Dissecting 5G New-Radio Latency: Interacting Layers and Latency-Generating Operations

Virgil Hamici-Aubert, Julien Saint-Martin, Renzo E. Navas, Georgios Z. Papadopoulos, Guillaume Doyen, Xavier Lagrange
IMT Atlantique, IRISA, UMR CNRS 6074, F-35700 Rennes, France. Email: firstname.lastname@imt-atlantique.fr

***Abstract*—The fifth-generation (5G) mobile network improvements in the Radio Access Network (RAN), including low latency, reliability, and massive connections, are key enablers for emerging technologies and applications, such as smart grids, factory automation, and metaverse. Latency on the RAN relies on a multiplicity of 5G parameters that can be configured as needed. Current End-to-End (E2E) latency diagnostic capabilities are unsatisfactory in understanding the RAN involvement, and exhibit the importance of improving our diagnostic capabilities for 5G RAN implementations. In this paper, we aim to dissect the Uplink (UL) RAN latency in order to understand the origins of latency generation. An extensive experimental campaign on a 5G open-source implementation (OpenAirInterface) allows us to study the UL E2E latency through several bitrates. We mark each packet in its payload, and use the identifiers of each New Radio (NR) layers to track its E2E latency at the protocol level. We identify the layers mainly involved in the RAN latency composition, exhibit their operations, and explain their origin. This work leverages the collected information and latency measurements through all NR protocol layers to improve the understanding and diagnostics of the UL E2E latency through the RAN. Last but not least, the source code and experimental data are publicly available.**

***Index Terms*—5G, Radio Access Network, Latency, OpenAirInterface, Diagnostic**

I. Introduction

Emerging technologies, including 5G, artificial intelligence, haptic and Extended Reality (XR) are contributing to the novel use of data, e.g., smart cities, and metaverse-like services [1]–[3]. The latency metric is critical for such applications [4]. 5G technology ensures the delivery of application data with the appropriate guarantees, e.g., coverage, massive connection, bandwidth, and low latency [5], [6].

End-to-End (E2E) latency depends on various components, including 5G components like the Radio Access Network (RAN). The 5G radio interface comprises a wide range of parameters and processes. Each process is configurable and interacts with each other. The change in one process parameter can affect another. Considering the multiplicity of 5G implementations and the multiple ways to configure them, it is challenging to identify the sources that affect latency. In addition, the proprietary aspect of certain 5G components, e.g., User Equipment (UE), 5G Core (5GC), RAN, and the New Radio (NR) protocol stack [7]–[9], makes them opaque and difficult to analyze. Understanding the sources of latency–particularly, the contribution of each 5G NR process–is crucial to minimize it.

ISBN 978-3-903176-72-0 © 2025 IFIP

In this paper, we extend LatSeq [10], an open-source latency measurement module developed for the 4G OpenAirInterface (OAI) implementation. The LatSeq base mode allows us to collect variable content at each layer level and associate them with a timestamp during Base Station (BS) packet processing. Our instrumentation enables tracking of a packet through all NR layers involved in its delivery on RAN in Uplink (UL), from UE to BS, and to consider each layer time operation as layer-specific latency. These layer-specific latencies provide an understanding of the composition of RAN latency. The collected metrics allow us to identify the latency-generating operations, and attribute events to an observed layer-specific latency.

We performed an extensive experimental campaign exploring the UL E2E latency of different bitrates using a parameter and continuous-flow generation.

Our main contributions are the following:

1) We provide a detailed study of UL E2E latency, with a focus on the 5G RAN layers. In particular, we study the layer's dependencies and the most latency-generating operations.
2) We diagnose the UL latency trend on an open-source-based 5G implementation based on OAI. Our experimental data and code are publicly available [11].

The rest of this paper is organized as follows. Sect. II presents the related work. Sect. III defines the latency collection method. In Secs. IV–V, we describe our experimental setup and results. Finally, Sect. VI concludes and offers future work perspectives.

II. Related Work

For the sake of brevity, we do not present the 5G radio interface. Readers can refer to [12]. Delivery of the UL application-level packets involves several mechanisms on the RAN. Identifying all latency-generating operations requires a complete view of the NR layers. In the following, we briefly review the literature on the 5G latency.

