DETERMINISTIC6G

Latency measurement data and characterization of RAN latency from experimental trials

The DETERMINISTIC6G project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 1010965604.

Latency measurement data and characterization of RAN latency from experimental trials

Grant agreement number: Project title: Project acronym: Project website: Programme:	101096504 Deterministic E2E communication with 6G DETERMINISTIC6G Deterministic6g.eu EU JU SNS Phase 1

Deliverable type:	Report
Deliverable reference number:	D4.3
Contributing workpackages:	WP4
Dissemination level:	PUBLIC
Due date:	30-04-2025
Actual submission date:	28-04-2025
Responsible organization:	KTH
Editor(s):	Gourav Prateek Sharma and James Gross
Version number:	V1.0
Status:	Final
Short abstract:	This deliverable presents updates to the measurement framework, which is used to collect delay measurement data in the 5G system, along with the final measurement results. It provides an overview of the enhancements made to the design and implementation of the framework, highlighting its improved capabilities. The deliverable also discusses various measurement scenarios corresponding to different traffic profiles and channel conditions, providing a detailed analysis of their impact on latency. The results obtained from these scenarios are presented, demonstrating the impact of varying traffic and network conditions on RAN delay and its subcomponents.
Keywords:	5G, 6G, software, latency, packet delay, retransmissions, PTP, measurements, RAN

Contributor(s):	Samie Mostafavi (KTH) Ahmad Traboulsi (KTH) Gourav Prateek Sharma (KTH) James Gross (KTH) Simon Egger (USTUTT) Frank Dürr (USTUTT)

Reviewer(s):	Simon Egger (USTUTT) Joachim Sachs (EDD)

Revision History

26/01/2025	Draft version for the first internal review
21/03/2025	Draft version for the PMT review
11/04/2025	Finalization of deliverable

Disclaimer

This work has been performed in the framework of the Horizon Europe project DETERMINISTIC6G co-funded by the EU. This information reflects the consortium's view, but the consortium is not liable for any use that may be made of any of the information contained therein. This deliverable has been submitted to the EU commission, but it has not been reviewed and it has not been accepted by the EU commission yet.

Executive summary

In recent years, packet delay, Packet Delay Variation (PDV) and reliability have emerged as critical Key Performance Indicators (KPIs) for networks, driven by the growing demand for time-sensitive applications in areas such as adaptive manufacturing, exoskeletons, XR, and smart farming. As we move from 5G to 6G, the need for ultra-reliable and low-latency communications has intensified to support these critical applications.

The DETERMINISTIC6G project aims to enable dependable time-critical communications in future 6G networks through a set of innovative enablers. A key aspect of this effort involves collecting extensive latency data from existing 5G networks. Traditional network measurement frameworks, however, fail to capture the complexity of end-to-end packet delays between application endpoints and the contributions of various 5G mechanisms to these delays.

To address these challenges, we previously proposed a measurement framework capable of conducting detailed latency measurements across both Commercial-off-the-Shelf (COTS) 5G setups and OpenAirInterface (OAI) platforms. This deliverable presents updates to the framework, including enhanced design features and capabilities, and describes the various scenarios used in the measurement campaign. These scenarios investigate the effects of traffic patterns, channel conditions, and background traffic on latency components. We also discuss the results and analysis derived from these scenarios, showcasing the framework's ability to provide deep insights into 5G system delay performance under different network conditions. The collected data serves two critical purposes in the project: (i) developing data-driven simulation models of 6G DetCom nodes and (ii) building a comprehensive dataset to train, validate, and test data-driven latency prediction models.

Contents

1 Introduction

With the evolution from 5G to 6G, the demand for ultra-reliable and low-latency communications has grown, driven by emerging time-sensitive applications such as extended reality (XR), autonomous vehicles, and adaptive manufacturing [DET23-D11]. These applications ask for stringent latency and reliability requirements on communication networks, necessitating enhanced measurement and analysis frameworks to optimize end-to-end (E2E) network performance.

DETERMINISTIC6G aims to enable dependable time-critical communications through a set of key enablers. As discussed in D4.2 [DET23-D42], one of the crucial aspects of this effort is the collection and characterization of latency measurements in existing 5G networks. The previously introduced latency measurement framework in D4.2 provided a structured approach for capturing and analyzing E2E delays, including breakdowns into core network and radio access network (RAN) components. However, even this measurement method fell short in fully capturing the complex relation of delay components within the RAN, particularly in dynamic network environments.

To address these challenges, this document presents D4.3, an extension of the latency measurement framework with enhanced methodologies. This deliverable builds upon D4.2 by incorporating an event-based measurement system that provides fine-grained tracking of packet transmission events, including segment attempts, retransmissions and scheduling decisions at the medium access control (MAC) and radio link control (RLC) layers. Furthermore, these improvements are enhanced towards an extensive measurement campaign that covers the impact of various network and channel conditions on delay.

The measurement campaign reported on in this document focuses on analyzing RAN latency characteristics. Multiple scenarios were designed to capture the influence of factors such as traffic load and channel conditions on various delay components. These scenarios include:

1. Baseline measurements: Establishing fundamental delay benchmarks for comparison.
2. Traffic impact measurements: Evaluating the effect of traffic related parameters such as packet sizes and transmission intervals and also the amount of the background traffic on latency .
3. Channel condition measurements: Investigating the impact of signal quality, and modulation schemes on the retransmission behavior.

By systematically studying these scenarios, D4.3 provides an in-depth understanding into how different traffic and network conditions affect RAN delay, enabling further optimization and predictive modeling efforts in other tasks in DETERMINISTIC6G.

The software components constituting the updated latency measurement framework can be found in the project's public Github repository. We also provide links to the Zenodo repositories for the measurements collected using the updated framework under different measurement scenarios. The links to the software and sample measurements are listed in Table 1-1.

Table 1-1 An overview of the software components and sample measurements relevant to the latency measurement framework and measurements.

Component name	License	Links
Latency Measurement Framework (Updated)	Apache License 2.0	Github Link Zenodo Link
Latency Measurement Data	Creative Commons Attribution 4.0 International	Zenodo Link

1.1 DETERMINISTIC6G Approach

Digital transformation of industries and society is resulting in the emergence of a larger family of time-critical services with needs for high availability presenting unique requirements distinct from traditional Internet applications like video streaming or web browsing. Time-critical services are already known in industrial automation; for example, an industrial control application that might require an end-to-end "over the loop" (i.e., from the sensor to the controller back to the actuator) latency of 2 ms and with a communication service requirement of 99.999% [3GPP16-22261]. But with the increasing digitalization similar requirements are appearing in a growing number of new application domains, such as extended reality and adaptive manufacturing [DET23-D11]. The general long-term trend of digitalization leads towards the incorporation of cyber-physical systems where the monitoring, control and maintenance functionality is moved from physical objects (like a robot, a machine or a tablet device) to a compute platform at some other location, where a digital representation – or digital twin – of the object is operated. Such Cyber Physical System (CPS) applications need a frequent and consistent information exchange between the digital and physical twins. Several technology developments in the ICT-sector drive this transition. The proliferation of (edge-) cloud compute paradigms provide new cost-efficient and scalable computing capabilities that are often more efficient to maintain and evolve compared to embedded compute solutions integrated into the physical objects. It also enables the creation of digital twins as a tool for advanced monitoring, prediction and automation of system components and improved coordination of systems of systems. New techniques based on Machine Learning can be applied in application design that can operate over large data sets and profit from scalable compute infrastructure. Offloading compute functionality can also reduce spatial footprint, weight, cost and energy consumption of physical objects, which is in particular important for mobile components, like vehicles, mobile robots, or wearable devices. This approach leads to an increasing need for communication between physical and digital objects and this communication can span over multiple communication and computational domains. Communication in this cyber-physical world often includes closed-loop control interactions, which can have stringent end-to-end KPI (e.g., minimum and maximum packet delay) requirements over the entire loop [WPP+22]. In addition, many operations may have high criticality, such as business-critical tasks or even safety relevant operations. Therefore, it is required to provide dependable time-critical communication which provides communication service-assurance to achieve the agreed service requirements.

Time-critical communication has in the past been mainly prevalent in industrial automation scenarios with special compute hardware like Programmable Logic Controller (PLC) and is based on a wired

communication system, such as EtherCat and Powerlink, which is limited to local and isolated network domains which is configured to the specific purpose of the local applications. With the standardization of Time-Sensitive Networking (TSN) and Deterministic Networking (DetNet), similar capabilities are being introduced into the Ethernet and IP networking technologies, which thereby provide a converged multi-service network allowing time critical applications in a managed network infrastructure allowing for consistent performance with zero packet loss and guaranteed low and bounded latency [TSN][DETNET]. The underlying principles are that the network elements (i.e. bridges or routers) and the PLCs can provide a consistent and known performance with negligible stochastic variation, which allows us to manage the network configuration to the needs of time-critical applications with known traffic characteristics and requirements.

