes and plays back the stream.

Given the real-time character of VCAs, stable and low latency and sustained network capacity are both essential to their performance. When the network cannot provide these, VCAs are left with three options. First, they can reduce the sending rate at the cost of reduced quality, hoping that this reduces congestion and jitter. Second,

[1] After Athena *Glaukopis* (lit. gleaming-eyed), Greek goddess of seeing.

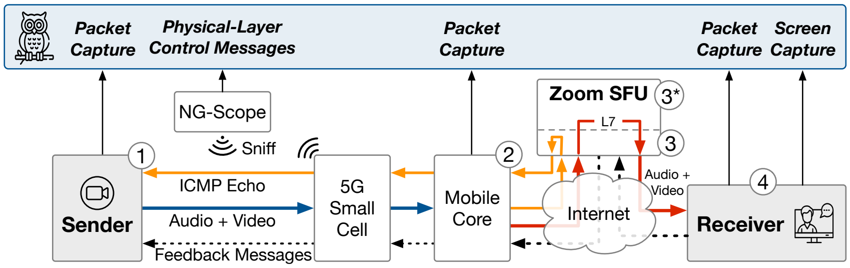


Fig. 2 — Athena's measurement framework targeting a Zoom session for a mobile device accessing the network via 5G Standalone.

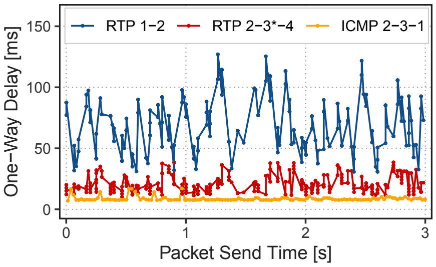


Fig. 3 — One-Way Delay in ICMP and Zoom RTP Media Traffic.

they can expand the jitter buffer at the cost of increased mouth-to-ear delay to better smooth out delay variations. Finally, they may not react and choose to accept a higher risk of stalls in order to maintain low end-to-end latency and high picture quality. Clearly, each option has pros and cons, so the choice of which to use depends on application requirements and user expectations.

To understand how 5G affects VCA QoE, we run a 20-minute two-party Zoom video call where the sender of the stream we monitor is connected to a private standalone 5G small cell [29], and the receiver is wired (Fig. 2), with all hosts NTP time synchronized. We inject a prerecorded video file, annotated frame-by-frame with QR codes, via a virtual camera device. At the receiver side, we capture the screen at 70 fps (slightly above the typical monitor refresh rate). Using this method, we determine if a particular frame was on the screen for longer than its intended (packetization) time given the current frame rate. Additionally, we compute picture quality by comparing each received frame with the corresponding sent frame and computing their structural similarity (SSIM) [41]. Cross traffic from six other cellular mobiles varies in throughput, from 0 to 14, 16, and finally 18 Mbps, in five-minute phases. Using this data, Athena computes sender-to-core (via the 5G RAN) and core-to-receiver (via the Zoom server) one-way delay (respectively, the red and blue lines in Fig. 2), effectively isolating the cellular uplink.

5G RAN uplink (only) jitters. Athena sees significant delay variation (jitter) on the 5G uplink in particular, ranging from 40 to 120 ms as Fig. 3 shows. Separating this delay into its audio and video components in Fig. 4, we see audio slightly less delayed but note a long tail of delay out to seconds. To ascribe the smaller jitter between the core and the receiver to the WAN or to Zoom itself, we concurrently ping the Zoom server from the core every 20 ms (orange lines in Figs. 2 and 3). **Takeways:** Athena sees that **(a)** the 5G uplink is the primary source of jitter, **(b)** the Zoom server's application-layer processing (not present in the ping probes) is a secondary source of jitter, and **(c)** the WAN, and importantly, the 5G RAN downlink provide low and stable delay.

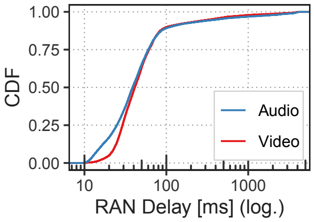


Fig. 4 — Zoom audio experiences lower delay than video.

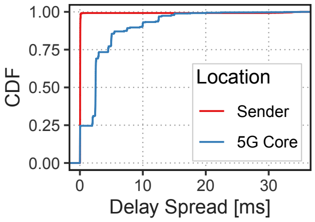


Fig. 5 — Delay spread introduced in the RAN Uplink.