Measurement of latency in a real world environment [7]–[9] and on proprietary materials (e.g., operators RAN) allows researchers to evaluate the 5G latency performance. They attribute the time-consuming latency to the processing capability of UE and the length of the wireline path (i.e., from the backhaul to the receiver).

The assessment of 5G mmWave for low latency requirements [13] highlights the suitability of Physical Layer (PHY) using high 5G numerology, but shows limitations focused on Medium Access Control (MAC) retransmissions and Radio Resource Control (RRC) mechanisms (e.g., handover).

In [14], [15], the authors consider RAN as an open box. They divide the UL path through the RAN into several components, each of which may comprise multiple layers. They identify the cost of the delay between the arrival of a packet to be transmitted and the possibility of using the allocation of radio resources received from BS, as well as the delay in allocation that increases Radio Link Control (RLC) buffering and the insufficient number of Physical Ressource Blocks (PRBs) per allocation, leading to packet segmentation.

A research team in our lab developed a measurement tool [10] for a 4G implementation from OAI[1]. This tool enables them to place points of logs inside the code to collect variable content during the binary's execution. Each point is associated with a timestamp that allows them to trace a packet through all layers and study its latency. They considered all the BS Long Term Evolution (LTE) layers for the UL and the Downlink (DL). They identified RLC as the latency-generating layer. Their other work [16] exhibits that the BS considers outdated Buffer Status Report (BSR) for the allocation, which does not allow BS to allocate enough bytes to UE at the right time, thus affecting latency.

The literature exposes RAN protocol aspects that affect UL latency, but focuses on MAC: resource allocation, MAC retransmissions, and BSR. However, measuring the latency of opaque material [7]–[9], grouping NR layers [15], or considering only the BS side [10] prevents a complete understanding of RAN latency. As we consider services that are highly sensitive to latency variation, we study in depth to observe the interactions between NR layers (MAC, RLC, and Packet Data Convergence Protocol (PDCP)) to find all possible latency-generating operations.

III. Latency Dissection

In this section, we describe the latency dissection through the UL NR layers and define the UL RAN latency for an application-level packet. We then explain the information we collect at each layer. This allows us to link the UE and the BS protocol stack to obtain the UL packet path for a packet over the RAN and its layer-specific latencies.

A. Implementation

We adapt the tool from [10], to this end we use the 5G implementation from OAI[2], which provides open-source-based UE, BS, and 5GC. We adapt LatSeq to 5G and place points at appropriate locations within the OAI code

[1]Orange open source - https://github.com/Orange-OpenSource/LatSeq
[2]OAI - 5G - v2.1.0 - https://gitlab.eurecom.fr/oai/openairinterface5g

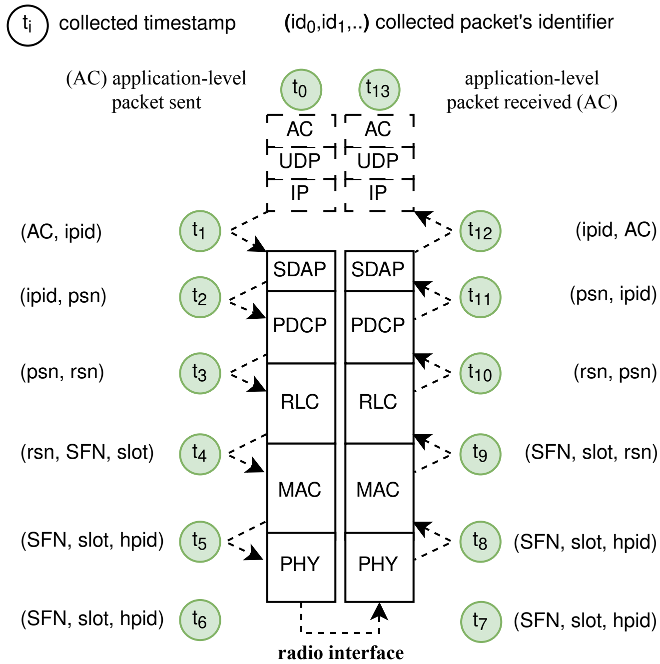


Fig. 1: Timestamps and packet's identifiers collection at each layer for one application-level packet, through the Uplink Radio Access Network protocol stack at UE and BS side. As we work with the IPv4, we consider the Internet Protocol (IP) identifier from RFC 6864, which does not exist in IPv6.

to measure the operation time of each layer. We use IPv4, but as our code checks the IP version in the header, our instrumentation is ready to include IPv6 if needed. Finally, we identify an application-level packet through the RAN by placing a 64-bit counter in the User Datagram Protocol (UDP) payload and refer to Application Counter (AC) in the rest of the paper.

