# An In-Depth Measurement Analysis of 5G mmWave PHY Latency and Its Impact on End-to-End Delay

Rostand A. K. Fezeu[1(✉)], Eman Ramadan[1], Wei Ye[1], Benjamin Minneci[1], Jack Xie[1], Arvind Narayanan[1], Ahmad Hassan[1], Feng Qian[1], Zhi-Li Zhang[1], Jaideep Chandrashekar[2], and Myungjin Lee[3]

[1] University of Minnesota, Twin Cities, USA
{fezeu001,ye000094,minne078,xie00056,hassa654,fengqian}@umn.edu,
{eman,arvind,zhzhang}@cs.umn.edu
[2] Interdigital, Wilmington, USA
Jaideep.Chandrashekar@InterDigital.com
[3] Cisco Systems, San Jose, USA
myungjle@cisco.com

**Abstract.** 5G aims to offer not only significantly higher throughput than previous generations of cellular networks, but also promises millisecond (ms) and sub-millisecond (ultra-)low latency support at the 5G physical (PHY) layer for future applications. While prior measurement studies have confirmed that commercial 5G deployments can achieve up to several Gigabits per second (Gbps) throughput (especially with the mmWave 5G radio), are they able to deliver on the (sub) millisecond latency promise? With this question in mind, we conducted to our knowledge the first in-depth measurement study of commercial 5G mmWave PHY latency using *detailed physical channel events and messages*. Through carefully designed experiments and data analytics, we dissect various factors that influence 5G PHY latency of both downlink and uplink data transmissions, and explore their impacts on end-to-end delay. We find that while in the best cases, the 5G (mmWave) PHY-layer is capable of delivering ms/sub-ms latency (with a minimum of 0.09 ms for downlink and 0.76 ms for uplink), these happen rarely. A variety of factors such as channel conditions, re-transmissions, physical layer control and scheduling mechanisms, mobility, and application (edge) server placement can all contribute to increased 5G PHY latency (and thus end-to-end (E2E) delay). Our study provides insights to 5G vendors, carriers as well as application developers/content providers on how to better optimize or mitigate these factors for improved 5G latency performance.

**Keywords:** mmWave · 5G · PHY Layer · Latency · Sub-millisec · End-to-end · Network measurement · 5G Latency Dataset · AWS WaveLength · AWS Local Zone · AWS Regional Zone

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. Brunstrom et al. (Eds.): PAM 2023, LNCS 13882, pp. 284–312, 2023.
https://doi.org/10.1007/978-3-031-28486-1_13

## 1 Introduction

The past few years have seen a rapid commercial deployment of 5G networks. With enhanced mobile broadband services (eMBB), 5G promises to offer much higher bandwidth than previous generations of cellular networks to consumers. Existing measurement studies [10,20,23,29,33] have found that 5G radio technologies can in general achieve higher throughput performance than 4G LTE. For example, with line of sight (LoS), mmWave 5G radio can deliver up to several Gbps of *downlink* (DL) bandwidth [20,29,33] and up to hundreds of Mbps *uplink* (UL) bandwidth [23], albeit their performance can fluctuate wildly.

**Motivations for this Study.** From the perspective of new applications which require mission critical communications, what is perhaps more exciting is the promise of 5G to offer millisecond (ms) or even sub-millisecond (PHY-layer) latency support to applications [Sect. 7.5 in [3]][1] *e.g.,* through the so-called *Ultra Reliable Low Latency Communication* (URLLC) services [Sect. 7.9 in [3]] [4,17,27]. These applications include but are not limited to, Autonomous Vehicles (AVs) and drones supported with edge-assisted cooperative driving/flying intelligence, Augmented/Virtual reality (AR/VR), and "metaverse", all which require extreme low latency and very high reliability to make crucial decisions.

**Background of 5G Measurement Studies and Research Gap.** Recently, several measurement studies have been conducted to assess the latency performance of current 5G deployments and their impact on applications [23–25,29,32–35,44]. These studies have shown that 5G E2E latency performance is affected by factors such as sporadic coverage, link quality disturbances due to User Equipment (UE) mobility, handovers, and poor interactions across the 5G network stack. Furthermore, they have focused solely on UL or DL separately, from an E2E perspective. However, they cannot be used to infer the latency of 5G in PHY-layer (*i.e.,* both UL and DL) and identify issues that could prevent 5G from delivering its expected latency performance on the PHY-layer nor what factors can significantly affect the delay in PHY-layer.

**Objectives of this Study.** In this paper, we present a measurement study of today's commercial mmWave 5G latency on the PHY-layer. Using AT&T and Verizon (VZW)'s mmWave 5G networks as case study, we seek to quantitatively answer the following critical, yet unaddressed questions: 1) Is today's commercial 5G network capable of delivering millisecond/sub-millisecond ($\leq$ 1ms) latency on the PHY-layer? If so, what is the best achievable PHY-layer latency in DL and UL? 2) Quantitatively, what are the important factors of the 5G Radio Access Network (RAN) that can significantly affect PHY-layer latency? 3) What factors are inherent in the design of the 5G RAN architecture, which may not be easily controlled or mitigated, and what factors are due to the current 5G network configuration or implementation of the cellular carriers, which may be further improved or even eliminated in future 5G deployments? 4) How do other factors

[1] The one-hope (UE to gNB) target for URLLC "should be 0.5ms for UL, and 0.5ms for DL".

such as the placement of the application server and packet payload affect the latency of 5G PHY-layer and therefore the E2E delay experienced by applications? We answer these questions through a close look analysis of 5G mmWave PHY-layer key performance indicators (KPIs) with the aim of quantifying the impact of various factors and configurations. Our approach is laid out as follows: First, we aim to quantitatively understand the PHY-layer latency and study it under the "best-case" scenario (Sect. 4). Second, we quantify the impact of several factors that impact the PHY-layer latency (Sects. 5 and 6). Lastly, we explore the latency benefits and drawbacks of deploying services on edge nodes supported by mmWave 5G (Sect. 7). Based on our knowledge, our paper is the first to answer the question, "Is sub-millisecond PHY-layer latency achievable with today's commercial 5G"? And what impact does several factors like 5G smartphone radio ON-OFF cycle and server placement have on the PHY-layer and E2E delays. Next, we summarize our key findings and contributions.

F1. **Today's Best Achievable PHY-layer Delay** (Sect. 4). Our analysis shows that the best achievable mmWave 5G PHY-layer latency is 0.85 ms which occurs about 2.27% of the time. Sub-millisecond ($\leq$ 1ms) PHY-layer latency is guaranteed only 4.42% of the time, with PHY-layer latency reaching up to 3.08 ms about 22.36% of the time (Sect. 4.1). This delay is limited by network side UL scheduling with control overhead contributing to the largest share (about 81%) compared to data overhead, as a result of scheduling requests and backoffs on the busy shared radio channel (Sect. 4.3).

F2. **Impact of Channel Conditions** (Sect. 5). A UE periodically (based on the configurations) reports the DL channel condition to the base station by calculating the value of the channel quality indicator (CQI), which is a number from 1 to 15, where 15 indicates the best channel condition. When the CQI value drops, transmitted data might be corrupted, requiring retransmission (ReTx). Our experiments show that: 1) The PHY-layer latency when exactly one ReTx occurs is 1.33 ms, making sub-millisecond ($\leq$ 1ms) PHY-layer latency not achievable. 2) As the number of ReTxs increases, the overhead of the PHY-layer data increases 3.5 times the overhead of the control (Sect. 5.1). 3) On average, there is a 2ms additional overhead delay on the PHY-layer when the CQI drops noticeably (Sect. 5.2).

F3. **Impact of Mobility and Handovers (HOs)** (Sect. 6). As mmWave is directional, highly susceptible to many impairment factors, and has shorter coverage ranges, mobility not only affects the channel condition experienced by a UE, but also causes HOs in some situations. All these further impact the latency on the PHY-layer. We find that: 1) When a UE is walking with good channel conditions (*i.e.,* high CQI value) and no HOs occur, the additional PHY-layer overhead due to mobility is 0.51 ms (Sect. 6.1). 2) When there is a HO, the minimum additional PHY-layer overhead is 2 ms (Sect. 6.2).

F4. **Impact of UE Sleep Cycle** (Sect. 7.2). As a way to reduce power consumption on 5G smartphones, 5G supports *discontinuous reception* (DRX). The operations of DRX modes depend on the UE's *state*. We focus only on the *connected* state (CDRX), namely, the UE has established a connection

with the base station. In such a state, the UE radio antennas go through *ON* and *OFF* cycles (*i.e.,* awake and asleep states). Two scenarios can occur; 1) The DL transmission occurs while the UE is awake, no additional delay is incurred (best case). 2) The network has data, but the UE is asleep (worst case). Our results show that there is an additional overhead of 6.4 ms (on average) to the PHY-layer latency in the worst case.

F5. **Impact of Packet Payload Size** (Sect. 7.3). We use PING packets to mimic different application payload sizes. We find that the packet payload size has little to no impact on the PHY-layer delay. Our results show that the same time is taken to transmit a ping packet with 100 bytes and 1200 bytes payload. This is because when the payload size of the PING packet increases, the network adopts more hybrid ARQ (HARQ) process IDs [1] that work in parallel to send and receive data between the UE and the base station.

C1. We present an in-depth and thorough analysis which allows for the quantitative revelation of the status quo of today's mmWave 5G PHY-layer delay, identifying carrier specific configurations and poor design choices which hinders 5G's promise of sub-millisecond PHY-layer delay.

C2. We study several factors that impact the latency on the PHY-layer and quantify them, showing that 5G network configurations and server placement decisions can significantly impact the PHY-layer delay and thus E2E latency.

C3. We make all our data as well as other artifacts used in our study publicly available to enable research continuation within the community: https://github.com/FarRoss/5gPHYLatency

**Ethical Considerations.**This study was carried out by paid and volunteer students. We purchased several dedicated smartphones for experiments only and several unlimited plans from AT&T and Verizon mmWave 5G carriers. No personal identifiable information (PII) was collected or used, nor were any human subjects involved. This study is consistent with the Wireless Network Customer Agreement.This work does not raise ethical issues.

## 2 Main Measurement Campaign and Challenges

In this section, we present our measurement methodology, experimental platform and setup, data collection approach, equipment, and tools used during this study.