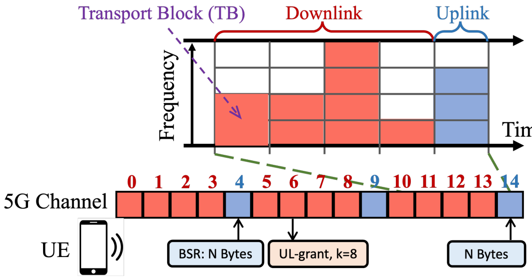
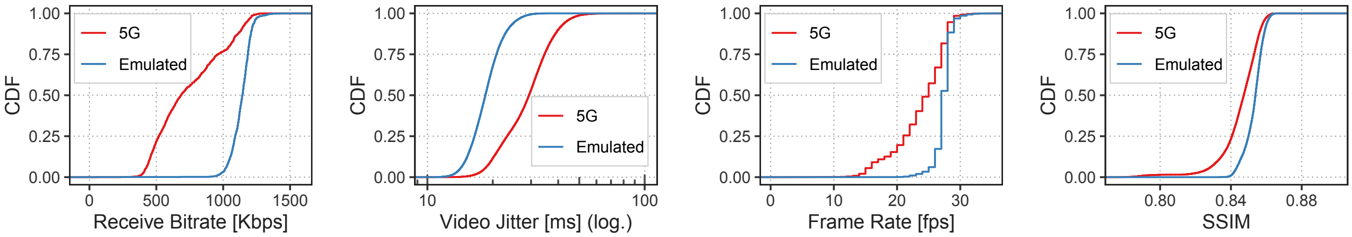


Fig. 6 — 5G frame structure: downlink and uplink switching pattern and BSR-based uplink transmission.

Drilling down into Athena's Zoom latency measurements, we observe that audio samples and video frames (usually consisting of multiple RTP packets) are sent in bursts. We calculate the *delay spread* — the time between the first and last packets of an audio sample or video frame — at the sender and in the 5G core, respectively, during a five-minute period without any cross



(a) Receive Media Bitrate **(b)** Frame-level Jitter **(c)** Frame Rate **(d)** Picture Quality

Fig. 7 — 5G degradation: Key QoE and performance metrics in 5G versus a wired Network with equal emulated capacity.

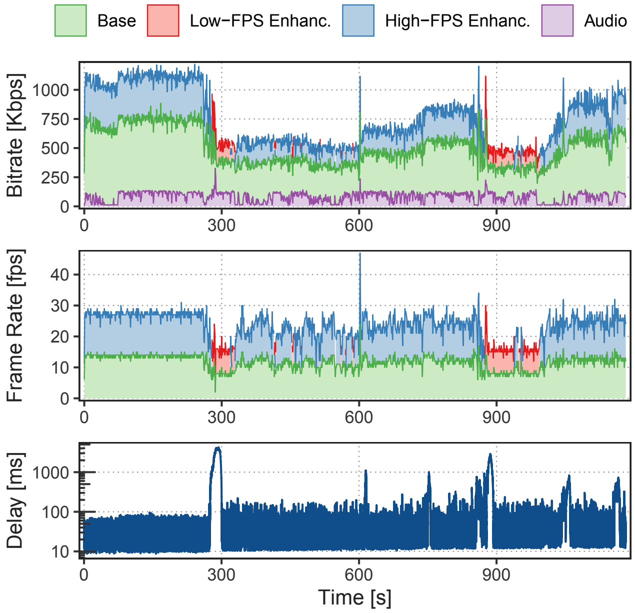


Fig. 8 — Zoom adaptation: Zoom reacts to both high absolute delay and high jitter primarily by adapting the frame rate.