It turns out that several elements in the digitalization journey introduce characteristics that deviate from the assumptions that are considered as baseline in the planning of deterministic networks. There is often an assumption for compute and communication elements and applications, that any stochastic behavior can be minimized such that the time characteristics of the element can be clearly associated with tight minimum/maximum bounds. Cloud computing provides efficient scalable compute, but introduces uncertainty in execution times; wireless communications provides flexibility and simplicity, but with inherently stochastic components that lead to packet delay variations exceeding significantly those found in wired counterparts; and applications embrace novel technologies (e.g. ML-based or machine-vision-based control) where the traffic characteristics deviate from the strictly deterministic behavior of old-school control. In addition, there will be an increase in dynamic behavior where characteristics of applications and network or compute elements may change over time in contrast to a static behavior that does not change during runtime. It turns out that these deviations of stochastic characteristics make traditional approaches to planning and configuration of end-to-end time-critical communication networks such as TSN or DetNet, fall short in their performance regarding service performance, scalability and efficiency. Instead, a revolutionary approach to the design, planning and operation of time-critical networks is needed that fully embraces the variability but also dynamic changes that come at the side of introducing wireless connectivity, cloud compute and application innovation. DETERMINISTIC6G has as objective to address these challenges, including the planning of resource allocation for diverse time-critical services end-to-end over multiple domains, providing efficient resource usage and a scalable solution [SPS+23].

DETERMINISTIC6G takes a novel approach towards converged future infrastructures for scalable cyber-physical systems deployment. With respect to networked infrastructures, DETERMINISTIC6G advocates (I) the acceptance and integration of stochastic elements (like wireless links and computational elements) with respect to their stochastic behavior captured through either short-term or longer-term envelopes. Monitoring and prediction of KPIs, for instance latency or reliability, can be leveraged to make individual elements plannable despite a remaining stochastic variance. Nevertheless, system enhancements to mitigate stochastic variances in communication and compute elements are also developed. (II) Next, DETERMINISTIC6G attempts the management of the entire end-to-end interaction loop (e.g. the control loop) with the underlying stochastic characteristics, especially embracing the integration of compute elements. (III) Finally, due to unavoidable stochastic degradations of individual elements, DETERMINISTIC6G advocates allowing for adaptation between applications running on top such converged and managed network infrastructures. The idea is to introduce flexibility in the application operation such that its requirements can be adjusted at runtime based on prevailing system conditions. This encompasses a larger set of application requirements that

(a) can also accept stochastic end-to-end KPIs and (b) that possibly can adapt end-to-end KPI requirements at run-time in harmonization with the networked infrastructure. DETERMINISTIC6G builds on a notion of time-awareness, by ensuring accurate and reliable time synchronicity while also ensuring security-by-design for such dependable time-critical communications. Generally, we extend a notion of deterministic communication (where all behavior of network and compute nodes and applications is pre-determined) towards dependable time-critical communication, where the focus is on ensuring that the communication (and compute) characteristics are managed in order to provide the KPIs and reliability levels that are required by the application. DETERMINISTIC6G facilitates architectures and algorithms for scalable and converged future network infrastructures that enable dependable time-critical communication end-to-end, across domains and including 6G.

1.2 Relation to other work packages

D4.3 establishes various interlinkages with other tasks within the DETERMINISTIC6G project, as illustrated in Figure 1.1. A detailed analysis of latency in 5G, previously presented in [DET23-D21], provided a breakdown of 5G user plane latency, which served as a foundation for designing the latency measurement framework proposed in D4.2. In this deliverable, we provide major updates to the latency measurement framework as well as describe the measurement campaign and obtained results.

The measurements collected using the latency measurement framework play a crucial role in two project tasks. First, they feed the development of data-driven latency prediction models. The framework, initially described in D4.2, together with the updates has now been utilized to deliver the results of experimental RAN latency measurements in this deliverable. The data obtained is used to train, test, and validate the latency predictors [DET23-D42][MSG+23]. The results are also intended to be integrated into the simulation model of 6G DetCom node.

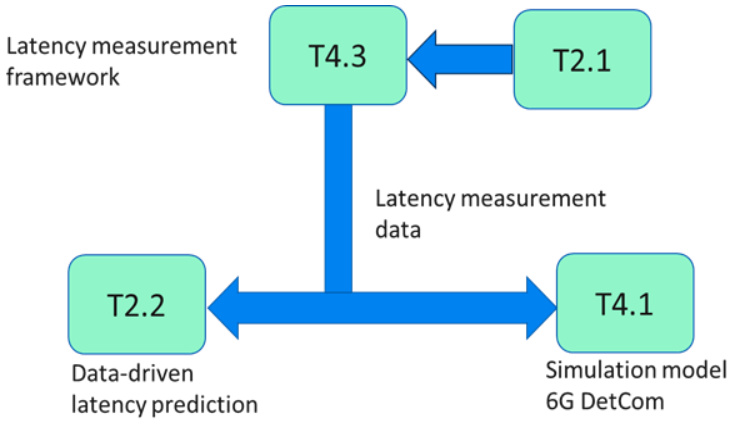

Figure 1.1 Relationship of task T4.3 with other tasks.

1.3 Objective of the document

The objective of this document is to present mainly the results obtained after experimental RAN latency measurements using the measurement framework initially introduced in D4.2. It also aims to describe the framework's design enhancements with respect to the original design. These improvements provide insights into the various scenarios used in the measurement campaign as demonstrated through the results and analyses presented.

1.4 Structure and scope of the document

The rest of this document is organized as follows:

Section 2 provides updates on the latency measurement framework, building upon the initial version presented in D4.2. This section highlights the enhancements made to the framework's design and implementation, focusing on its improved capabilities for capturing detailed latency metrics. Additionally, it discusses the essential components and mechanisms of the framework that enable comprehensive latency analysis.

Section 3 describes the various scenarios explored as part of the measurement campaign. It elaborates on the configurations and conditions used to investigate the impact of factors such as traffic profiles, channel conditions and background traffic on latency components. This section provides insights into how these scenarios were designed to simulate real-world network conditions.

Section 4 presents the results obtained from the measurement scenarios and offers a detailed analysis of the findings. This section shows how the updated framework enables a deep understanding of 5G system delay behavior and highlights its ability to capture the relationship of various delay components with different channel and traffic conditions.

Finally, Section 5 concludes the document by summarizing the key updates, results, and insights presented. It also discusses prospective future directions for advancing the latency measurement framework, including potential enhancements and new use cases for 5G and 6G networks.

2 Latency measurement framework overview and updates

The latency measurement framework, initially presented in D4.2, was designed to facilitate a comprehensive analysis of network performance with respect to end-to-end (E2E) delay in 5G networks [DET23-D42][SSG+23]. Building upon OpenAirInteface 5G, the primary objective of the framework was to capture timestamps and key metadata from a packet's journey, starting from the client application (UE side) to the server application (behind the core network) [OAI5G]. The framework enabled a detailed decomposition of E2E delay into its core components, including RAN and core network delays and further dissected RAN delays into queuing, segmentation, retransmission and processing delays [MTS+24]. By incorporating metadata such as Transport Block Size (TBS), Modulation and Coding Scheme (MCS) indices and channel quality indicators, the framework provided valuable insights into the factors contributing to latency and supported advanced latency optimization.

The updated version of the framework introduces a richer, event-based measurement and analysis methodology. While the original framework relied on timestamp alignment to estimate delay components, the enhanced version tracks granular events that drive packet transmissions. Key features include timestamping segment events, tracking retransmissions, and capturing scheduling and acknowledgment or negative acknowledgement (ACK/NACK) decisions at the MAC and RLC layers. This event-based approach offers a more detailed view of packet dynamics, such as associating Hybrid Automatic Repeat request (HARQ) attempts with specific segment transmissions and recording scheduling handshake parameters.

The data is stored in a time-series database, enabling efficient retrieval and querying of historical metrics and network events. This structure allows seamless access to both raw timestamps and

processed labels (e.g., delay segments, retransmission counts, scheduling decisions). These detailed records provide a robust foundation for data-driven analysis, predictive modeling, and machine learning tasks, showcasing the framework's potential to optimize network performance and support advanced research for 5G and future 6G networks.

High-level features in the updated version:

Modular Data Processing:

The updated latency measurement framework employs a modular data processing approach to simplify the analysis of end-to-end latency in 5G networks. By dividing processing into different modules: packet analysis, channel analysis and schedule analysis, the framework isolates key contributors to latency, such as queuing, segmentation and retransmissions. This modularity enables a focused and detailed understanding of each component's impact on delay while providing the flexibility to integrate new features in the future.