**Commercial 5G Networks.** We judiciously select two urban areas in two densely populated large metropolitan cities in the U.S., which are two cities with the first mmWave 5G deployments launched in April 2019. **Area 1)** A four-way intersection with three dual-panel faced 5G towers. **Area 2)** A four-blocks loop near the U.S. Bank Stadium in downtown Minneapolis with three 5G base stations. Each block is about 90 m. These two outdoor urban areas are very busy with heavy traffic, several restaurants, coffee shops, railroad crossings, and outdoor parks. At the time of this study, high band/mmWave (24.25–27.5 GHz) 5G deployment is supported by three major U.S. cellular carriers (AT&T, T-Mobile,

and Verizon (VZW)) using Non-Standalone mode (NSA) [5]. NSA adopts a dual connection mode in which 4G acts as an anchor for the control plane functionality and to ensure continuous data connectivity. On the other hand, Standalone mode (SA) relies on 5G for all control and data plane activities. Since mmWave deployments are not continuous and have coverage holes, using mmWave with SA 5G can lead to loss of connectivity during mobility. Additionally, any future SA mmWave 5G deployments will most likely use the same 5G RAN technologies. Thus, we believe that our finding will also be valid for future mmWave SA 5G deployments. Mid-bands (3.3-3.8 GHz) and low-bands (700 MHz, n28) have not been deployed yet, thus, beyond the scope of this study. Refer to recent work [22] for a study of the mid-band 5G in Europe. Most of our controlled experiments are focused specifically on **Area 1**.

**5G UE and Measurement Tools.** We use four phones, two S20s (Exynos 990 Qualcomm SM8250 Snapdragon 865 5G) and two S21 Ultras (Exynos 2100 Qualcomm SM8350 Snapdragon 888 5G) [8]. We believe that these phones represent the state-of-the-art 5G smartphones at the time we conducted the measurement study with powerful communication modems, Mali-G77 MP11 and Mali-G78 MP14, respectively. Moreover, smartphone chip-sets do not affect the network performance at the TCP and application layers [44].

To access the 5G New Radio (NR) stack and PHY-layer KPIs from chipset's diagnostic interfaces (*Diag*), we use a professional tool called *XCAL* [6]. XCAL runs on a laptop connected to smartphones via USB or USB-C (Fig. 1). It monitors, decodes, and deciphers signaling messages and the 5G RAN protocol stack interactions between the UE and gNB following the 3GPP Rel-15 standard. For our controlled experiments, we choose traceroute and ICMP-based PING packets of 32 bytes because of two reasons; 1) It is readily available in Android smartphones and does not require rooting devices. 2) To avoid any limitations due to lack of radio resources using bigger packet sizes. However, we also study the impact of larger packet sizes on PHY-layer and E2E delay (See Sect. 7).

**Cloud Server.** To explore the benefits of deploying services on the edge, we perform our latency measurements using the Amazon Web Services (AWS) cloud platform [2]. We selected three AWS *nodes* to interact with the UE as shown in Fig. 1 (i) An AWS Wavelength (WL) node is the nearest edge and is directly connected to the VZW's 5G core network. It provides a commercially available 5G edge cloud service through VZW's 5G in the same geographical location as the UE. (ii) An AWS Local Zone (LZ) node is the second-nearest edge located in the same geographical location as the UE. Unlike WL, LZ is not directly connected to VZW's 5G core network. (iii) An AWS Regional (RG) node is the farthest away from a UE but is also located in the same geographical region as the. UE[2] Other main operators, like AT&T and T-Mobile are not directly connected to an edge platform. Therefore, we use VZW to measure the latency for the best-case scenario using a WL node.

[2] Our definition of region in this paper is as per AWS, and it is a cluster of a minimum of 3 data centers.

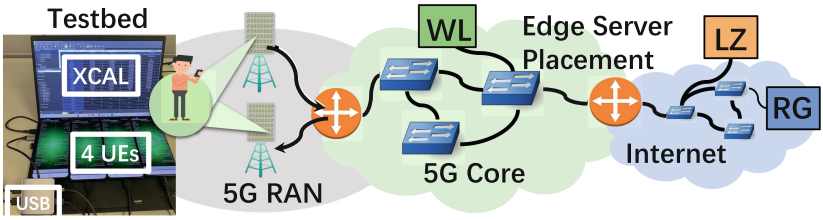



**Fig. 1.** Measurement Setup and Edge Server Placement.

**Challenges.** In this study, we face three main challenges; **[C1]** Internet-side buffering, congestion, and data transportation policies of the carrier network can negatively affect the E2E round-trip time (RTT). We minimize this impact by using a WL node. To ensure high-speed connectivity, we conducted several test runs using the Ookla speed test [9], and 5G Tracker [31] to measure the 5G performance. We validate the results are within the expected 5G performance before we start each experiment. **[C2]** We have no visibility into the commercial cellular carrier network. We use XCAL to overcome this challenge. The major advantage of XCAL compared to other wireless network analyzers such as MobileInsight [26] and 5G-Tracker [31] is its ability to decode 5G signaling messages. **[C3]** We need to monitor and trace a PING packet in the 5G RAN stack of the UE down the PHY-layer to the gNB and identify when the gNB sends a packet to the UE. To do this, we leverage consecutive PING echo request intervals. Specifically, 1) we monitor the PHY-layer activities with and without data transfer and 2) we enforce the reception of the PING echo reply from the server between consecutive PING echo requests. Unless otherwise mentioned, we use 1000 ms as the PING intervals. This approach also avoids the case when two or more PING echo replies are sent to the UE at once due to network-side buffering/congestion. During no data periods, our observation of the PHY-layer control channels show that, based on the network configurations, the UE sends (periodic, semi-periodic, or aperiodic) reports to the network which aid in resource allocation and scheduling decisions [Sect. 5.2 in [14]]. Simply put, this approach is like a heartbeat with varying beat intervals, where the corresponding echo requests/responses are the beats. This helps establish the time spent in each phase, as explained later in Fig. 5. Another issue we faced is that XCAL reports the data per channel. Since UE and gNB communicate using several channels, domain knowledge is required to correlate the different events and establish the timeline to trace the UL vs. DL packets. We discuss this in more details later (See Sect. 3).

**Experiments and Data Collection.** With the above methodology, we conducted several controlled experiments on 5G, resulting in 192+ hours of experiments. Our experiments span different hours (morning, rush hours, night) and days (including weekends). The state of UE Radio Resource Control (RRC) [Sect. 5 in [15]] may further skew the measurement results [33], *i.e.,* if the UE is in RRC_IDLE or RRC_Inactivity state when sending a packet, an additional delay is incurred to transition to RRC_Connected before sending the PING request. The UE

will always be in RRC_Connected state when receiving the PING echo reply, as the length of RRC_Connected is 320 ms [33] which is far greater than the worst RTT (100 ms) observed in our experiments. Before each experiment, we close/stop all background apps, disable background-app refresh, and turn off the WiFi interface. To avoid delay overhead during transitions from RRC_IDLE or RRC_Inactivity to RRC_Connected state, we first play a random YouTube video for 30 s, then immediately close the YouTube app, wait 2 s, and then start the experiment. This ensures that the UE is in the RRC_Connected state before sending the echo request. To minimize the UE-side factors that may affect our measurements, we placed the smartphones on a flat surface during stationary experiments and kept them attached to a car phone holder for driving experiments.

## 3 5G PHY Processing and Factors

In this section, we introduce the 5G NR, 5G RAN, and zero in on the 5G PHY-layer, and outline its key operations. The goal is two-fold: 1) introduce the key PHY-layer interactions used in 5G NR defined by the 3GPP standards that are most relevant to our study to justify our results and insights; and perhaps more importantly, 2) dissect the various components of 5G PHY processing, and identify the major factors which may influence 5G PHY latency, and consequently the E2E latency experienced by applications running on a UE or a remote server.

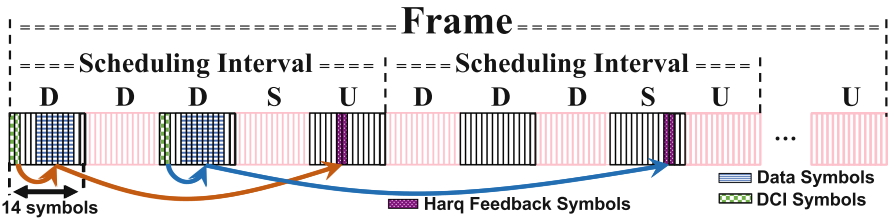



**Fig. 2.** Illustration of Frame and Scheduling Intervals.

Like 4G and its predecessors, 5G is a *scheduling* system: when a UE can receive or transmit data is completely controlled by the base station (4G eNB or 5G gNB) through Medium Access Control (MAC) scheduling. The MAC layer multiplexes and segments the upper layer data (*e.g.,* user traffic) into transport blocks [Sect. 6.1.1.1 in [14]] of *dynamic* sizes (See Sect. 4.3 for more details). Then it passes the transport blocks down the PHY-layer to be transmitted through dedicated DL and UL transport channels.[3] 5G NR introduces *flexible* subcarrier spacing (SCS), from 15 kHz (same as in 4G LTE), to 30 kHz (mid-band), and 120 kHz (mmWave), to accommodate diverse UE capabilities and meet varying

[3] The primary physical channel for the DL transmissions (base station to UE) is PDSCH (physical downlink shared channel), and for the UL transmissions (UE to base station) is PUSCH (physical uplink shared channel).

bandwidth and latency requirements of applications. The wider SCS not only allows for higher channel bandwidth, but also enables lower latency through a shorter *slot* time, *i.e.,* from 1 ms in 15 kHz down to 0.125 ms in 120 kHz (mmWave). A *slot* is defined as the basic (time) unit in which radio transmissions are commonly scheduled [Sect. 4.3.1 in [12]] (See Sect. 4). Our study focuses on 5G mmWave, as it can (potentially) provide both high bandwidth and low latency.

During each *slot*, one data chunk[4] is transmitted over the radio interface to/from the UE. The scheduling configurations are exchanged via the *downlink control information* (DCI)/the *uplink control information* (UCI) carried in the Physical DL Control Channel (PDCCH)/Physical UL Control Channel (PUCCH) respectively, as part of the PHY-layer control signaling (See Fig. 2). 5G mmWave uses *time division duplex* (TDD) which means both the DL and UL share the same carrier frequency (physical transport channel) [16]. However, the transmissions of DL and UL are scheduled at different times, *e.g.,* using different *slots* on the same frequency. We expand on these points below.