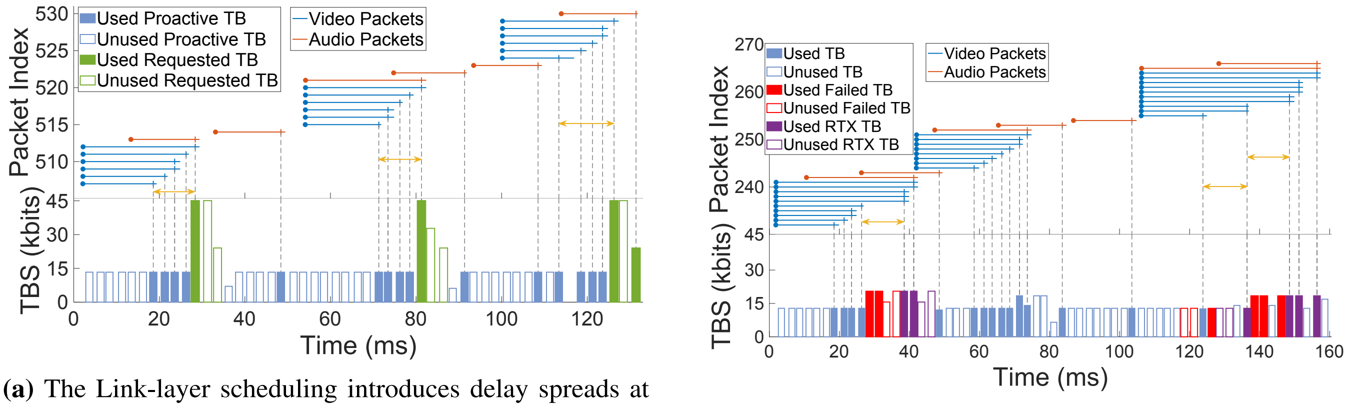
traffic on the cellular network. We observe (Fig. 5) that the RAN uplink spreads out the one-way delay of samples and frames at the receiver in increments of 2.5 ms. **Takeaway:** Delay spread accounts for the difference in packet-level one-way delay between audio and video in Fig. 4, as audio samples rarely span multiple packets and are thus only delayed when sent in conjunction with a video frame. We explain this effect that stems from RAN scheduling in Section 3.1.

5G Impairs QoE. We next use Athena to isolate the effect of the aforementioned 5G delay and jitter on Zoom itself. We create a baseline with a fixed 15 ms latency that emulates the cellular network's capacity (calculated from the physical transport block sizes) using Linux traffic control (tc) over a wired network. Figure 7 compares key QoE and performance metrics between the two networks. **Takeaway:** We see that 5G consistently delivers lower quality both with respect to bitrate and media-level jitter, as well as user-centric metrics such as frame rate and picture quality.

How Zoom Adapts. To understand how Zoom adapts to delay variations, we plot Zoom's frame rate and bit rate as a timeseries in Fig. 8. Zoom uses the temporal scaling dimension of Scalable Video Coding (SVC) [18, 37], including a *base layer* at either seven or 14 fps, and adding enhancement layers to reach 14 or 28 fps, respectively. When the target frame rate is 14 fps, Zoom uses a different identifier for the enhancement layer, denoted "Low-FPS Enhancement" in the figure. The layer identifier is included in the RTP header extensions. We spoke with Zoom engineers to confirm that Zoom indeed uses this type of media scalability. We can see that Zoom reacts to very high absolute delay (above one second) by changing the set of SVC layers and more permanently reducing the frame rate to 14 fps. If the jitter is high, Zoom appears to transiently skip frames, reducing to rates around 20 fps. The upper plot shows the impact on the overall bandwidth utilization the two adaptation strategies have. Either adaptation leads to impaired user experience, as summarized in Fig. 7.

3 SHEDDING LIGHT ON THE 5G RAN

The 5G network introduces significant delay variations in the uplink direction, as illustrated in Figs. 3, 4 and 8. We now explain the two main causes of these variations in detail: link-layer scheduling and link-layer retransmissions.



(a) The Link-layer scheduling introduces delay spreads at frame level, in increments of 2.5 ms, denoted as yellow double arrows.

(b) The Link-layer retransmissions inflate the packet delay by 10 ms, denoted as yellow double arrows.

Fig. 9 — Time series examples of the Athena Zoom traffic trace, incorporating both transport layer packet information and physical layer TB information. The packets are synchronized with the TBs that carry them, where dashed lines are used to connect each packet to its corresponding TB.

3.1 Link-Layer Scheduling

Our private 5G small cell operates in Time Division Duplexing (TDD) mode, as shown in Fig. 6, which divides time into periodic downlink and uplink slots. In our cell's configuration, downlink slots occur four times as frequently as uplink slots, with uplink slots appearing every 2.5 milliseconds.