B. Uplink Radio Access Network Packet Latency Definition

We extend the UL RAN packet latency definition from [10] to include both sides of the NR protocol stack (i.e., UE and BS). For a unique application-level packet, we consider its UL latency on RAN from UE Service Data Adaptation Protocol (SDAP) to BS SDAP. Thus, the UL RAN latency for a considered application-level packet means to consider the complete transmission of all its RLC segments and their MAC retransmissions.

C. Uplink Timestamps Collection

Fig. 1 presents our information collection through the layers involved in the IP packet sending from the UE application-level output (t_0) to the UE application-level input (t_{13}). The layers identifiers in the bracket allow us to follow the packet through its processing in each layer. Each timestamp collection occurs when the current layer submits the Packet Data Unit (PDU) to the next layer, marking the end and the beginning of treatments at the current layer and

TABLE I: RAN layer-specific latencies and latency-generating operations (see Fig. 1, for the t_i definitions).

layer	side	layer-specific latencies		latency-generating operation
		name	value(ms)	
IP	application	Lat_{E2E}	$(t_{13} - t_0)$	∅
SDAP	UE	Lat_{ue_sdap}	$(t_2 - t_1)$	∅
PDCP	UE	Lat_{ue_pdcp}	$(t_3 - t_2)$	∅
RLC	UE	Lat_{ue_rlc}	$(t_4 - t_3)$	buffering, retransmission
MAC	UE	Lat_{ue_mac}	$(t_5 - t_4)$	retransmission
PHY	UE	Lat_{ue_phy}	$(t_6 - t_5)$	∅
∅	hardware	$Lat_{hardware}$	$(t_7 - t_6)$	∅
PHY	BS	Lat_{bs_phy}	$(t_8 - t_7)$	∅
MAC	BS	Lat_{bs_mac}	$(t_9 - t_8)$	∅
RLC	BS	Lat_{bs_rlc}	$(t_{10} - t_9)$	segments reassembly
PDCP	BS	Lat_{bs_pdcp}	$(t_{11} - t_{10})$	SN reordering
SDAP	BS	Lat_{bs_sdap}	$(t_{12} - t_{11})$	∅

next layer, respectively; except some particular mechanisms explained in our shared resources.

D. Uplink Application-Level Packet Path

The application-level packet RLC Sequence Number (SN) (`rlc_sn`) links the MAC and the air interface frame information, i.e., the HARQ Process Identifier (HPID), the System Frame Number (SFN), and the slot. As RLC does not concatenate packets and the SFN is on 10 bits, it is unique over 10 seconds. That allows us to identify a unique application-level packet and all its segments on the air interface. We use this logic on both sides to link the packet at PHY, and trace it through all layers from its RAN input (i.e., UE SDAP) to its output (i.e., BS SDAP).

E. Application-Level Packet Latency Through the RAN

Table I describes the E2E latency (Lat_{E2E}) and all other layer-specific latencies on the UE and BS NR protocol stack. It also shows the latency-generating operations that could impact the packet latency. Those operations are the RLC buffering and the MAC retransmission on the transmitter side. In addition, RLC and PDCP must also wait for all segments to transmit the entire packet to PDCP (i.e., reassembling) and maintain the SN order before submitting to SDAP (i.e., reordering), respectively, at the receiver side.

IV. Experimental Setup

We have deployed a Software Defined Radio (SDR)-based platform to measure and analyze the latency over the air. In this section, we introduce our experimental testbed and the metrology we used to analyze RAN latency.