**Slots and Scheduling.** The 3GPP standards allow flexible scheduling of which *slots* are dedicated for DL vs. UL transmissions [Sect. 5 in [16]]. However, we find that current commercial 5G deployments still use a "fixed" pattern. For example, as illustrated in Fig. 2, VZW mmWave 5G uses a 5-*slots* pattern, `DDDSU` for DL/UL transmission scheduling: The first three *slots* ("DDD") are reserved for DL transmission only, the last *slot*, ("U") is reserved for UL transmission only, while the fourth *slot*, ("S") is *flexible* – it can be used either for DL or UL transmission, or both. *For DL Transmission (data sent from gNB):* the scheduling information carried in the DCI specifies which symbols within "D" (and "S") *slots* are used to carry data; it also indicates which symbols in the "U" (and "S") *slots* may be used to carry UL transmissions, including UCI. DCI is typically carried in the first 1-3 symbols in a "D" or "S" *slot*, while UCI is carried in the last symbol in a "U" or "S" *slot*. While the UE is *active* in a "Connected" state, it monitors the physical channels to see if there is DL data and/or control traffic for it. *For UL Transmission (data sent from UE):* the UE first sends a *scheduling request* in either the "U" or "S" *slot* which only informs the network that the UE has data to transmit. The UE later sends the Buffer State Report (BSR) [Sect. 5.4.5 in [13]], which informs the network the UL data volume. With the BSR information, the network then explicitly *grants* the UE resources. Lastly, the UE prepares and transmits the data using the scheduled future UL *slots*. As a result, we can deduce that this configuration enables asymmetric traffic between UL and DL demands. Thus, UL transmissions likely incur longer latency than DL, which is also confirmed by our results in Sect. 4.

**Channel Conditions (CQI), Modulation and Coding Schemes (MCS).** A UE periodically reports to the gNB the DL channel condition using the channel quality indicator (CQI), a number from 1 to 15, where 15 indicates the best

[4] Assuming no spatial multiplexing, which is the case of VZW 5G mmWave. However, with spatial multiplexing, at most 2 Transport Blocks can be transmitted per *slot*.

channel condition [Sect. 5.1.6 in [14]]. The gNB uses this CQI value to determine which modulation (*e.g.,* QPSK, 32QAM, or 64QAM) and coding rate (*e.g.,* the number of redundant bits) to use to encode the data. This is collectively referred to as the Modulation and Coding Schemes (MCS) [Sect. 5.1.3 in [14]]. The MCS value informs a UE on how to decode a DL transmission or how to encode a UL transmission. The main take-away is the following: higher CQI generally leads to higher MCS – *if* there is sufficient data buffered to warrant it; and higher MCS means more information bits (*i.e.,* more data from the upper layer) is carried per *slot*. As the MAC layer multiplexes data from multiple "logical" channels (*e.g.,* RRC messages, multiple concurrent user sessions), an IP packet from an application server to a UE (or vice versa) can be segmented into multiple data chunks, therefore requiring multiple *slots* for the packet to be delivered to the user (or server), incurring longer latency even under "ideal" channel conditions.

**Hybrid ARQ (HARQ) Re-transmission Processes.** As in 4G, 5G employs a hybrid ARQ (HARQ) mechanism that combines forward error correction (FEC) coding and automatic re-transmission (ReTx) request (ARQ) to recover errors. At either the gNB or UE, the MAC layer is responsible for re-transmitting a data chunk upon receiving a negative acknowledgment (NACK). For DL, a UE has to explicitly ACK or NACK every transmission. For UL, the gNB implicitly "NACKs" corrupted received data for the UE to (re)transmit (Sect. 5.1). Under poor channel conditions, transmitted data chunks are likely to be corrupted, and require ReTxs. This is reflected by the *block-level error rate* (BLER) [13]. As ReTxs require additional *slots*, poor channel conditions and higher BLERs can significantly increase the latency experienced by users.

**DRX Mode in *Connected State: CDRX.*** Both 4G and 5G support *discontinuous reception* (DRX) for the UE power management. The operations of DRX modes depend on which *state* the UE is in. We focus only on the *Connected* state (CDRX), namely, the UE has established a connection with the gNB [1]. In such state, the UE goes through *active* and *sleep* cycles to save power. Only when active, the UE searches for data, receives, or transmits data. Therefore, if data from an application session arrive at the gNB while the UE is asleep, the gNB has to wait until the next active cycle to signal the UE and allocate DL radio resources for DL transmission, which further increases the latency (Sect. 7.2).

**Mobility and Handovers.** mmWave 5G is directional, highly susceptible to many impairment factors, and has shorter coverage ranges. Therefore, UE mobility not only affects the channel conditions experienced, but also causes handovers (HOs) in some situations, further affecting the E2E latency experienced by users/applications (Sect. 6).

## 4 5G PHY-layer Latency: Best Cases

Throughout this section, we define the best-case as: the UE is stationary, in RRC_Connected state, and facing a 5G base station. This is because the channel

conditions *i.e.,* CQI values $\geq 12$ which indicates high MCS [Sect. 5.2.2 in [14]] (See Sect. 5) and no ReTxs occur. We summarize all the latency definitions in Table 1.

### 4.1 Quantifying Best-Case PHY Latency

PHY-layer latency, $T_{Phy}$ is defined as the time taken to send a PING echo request in the UL, ($T_{UL}$) and receive the corresponding echo reply in the DL, ($T_{DL}$) on the physical layer. *i.e.,* $T_{Phy} = T_{UL} + T_{DL}$. To compute $T_{Phy}$, we carefully trace every PING packet on the UE side down the 5G RAN stack. Based on the data collected on the different radio channels, we use domain knowledge to: 1) isolate the PING packet from other noisy data such as beam management-related control plane messages, 2) correlate the different transport channel PING related messages, and 3) synchronize (and group) the different channel events in UL and DL. Furthermore, we compute i) the time taken to send the PING data on the physical transport data channel, $T_{Phy}^{Data}$ and ii) the time taken to send related control messages on the physical transport control channels, $T_{Phy}^{Ctrl}$.

**Table 1.** Summary of the Definitions for the Different Latency Terms Used

| Delay<br>⇓Quantity | Delay<br>⇓ Definition/Breakdown | Delay in terms<br>⇓ UE-gNB Interactions |
|---|---|---|
| $T_{UL}^{Ctrl}$ | UL Control delay in the PHY-layer | $T_{UL}^{Ctrl} =$ (U1) + (U2) |
| $T_{UL}^{Data}$ | UL Data delay in the PHY-layer | $T_{UL}^{Data} =$ (U3) |
| $T_{UL}$ | UL delay in the PHY-layer,<br>$T_{UL} = T_{UL}^{Ctrl} + T_{UL}^{Data}$ | $T_{UL} =$ (U1) + (U3) + (U3) |
| $T_{DL}^{Ctrl}$ | DL Control delay in the PHY-layer,<br>$T_{DL}^{Ctrl} = T_{DL}^{Ctrl1} + T_{DL}^{Ctrl2}$ | $T_{DL}^{Ctrl} =$ (D1) + (D3) |
| $T_{DL}^{Data}$ | DL Data delay in the PHY-layer | $T_{DL}^{Data} =$ (D2) |
| $T_{DL}$ | DL delay in the PHY-layer,<br>$T_{DL} = T_{DL}^{Ctrl} + T_{DL}^{Data}$ | $T_{DL} =$ (D1) + (D2) + (D3) |
| $T_{Phy}$ | UL and DL delay in the PHY-layer,<br>$T_{Phy} = T_{DL} + T_{UL}$ | $T_{Phy} =$ (U1) + (U3) + (U3)<br>+ (D1) + (D2) + (D3) |
| $T_{5G\_RAN}$ | Delay in the PHY-layer including 5G RAN delay of the UE | See Fig. 20 |
| $T_{5G\_Core+Inet}$ | Delay from (U1) to cloud server to (D1) | $T_{5G\_Core+Inet} =$ (U1)<br>+ wired delay + (D1) |
| $T_{E2E\_RTT}$ | Round Trip Time<br>from the applications | $T_{E2E\_RTT} = T_{5G\_Core+Inet} +$<br>$T_{5G\_RAN}$ |
| $T_{Phy\_RTT}$ | Round Trip Time<br>from the PHY-layer | See Fig. 20 |

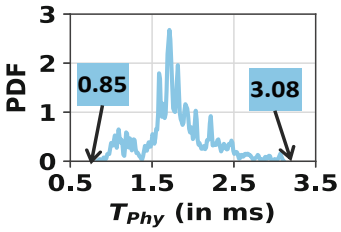



**Fig. 3.** Best Case $T_{Phy}$ showing min and max achievable delays.

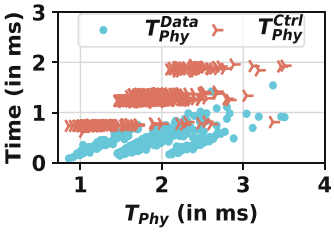



**Fig. 4.** Breakdown of $T_{Phy}$ into Control and Data delays.

**Results.** We make the following observations. (1) In the best case, today's $T_{Phy}$ delay scale, can be as low as 0.85 ms and as high as 3.08 ms (See Fig. 3). (2) Interestingly, *only* 4.43% of all our dataset samples have delays $\leq$ 1 ms. In other words, sub-millisecond latency occurs about $\leq$ 5% of the time. Most delays fall between 1 ms and 2.5 ms (*i.e.,* 87.69%), and 7.83% have delays between 2.5 ms and 3.08 ms. The maximum best case $T_{Phy}$ latency is largely unsurprising: previous studies have calculated this delay to be between 2.19±0.36ms [44]. Nevertheless, our results provide insight into today's *expected* delay scale, which can inspire new design opportunities. For example, to ensure that 5G can support latency-critical applications, sub-milliseconds PHY-layer transmission is a must. In particular, Rel 15 38.913 [3] standardized the 5G first hop (*i.e.,* PHY-layer) delay for URLLC to 1 ms. (3) A breakdown of the best case $T_{Phy}$ delay into the control ($T^{Ctrl}_{Phy}$) and data ($T^{Data}_{Phy}$) overhead shows that the control overhead is on average 3.78 times more than the data overhead (See Fig. 4). Thus, it is clear that, today's mmWave 5G PHY-layer latency is far from enabling latency-critical applications. The question now remains, *what are the design opportunities or improvements which can favor the majority of the delay to fall below 1 ms?* To answer this question, we use Fig. 5 to dissect $T_{Phy}$ into DL and UL delays.