The base station allocates uplink resources to the mobile via uplink grants, which specify the transport block size for each uplink slot. There are two types of uplink grants: requested and proactive. For requested grants, a 5G mobile reports the amount of data in its transmission buffer using a Buffer Status Report (BSR) [1], as illustrated in Fig. 6. Based on the BSR, the base station allocates uplink grants to match the mobile's traffic demand. However, there is a scheduling delay between the time a mobile sends a BSR and when it receives and utilizes the uplink grant [38], *ca.* 10 ms in our Private 5G network.

To mitigate this BSR scheduling delay, some base stations use proactive grants, which pre-allocate uplink resources to the mobile before receiving any BSR. Proactive grants can consistently reduce delay by around 10 ms for sporadic packets. They, however, come at the cost of potentially wasting bandwidth (if remain unused) and require extra computing resources. Additionally, this scheduling does not fit in well with bursty traffic patterns as present in VCA traffic.

To investigate link-layer scheduling on VCA traffic, we use Athena to drill down into our collected trace (Section 2) and present a time series in Fig. 9(a). In the upper part of the figure, each horizontal line represents a packet, where the left and right edges indicate the timestamps when the packet is sent at the sender and when we capture it at the mobile core, respectively, as shown in Fig. 2. The length of the line represents the one-way delay between the sender and the mobile core. The lower part of the figure shows the physical layer transport block (TB) sizes within the same time period.

Multiple packets, sent in a burst, comprise each video frame—when these are ready at the mobile, a proactive TB can carry only one or two of them. Given the 2.5 ms downlink-uplink period, another proactive grant arrives 2.5 ms later, allowing the mobile to send another one or two packets. This process continues until the BSR-requested grant arrives, typically around 10 ms after the initial packet transmission, at which point all remaining packets in the UE's buffer are delivered by the BSR-requested TBs. This scheduling approach results in the previously discussed delay spread at the frame level, which is denoted as yellow double arrows in Fig. 9(a).

Proactive grants also cause over-granting issues. As shown in Fig. 9(a) (green bars), the BSR-requested TB size is based on the buffer status at the time the UE sends the BSR. However, once the BSR-requested TB becomes available to the mobile, 10 ms later, the remaining data in the mobile's buffer has decreased, because proactive TBs have already delivered some packets during the BSR scheduling delay period. This over-granting results in some requested TBs being wasted without transmitting any data (the unfilled green bars in Fig. 9(a)), ultimately leading to a waste of bandwidth.

3.2 Link-Layer Retransmissions

5G link-layer retransmissions happen due to mobility and dynamic channel conditions, which cause errors or data loss in the transmitted TB. These retransmissions occur frequently, particularly in environments with high interference or signal variability. As a result, retransmissions introduce additional delay to the packets they carry, impacting overall network latency. In our configuration, retransmission delay is 10 ms.

Fig. 9(b) illustrates another time series example, highlighting failed and retransmitted TBs in red and purple, respectively. In instances where retransmissions occur for TBs containing packets, the packet delay (the length of the horizontal lines in Fig. 9(b)) is typically inflated by 10 milliseconds, indicated with yellow double arrows. If the retransmitted TBs fail again at the base station, it leads to multiple rounds of retransmissions of the same TB, further inflating the packet delay by multiples of 10 milliseconds. This introduces additional variations to the network latency. Additionally, Athena's observations reveal that the base station also mandates the UE to retransmit empty proactive and requested TBs, which results in unnecessary bandwidth consumption.

4 A DELAY-BASED SOLUTION?

It is well known that loss-based congestion control is poorly-suited for low-latency video conferencing applications because it intentionally creates network buffering to probe network capacity. When packet loss occurs, it indicates that the buffer is full and the network is already congested, leading to increased delay and a degraded user experience.

Delay, on the other hand, is widely recognized as the earliest indicator of network congestion. Consequently, delay-based congestion control algorithms, such as SCREAM [20], NADA [45], and GCC [7], have been widely adopted by video conferencing applications to provide the most responsive performance to network capacity fluctuations. Among these, GCC is the default congestion control algorithm for WebRTC [3, 10], a widely used real-time communication standard that powers, for example, Google Meet [15]. Since the congestion control used in Zoom is proprietary and the details are unknown, we use GCC as an example to demonstrate the potential impact of RAN-induced delay variations on the design of a delay-based congestion control protocol.