A. Experimental Testbed

1) Material: Fig. 2 illustrates our experimental testbed, including two main machines for UE (i.e., PC1) and BS plus the core (i.e., PC2). PC3 sends the commands to the two first, as they are dedicated to the 5G implementation operating and configured for real-time. The UE execution is in a Dell precision 3650 Tower with an Intel(R) Xeon(R) W-1270 CPU @ 3.40GHz and 32GB RAM with Ubuntu 22.04.5 LTS. The BS and the core execution are in a Dell precision 5820 Tower

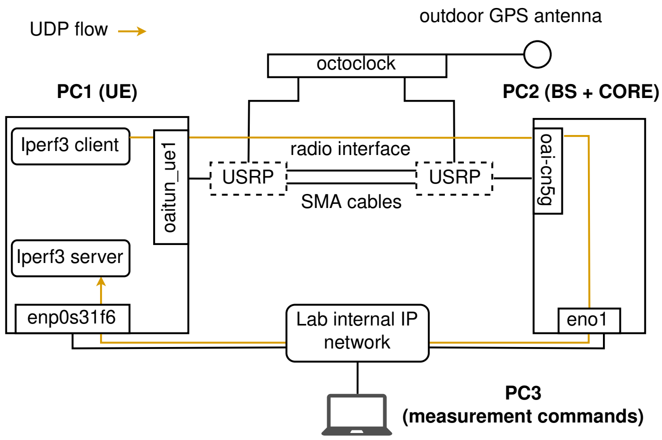


Fig. 2: Experimental testbed.

with an Intel(R) Xeon(R) W-2245 CPU @ 3.90GHz and 32 GB RAM with Ubuntu 20.04.6 LTS. PC3 is an HP ZBook Studio G3 with an Intel(R) Xeon(R) CPU E3-1505M v5 @ 2.80GHz with Ubuntu 24.04.1 LTS. The radio modules are universal software radio peripherals (USRPs) Ettus x300[3]. SMA cables carry radio frequencies that allow us to operate in a closed environment without radio interferences and in conformity with the regulation. We synchronize the reference signal and the pulse-per-second of each USRP with an Ettus Octoclock[4].

2) 5G network configuration: Our experimental setup uses the default radio bearer, we activate SDAP, RLC is in Acknowledged Mode (AM), and MAC uses only the padding BSR. We use the default amount PRBs from OAI, where the minimum by allocation is 5. The UE and BS Modulation and Coding Scheme (MCS) are dynamic. The indicators rely on radio quality (i.e., the Block Error Rate (BLER) and the Signal-to-Noise Ratio (SNR)) are in good condition. The experimental testbed operates in band 41; the absolute frequency SSB is 2593.35 MHz and the absolute frequency point A is 2574.24 MHz; the bandwidth is 40 Mhz; the radio frame organization is in Time Division Duplex (TDD) with 30 KHz subcarrier spacing and its configuration is that chosen by the French regulation authority (i.e., two DDDSU cycles on 10 milliseconds). Thus, 106 PRBs are available.

3) Uplink traffic generation: We generate a UDP flow with Iperf3[5], a network diagnostic tool. Iperf allows us to place a 64-bit counter in the UDP payload to identify a packet at the application-level (see AC from Sect. III-A).

4) Uplink latency measurement: A closed loop allows us to observe the traffic output and input on the same machine (PC1). Tshark probes allow us to tag the application-level packets sent (oaitun_ue1 in Fig. 2) and received (enp0s31f6 in Fig. 2) to measure the UL latency. The IP lab link can add between 0.5 ms and 1.5 ms to our Lat_{E2E} samples, which does not affect our observations. We use the Network

[3]Ettus USRP x300 - https://www.ettus.com/all-products/x300-kit
[4]Ettus Octoclock - https://www.ettus.com/all-products/octoclock/
[5]Iperf3 - https://iperf.fr/

Time Protocol (NTP) to synchronize PC1 and PC2 locally, ensuring accuracy lower than 1 millisecond.

B. Metrology

TBS_i and *BSRi* denote the i-th collected value Transport Block Size (TBS) assigned by BS to UE and BSR sent by UE, respectively.

Including all the headers and the payload sizes, we define the necessary size in bytes to send a packet without segmentation as *Basic Packet Unit (BPU)* by:

$$BPU = 1059 \text{ Bytes} \quad (1)$$

Let T_i^A, be the time at which packet i arrives in the RLC buffer, the packet inter-arrival Δt_i time is given by:

$$\Delta t_i = T_i^A - T_{i-1}^A \quad (2)$$

Let T_i^O, be the time at which packet i leaves the RLC buffer, the buffer latency W_i is given by:

$$W_i = T_i^O - T_i^A \quad (3)$$

The Buffer Occupancy Indicator (BOI) is given by:

$$BOI_i = \frac{W_i}{\Delta t_i} \quad (4)$$

Since the UE RLC instance can be viewed as a queuing system, we can apply the Little Law [17]. Averaging BOI_i give the average number of RLC Service Data Unit (SDU) that are queued in the UE RLC instance.