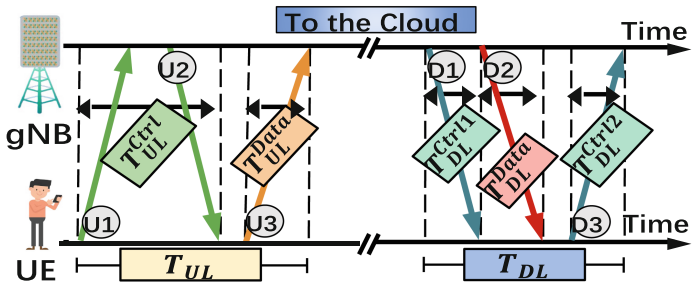
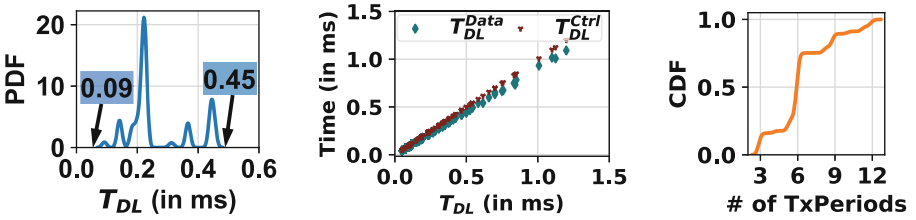



**Fig. 5.** PHY-layer Interaction between UE and gNB.

### 4.2 Dissecting DL PHY Latency

**DL Transmission** : As shown in Fig. 5, when data arrives at the gNB destined for a UE, the gNB first sends the data schematics via a control plane message in step (D1).[5] We calculate the time taken to send this control message to the UE as $T_{DL}^{Ctrl1}$. The (D1) message contains information for the UE to successfully decode and consume the data. This control plane message tells the UE when exactly it can expect data (*i.e.,* in which *slot* (s)), the data encoding format to decode the data, which *slot* the UE would use to send the ACK/NACK when it has successfully decoded the data, and other related information. The actual data transmission happens at step (D2), and lasts $T_{DL}^{Data}$ long. Finally, in step (D3), the UE sends the ACK/NACK control plane message for the received data. This time lasts $T_{DL}^{Ctrl2}$ long. The total DL time, $T_{DL} = T_{DL}^{Ctrl1} + T_{DL}^{Data} + T_{DL}^{Ctrl2}$ refers to the DL delay during which the gNB schedules DL resources and sends the data to the UE on the common channel.




**(a)** Best Achievable $T_{DL}$.

**(b)** Breakdown of $T_{DL}$ into Control and Data delays.

**(c)** Nrof *slots*.

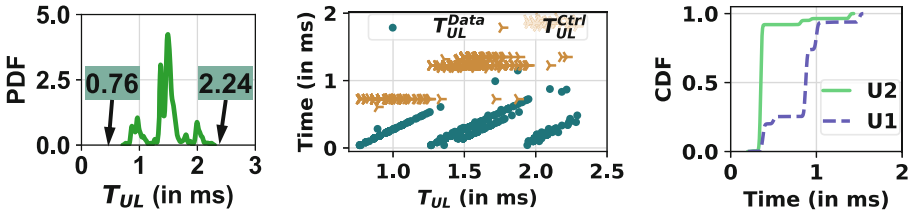
**Fig. 6.** Dissecting the Best Case PHY-layer DL Latency.

**DL Latency Results.** We find that the *best* (*i.e.,* min) DL delay $T_{DL}$ is 0.09 ms, which occurs 1.95% of the time (See Fig. 6a). This implies that (D1), (D2), and (D3) can occur within one *slot* ($\leq$ 0.125 ms), the ***S*** *slot* in DDDSU. However, we can see that $T_{DL}$ has multiple peaks such as 0.17, 0.22, and 0.45 ms. This is due to scheduling the 3 predefined tasks (D1), (D2), and (D3) across *slots* and varying number of OFDM symbols within each *slot* (refer to Fig. 2). For example, when $T_{DL}$ = 0.45 ms, (D1), (D2), and (D3) span 3.6 *slots* (*i.e.,* 0.45 ms $\div$ 0.125 ms). We also find that, more than 50% of the time, the network configures the UE to wait at least 6 *slots* (0.75 ms) before it can send the ACK control message in (D3) (See Fig. 6c). This time includes the processing delay on the UE side.

**Impact of Physical DL Control Overheads.** Figure 6b shows the breakdown of $T_{DL}$ into control and data latency. We can notice that $T_{DL}$ is evenly split between the control, $T_{DL}^{Ctrl}$ and the data, $T_{DL}^{Data}$ delays. This behavior is

[5] This data schematics corresponds to the DCI as shown in Fig. 2.

irrespective of the packet payload size (See Sect. 7.3) and is due to the fact that; 1) Today's mmWave 5G implements same *slot* scheduling, *i.e.,* (D1) and (D3) are in the same *slot* (as shown in Fig. 2) and, 2) the DL control ((D1) and (D3)) and DL data ((D2)) messages occupy two-to-eight and one-to-nine OFDM symbols respectively.




**(a)** Best case $T_{UL}$.

**(b)** Breakdown of $T_{UL}$ into Control and Data Delays.

**(c)** (U1) vs. (U1).

**Fig. 7.** Dissecting the Best Case PHY-layer UL Latency.

### 4.3 Dissecting UL PHY Latency

**UL Transmission.** As shown in Fig. 5, when a UE has data to transmit, it sends a *scheduling request* to procure access to the *busy shared* radio channel as in step (U1), and waits for an explicit *grant* in step (U2). We refer to this combined time as $T_{UL}^{Ctrl}$, which can involve multiple unsuccessful *scheduling request* attempts due to back-offs on the *busy shared* channel. Afterward, the UE prepares and sends the data in step (U3). We refer to this time as $T_{UL}^{Data}$. The total time $T_{UL}$ $= T_{UL}^{Ctrl} + T_{UL}^{Data}$ is the UL delay in the PHY-layer.

**UL Latency Results.** Theoretically, as per the cyclic "fixed" *slots* pattern per radio frame, the lower bound UL *slots* combination is "**UDDDS**" *i.e.,* 0.125 * 5 = 0.625 ms (See Fig. 2). This is because, the UE can request access to the *busy shared* channel ((U1)) in the U *slot*, waits to be *granted* access ((U2)) in one of the three D *slots* (-DDD-), and then sends the UL data ((U3)) in the last S *slot*. In our experiments, we find that the "best" (*i.e.,* min) PHY-layer UL delay is 0.76 ms, which corresponds to the *slots* combination U D D D S U which needed one extra slot than the theoretical bound mentioned above (See Fig. 7a for $T_{UL}$ distribution). We can see multiple peaks in the figure, the percentage of achieving 0.76 ms is 7.46%, for 2.24 ms is 45.496%, and the mean $T_{UL}$ = 1.46 ms. The reason for multiple peaks is two folds: 1) Within a *slot*, the UE may be scheduled varying number of OFDM symbols, and 2) UL scheduling overhead due to back-offs on the busy shared channel as we explain next.

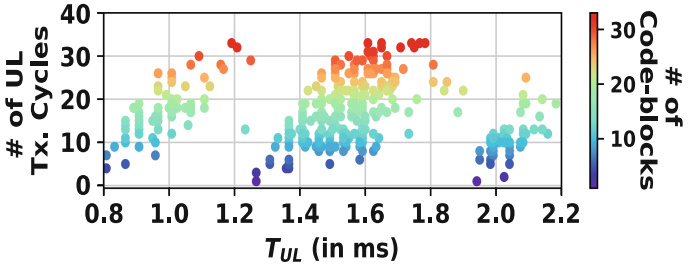



**Fig. 8.** Linear Relation Between the UL Latency ($T_{UL}$) and code-blocks.

**UL Ctrl and Data Latency.** We further break $T_{UL}$ down and characterize the cost on each network communication group, *i.e.,* the control ($T_{UL}^{Ctrl}$) and data ($T_{UL}^{Data}$) overheads. Figure 7b shows that, considering a $T_{UL}$ time of 1.5 ms as an example, the control overhead $T_{UL}^{Ctrl}$ accounts for approximately 81% *i.e.,* 1.7 ms. Simply put, the control overhead ((U1) + (U2)) is responsible for the lion share of the UL delay, unlike the case for DL. This shows that the UL control overhead $T_{UL}^{Ctrl}$ ((U1) + (U2)) takes much longer than data transmission $T_{UL}^{Data}$ ((U3)) in the UL. This is because of two reasons: 1) We find that the UE takes more time waiting to be granted access to the busy shared channel ((U1)) than the actual grant time ((U2)) as shown in Fig. 7c. 2) A single UL transport block gets split into multiple code-blocks [Sect. 6.1.1.1 in [14]] in the UE MAC layer, which are then transmitted on the PHY-layer, and reassembled in the gNB MAC layer. In the "best" case, all the code-blocks are transmitted in one UL transmission cycle (Tx Cycle), as warranted by the allocation of network resources as specified in (U2). We define a Tx Cycle as one round of (U1), (U2), and (U3). However, when the (U2) resource allocated "grant" size is insufficient, each code-block goes through a separate UL Tx Cycles. Thus, a single UL packet can go through multiple *slots* before being completely transmitted on the PHY-layer. Figure 8 shows that $T_{UL}$ increases linearly as the number of code-blocks increases. The number of UL Tx cycles is less than or equal to the number of code-blocks. The jumps in the figure are due to varying the number of OFDM symbols within each *slot*.