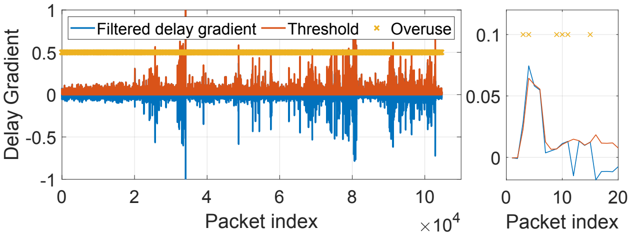


Fig. 10 — GCC running at a mobile connected via a Private 5G network detects frequent network overuse based on its filtered packet one-way delay gradient estimate.

GCC leverages the *one-way delay gradient* to detect the status of network usage, which is defined as $d_m = (T_i - T_{i-1}) - (t_i - t_{i-1})$ where the t_i and T_i are the sending and receiving timestamps of the i-th packet, respectively. GCC then smooths the delay gradient by applying a trendline filter to obtain the filtered delay gradient. If the filtered delay gradient is positive and exceeds a certain threshold, the network is identified as being overused. Conversely, if the delay gradient is negative and falls below a negative threshold, the network is considered underused. To demonstrate the impact of RAN-induced delay variations, we measure the filtered one-way delay gradient of the packets transmitted within one video conference session inside an idle 5G network within which our mobile is the only user. We plot the filtered delay gradient, the threshold and the detected network overuse in Fig. 10, from which we can observe significant fluctuations in the gradient, which could potentially result in frequent identification of network overuse and underuse, while the network is consistently idle and underused. Such frequent misidentification of overuse could severely mislead GCC, causing it to falsely react to phantom network fluctuations.

5 ATHENA LOOKING FORWARD

Athena's analysis of latency artifacts reveals a physical-layer scheduling and retransmission cause. This naturally motivates future measurement studies Athena's first-in-kind cross-layer methodology enables, and leads to many opportunities for the mitigation of latency inflation and its detrimental effects on video-conferencing applications.

5.1 Cross-Layer Measurements

While we demonstrate the potential of the Athena measurement framework in the context of video conferencing and 5G, our methodology is also a blueprint for future measurement. In general, there are more and more diverse applications that exhibit various traffic patterns (*e.g.*, short video [26], video on demand, web browsing, interactive applications) and an ever-growing set of physical and link-layer technologies (e.g., 4G and 5G with a wide range of multiple-access and duplexing strategies, Wi-Fi, satellite networks, and Bluetooth). All underlying networks introduce different artifacts that are of varying importance to the different classes of applications. A challenge here is to find a generic way how these diverse physical-layer technologies can match and interact with application-layer demands to maximize performance. We call for a more frequent, principled interchange of information between layers enabled through continuous, fine-grained measurement—the Athena framework enables exactly this.

To gain deeper insights and gather more data points for this vision, in the future, we plan to use Athena to further measure Google Congestion Control (GCC) and work toward a GCC simulator that evaluates video-conferencing behavior in various physical-layer contexts. For example, in the context of cellular networks, different base stations use different duplexing strategies. Also, the wireless spectrum can be divided along multiple axes. Time slicing (as in TDD) is done using different slice lengths in differing frequency bands, and some cellular networks use Frequency Division Duplexing (FDD) for uplink and downlink, resulting in differing impacts on application-layer latencies (cf. Section 2).

5.2 A More Application-Aware RAN?

Video conferencing and other real-time communication applications exhibit a very predictable traffic pattern: a video frame is sent approximately every 33 ms (at 30 fps) or every 66 ms (at 15 fps). The size of the frames also rarely changes significantly as VCAs typically do not use I-frames but rather transmit all video as a series of P-frames that only encode the difference from the previous picture [28]. In Section 3.1, we show that the 5G TDD uplink-grant scheduling mechanism delivers the majority of packets using small proactive grants, leading to delay spread at the frame level, while reactive grants typically arrive too late and often remain unused, wasting resources for other users (Fig. 9(a)).

Given the predictability of VCA traffic, there are ample opportunities for the RAN to issue uplink grants in a more informed, application-aware way. This can be realized in two ways. First, video-conferencing packets can be annotated (e.g., through RTP extensions) with media-level metadata. This information could include the number of streams originating at a particular sender, together with data about their sampling rates (in the case of audio) or frame rates in the case of video, together with a periodically updated estimate for the current frame size as this may depend on multiple factors. Using this information, the base station can issue grants exactly at the right times when a sample or frame is generated and ready for transmission. Second, the base stations can use machine learning to learn the current transmission patterns, and predict future traffic demands to precisely issue grants.