- Packet Analyzer: Extracts packet-level metrics, including delays and retransmissions.
- Channel Analyzer: Examines channel conditions, such as MCS indices and HARQ performance.
- Schedule Analyzer: Analyzes scheduling decisions and their effects on latency.

Enhanced Visualization Accompanying the Schedule Analyzer (see visualizations below):

- Provides detailed, per-packet visualizations of segmentation process with timestamps for arrival of RLC segments in the MAC of the UE and the departure from the MAC layer of the gNB.
- Includes timelines of the scheduling process, showing events such as scheduling requests, schedule grants, buffer status reports and other key messages with their respective timestamps.
- Offers an intuitive interface for understanding the timing and sequence of scheduling activities, aiding in the identification of bottlenecks and inefficiencies in packet scheduling.

These improvements ensure the framework is equipped to provide detailed insights into delay and its subcomponent and their sensitivity to various traffic and channel conditions on the delay.

3 Measurement scenarios

To comprehensively understand the RAN delay characteristics of our 5G system, we have organized the measurement campaign into distinct scenarios using OpenAirInterface (OAI) 5G and Software-Defined Radios (SDRs) [EXPO]. To this end, a testbed at KTH under the name ExPECA [MMR+23] serves as a suitable testbed for running extended measurements in an isolated environment. More information on the ExPECA textbed can be found at [EXPM], [EXPU]. This approach allows us to isolate and analyze the impact of key factors on specific delay components and also to illustrate how different conditions influence end-to-end latency. We refer the reader to the previous deliverable D4.2 for the definition of various delay components [DET23-D42]. Focusing on a well-defined scenario lets us investigate the impact of key factors impacting delay, such as traffic load, channel conditions and network configurations. Below, we outline the scenarios included in our measurement campaign and the objectives of each.

Scenario 0: Baseline delay: This scenario establishes the baseline for evaluating delays in subsequent scenarios. The primary goal of this scenario is to understand the fundamental delay contributions of the system itself, providing a baseline delay benchmark. Figure 3.1 shows the illustration for the setup

for baseline delay measurements and the corresponding configuration is shown in Table 3-1. For the synchronization of the different hosts, PTP-based time synchronization in an out-of-band wired network is used to provide time reference to different nodes in the setup [PTP4L][PHCS], as discussed in [DET23-42].

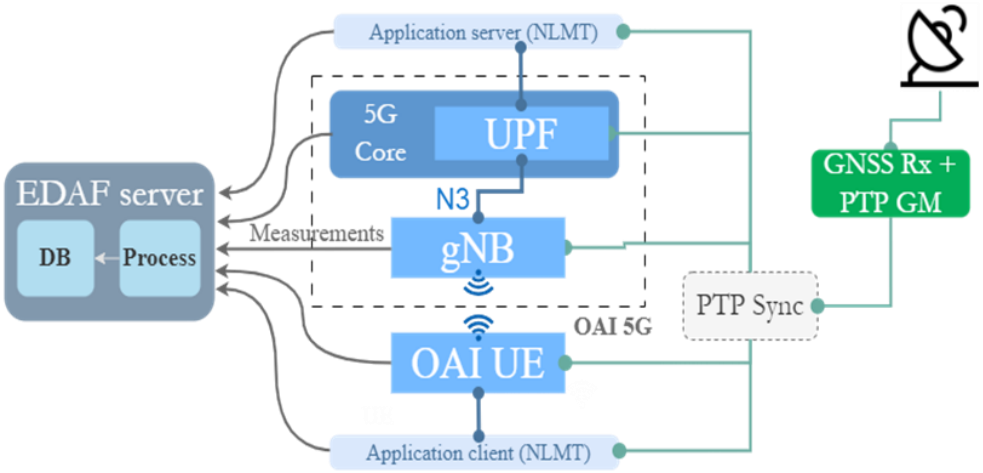

Figure 3.1 Measurement setup for scenario 0 (baseline delay).

Table 3-1 Default parameter configuration for scenario 0 (baseline delay).

Parameter	Values/Range
Measurement flow payload	100 B
Measurement flow packet interval	50 ms
Band	NR77 (3.5 GHz)
Bandwidth	40 MHz (106 PRBs)
RSRP	-95 dBm
Sub-carrier spacing	30kHz ($\mu = 1$)
UL:DL ratio	1:3 (DDDSU)
TDD periodicity	2.5 ms
Total number of packets	40,000

For the delay measurements in slotted communication systems such as 5G, it is typical that frame-alignment delay contributes to a non-negligible portion of end-to-end delays. Moreover, if there is small drift between the clocks of the packet source and the 5G system, the RAN delay shows a sawtooth behavior with respect to time as shown in Figure 3.2. The plot shows the end-to-end delay for 200 consecutive packets. It can be observed that the deviation between the maximum and minimum frame-alignment delay is around 5 ms and the period is 4 s (~ 80 packet intervals at 50 ms). In general, the larger the drift between the two clocks the higher the frequency of the sawtooth. It is obvious that the varying frame-alignment delay from packet to packet contributes to a PDV that is independent of traffic or channel conditions. Therefore, for the comparison of two distinct measurements runs it is useful to isolate the frame-alignment delay from the end-to-end delay. To

that end, for a given packet, we obtain RAN delay (in UL) without frame-alignment delay, by subtracting frame-alignment delay from the RAN delay. The UL traffic was generated using a tool called Network Latency Measurement Tool [NLMT]. Here, RAN delay is defined as the difference between the time when packet leaves the RLC layer in the gNB and the time when it enters the RLC layer in the UE. Figure 3.2 also shows the frame-alignment along with RAN delay in the UL direction.

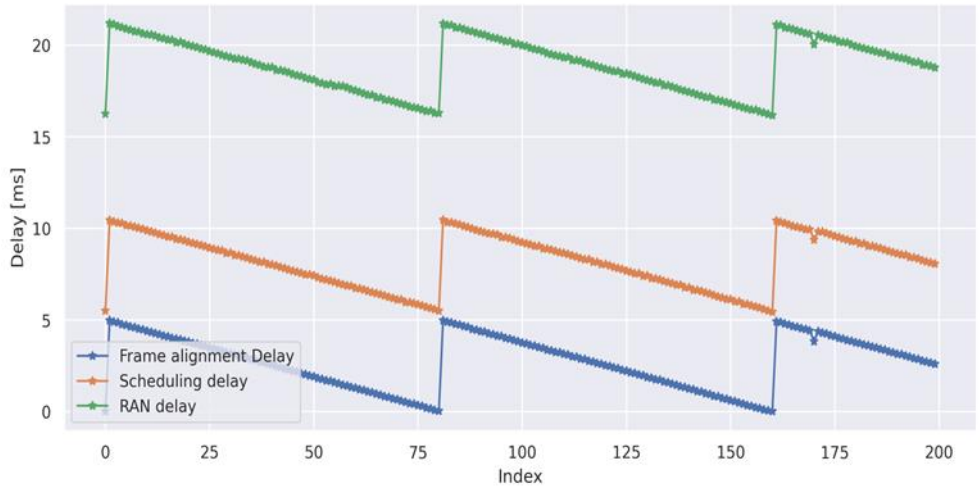

Figure 3.2 Time-series of UL RAN delay, frame-alignment delay and scheduling delay for a sequence of 200 packets sent in UL.

Furthermore, the scheduling delay can also be isolated from the remaining RAN delay. Here, scheduling delay is defined as the time between the moment packets send their scheduling request and the moment the UE sends its first RLC segment. This delay includes the time taken by gNB to grant UL resources to a UE. The updated delay measurement framework allows us to systematically analyze