***Summary and Implications:*** *In the best-case scenario, PHY-layer latency satisfies the sub-millisecond requirement (≤ 1ms) only 4.43% of the time. It can reach up to 3.08 ms [22.36% of the time]. The average PHY-layer latency is 1.79 ms. These results imply that sub-millisecond PHY-layer transmission is indeed achievable in today's commercial mmWave 5G networks. However, this minimum latency is limited by the UL scheduling in the RAN and is largely dominated by the control overhead. We believe that our results provide two incentives for enhancements or perhaps protocol re-design; 1) Implementing and adapting all 61 proposed slots scheduling interval configurations as per 3GPP* [Sect. 7.3.1 in [11]], *and dynamically adapting specific slot patterns for UL and DL heavy transmissions for different use cases will further reduce this latency. 2) For*

*UL-centric apps with heavy UL traffic demands like AR, the cyclic fixed slot configuration means that, the network is not aware of the UE-side heavy traffic demands. Therefore, we claim that, offloading some UL functions to the UE will help cap the lion share control plane overhead and further reduce latency. For example, introducing a mechanism by which a UE can signal heavy UL traffic to the network and request a UL specific slot configuration or implementing a true cross-layer signaling mechanism to anticipate and signal specific application PHY-layer latency requirements could be ways to achieve this. This might also help address variations (or instabilities) in latency, although these instabilities are largely due to channel conditions (see below).*

## 5 Impact of Channel Conditions

Taking into account the invisibility of the network side information, we use CQI in the UL to study the impact of PHY-layer radio conditions on latency. Recall from Sect. 4 that MCS determines the number of useful bits transmitted per *slot*. A lower MCS leads to more redundant bits and fewer useful bits transmitted per *slot*, and vice versa. Fig. 9 shows the impact of CQI on MCS. On one hand, when the UE reports a high CQI value, which implies good channel conditions, the network generally selects a high MCS to be used for data encoding. On the other hand, Fig. 10 shows that a lower CQI value results in corrupted data, which leads to more ReTxs on the PHY-layer captured by the BLER. These ReTxs are transparent to the application layer, but can further increase the E2E RTT. Therefore, we quantify the impact of CQI and ReTxs on the PHY-layer latency, and further explain its impact on the E2E application perceived latency.

**Methodology:** Previous studies have shown the impact of HO on E2E RTT and have found that HO patches[6] occur in well-defined areas around 5G towers [30]. We leverage these findings to improve the credibility of our results by minimizing the number of HOs during our experiments: First, we conduct repeated experiments to identify the HO patches around our chosen areas. Second, we conduct controlled LoS walking experiments and do not walk beyond identified potential

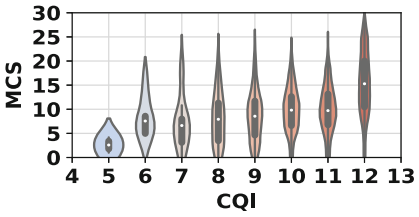



**Fig. 9.** Impact of CQI on Modulation Coding Scheme (MCS).

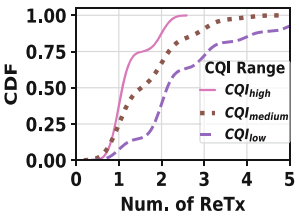



**Fig. 10.** Impact of CQI on Number of ReTxs.

[6] Defined as the area between two 5G towers A and B where HO occurs from tower A to B or vice versa.

HO areas. Third, despite these measures to ensure no HO, we still observe and discard experiments with any HO occurrences. As a way to quantify the impact of the CQI on latency, we divide the CQI values into $CQI_{low}$ = (6, 9], $CQI_{medium}$ = (9, 12], and $CQI_{high}$ = (12, 15], and refer to it as such hereafter. Note that even when the UE is in $CQI_{high}$, the CQI value can still change slightly between 12+ and 15, and ReTxs may occur. Thus, during our experiments, we fix the CQI range, keep all other factors constant, and investigate the impact of slight CQI changes on the PHY-layer latency.

### 5.1 Understanding the Impact of ReTxs on $T_{Phy}$

**A single ReTx can Defeat the 1 ms PHY-Layer Delay:** Previously, we showed that sub-millisecond $T_{Phy}$ is indeed achievable in the best case scenario, *i.e.,* $T_{Phy}$ = 0.85 ms (Sect. 4.1). However, Fig. 11 shows that when exactly one ReTx occurs (Num. ReTx = 1), the best case (*i.e.,* min) $T_{Phy}$ is 1.33 ms and about 2.27% of the PING packets experience only 1 ReTx. We find that, the network "NACKs" corrupted received data (*i.e.,* undelivered (U3) message) by implicitly granting the UE access to the radio channel (*i.e.,*(U2)) without an explicit channel request from the UE (*i.e.,*(U1)). Practically, an example of such interaction can be: Assume the UE sends the initial corrupted data in the "U" *slot* of the previous schedule interval (*i.e.,* "**DDDSU** — DDDSU"). It has to wait and receive the implicit grant in one of the three "D" *slots* of the next scheduled interval (*i.e.,* "DDDSU — **DDD**SU") and re-transmits the data in the "S" *slot* (*i.e.,* "DDDSU — DDD**S**U"). Theoretically, this will incur an additional lower bound overhead of 0.375 ms (*i.e.,* 0.125 ms x 3 (*slots*)). Therefore, $T_{Phy}$ = 0.85 ms + 0.375 ms = 1.225 ms. However, our experiments show that the actual PHY-layer delay with one ReTxs is 1.33 ms, 0.105 ms higher than the theoretical, but lower than one *slot* (0.125 ms).

**Characterizing the Cost of ReTxs:** To characterize the cost of ReTxs in UL and DL, we plot the latency for different numbers of ReTxs. Figure 12 shows that, unlike DL transmissions, ReTxs have a significant impact on UL transmissions due to the same theoretical analysis as explained above. Furthermore, Fig. 13

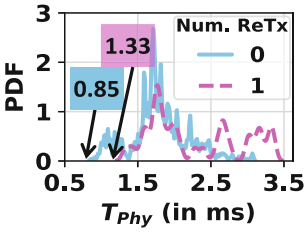



**Fig. 11.** 1ms $T_{Phy}$ is defeated with one ReTx.

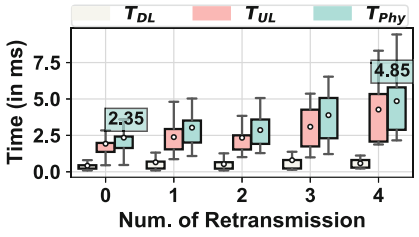



**Fig. 12.** Impact of Retransmissions on $T_{DL}$, $T_{UL}$, and $T_{Phy}$.

shows the impact of ReTxs on $T_{Phy}^{Ctrl}$ and $T_{Phy}^{Data}$. We find that, as the number of ReTxs increases, $T_{Phy}^{Data}$ increases much faster than $T_{Phy}^{Ctrl}$; slope of line $l3$ $m_{Phy}^{Data} = 0.2072$, slope of $l1$ $m_{Phy}^{Ctrl} = 0.0219$. Hence, $T_{Phy}^{Data}$ grows at $\approx$ 9.5x the rate of $T_{Phy}^{Ctrl}$ when the number of ReTxs increases. More specifically, $T_{Phy}^{Ctrl}$'s dominance in $T_{Phy}$ (as shown in Sect. 4.3) decreases significantly from 79.2% to 60.1% then to 45.9% when the number of ReTxs increases from 0 to 3 to 6, respectively. This is due to two reasons; 1) for the control overhead: implicit "NACKs" from the gNB eliminates (U1) from subsequent ReTxs and (U2) $\ll$ (U1) (See Fig. 7c), and 2) for the data overhead: we find that (U3) usually takes between 0.0625 ms to 0.125 ms, (U2) takes on average 0.018 ms. Hence, $T_{Phy}^{Data}$ ((U3)) overhead increases by [3.5X, 7X] faster than $T_{Phy}^{Ctrl}$ ((U2)).

### 5.2 Impact of CQI on $T_{UL}$ and $T_{Phy}$

We study the impact of $CQI_{low}$, $CQI_{medium}$ and $CQI_{high}$ with a fixed number of ReTxs. Fig. 14 shows that when there is no ReTxs, there is at least an additional 2 ms overhead on $T_{Phy}$ with poor channel conditions (*i.e.,* CQI changes from $CQI_{high}$ to $CQI_{low}$). A similar conclusion is observed for $T_{UL}$ (See Fig. 15). This overhead is due to a lower MCS when the CQI drops to $CQI_{low}$. This will cause a decrease in the code rate *i.e.,* less useful bits are transmitted per *slot*, resulting in more time to transmit an entire transport block. The impact of CQI on $T_{DL}$ is rather insignificant.

***Summary and Implications:*** *The HARQ process is primarily used to speed up ReTxs. The sender stores all transmitted data in its buffer and discards them only after receiving an ACK from the receiver. The receiver also stores all erroneous packets and uses them to improve decoding* [Sect. 5.4.2 in [13]]. *This may cause unavoidable latency overhead, particularly when the channel conditions change very suddenly from $CQI_{high}$ to $CQI_{low}$. This is because UL data transmission that is encoded with an MCS value suitable for the current reported CQI value may not be suitable at a later time when there is a ReTx and the CQI value drops. This can result in more ReTxs and higher latency, which explains*

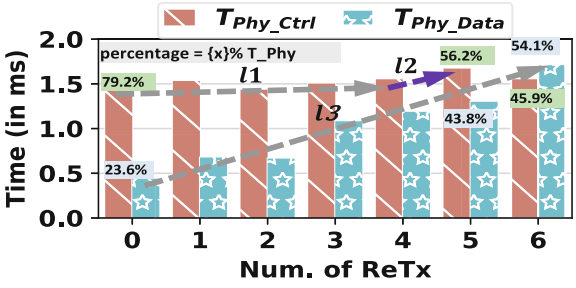



**Fig. 13.** Impact of ReTxs on $T_{Phy}^{Ctrl}$ and $T_{Phy}^{Data}$.

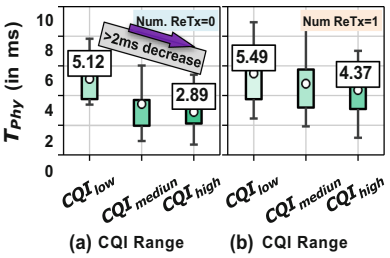



**Fig. 14.** Impact of CQI and Num. ReTx on $T_{Phy}$. We consider Num. ReTx=0 and Num. ReTx=1.

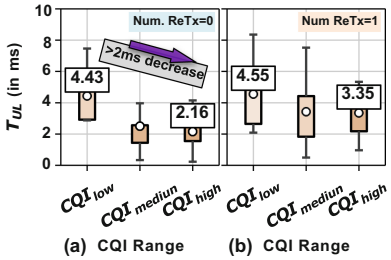



**Fig. 15.** Impact of CQI and Num. ReTx on $T_{UL}$. We consider Num. ReTx=0 and Num. ReTx=1.

*why exactly one ReTx with PHY-layer latency 1.33 ms occurs about 2.27% of time. Given this, we conclude that improving the HARQ process to account for CQI to MCS mismatch, especially when channel conditions drop, can provide a remedy and perhaps eliminate the additional overhead due to more ReTxs. In the practical sense, this calls for an extensive re-design of mmWave PHY-layer operations.*

## 6 Impact of Mobility

In this section, we address two key questions: First, what is the additional PHY-layer overhead due to UE-side activity (*i.e.,* mobility) in mmWave 5G? and second, how does mobility influence the PHY-layer latency in UL and DL?