The Open-RAN Alliance specifies the RAN Intelligent Controller (RIC) as a software component that provides centralized control and optimization of radio network functions [33, 35]. Network operators can use the RIC to apply customized algorithms to various RAN operations, including resource allocation. Specifically, a Real-Time RIC [21] can be employed to implement such an intelligent traffic learning algorithm and subsequent grant scheduling. Either approach has the potential to cut the delay inflation experienced by frames in half. Note that the frame-level delay (*i.e.*, from the transmission of the first packet of a frame to the reception of the last packet of the frame) is extremely relevant as a frame cannot be rendered until all of its packets have been received.

5.3 More RAN-Aware Applications?

Conversely to the RAN becoming application-aware, there is a clear need for applications and transport protocols to be better informed of the RAN's state. Here, the key architectural challenge is to define the congestion protocol for application and transport layer senders. How should the wireless access network abstract its complexities to higher-layer senders? How should this information be communicated to senders across the wide area network?

Generally, the RAN could give applications finer insights into its operation through telemetry that, for example, conveys the cause for a particular delay increase. Alternatively, the RAN could mask RAN-induced delays through the congestion-control feedback channel by modifying per-packet delay information as reported by, for example, RTCP transport-wide congestion-control messages in GCC. As a protocol, L4S [4–6] is attractive, as it adopts ECN bits in the IP header to accelerate or brake the sender (*cf.* ABC [16]), which stands a good chance at practical and incremental deployability, topics under close consideration in the IETF. But challenges remain here, too: how should control of the accelerate-brake signal be defined in the presence of retransmissions due to (unpredictable) loss versus the more predictable delay spikes and spreads that we observe with Athena?

6 RELATED WORK

There is a line of research designing cellular PHY-layer monitoring tools, as in [23, 25, 43]. Additionally, [40, 44] integrates 4G/5G PHY-layer measurements with the transport layer to enhance congestion control algorithms. Previous studies have also investigated how cellular network affects network latency, for sporadic and small traffic applications [38], and in high user mobility scenarios [31]. Our work differs by correlating information across the physical, link, network, and application layers, and providing a in-depth analysis revealing the root cause of the impaired VCA QoE under 5G.

7 CONCLUSION

The Athena measurement framework is the first of its kind to deeply look across all layers of the network stack, an approach whose time has come given the accelerating pace of innovation at the high and low ends of the stack. In this paper, we have reported a proof of concept of the Athena approach for Zoom, and have scanned the horizon of new work that Athena enables.

ACKNOWLEDGEMENTS

This work has been supported by the National Science Foundation under Grant Nos. AST-2232457 and CNS-2223556, and by DARPA under contract HR001120C0107.

REFERENCES

[1] 3GPP. TS138.321: 5G NR Medium Access Control (MAC) protocol specification.

[2] S. A. Baset, H. G. Schulzrinne. An analysis of the skype peer-to-peer internet telephony protocol. *IEEE INFOCOM 2006*, 1–11. IEEE, New York, NY, USA, 2006. doi:10.1109/INFOCOM.2006.312.

[3] N. Blum, S. Lachapelle, H. Alvestrand. Webrtc: Real-time communication for the open web platform. *Communications of the ACM*, **64**(8), 50–54, 2021. ISSN 0001-0782. doi:10.1145/3453182.

[4] B. Briscoe, K. D. Schepper, M. Bagnulo, G. White. Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture. Request for Comments RFC 9330, Internet Engineering Task Force, 2023. doi:10.17487/RFC9330. Num Pages: 36.

[5] B. Briscoe, K. D. Schepper, O. Tilmans, M. Kuhlewind, J. Misund. Implementing the 'Prague Requirements' for Low Latency Low Loss Scalable Throughput (L4S).

[6] D. Brunello. *L4S in 5G networks*, 2020.

[7] G. Carlucci, L. De Cicco, S. Holmer, S. Mascolo. Analysis and design of the google congestion control for web real-time communication (webrtc). *Proceedings of the 7th International Conference on Multimedia Systems*, MMSys '16. Association for Computing Machinery, New York, NY, USA, 2016. ISBN 9781450342971. doi:10.1145/2910017.2910605.