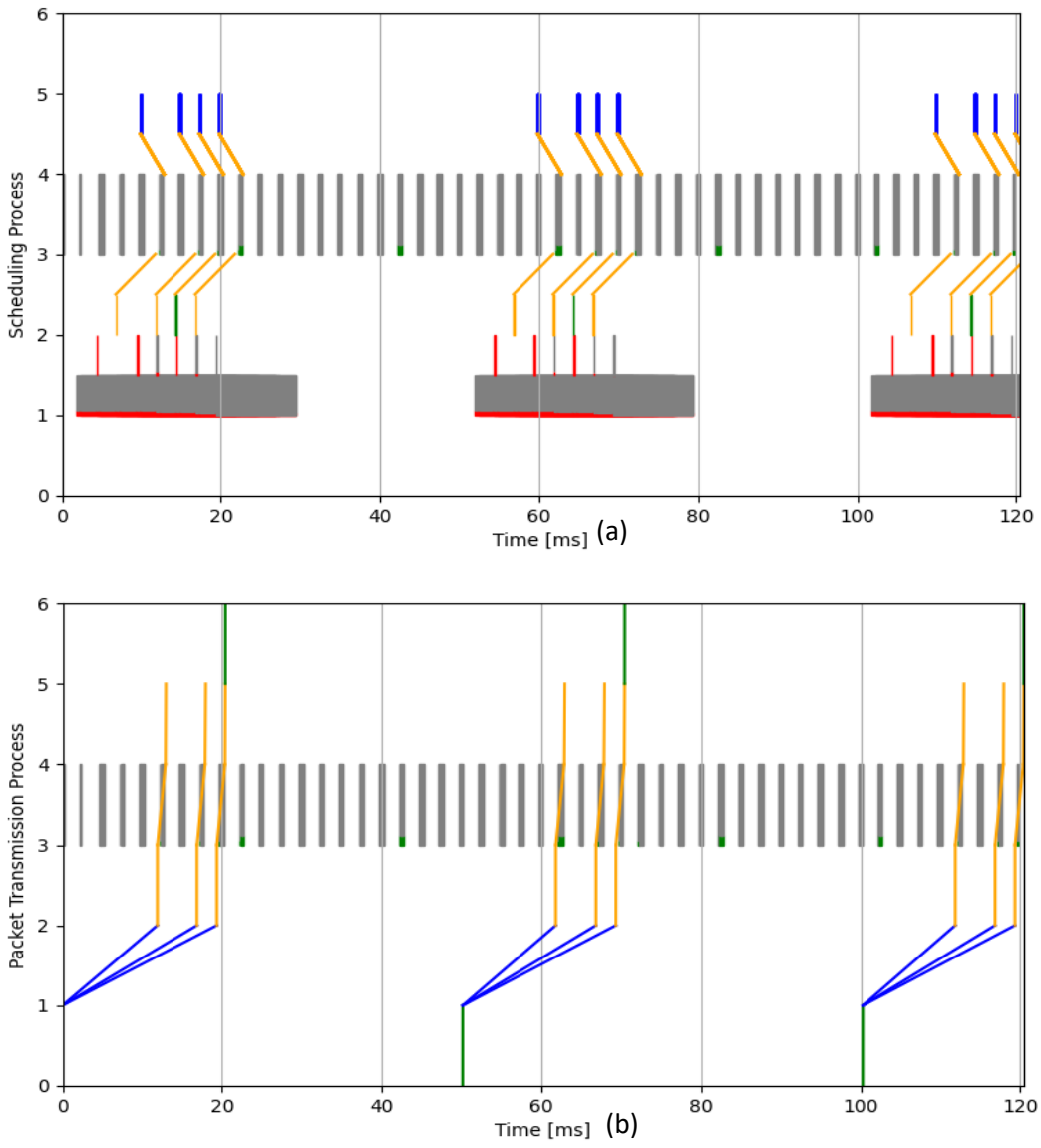

Figure 3.3 Timeline of the (a) scheduling process and (b) arrival/service times of RLC segments of three consecutive UL packets.

scheduling delay. Figure 3.3 (a) shows the process of scheduling of a packet in UL direction and Figure 3.3 (b) shows the arrival times of the RLC segments in the MAC layer of the UE and their corresponding arrivals at the MAC layer of the gNB. The red color indicates the proportion of buffer occupancy as the segments depart the UE; each packet is segmented into three RLC (shown in yellow in (b)) segments. The green line in (b) indicates the assembly of the RLC segments as the third segment arrives. The resulting scheduling delay, i.e., the delay between the arrival of the packet and its first segments ready

to be sent by the MAC layer, is around 11 ms. After isolating frame-alignment and scheduling delay from the RAN delay the remaining delay can be used as a baseline reference when performing comparisons with other scenarios. The resulting RAN delay after removing frame-alignment delay and scheduling delay is shown in Figure 3.4.

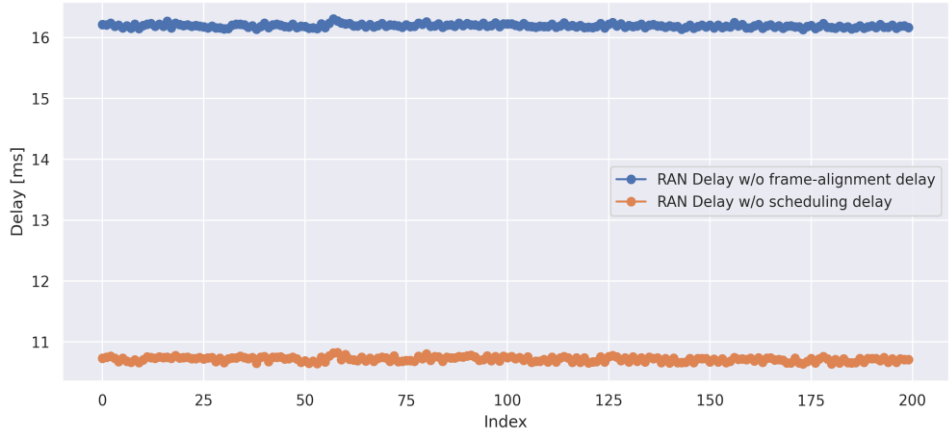

Figure 3.4 UL RAN delay after isolating frame-alignment delay and scheduling delay.

Scenario 1: Impact of traffic: This scenario examines the impact of varying traffic profiles on the 5G system's delay components. By employing different traffic profiles, we quantify the delays that accumulate as traffic intensity increases. Specifically, three aspects are analyzed: (1) the effect of packet payload length (scenario 1.1), (2) the effect of packet transmission intervals (scenario 1.2) and (3) the effect of 'background flow' traffic on a fixed 'measurement flow' (scenario 1.3). The setup for Scenarios (I) and (2) is like that shown in Figure 3.1, whereas for Scenario (3) another UE was connected to the OAI5G. In this sub scenario, one COTS UE is used to generate background traffic in the 5G system whereas the measurements (timestamping) are performed on the other UE. Key metrics studied in this scenario include the RLC queuing delays with increasing traffic and its effect on segmentation and link delays. Additionally, the relation between higher traffic loads and resource availability, such as scheduling delays and transmission grants, is also examined. This analysis offers valuable insights into how traffic dynamics affect end-to-end latency and helps optimize 5G system operations. The findings are crucial for ensuring delay guarantees in dynamic network environments.

Scenario 2: Impact of channel conditions: This scenario explores how varying channel conditions influence delay components, in particular, retransmission delays and also the overall latency. The goal of this scenario is to understand how channel conditions and link adaptation mechanisms influence delay and to identify strategies for mitigating latency under challenging wireless conditions. Two distinct cases are examined: one with a static MCS configuration (scenario 2.1), which avoids dynamic adaptation to channel conditions and another where the system dynamically adjusts MCS for a given block error rate (BLER) target (scenario 2.2) to optimize performance. Key metrics studied under this scenario include the number of retransmissions, time for each re-transmission and the distribution of MCS indices.

4 Results and analysis

In the following section, we describe a subset of measurement and provide analysis.

4.1 Baseline delay

For baseline delay, we considered the setup and configuration described in Section 3. Here, one OAI UE is connected to the OAI 5G setup. Figure 4.1 shows the complementary cumulative distribution function (CCDF) of uplink RAN delay without scheduling delay for a sequence of 40,000 packets. It is worth noting that packet delays exceeding the value of 35 ms were not considered for the analysis here as these are outliers that are typically caused by one or more RLC retransmissions.

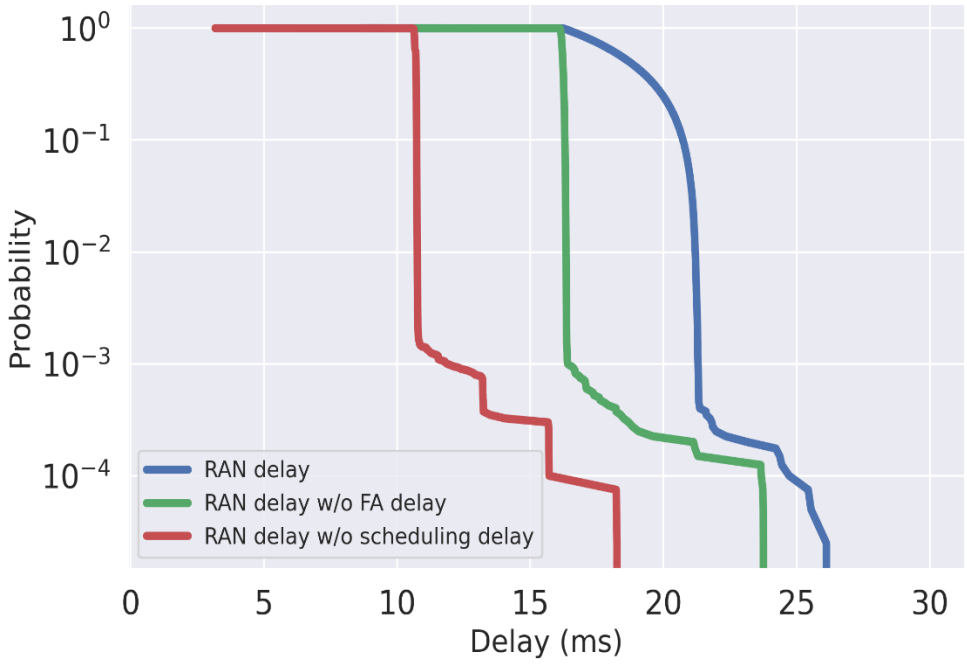

Figure 4.1 CCDFs of UL RAN delay, RAN delay without frame-alignment delay and RAN delay without scheduling delay.