C. Experiment Configuration and Parameters

Each experimental test is 60 seconds of UL communication. The payload size UDP is the power of 2 closest to the Ethernet Maximum Transmission Unit (MTU) (i.e., 1500 bytes) with 1024 bytes. Iperf option allowed us to place a counter of 64 bits to tag each application-level packet. For all experiments, the bitrate parameter takes its values in power of two from 0.5 to 8 Mbps, to be lower than our observed testbed limit (i.e., 14 Mbps) and avoid throughput saturation.

V. Latency Evaluation

In this section, we expose the results of our experimental campaign. We present an experiment through three interpretations and analyses of the results. Subsection IV-B defines all the metrics and indicators that we explore in the following.

A. Results Computation

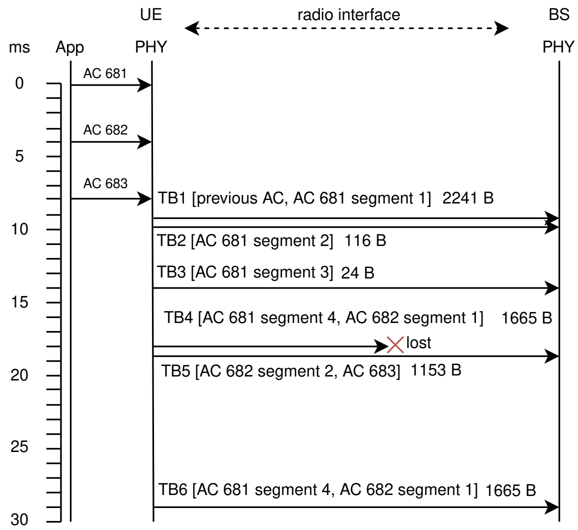
We extracted 30-second period of stable communication from each test to not take into account the occasional and abnormal latency peak. We did at least 10 tests for each bitrate studied, and each test comprises at least 915, 1830, 3659, 7321, 14642 and 29291 latency samples, respectively, for 0.25 to 8 Mbps. Observation of the confidence interval ensured that there were no uncontrolled random phenomena that could bias our results. All latency averages studied were computed with the average of each experimental test for each bitrate to smooth experimental noise. Taking into account the definition of UL latency in Sect. III-B, for a RLC packet or segment transmission considered, if MAC retransmission occurs on the transmitter side, we include additional delays Lat_{ue_phy} and $Lat_{hardware}$ in Lat_{ue_mac}. For Lat_{ue_rlc} and Lat_{bs_rlc} we consider only the final segment in sequence for a packet on the transmitter and receiver side.

B. Experimental Testbed Latency Analysis

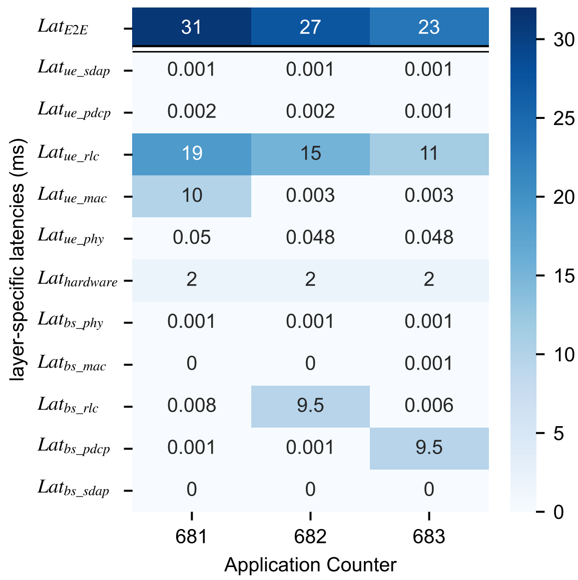
In this experiment, we analyze the testbed's UL latency (Lat_{E2E}). We start by focusing on an example from a 2Mbps test to show the interacting layers that trigger latency-generating operations. That allows us to identify the layer-specific latencies that influence Lat_{E2E}. Then, we generalize those observations through the overall bitrate studied to exhibit the proportion of each of the layer-specific latencies. Finally, we explore the reasons behind the more affecting layer-specific latency to explain our testbed Lat_{E2E} trend.