**Methodology:** Similar to our experimental setup in Sect. 5, we minimize HOs and conduct clear LoS walking experiments and do not walk beyond identified potential HO patches. We study the best case *i.e.,* the UE is in $CQI_{high}$ with slight CQI fluctuations and no ReTxs.

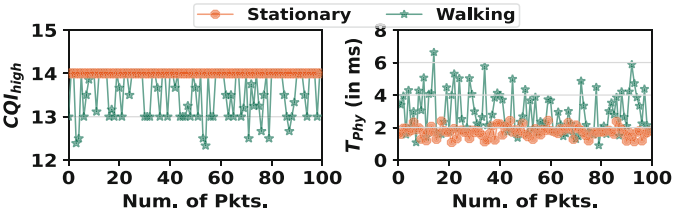



**Fig. 16.** Variability in $T_{Phy}$ caused by mobility when UE is in $CQI_{high}$ and no ReTxs.

### 6.1 Impact of Mobility (No HOs) on $T_{Phy}$

Mobility causes rapid signal quality fluctuations in mmWave which has a direct impact on $T_{Phy}$. In Fig. 16, the left Fig. shows the CQI fluctuations when

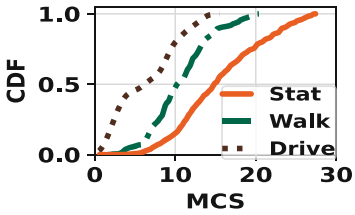



**Fig. 17.** Impact of mobility on MCS.

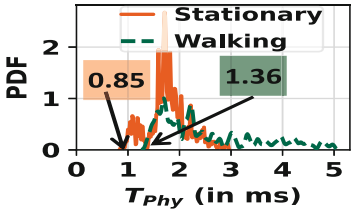



**Fig. 18.** Impact of mobility on $T_{Phy}$.

the UE is in $CQI_{high}$ while walking and stationary and the Fig. on the right shows $T_{Phy}$ while walking and stationary. We see that, even in $CQI_{high}$, the CQI values fluctuates frequently when the UE is walking. This is because, as shown in Fig. 17, the network adopts a lower MCS values during mobility as a way to minimize the number of ReTxs and meet the target BLER rate of <10% [Table 8.1.1-1 in [12]]. However, adopting lower MCS increases the best case (*i.e.,* min) $T_{Phy}$ from 0.85 ms to 1.36 ms between stationary and walking, respectively, shown in Fig. 18. A difference of 0.51 ms, about 5 *slots*.

### 6.2 Quantifying the Impact of HOs on $T_{Phy}$

We aim to quantify the minimum PHY-layer latency overhead due to HOs, $T^{HO}_{Phy}$. Unlike the previous section, which focuses on mobility without HOs, we now study the impact of HO on PHY-layer. We conduct walking and driving experiments, ensuring that we move across 5G towers to trigger HOs. We find that the minimum additional latency overhead due to HO from one 5G tower to another 5G tower (5G → 5G HO) is 2 ms, which corresponds to 16 *slots* (See Fig. 19). Additionally, we see that when the user is driving, approximately ≥ 50% of $T^{HO}_{Phy}$ takes at least 3 ms compared to 2 ms while walking.

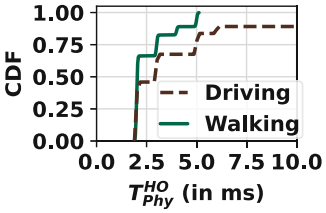



**Fig. 19.** Impact of 5G–>5G HOs on $T_{Phy}$ during walking vs. Driving.

**Summary and Implications:** *Although, the effect of mobility causes fast and frequent instability in PHY-layer latency which are problematic for latency-critical applications like AR/VR, we argue that, it can be avoided to some extent. Here, we discuss two cases. Case 1: The additional latency due to mobility can be minimized from 1) the UE side by actively sensing and predicting blockage* [21] *and/or requesting more slots when blockage is unavoidable. The later approach requires more investigation and has not yet being studied. 2) from the network side by taking into account the UE-side contextual factors and/or upper layer Quality of service (QoS) when making scheduling decisions. Practically, this might require leveraging signalling messages, camera data, and cross-layer communication to*

*develop mobility-aware applications. Case 2: Given mmWave's directional propagation and high sensitivity to obstruction, the additional delay is not avoidable in few cases. For instance, when the obstruction is due to factors beyond the control of UE or network e.g., moving vehicles, people and tall building etc. Dense mmWave cell tower deployments can help in this case, however, such deployments are costly and may not be the first choice for commercial carriers.*

## 7 E2E Application Latency

Here we break the E2E delay into the 5G RAN, including $T_{Phy}$ and the 5G Core + Internet latency, and study the impact of the PHY-layer on the E2E delay. We aim to understand: 1) the role of server placement on E2E delay, 2) how the UE sleep cycle (*i.e.,* CDRX) incurs additional delay?, and lastly, 3) what impact do various packet payload sizes have on the PHY-layer and E2E latency?

**Methodology:** We deploy three VMs, each running on AWS WL, LZ, and RG edge nodes. We have verified these VMs placement relative to a UE in our two chosen locations by conducting a simple PING and traceroute experiment over mmWave 5G. The traceroute experiment reveals that, the UE is 8, 19, and 22 hops away from the WL, LZ, and RG servers respectively. A geolocation PING shows the WL and LZ in the same region as the UE. We conduct stationary clear LoS experiments, *i.e.,* the UE is in $CQI_{high}$ with no ReTxs as follows: Three UEs send PING echo requests to the three VMs at various PING intervals (5, 8, 10, and 15 ms) using varying PING payload sizes (*i.e.,* 32, 100, 400, 900, and 1200 bytes). We enforce the reception of the PING echo reply from the server before consecutive PING echo requests. This lets us dissect the E2E delay by isolating each PING and studying the UE sleep cycle timers. We adopt varying PING payloads to mimic different application traffic patterns.

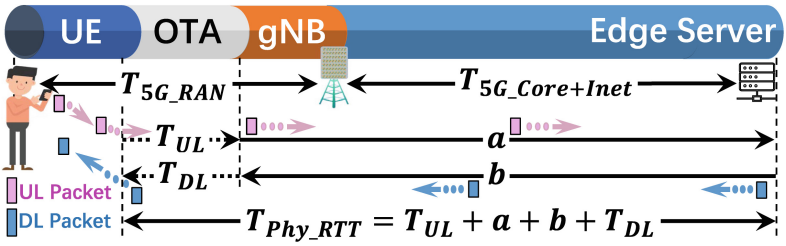



**Fig. 20.** Dissecting the E2E RTT into $T_{5G\_RAN}$ and $T_{5G\_Core+Inet}$.

### 7.1 Role of Server Placement

**Dissecting the E2E Application Perceived Latency.** We divide the E2E RTT delay into two components: i) The 5G RAN delay, $T_{5G\_RAN}$ defined as, the packet time spent on the PHY-layer including the processing time by the

5G RAN upper layers in the UE and ii) the 5G Core + Internet delay, *i.e.*, $T_{5G\_Core+Inet}$ defined as the time from when the UE sends the PING echo request in (U3) to when it receives the PING echo reply from the edge server on the PHY-layer in (D1). Therefore, $T_{E2E\_RTT} = T_{5G\_RAN} + T_{5G\_Core+Inet}$ (See Fig. 20). To divide the E2E RTT, we compute $T_{Phy\_RTT}$, the physical layer RTT including $T_{5G\_Core+Inet}$ as shown in Fig. 20. Then, $T_{5G\_Core+Inet} = T_{Phy\_RTT} - (T_{UL} + T_{DL})$. From $T_{5G\_Core+Inet}$, we calculate $T_{5G\_RAN} = T_{E2E\_RTT}$- $T_{5G\_Core+Inet}$.

**Results.** As shown in Fig. 21 and Table 2, the 5G RAN delay takes on average 7.32 ms regardless of the server location. However, as the distance between the UE and the server increases, $T_{5G\_Core+Inet}$ increases dramatically to be 10 ms, 30 ms, and 35 ms (on average) across the WL, LZ, and RG servers, respectively. This signifies the importance of edge server placement on RTT. Next, we demonstrate the benefit of deploying applications on the WL, and setbacks of deploying applications on the LZ and RG servers w.r.t. a UE location.

**Table 2.** E2E RTT Delay Breakdown Across Edge Servers

| Delay Components ⇒<br>⇓ Edge Server | $T_{E2E\_RTT}$<br>Mean ±**std. dev.** | $T_{5G\_RAN}$<br>Mean ±**std. dev.** | $T_{5G\_Core+Inet}$<br>Mean ±**std. dev.** |
|---|---|---|---|
| WL | 17.27 ms±**1.31 ms** | **7.02 ms** ±**3.77 ms** | 10.25 ms±**3.84 ms** |
| LZ | 38.15 ms±**1.83 ms** | **7.34 ms** ±**4.25 ms** | 30.81 ms±**4.41 ms** |
| RG | 44.08 ms±**3.04 ms** | **7.60 ms** ±**4.72 ms** | 35.82 ms±**4.55 ms** |

### 7.2 Impact of CDRX on $T_{Phy}$

mmWave 5G makes use of CDRX to achieve UE power management for efficient energy consumption and to synchronize UE wake-up timing with DL data transmission [33]. While in RRC_Connected state, the gNB configures the UE to go through active and sleep cycles. The UE CDRX behavior is determined using several timers, which we explain below. We serendipitously employ the CDRX cycles to estimate and bound the "wired" part (between the gNB and the edge server) of the E2E latency.

**CDRX Sleep Timers.** The CDRX cycle is controlled by the CDRX_ON and the CDRX_OFF timers – The CDRX_ON timer determines how long the UE will stay ON and the CDRX_OFF timer dictates the duration the UE will stay OFF. The CDRX ON/OFF duration cycles may be extended further on the basis of the CDRX_Inactivity timer. The CDRX_Inactivity timer determines how long the UE MUST stay ON upon reception of access to the busy shared channel (*i.e.*,(U2)), which will further extend the duration of the UE ON [1]. We observe that, both VZW and AT&T configure the CDRX _ON and CDRX _Inactivity duration as 8 ms and 30 ms, respectively.