4.2 Impact of traffic parameters

Next, we consider the impact of traffic parameters, namely, packet size, packet transmission interval and the amount of background traffic in the 5G system, on the RAN delay characteristic.

4.2.1 Packet size

For this sub-scenario, packet payload length was varied, and the RAN delay and its components were recorded.

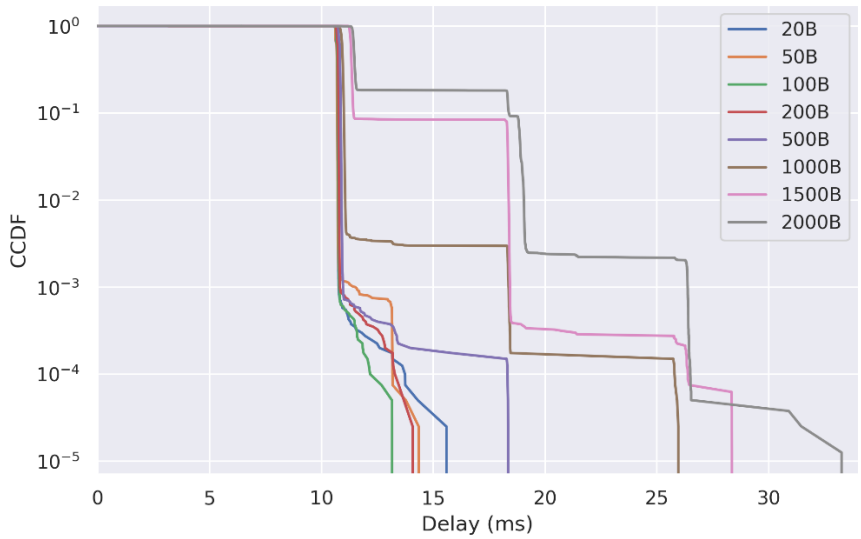

Figure 4.2 CCDF showing the impact of payload length on UL RAN delay without scheduling delay.

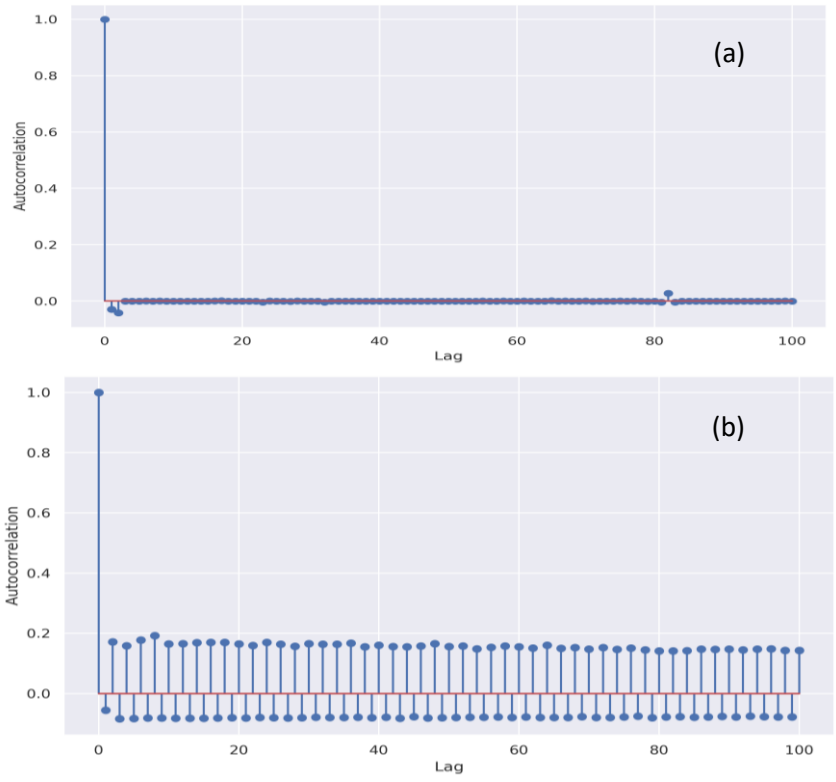

Figure 4.3 Autocorrelation plot of the UL RAN delay without scheduling delay for (a) 100B and (b) 1500B.

The CCDF plot in Figure 4.2 illustrates the RAN delay excluding scheduling delay for various payload sizes. It shows that smaller payloads (until 500B) experience consistently lower delays, especially at higher quantile values and reflecting minimal queuing or segmentation effects. As the payload size increases (e.g., 1000B and 2000B), the delays grow significantly, especially at higher percentiles,

indicating perhaps increased segmentation delays and resource contention. Figure 4.3 shows the autocorrelation plot of RAN delay without scheduling delay for payload of 100B and 1500B, respectively. The autocorrelation of 100B packet delays drops quickly showing minimal correlation and largely independent delays over time. In contrast, the 1500B packet delays show periodic autocorrelation, suggesting systematic patterns likely caused by network effects like queueing. This highlights that larger packet sizes are more influenced by network dynamics, while smaller packets experience more random delays. As shown in Figure 4.4, where packets with payload length > 1000B seem to have a significant proportion of four segments, the large number of segments served over multiple time slots might lead to delay dependence between consecutive packets. This highlights the role of payload size in influencing RAN latency characteristics.

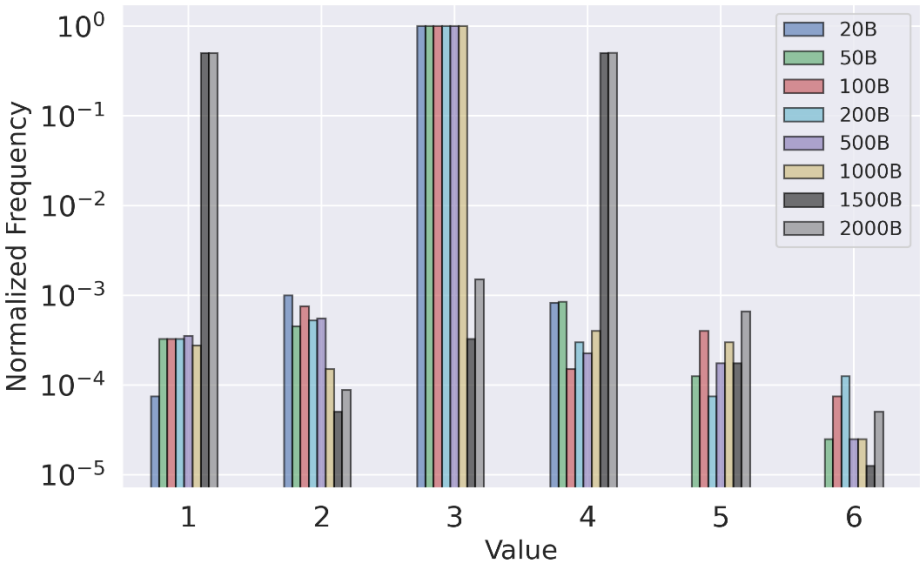

Figure 4.4 Histogram showing the distribution of number of RLC segments for different payload lengths.

4.2.2 Packet transmission interval

For this sub-scenario, packet transmission interval varied, and the RAN delay and its components were recorded.