[8] G. Carlucci, L. De Cicco, S. Holmer, S. Mascolo. Congestion control for web real-time communication. *IEEE/ACM Transactions on Networking*, **25**(5), 2629–2642, 2017. doi:10.1109/TNET.2017.2703615.

[9] H. Chang, M. Varvello, F. Hao, S. Mukherjee. Can you see me now? a measurement study of zoom, webex, and meet. *ACM Internet Measurement Conference*, 216–228. ACM, New York, NY, USA, 2021. ISBN 9781450391290.

[10] T. W. W. W. Consortium. W3C recommendation: WebRTC: Real-time communication in browsers, 2023. Retrieved April 6, 2023, from https://www.w3.org/TR/2023/REC-webrtc-20230306.

[11] S. Dhawaskar Sathyanarayana, K. Lee, D. Grunwald, S. Ha. Converge: QoE-driven Multipath Video Conferencing over WebRTC. *Proceedings of the ACM SIGCOMM 2023 Conference*, 637–653. ACM, New York NY USA, 2023. ISBN 9798400702365. doi:10.1145/3603269.3604822.

[12] M. Dong, Q. Li, D. Zarchy, P. B. Godfrey, M. Schapira. PCC: Re-architecting congestion control for consistent high performance. *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, 395–408. USENIX Association, Oakland, CA, 2015. ISBN 978-1-931971-218.

[13] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, G. Smaragdakis. The lockdown effect: Implications of the covid-19 pandemic on internet traffic. *ACM Internet Measurement Conference*, 1–18. ACM, New York, NY, USA, 2020. ISBN 9781450381383. doi:10.1145/3419394.3423658.

[14] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, D. J. Leith. srsLTE: an open-source platform for LTE evolution and experimentation. *ACM WiNTECH*, 2016.

[15] Google meet, 2023. Retrieved April 14, 2023, from https://meet.google.com.

[16] P. Goyal, M. Alizadeh, H. Balakrishnan. ABC: A Simple Explicit Congestion Controller for Wireless Networks. *Proceedings of the 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI '20)*, 2020.

[17] T. D. Hoang, C. Park, M. Son, T. Oh, S. Bae, J. Ahn, B. Oh, Y. Kim. LTESniffer: An Open-source LTE Downlink/Uplink Eavesdropper. *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 43–48. Association for Computing Machinery, New York, NY, USA, 2023. ISBN 978-1-4503-9859-6. doi:10.1145/3558482.3590196.

[18] ISO/IEC. Standard 14496-10:2022.

[19] V. Jacobson. Congestion avoidance and control. *Symposium Proceedings on Communications Architectures and Protocols*, SIGCOMM '88, 314–329. Association for Computing Machinery, New York, NY, USA, 1988. ISBN 0897912799. doi:10.1145/52324.52356.

[20] I. Johansson. Self-clocked rate adaptation for conversational video in lte. *Proceedings of the 2014 ACM SIGCOMM Workshop on Capacity Sharing Workshop*, CSWS '14, 51–56. Association for Computing Machinery, New York, NY, USA, 2014. ISBN 9781450329910. doi:10.1145/2630088.2631976.

[21] W.-H. Ko, U. Dinesha, S. Shakkottai, D. Bharadia. EdgeRIC: Empowering Realtime Intelligent Optimization and Control in NextG Cellular Networks. *NSDI*, 2024.

[22] Y. Kong, H. Zang, X. Ma. Improving tcp congestion control with machine intelligence. *Proceedings of the 2018 Workshop on Network Meets AI & ML*, NetAI'18, 60–66. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450359115. doi:10.1145/3229543.3229550.

[23] S. Kumar, E. Hamed, D. Katabi, L. Erran Li. LTE Radio Analytics Made Easy and Accessible. *SIGCOMM Comput. Commun. Rev.*, **44**(4), 211–222, 2014. ISSN 0146-4833. doi:10.1145/2740070.2626320.

[24] I. Lee, J. Lee, K. Lee, D. Grunwald, S. Ha. Demystifying commercial video conferencing applications. *ACM International Conference on Multimedia*, 3583–3591. ACM, New York, NY, USA, 2021. ISBN 9781450386517.

[25] Y. Li, C. Peng, Z. Yuan, J. Li, H. Deng, T. Wang. Mobileinsight: Extracting and analyzing cellular network information on smartphones. *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, 202–215, 2016.