1) Highlighting the interacting layers triggering the latency-generating operations: We present a scenario from experimental tests on 2 Mbps to observe the interacting layers responsible for latency-generating operations and to exhibit their layer-specific latencies. Fig. 3a shows an example where an application on the UE side sends three packets with $4ms$ inter-arrival (Δt). This example exhibits the allocation constraints for UE in time (i.e., the communication begins $9ms$ after the first reception of AC) and in quantity (i.e., UE sends all ACs in several segments except AC 683). Both allocation limits and a unique MAC retransmission create disorder between segments and the ACs' reception. Fig. 3b shows that the higher layer-specific latency for the three ACs studied is Lat_{ue_rlc}. We observe that MAC retransmission directly impacts Lat_{ue_mac}. In addition, the disorder in the reception of the segments for AC 682 increases Lat_{bs_rlc} by $10ms$ milliseconds (i.e., from the reassembly operation). Finally, reception of AC 683 before AC 681 and 682 creates a delay of around $10ms$ in Lat_{bs_pdcp} (i.e., from the reordering operation). We checked the whole experimental test and confirmed that this scenario is representative. Lat_{E2E} is mostly composed of Lat_{ue_rlc}, and MAC retransmission is a regular event that triggers Lat_{ue_mac}, Lat_{bs_rlc}, and Lat_{bs_pdcp}. The $Lat\_hardware$ is constant and all other layer-specific latencies are insignificant.

2) Exhibiting the Uplink E2E latency composition: We generalize our analysis on the layer-specific latencies previously identified to the overall bitrates studied (i.e., from 0.25 Mbps to 8 Mbps). Fig. 4 shows that Lat_{ue_rlc} is at least higher than half of Lat_{E2E}. The visible averages between $300\mu s$ and 1.2 ms for Lat_{bs_rlc}, and between $570\mu s$ and 1.42 ms for Lat_{bs_pdcp} appear to be not costly in our testbed. However, these latencies could lie in breaking the low latency requirement of services highly sensitive to latency increase. In addition, we observe that the 9th decile of Lat_{ue_mac} often reaches around 10 ms. That exhibit a more significant impact on the low latency requirement. The Lat_{E2E} trend from 0.5 to 8 Mbps follows the Lat_{ue_rlc}, except for 0.25 Mbps



(a) Transmission and retransmission example over a shortened New Radio protocol stack and air interface. Each TB is followed by its Transport Block Size in bytes.



(b) Latency heatmap for the AC observed in Fig. 3a, with the Lat_{E2E} on the top and all its layer-specific latencies composition.

Fig. 3: Exhibition of the interactions between layers through a representative example with 2 Mbps bitrate.

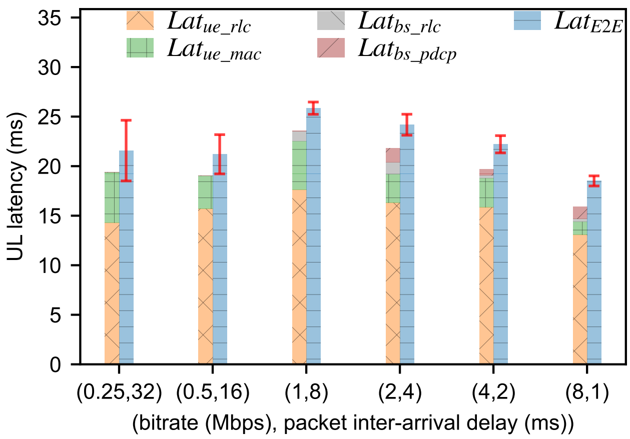


Fig. 4: Significant layer-specific latencies compared to End-to-End latency for all observed bitrate.

due to its Lat_{ue_mac}, however, we do not observe a clear Lat_{ue_rlc} trend with an increase from 0.25 to 1 Mbps and a decrease from 1 to 8 Mbps. The averages of the stacked layer-specific latencies are close to those of Lat_{E2E}, where the difference is almost equal to the $Lat_{hardware}$, and confirm the identification of layer-specific latencies. The confidence interval variation relies on the different amount of MAC retransmissions in each test for the same bitrate.