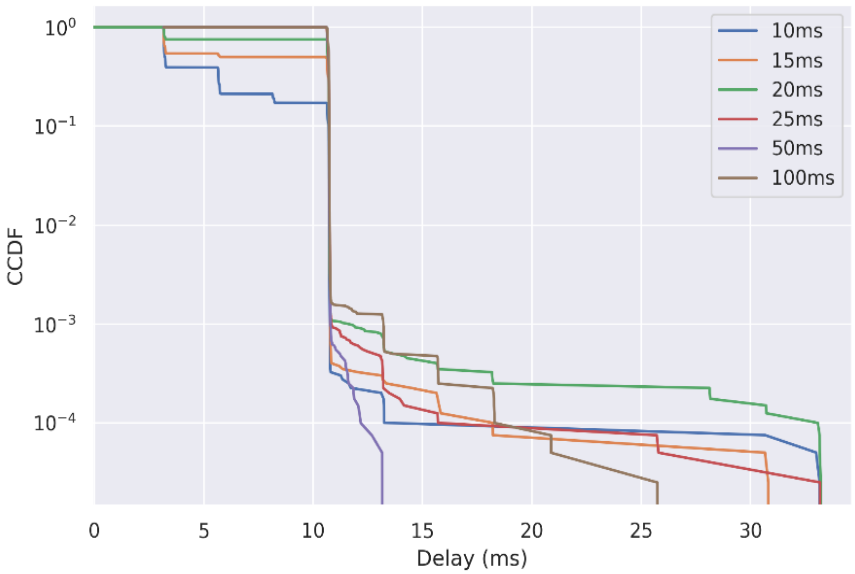

Figure 4.5 CCDF showing the impact of packet transmission interval on UL RAN delay without scheduling delay.

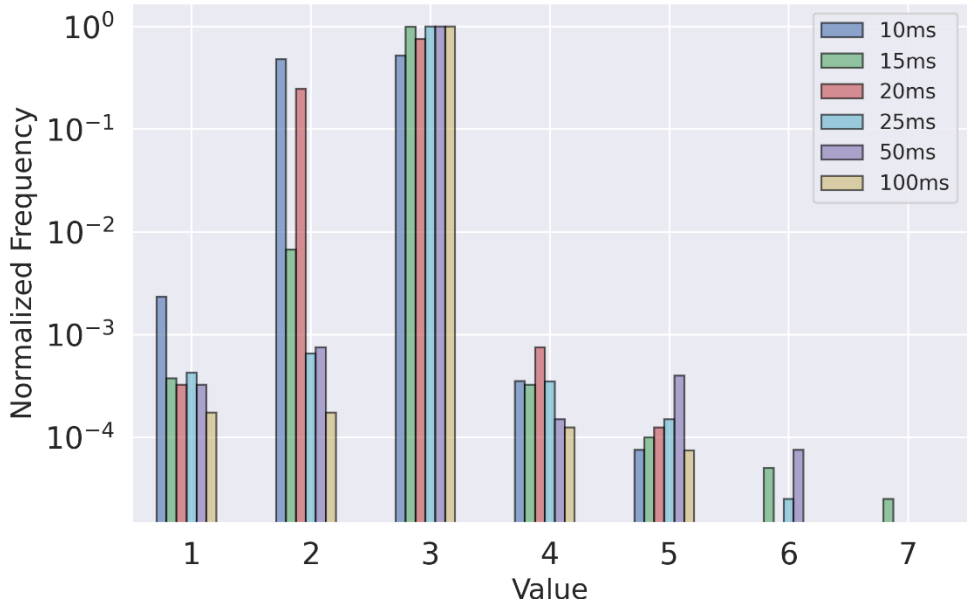

Figure 4.6 Histogram showing the distribution of number of RLC segments for different packet transmission intervals.

The CCDF plot in Figure 4.5 illustrates the RAN delay without scheduling delay for varying packet transmission intervals. Interestingly, the RAN delay appears to decrease for smaller packet transmission intervals, which is counterintuitive, as shorter intervals typically increase queuing and resource contention. The explanation for this is, for now, a subject of inquiry and is under investigation. Additionally, as shown in Figure 4.6, this leads to a reduction in the average number of RLC segments per packet with shorter transmission intervals. Notably, the percentage of packets requiring two segments is higher for transmission intervals below 20 ms.

4.2.3 Background traffic

In this sub-scenario, the amount of background traffic in the 5G was varied and the RAN delay and its components were recorded. The amount of background traffic, which is also periodic traffic, varies by varying the payload length. For instance, 15.6 kB payload with a periodicity of 50 ms results in a 2.5 Mbps traffic load while 62.5 kB results in a traffic load of 10 Mbps. The CCDF plot in Figure 4.7 illustrates the RAN delay without scheduling delay as background traffic is gradually increased in a 5G system using another UE. First, even with a small amount of background traffic (0.01 Mbps) there is an increase in delays as compared to the case when there was no UE connected (shown in blue). At no background or low background traffic levels (e.g., 0.01 Mbps), delays are minimal, with the curve dropping off quickly, indicating low contention for resources. As the background traffic increases to higher levels (i.e., above 2.5 Mbps), the delay CCDF shifts towards the right, indicating more packets experiencing higher delays. The histogram showing the distribution of the number of RLC segments for different background traffic is shown in Figure 4.8. The gradual rise in delays at higher traffic levels shows the sensitivity of RAN latency to background traffic intensity. This highlights the need for efficient resource management and prioritization under a realistic multi-UE scenario.

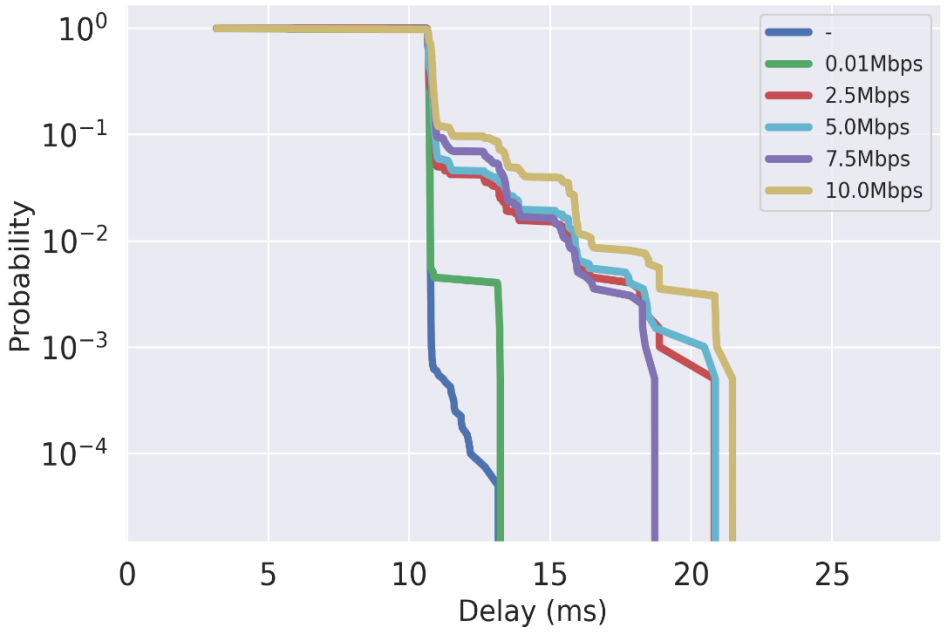

Figure 4.7 CCDF showing the impact of background traffic (generated using a separate UE) on UL RAN delay without scheduling delay.

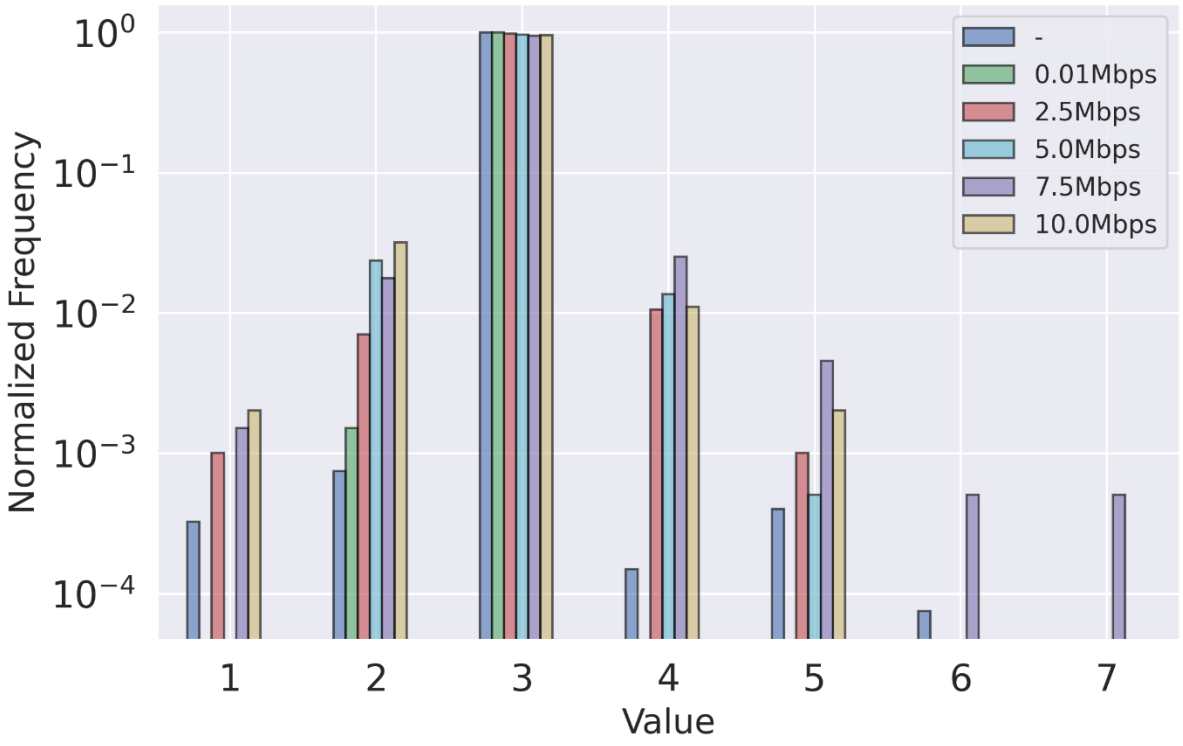

Figure 4.8 Histogram showing the distribution of number of RLC segments for different background traffic.