[26] Z. Li, Y. Xie, R. Netravali, K. Jamieson. Dashlet: Taming Swipe Uncertainty for Robust Short Video Streaming. *NSDI*, 2023.

[27] K. MacMillan, T. Mangla, J. Saxon, N. Feamster. Measuring the performance and network utilization of popular video conferencing applications. *ACM Internet Measurement Conference*, 229–244. ACM, New York, NY, USA, 2021. ISBN 9781450391290.

[28] O. Michel, S. Sengupta, H. Kim, R. Netravali, J. Rexford. Enabling passive measurement of zoom performance in production networks. *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, 244–260. Association for Computing Machinery, New York, NY, USA, 2022. ISBN 9781450392594. doi:10.1145/3517745.3561414.

[29] Mosolab. Mosolab Canopy Small Cell, 2023.

[30] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, Z.-L. Zhang. A First Look at Commercial 5G Performance on Smartphones. *Proceedings of The Web Conference 2020*, 894–905. ACM, Taipei Taiwan, 2020. ISBN 978-1-4503-7023-3. doi:10.1145/3366423.3380169.

[31] Y. Ni, F. Qian, T. Liu, Y. Cheng, Z. Ma, J. Wang, Z. Wang, G. Huang, X. Liu, C. Xu. {POLYCORN}: Data-driven cross-layer multipath networking for high-speed railway through composable schedulerlets. *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 1325–1340, 2023.

[32] A. Nistico, D. Markudova, M. Trevisan, M. Meo, G. Carofiglio. A comparative study of rtc applications. *IEEE International Symposium on Multimedia*, 1–8. IEEE, New York, NY, USA, 2020.

[33] Open ran alliance, 2024. Retrieved June 27, 2024, from https://www.o-ran.org.

[34] I. E. Richardson. *The H.264 Advanced Video Compression Standard*. Wiley Publishing, 2nd edn., 2010. ISBN 978-0-470-51692-8.

[35] R. Schmidt, M. Irazabal, N. Nikaein. FlexRIC: an SDK for next-generation SD-RANs. *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*, 411–425. ACM, Virtual Event Germany, 2021. ISBN 978-1-4503-9098-9. doi:10.1145/3485983.3494870.

[36] H. Schulzrinne, S. L. Casner, R. Frederick, V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC 3550, 2003. doi:10.17487/RFC3550.

[37] H. Schwarz, D. Marpe, T. Wiegand. Overview of the scalable video coding extension of the h.264/avc standard. *IEEE Transactions on Circuits and Systems for Video Technology*, **17**(9), 1103–1120, 2007. doi:10.1109/TCSVT.2007.905532.

[38] Z. Tan, J. Zhao, Y. Li, Y. Xu, S. Lu. {Device-Based}{LTE} latency reduction at the application layer. *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, 471–486, 2021.

[39] J.-M. Valin, K. Vos, T. B. Terriberry. Definition of the Opus Audio Codec. Request for Comments RFC 6716, Internet Engineering Task Force, 2012. doi:10.17487/RFC6716. Num Pages: 326.

[40] H. Wan, K. Jamieson. Evolving Mobile Cloud Gaming with 5G Standalone Network Telemetry, 2024. doi:10.48550/ARXIV.2402.04454. Version Number: 1.

[41] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**(4), 600–612, 2004. doi:10.1109/TIP.2003.819861.

[42] K. Winstein, H. Balakrishnan. Tcp ex machina: computer-generated congestion control. *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, 123–134. Association for Computing Machinery, New York, NY, USA, 2013. ISBN 9781450320566. doi:10.1145/2486001.2486020.

[43] Y. Xie, K. Jamieson. Ng-scope: Fine-grained telemetry for nextg cellular networks. *Abstract Proceedings of the 2022 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS/PERFORMANCE '22, 27–28. Association for Computing Machinery, New York, NY, USA, 2022. ISBN 9781450391412. doi:10.1145/3489048.3522652.

[44] Y. Xie, F. Yi, K. Jamieson. Pbe-cc: Congestion control via endpoint-centric, physical-layer bandwidth measurements. *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 451–464, 2020.

[45] X. Zhu, R. Pan. Nada: A unified congestion control scheme for low-latency interactive video. *2013 20th International Packet Video Workshop*, 1–8, 2013. doi:10.1109/PV.2013.6691448.

[46] Zoom, 2023. Retrieved April 14, 2023, from https://zoom.us.