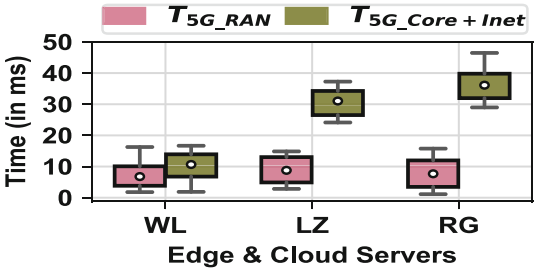



**Fig. 21.** Impact of Server Placement on E2E RTT Delay Breakdown.

**Delay Due to CDRX.** Since the CDRX_Inactivity timer starts when the UE acquires access to the busy shared channel, we therefore compute $T_{CDRX\_Overhead}$ = $T_{PhyRTT\_CDRX}$ - 30 (CDRX_Inactivity duration), where $T_{PhyRTT\_CDRX}$ is the time between when a UE acquires access to the busy shared channel (U2) and receives the DCI which indicates an echo PING reply on the PHY-layer (D1), *i.e.,* the time from U2—>edge server—>D1 in Fig. 5. We find that, in the WL case, the UE will never go to sleep before receiving the echo PING reply from the server. This is because, in the WL, $T_{PhyRTT\_CDRX}$ << 30ms (CDRX_Inactivity). However, in the LZ and RG cases, the UE goes into sleep mode (CDRX_OFF) about 60% and 97% of the time respectively before receiving the PING echo reply (See Fig. 22a). We show a detailed illustration of this behavior for each server in Fig. 23 by showing the arrival time for three sample PING echo replies w.r.t. the UE status CDRX ON/OFF. We further compute $T_{CDRX\_Overhead}$, the additional time taken before the network sends the PING echo reply to the UE when the UE is asleep (CDRX_OFF) because the CDRX_Inactivity timer has expired. We find that, $T_{CDRX\_Overhead}$ = 6.4 ms (on average) (See Fig. 22b).

### 7.3 Impact of Packet Payload Size

By varying the PING payload size, we can understand how the amount of data sent and received affects PHY-layer latency and E2E RTT. We find that the payload size has little to no impact on $T_{Phy}$ (See Fig. 24). We notice that, when the payload size increases, the network may schedule multiple HARQ processes that work simultaneously to carry the UE data during specific *slots*. The number of scheduled HARQ processes is sent to the UE in U2. The UE then uses the assigned processes during scheduled UL *slot*. Simply put, when the number of HARQ processes increases, more bytes can be sent in the same *slot* without increasing the latency. We observe a maximum of 16 HARQ processes in VZW mmWave 5G, which conforms with 3GPP's specification [Sect. 5.4.2 in [13]]. This mmWave 5G design has little to no impact on the control overhead, $T_{Phy}^{Ctrl}$, as only one U1 message is needed to report the UE buffer status when the data size increases. This will also have little to no impact on $T_{Phy}^{Data}$. However, we observe

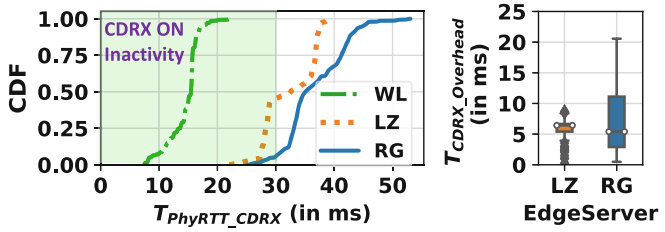



**Fig. 22.** Impact of CDRX and server placement on PHY-layer. a) [1] In the WL case, the UE will NEVER go to sleep. [2] In the LZ case, the UE goes to sleep 60% of the time, while [3] in the RG case, the UE will go to sleep 97% of the time. b) Additional 6.4 ms delay (on average) overhead due to CDRX.

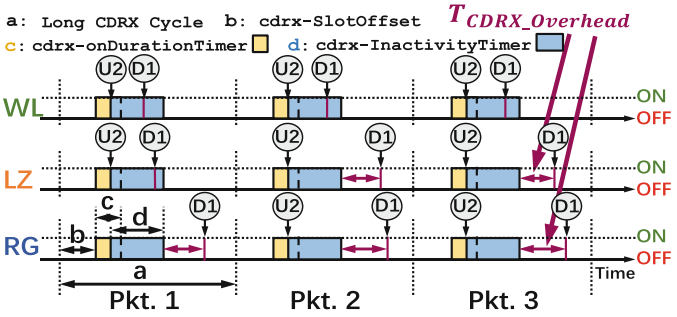



**Fig. 23.** Detailed illustration of how the CDRX and server placement impact the E2E RTT. Sever placement causes an additional delay due to CDRX, $T_{CDRX\_Overhead}$ in the LZ and RG edger server.

an insignificant increase in the E2E RTT (See Fig. 25). This is because both the UE and the gNB will take more time to reassemble the data chunks from all processes before forwarding it to the RAN upper layer for processing.

***Summary and Implications.*** *Although the role of CDRX in the management of UE power is paramount [24], our experiments show that there is a trade-off with the E2E latency in the LZ and RG edge nodes. Without devaluing the CDRX benefits, our experiments reveal that, the additional overhead due to CDRX (i.e., $T_{PhyRTT\_CDRX}$ = 6.4 ms) is primarily due to the network side CDRX sleep timer configurations. We claim that adopting dynamic context-aware CDRX timer configuration may significantly reduce or perhaps even eliminate the latency effect due to CDRX especially in far edge nodes. For example, increasing the CDRX_Inactivity timer from 30 to 35 or 40 ms can potentially reduce the perceived latency of the E2E application by 6.4 ms on average. Additionally, it will be beneficial to customers with limited monetary resources as deploying applications on the closest edges, such as the WL node, is very expensive [45]. However, achieving this context-aware CDRX timer configurations requires a truly 5G NR*

*cross-layer design which perhaps calls for a protocol redesign. This approach is particularly difficult and have not yet been studied in the literature.*

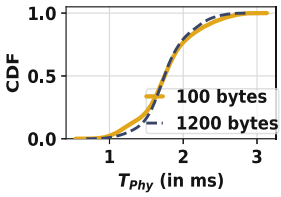



**Fig. 24.** Impact of Payload size on $T_{Phy}$.

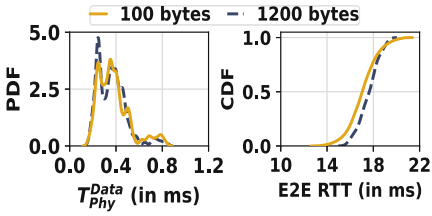



**Fig. 25.** Payload size has little to no impact on $T_{Phy}^{Data}$ and E2E RTT.

## 8 Related Work

We discuss the related work in two categories: **Commercial 5G Network Measurements.** Researchers have conducted several studies on commercial 5G networks since their debut in 2019. Among them, Narayanan *et al.* examines for the first time the performance of mmWave 5G on smartphones [29]. The same team also investigates 5G performance prediction [30], application QoE, and device power consumption [33]. Xu *et al.* study the coverage, performance, and energy consumption of sub-6Ghz 5G in China [44]. Rochman *et al.* compare 5G deployment in Chicago and Miami [37]. Rischke *et al.* measure 5G campus networks [36]. Pan and Claudio *et al.* examine the 5G performance on high-speed trains and in public bus transit systems respectively [22,34]. Compared to all the above studies, our work focuses on the latency of 5G networks in the context of 5G last-mile latency support for edge computing [7]– an important but under explored topic.

**5G Physical Layer.** There are a plethora of works on the PHY-layer foundations of 5G, including mmWave [40,42], signal propagation [41,43], beam forming [18,38], and massive MIMO [39,46], to name a few. Compared to the above works that solely tackle the E2E latency, [19,23,24,28,29,33,44] also quantify the PHY-layer UL and DL latency separately, but not both from different points of view. Almost in line with our work, Xu *et al.* quantify the latency of 5G mmWave PHY-layer in China to be 2.19±0.36 ms [44]. However, they do not state or show whether <1ms PHY-layer latency is achievable with today's mmWave 5G NR deployments. Additionally, factors that can further increase PHY-layer latency were not explored. Thus, to our knowledge, our paper is the first to systematically study and quantify the impact of several factors on PHY-layer latency, and the impact of server placement and CDRX on E2E latency. We are also the first to answer the question "Is sub-millisecond PHY-layer latency feasible with today's commercial 5G". Additionally, our paper provides insights to network operators to capitalize on which other related works lack on.

## 9 Discussion and Future Work

Throughout this study, we took a careful approach to quantify the impact of each factor in today's mmWave PHY-layer latency. In each section, we controlled (as best as we could) one factor at a time and carefully designed experiments to study the factor under investigation. Our approach to quantify the additional overhead per factor "in its best case scenario" revealed that, although sub-millisecond PHY-layer transmission is indeed possible in today's mmWave 5G, any slight change in each factor certainly defeats the sub-millisecond promise of mmWave 5G and the combined impact of all factors leads to a wide variability in the E2E RTT perceived by the applications. Thus, the main message is that current 5G wireless radio technology still has a long way to go to be able to achieve sub-millisecond latency.

Our results also highlighted several aspects for 5G cellular carriers to consider in order to overcome this poor latency performance such as: i) implementing and adapting all 61 proposed *slots* scheduling interval configurations as per 3GPP standards, ii) dynamically adapting specific *slot* patterns for UL and DL heavy transmissions for different use cases, and iii) improving the HARQ process to account for CQI to MCS mismatch, especially when the channel conditions drop.

However, we believe that implementing a true cross-layer designed is called for to further improve the latency performance. This cross-layer design can allow the anticipation and signaling of specific application PHY-layer latency requirements to be adapted accordingly by carriers such as: i) using a dynamic *slot* configuration based on the application traffic demand instead of a fixed configuration, ii) requesting more *slots* when blockage is sensed and predicted by the UE side, and iii) adopting dynamic context-aware CDRX timer configuration when applications are deployed on far edge nodes from the UE.

Our study is limited to PING packets and today's mmWave 5G NSA deployments. We believe that future deployments of SA will most likely use the current 5G radio access network technologies. With that assumption, 5G SA deployments might reduce the 5G Core + Internet latency, but may not affect the 5G RAN. Since our study is focused on the radio side, we believe the insights of this work reveals that the physical layer's impact on latency will still unfortunately be present in future mmWave SA 5G Deployments. Our work also sheds the light on several research directions to explore including the impact of additional factors such as: i) application traffic patterns on 5G latency and ii) the number of users within the communication range of one 5G base station or across multiple 5G base stations (given cellular carrier collaboration).