4.3 Impact of channel parameters

Next, we consider the impact of parameters relevant to the channel conditions in a 5G system. In this scenario, we specifically focus on retransmission delay, which is a key metric reflecting network performance under varying channel conditions. Retransmission delay quantifies the time taken for a packet to be successfully delivered after experiencing one or more retransmissions. Channel conditions together with the link-adaption algorithm play a crucial role in the HARQ process, as they directly influence probability of packet errors and the need for retransmissions.

4.3.1 Target BLER and static MCS

In this sub-scenario, we considered at one time a COTS UE is connected to the OAI 5G system as described in Section 3 and a periodic UL traffic of 100B payload was sent with a periodicity of 100 ms.

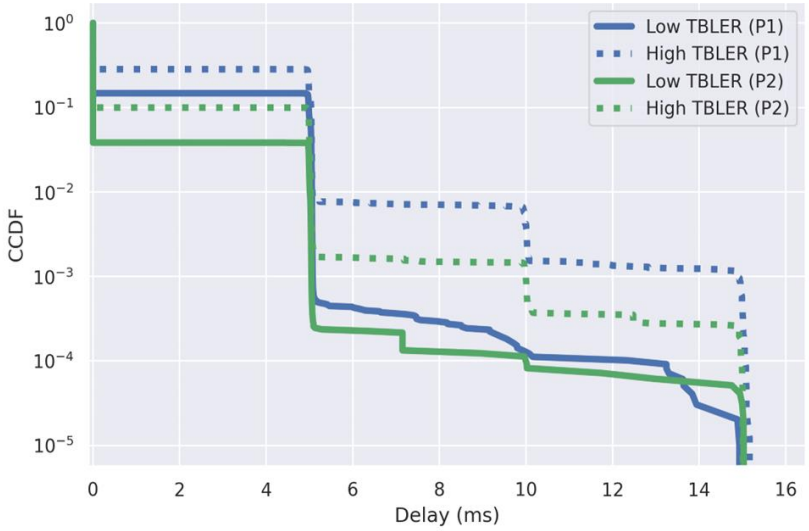

Figure 4.9 Impact of target BLER on the CCDF of UL retransmission delay.

Figure 4.9 shows the CCDF of the retransmission delay plotted for two different UE - gNB connection pairs. P1 and P2 refers to the connections in NR bands 41 and 48, respectively. For each pair, measurements were collected in two different Target Block Error Rate (TBLER) settings. Low TBLER refers to the range of [0.05-0.10] while high TBLER refers to the range of [0.15-0.25]. First, as expected, the higher TBLER range leads to more retransmissions and second, we observe that band 48 is slightly better in terms of number of re-transmissions for unknown reasons.

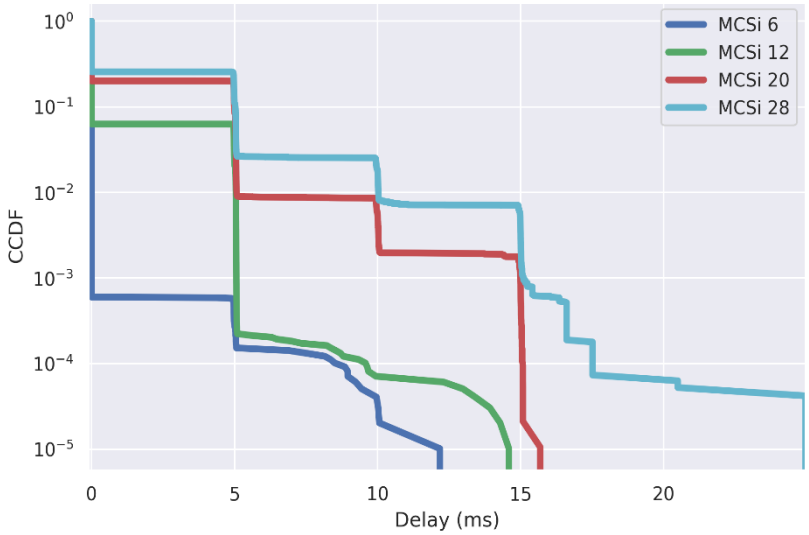

Figure 4.10 Impact of static MCS indices on the CCDF of UL retransmission delay.

Figure 4.10 shows the impact of setting the MCS index to a fixed value on the retransmission delay CCDF. It is clear from the figure that the retransmission delay increases with the increasing MCS index. This occurs because higher MCS indices correspond to less robust modulation and coding. Therefore,

higher MCS indices result in a greater number of bit errors, requiring more retransmissions and thus increasing the overall retransmission delay.

4.3.2 UE type

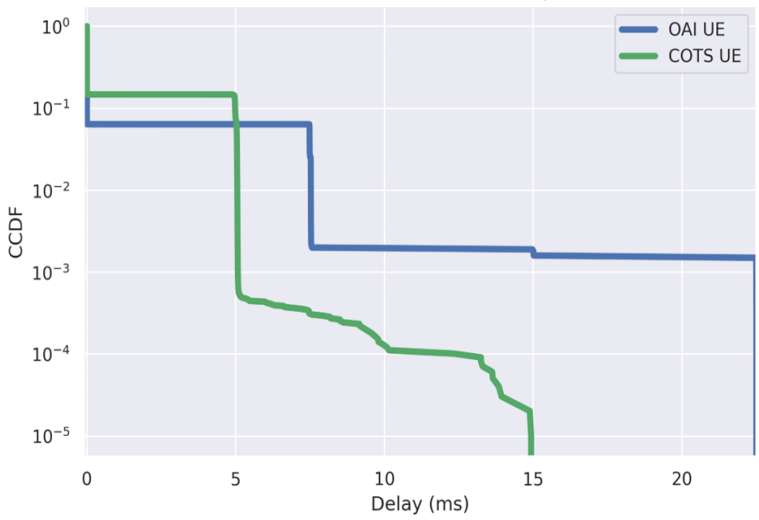

Figure 4.11 Comparison of the CCDF delay of UL retransmission delay in OAI UE and COTS UE.

Figure 4.11 shows the comparison of the CCDF delay of retransmission delay of two types of UE: OAI UE and COTS UE when both are connected to the OAI 5G system. The figures show that the OAI UE experiences longer retransmission delays compared to COTS UE, particularly in the tail of distribution. Additionally, the retransmission intervals are different: approximately 5 ms for COTS UE and 7.5 ms for OAI UE. This difference in performance is likely due to the COTS UE's superior processing capabilities, which enable a faster HARQ loop compared to the OAI UE.

4.4 UE transmission gain

Figure 4.12 shows the impact of UE transmission gain on the CCDF of retransmission delay. UE transmission gain refers to the amount of amplification (in dB) applied to the signal before it is transmitted by the UE. As the transmission gain increases from -15 dB to -5 dB, the retransmission delay improves slightly in the bulk of the distribution, with moderate reductions in delay for most packets. However, a slight reduction in retransmission delays is observed in the tail of the distribution, indicating that extreme delays become much less frequent with higher transmission gain.

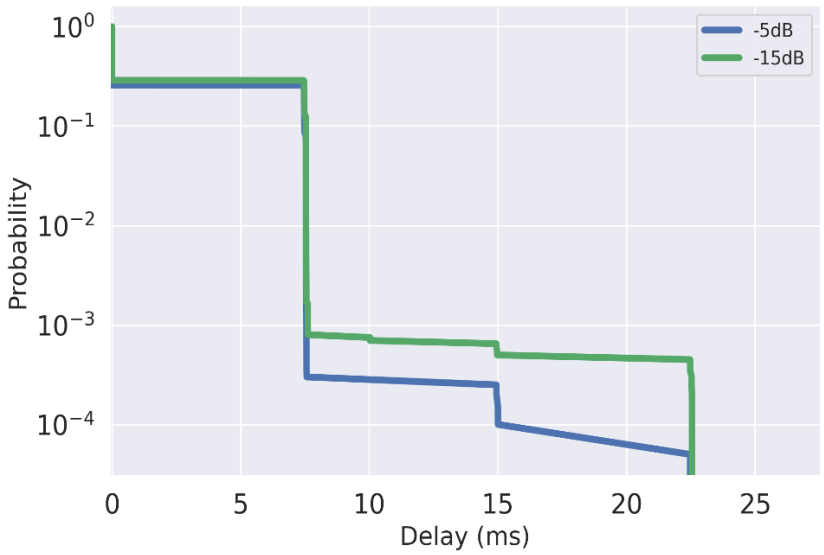

Figure 4.12 Impact of UE transmission gain on UL retransmission delay CCDF.