3) Analyzing the testbed Uplink applicative latency trend: Table II shows that the *BOI*'s average, and several modes for *BSR* and *TBS* in *BPU*. *BOI* average starts to be more than one by 1 Mbps. We checked the *BOI*s' ECDF for all bitrates, the buffer is always empty before receiving a new packet for 0.25 Mbps and half of the time for 0.5 Mbps, and from 1 Mbps RLC buffer is never drained before receiving the next packet. The inefficiency in draining the buffer RLC increases the scenario occurrences explained in V-B1 and explains that Lat_{bs_rlc} and Lat_{bs_pdcp} appear from 1 Mbps in Fig. 4. The *BSR* modes increase as a function of bitrate. As we generate constant flows, we expected to observe constant *BSR*. However, we observe several modes from 1 Mbps for each distribution, which denotes the *BSR* irregularity. As exposed in 4G from [16] the delay between UE BSR and its processing on the BS side leads BS to allocate radio resources based on outdated BSR and increases latency; that is also the case in 5G. Those observations explain the increases in Lat_{ue_rlc} from 0.25 Mbps to 1 Mbps. The *TBS* modes over one *BPU* begin at 1 Mbps and increase as a function of bitrate. The capability to send several *BPU* in the same Transport Block (TB) starts from 2 Mbps that smooth the latency average, that explains why the Lat_{ue_rlc} decreases. By extension, Lat_{E2E} decreases from 2 Mbps to 8 Mbps. Thus, we can divide the Lat_{E2E} trend into three parts. The first is a flat spot caused by MAC retransmissions on 0.25 Mbps side, and the second is between 0.5 and 1 Mbps caused by the inefficiency to drain the RLC buffer and the outdated *BSR*. The third is when more significant allocations allow MAC to send several *BPU*s in the same PDU (i.e., the same TB), from 2 Mbps to 8 Mbps.

C. Limitations

Our previous experiment and analysis demonstrate that our instrumentation enables us to observe, describe and explain our experimental testbed RAN latency trend. However, our contributions bring about some limitations to our assumptions. Indeed, we conducted our experimental campaign on

TABLE II: *BOI* mean, $BSR(BPU)$ and $TBS(BPU)$ (statistics) modes for the overall bitrate. *BOI* mean denoting the average of SDU in the Radio Link Control transmission buffer.

bitrate (Mbps)	0.25	0.5	1	2	4	8
BOI mean	0.4	0.9	2.2	4	7.6	22.8
$BSR(BPU)$ modes	0	0	(0,1,2)	(3,1,2)	(5,9,3,7)	(9,13,7)
$TBS(BPU)$ modes	0	0	(0,1.5)	(0,1.5,2)	(0,1.5,2.5,4,6)	(1.5,0,5.5)

RLC in AM, that it should be the same in Unacknowledged Mode (UM), but that needs to be verified. In addition, we did not measure the overhead of our instrumentation and assume that it is the same cost reported on 4G from [10].

VI. Conclusions and Future Work

In this work, we explored and analyzed the UL E2E latency on the RAN by placing probes in each 5G NR layer. Our results exhibit the interactions between layers that trigger the latency-generating operations. Where inefficient RLC buffer draining, packet segmentation, sending of segments for different packets in the same TB and MAC retransmissions, generate buffering and reassembly RLC, and reordering PDCP. Those latency-generating operations impact the RLC and MAC latency on the transmitter side, and the RLC and PDCP latency on the receiver side. Our latency decomposition studied at different bitrates confirms these observations and exposes that the UL RAN latency is composed of RLC and MAC on the transmitter side and RLC and PDCP on the receiver side. We also diagnose our experimental testbed UL latency trend through the considered bitrates range. As a function of bitrate, the decrease in packet inter-arrival associated to outdated BSR does not allow UE to drain its RLC buffer efficiently, which increases latency, whereas on higher bitrates larger allocations allow UE to send several packets in the same TB, smoothing out the average latency.