## 10 Conclusion

Using a commercial 5G tool to extract detailed physical channel events and messages, this study presents a first-of-a-kind comprehensive in-depth measurement study of mmWave 5G latency performance on the PHY-layer. Our findings show that the current 5G RAN-induced latency is limited by both UL scheduling and

carrier configurations. To summarize Today's mmWave status quo latency: (1) In the best case scenario, the best achievable mmWave PHY-layer latency is around 0.85 ms. (2) changing any factor affecting this best case scenario, even slightly, leads the PHY-layer latency to be more than 1 ms. (3) These factors combined with the Internet (buffering and congestion) result in a wide variability in the E2E RTT perceived by the applications. (4) Finally, our study and analysis of PHY-layer latency suggest that 5G NR is indeed capable of delivering (sub)ms latency performance. However, due to inefficiencies at the 5G NR sub-layers (combined with the network stack and above), these low-latency benefits are not reflected at the application layer.

**Acknowledgements.** This research was supported in part by NSF under Grants CNS-1901103, CNS-1915122, CNS-2038559, CNS-21544078, CNS-2128489, CNS-2220286, CCF-2212318 and CNS-2220292 as well as a Cisco Research Award and an InterDigital gift.

## References

1. 5G NR: Connected Mode DRX. https://howltestuffworks.blogspot.com/2021/04/5g-nr-connected-mode-drx.html. Accessed Nov 2022
2. Amazon web services (aws). https://aws.amazon.com/
3. 5G; study on scenarios and requirements for next generation access technologies (3gpp tr 38.913 version 14.3.0 release 14) (2017). https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/15.00.00_60/tr_138913v150000p.pdf
4. https://www.gsma.com/futurenetworks/wiki/cloud-ar-vr-whitepaper/ (2019)
5. 5G SA vs 5G NSA: What are the differences? https://www.alepo.com/5g-sa-vs-5g-nsa-what-are-the-differences/ (2022). Accessed Nov 2022
6. Accuver XCAL. https://www.accuver.com/sub/products/view.php?idx=6&ckattempt=2 (2022). Accessed Nov 2022
7. AWS Wavelength. https://aws.amazon.com/wavelength/ (2022). Accessed Nov 2022
8. Samsung galaxy S21 5G featuring a Qualcomm snapdragon 888 5G mobile platform. https://www.qualcomm.com/snapdragon/device-finder/samsung-galaxy-s21-5g (2022). Accessed Nov 2022
9. Speedtest by Ookla. https://www.speedtest.net/ (2022). Accessed Nov 2022
10. T-Mobile hits 3 Gbps 5G speeds without mmWave in world record production test. https://9to5mac.com/2022/06/14/t-mobile-3-gbps-5g-speeds/ (2022). Accessed Nov 2022
11. 3GPP: 5G; NR; Multiplexing and channel coding (3GPP TS 38.212 version 15.2.0 Release 15) (2018). https://www.etsi.org/deliver/etsi_ts/138200_138299/138212/15.02.00_60/ts_138212v150200p.pdf. Accessed Nov 2022
12. 3GPP: 5G; NR; Requirements for support of radio resource management (3GPP TS 38.133 version 15.3.0 Release 15) (2018). https://www.etsi.org/deliver/etsi_ts/138100_138199/138133/15.03.00_60/ts_138133v150300p.pdf. Accessed Nov 2022
13. 3GPP: 5G NR: Medium Access Control (MAC) protocol specification (3GPP TS 38.321 version 15.5.0 Release 15) (2019–05). https://www.etsi.org/deliver/etsi_ts/138300_138399/138321/15.05.00_60/ts_138321v150500p.pdf. Accessed Nov 2022

14. 3GPP: 5G; NR; Physical layer procedures for data (3GPP TS 38.214 version 16.2.0 Release 16). https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/16.02.00_60/ts_138214v160200p.pdf (2020). Accessed Nov 2022
15. 3GPP: 5G; NR; Radio Resource Control (RRC); Protocol specification (3GPP TS 38.331 version 16.2.0 Release 16) (2020). https://www.etsi.org/deliver/etsi_ts/138300_138399/138331/16.02.00_60/ts_138331v160200p.pdf. Accessed Nov 2022
16. 3GPP: 5G; NR; Radio Link Control (RLC) protocol specification (3GPP TS 38.322 version 16.2.0 Release 16) (2021). https://www.etsi.org/deliver/etsi_ts/138300_138399/138322/16.02.00_60/ts_138322v160200p.pdf. Accessed Nov 2022
17. Admin, G.: News & events (2017). https://www.3gpp.org/news-events/3gpp-news/sa1-5g
18. Ahmed, I., et al.: A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives. IEEE Commun. Surv. Tutorials **20**(4), 3060–3097 (2018)
19. Corneo, L., Eder, M., Mohan, N., Zavodovski, A., BayhanZ, S.: Surrounded by the clouds. In: The Web Conference (2021)
20. Dinh, P., Ghoshal, M., Koutsonikolas, D., Widmer, J.: Demystifying resource allocation policies in operational 5G mmwave networks. In: 2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–10 (2022). https://doi.org/10.1109/WoWMoM54355.2022.00016
21. Fang, Z., Wang, G., Xie, X., Zhang, F., Zhang, D.: Urban map inference by pervasive vehicular sensing systems with complementary mobility. Proceed. ACM Inter. Mobile Wearable Ubiquit. Technol. **5**(1), 1–24 (2021)
22. Fiandrino, C., Juárez Martínez-Villanueva, D., Widmer, J.: Uncovering 5G performance on public transit systems with an app-based measurement study. In: Proceedings of the 25th International ACM Conference on Modeling Analysis and Simulation of Wireless and Mobile Systems, pp. 65–73 (2022)
23. Ghoshal, M., et al.: An in-depth study of uplink performance of 5g mmWave networks, pp. 29–35. 5G-MeMU 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3538394.3546042
24. Hassan, A., et al.: Vivisecting mobility management in 5G cellular networks. In: Proceedings of the ACM SIGCOMM 2022 Conference, pp. 86–100. SIGCOMM 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3544216.3544217
25. Hassan, A., et al.: Vivisecting mobility management in 5G cellular networks. In: Proceedings of the ACM SIGCOMM 2022 Conference. pp. 86–100. SIGCOMM 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3544216.3544217
26. Li, Y., et al.: Experience: a five-year retrospective of mobileInsight. In: Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, pp. 28–41 (2021)
27. McLaughlin, R.: 5G low latency requirements (2021). https://broadbandlibrary.com/5g-low-latency-requirements/
28. Mohan, N., Corneo, L., Zavodovski, A., Bayhan, S., Wong, W., Kangasharju, J.: Pruning edge research with latency shears. In: Proceedings of the 19th ACM Workshop on Hot Topics in Networks, pp. 182–189 (2020)
29. Narayanan, A., et al.: A first look at commercial 5G performance on smartphones. In: Proceedings of The Web Conference 2020, pp. 894–905 (2020)

30. Narayanan, A., et al.: Lumos5G: mapping and predicting commercial mmWave 5G throughput. In: Proceedings of the ACM Internet Measurement Conference, pp. 176–193. IMC 2020, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3419394.3423629
31. Narayanan, A., Ramadan, E., Quant, J., Ji, P., Qian, F., Zhang, Z.L.: 5G tracker: a crowdsourced platform to enable research using commercial 5G services. In: Proceedings of the SIGCOMM2020 Poster and Demo Sessions, pp. 65–67 (2020)
32. Narayanan, A., et al.: A comparative measurement study of commercial 5G mmWave deployments. In: IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, pp. 800–809 (2022). https://doi.org/10.1109/INFOCOM48880.2022.9796693
33. Narayanan, A., et al.: A variegated look at 5g in the wild: performance, power, and qoe implications. In: Proceedings of the 2021 ACM SIGCOMM 2021 Conference, pp. 610–625. SIGCOMM 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3452296.3472923
34. Pan, Y., Li, R., Xu, C.: The first 5G-LTE comparative study in extreme mobility. Proceed. ACM Measure. Anal. Comput. Systems **6**(1), 1–22 (2022)
35. Ramadan, E., Narayanan, A., Dayalan, U.K., Fezeu, R.A., Qian, F., Zhang, Z.L.: Case for 5G-aware video streaming applications. In: Proceedings of the 1st Workshop on 5G Measurements, Modeling, and Use Cases, pp. 27–34 (2021)
36. Rischke, J., Sossalla, P., Itting, S., Fitzek, F.H., Reisslein, M.: 5G campus networks: a first measurement study. IEEE Access **9**, 121786–121803 (2021)
37. Rochman, M.I., et al.: A comparison study of cellular deployments in Chicago and Miami using apps on smartphones. In: Proceedings of the 15th ACM Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization, pp. 61–68 (2022)
38. Roh, W., et al.: Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results. IEEE Commun. Mag. **52**(2), 106–113 (2014)
39. Shepard, C., Blum, J., Guerra, R.E., Doost-Mohammady, R., Zhong, L.: Design and implementation of scalable massive-Mimo networks. In: Proceedings of the 1st International Workshop on Open Software Defined Wireless Networks, pp. 7–13 (2020)
40. Singh, V., Mondal, S., Gadre, A., Srivastava, M., Paramesh, J., Kumar, S.: Millimeter-wave full duplex radios. In: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, pp. 1–14 (2020)
41. Solomitckii, D., Orsino, A., Andreev, S., Koucheryavy, Y., Valkama, M.: Characterization of mmWave channel properties at 28 and 60 GHZ in factory automation deployments. In: 2018 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6. IEEE (2018)
42. Sur, S., Pefkianakis, I., Zhang, X., Kim, K.H.: Towards scalable and ubiquitous millimeter-wave wireless networks. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, pp. 257–271 (2018)
43. Sur, S., Venkateswaran, V., Zhang, X., Ramanathan, P.: 60 GHZ indoor networking through flexible beams: a link-level profiling. In: Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 71–84 (2015)

44. Xu, D., et al.: 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, pp. 479–494 (2020)
45. Xu, M., et al.: From cloud to edge: a first look at public edge platforms, pp. 37–53. IMC 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3487552.3487815. https://doi-org.ezp1.lib.umn.edu/10.1145/3487552.3487815
46. Zhao, R., Woodford, T., Wei, T., Qian, K., Zhang, X.: M-cube: a millimeter-wave massive mimo software radio. In: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, pp. 1–14 (2020)