5 Conclusion

In D4.2, we presented the initial design and implementation of the latency measurement framework, which was proposed to perform comprehensive packet delay and associated metadata measurements in 5G networks. In this document, D4.3, we have provided updates to the framework, highlighting enhancements that improve its capabilities for capturing detailed delay metrics and supporting advanced analysis. We discussed how the analysis is now split into three modules: (i) packet analysis, (ii) scheduling analysis, and (iii) channel analysis.

We also described the measurement campaign in detail, outlining the various scenarios, setups and configurations used to investigate the impact of key factors such as traffic profiles, channel conditions and background traffic on delay components. Sample measurements collected from these scenarios were presented, along with an analysis of the results, showcasing the framework's ability to provide valuable insights into 5G system delay performance under various conditions. Furthermore, to support transparency and reproducibility, we have hosted all collected measurements on Zenodo, providing open access to the datasets for the broader research community to enable further analysis and development.

For future work, the framework can be extended to perform online monitoring and analysis, enabling real-time assessment of delay performance in operational 5G networks. Additionally, the measurement campaign could be expanded to include more scenarios, such as mobile User Equipment (UE), handovers and 5G QoS measurements, to further investigate delay behavior under diverse and dynamic conditions. These enhancements would broaden the measurement framework applicability.

6 Reference

[3GPP16-22261]	3GPP TS 22.261, “Service requirements for the 5G system,” v19.4.0
[3GPP18-R94]	3GPP TSG-RAN WG1 Meeting 94 R1-1809277, “IMT-2020 self-evaluation: UP latency in NR”
[DET23-D11]	DETERMINISTIC6G, Deliverable 1.1, “DETERMINISTIC6G use cases and architecture principles,” Jun. 2023, https://deterministic6g.eu/index.php/library-m/deliverables
[DET23-D21]	DETERMINISTIC6G, Deliverable 2.1, “First report on 6G centric enablers,” Jun. 2023, https://deterministic6g.eu/index.php/library-m/deliverables
[DET23-D22]	DETERMINISTIC6G, Deliverable 2.2, “First Report on the time synchronization for E2E time awareness,” Dec. 2023, https://deterministic6g.eu/index.php/library-m/deliverables
[DET23-D31]	DETERMINISTIC6G, Deliverable 3.1, “Report on 6G convergence enablers towards deterministic communication standards,” Dec. 2023, https://deterministic6g.eu/index.php/library-m/deliverables
[DET23-D41]	DETERMINISTIC6G, Deliverable 4.1, “Digest on First DetCom Simulator Framework Release,” Dec. 2023, https://deterministic6g.eu/index.php/library-m/deliverables
[DET23-D42]	DETERMINISTIC6G, Deliverable 4.2, “Digest on Latency measurement framework,” Feb. 2024, https://deterministic6g.eu/index.php/library-m/deliverables
[DETNET]	Deterministic Networking (DetNet) Working Group, [Online]. Available: https://datatracker.ietf.org/wg/detnet/about/
[EXPM]	ExPECA Testbed Map, [Online]. Available: https://expeca.proj.kth.se/map/
[EXPO]	ExPECA OpenAirInterface Setup guide, [Online]. Available: https://gitlab.eurecom.fr/samiemostafavi/openairinterface5g-edaf
[EXPU]	ExPECA User Guide, [Online]. Available: https://expeca.proj.kth.se/docs/
[MMR+23]	S. Mostafavi, V.N. Moothedath, S. Ronngren, N. Roy, G.P. Sharma, S. Seo, M.O. Muñoz and J. Gross, “ExPECA: An Experimental Platform for Trustworthy Edge Computing Applications,” Nov. 2023, doi: 10.48550/arXiv.2311.01279
[MSG+23]	S. Mostafavi, G. P. Sharma, and J. Gross, "Data-Driven Latency Probability Prediction for Wireless Networks: Focusing on Tail Probabilities," IEEE Globecom, doi: 10.1109/GLOBECOM54140.2023.10437281, 2023
[MTS+24]	S. Mostafavi, M. Tillner, G. P. Sharma, and J. Gross, "EDAF: An End-to-End Delay Analytics Framework for 5G-and-Beyond Networks," IEEE Infocom Workshop, doi: 10.1109/INFOCOMWKSHPS61880.2024.10620853, 2024.
[NLMT]	NLMT, [Online]. Available: https://github.com/samiemostafavi/nlmt
[OAI5G]	OpenAirInterface 5G Implementation: https://gitlab.eurecom.fr/oai/openairinterface5g
[PHCS]	Phc2sys, [Online]. Available Online: https://linux.die.net/man/8/phc2sys
[PTPL]	ptp4l, [Online]. Available: https://linux.die.net/man/8/ptp4l
[RFH+21]	F. Ronteix–Jacquet, A. Ferrieux, I. Hamchaoui, S. Tuffin and X. Lagrange, "LatSeq: A Low-Impact Internal Latency Measurement Tool for

	OpenAirInterface," 2021 IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China, 2021, pp. 1-6, doi: 10.1109/WCNC49053.2021.9417345.
[RSI+21]	J. Rischke, P. Sossalla, S. Itting, F. H. P. Fitzek and M. Reisslein, "5G Campus Networks: A First Measurement Study," in IEEE Access, vol. 9, pp. 121786-121803, 2021, doi: 10.1109/ACCESS.2021.3108423.
[SPS+23]	G. P. Sharma, D. Patel, J. Sachs, M. De Andrade, J. Farkas, J. Harmatos, B. Varga, H. -P., Bernhard, R. Muzaffar, M. Ahmed, F. Duerr, D. Bruckner, E.M. De Oca, D. Houatra, H. Zhang and J. Gross, "Toward Deterministic Communications in 6G Networks: State of the Art, Open Challenges and the Way Forward," in IEEE Access, vol. 11, pp. 106898-106923, 2023, doi: 10.1109/ACCESS.2023.3316605
[SSG+23]	S. S. Mostafavi, G. P. Sharma and J. Gross, “DETERMINISTIC6G COTS 5G latency measurements”. Zenodo, Dec. 15, 2023. doi: 10.5281/zenodo.10390211.
[TSN]	Time-Sensitive Networking (TSN) Task Group, [Online]. Available: https://1.ieee802.org/tsn/
[WPP+22]	G. Wikström et al, “6G – Connecting a cyber-physical world”, Ericsson white paper, GFTL-20:001402, February 2022, https://www.ericsson.com/4927de/assets/local/reports-papers/white-papers/6g--connecting-a-cyber-physical-world.pdf

7 List of abbreviations

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
5G-Adv	5G Advanced
AI	Artificial Intelligence
CCDF	Complementary Cumulative Distribution Function
CDF	Cumulative Distribution Function
COTS	Commercial Off-The-Shelf
CPS	Cyber-Physical Systems
DB	Database
DetCom	Deterministic Communications
DetNet	Deterministic Networking
DVP	Delay Violation Probability
GM	Grandmaster
gNB	Next Generation NodeB
GNSS	Global Navigation Satellite System
INT	In-band Network Telemetry
IP	Internet Protocol

KPI	Key Performance Indicator
MAC	Media Access Control
MCS	Modulation and Coding Scheme
ML	Machine Learning
NIC	Network Interface Card
NLMT	Network Latency Measurement Tool
NR	New Radio
OAI	OpenAirInterface
OSI	Open Systems Interconnection
OWD	One-Way Delay
P4	Programming Protocol-Independent Packet Processors
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PHY	Physical Layer
PRB	Physical Resource Block
PTP	Precision Time Protocol
RLC	Radio Link Control
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RTT	Round-Trip Time
SCS	Sub-Carrier Spacing
SDAP	Service Data Adaptation Protocol
SDR	Software-Defined Radio
TB	Transport Block
TBS	Transport Block Size
TCP	Transmission Control Protocol
TDD	Time Division Duplexing
TSN	Time-Sensitive Networking
UDP	User Datagram Protocol
UL	Uplink
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communications
XR	Extended Reality