Future work includes extending our instrumentation to measure DL latency and generalizing LatSeq to other 5G implementations. Furthermore, we plan to analyze the change in system behavior by adapting the experiment parameters to emulate future 5G flow requirements (e.g., a Virtual Reality (VR)-based application). In addition, placing the experiment on scale with several UE should reflect a more realistic condition.

Acknowledgment

This work was carried out in the context of 5GMetaverse, a project funded by the French government as part of the economic recovery plan, namely “France Relance” and the investments for the future program.

References

[1] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, “Business case and technology analysis for 5g low latency applications,” *IEEE Access*, vol. 5, pp. 5917–5935, 2017.

[2] S.-M. Park and Y.-G. Kim, “A metaverse: Taxonomy, components, applications, and open challenges,” *IEEE Access*, vol. 10, pp. 4209–4251, 2022.

[3] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, “Latency critical iot applications in 5g: Perspective on the design of radio interface and network architecture,” *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[4] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, “A survey on low latency towards 5g: Ran, core network and caching solutions,” *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[5] M. Attaran, “The impact of 5g on the evolution of intelligent automation and industry digitization,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 5977–5993, May 2023.

[6] E. J. Oughton, W. Lehr, K. Katsaros, I. Selinis, D. Bubley, and J. Kusuma, “Revisiting wireless internet connectivity: 5g vs wi-fi 6,” *Telecommunications Policy*, vol. 45, no. 5, p. 102127, 2021.

[7] 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 Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 479–494.

[8] I. Khan, M. Ghoshal, J. Angjo, S. Dimce, M. Hussain, P. Parastar, Y. Yu, C. Fiandrino, C. Orfanidis, S. Aggarwal, A. C. Aguiar, O. Alay, C. F. Chiasserini, F. Dressler, Y. C. Hu, S. Y. Ko, D. Koutsonikolas, and J. Widmer, “How mature is 5g deployment? a cross-sectional, year-long study of 5g uplink performance,” in *2024 IFIP Networking Conference (IFIP Networking)*, 2024, pp. 276–284.

[9] Z. Zhang, H. Wang, H. Lv, J. Sun, G. Li, and X. Han, “Chat: Accurate network latency measurement for 5g e2e networks,” *IEEE/ACM Transactions on Networking*, vol. 31, no. 6, pp. 2854–2869, 2023.

[10] F. Ronteix–Jacquet, A. Ferrieux, I. Hamchaoui, S. Tuffin, and X. Lagrange, “Latseq: A low-impact internal latency measurement tool for openairinterface,” in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1–6.

[11] V. Hamici-Aubert, J. Saint-Martin, R. Navas, G. Z. Papadopoulos, G. Doyen, and X. Lagrange, “Dissecting 5g new-radio latency: Interacting layers and latency-generating operations: Code and data.” May 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15377254

[12] E. Dahlman, S. Parkvall, and J. Sköld, “Chapter 6 - radio-interface architecture,” in *5G NR: the Next Generation Wireless Access Technology*, E. Dahlman, S. Parkvall, and J. Sköld, Eds. Academic Press, 2018, pp. 73–102.

[13] R. A. K. Fezeu, E. Ramadan, W. Ye, B. Minneci, J. Xie, A. Narayanan, A. Hassan, F. Qian, Z.-L. Zhang, J. Chandrashekar, and M. Lee, “An in-depth measurement analysis ofnbsp;5g mmwave phy latency andnbsp;its impact onnbsp;end-to-end delay,” in *Passive and Active Measurement: 24th International Conference, PAM 2023, Virtual Event, March 21–23, 2023, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 284–312.

[14] A. Maghsoudnia, E. Vlad, A. Gong, D. M. Dumitriu, and H. Hassanieh, “Ultra-reliable low-latency in 5g: A close reality or a distant goal?” in *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, ser. HotNets ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 111–120.

[15] S. Mostafavi, M. Tillner, G. P. Sharma, and J. Gross, “Edaf: An end-to-end delay analytics framework for 5g-and-beyond networks,” in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2024, pp. 1–6.

[16] F. Ronteix-Jacquet, X. Lagrange, I. Hamchaoui, and A. Ferrieux, “Rethinking buffer status estimation to improve radio resource utilization in cellular networks,” in *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, 2022, pp. 1–5.

[17] L. Kleinrock, *Theory, Volume 1, Queueing Systems*. USA: Wiley-Interscience, 1975.