# POLITECNICO DI TORINO

**Master's Degree in Communications and Computer Networks Engineering**


**Master's Degree Thesis**

# Evaluating Latency in a 5G Infrastructure for Ultralow Latency Applications

**Supervisors**

**Prof Cristina ROTTONDI**

**Prof. Andrea BIANCO**

**Dr. German SVIRIDOV**

**Candidate**

**Omid AKBARZADEH**

**Fall 2021**

**Abstract**

It's predicted that 5G, the fifth generation of mobile network, would be commercially available worldwide shortly as 5G currently deployed has limited coverage. There will be a substantial boost in performance and reliability compared to the current generation - 4G mobile network. The packet latency of 5G is roughly ten-fold less than that of existing 4G networks. For several applications of 5G, packet latency must be at about one millisecond. For emerging new services such as virtual reality (VR), live streaming using mobile networks, autonomous vehicles, and tactile Internet with the capability of machines and equipment remote operation and high sensitivity via mobile network, significantly low latency is required. With my thesis, I have attempted to understand how real-world variables impact packet delay. The lack of comprehensive research on the particular aspects affecting packet delay in a 5G network in a realistic condition in contrast with laboratory conditions can be observed, where the reported packet latency would be pretty low. Several elements impact packet delay in a practical 5G network, and the outcome of this research has helped to recognize them. To determine how each part contributes to overall packet delay, this information is collected. The future extension of this study can be testing the 5G network's latency reduction methods effectiveness. Identifying how 5G network packet latency is caused and affected was required before reducing packet latency to meet 5G latency objectives. To reach this aim, this effort was undertaken. It was determined that a 4G network's lower latency improvement threshold was achieved by evaluating the latency reduction strategies in their optimal configuration. Therefore, to meet 5G latency objectives, 4G radio access and core network technology must be changed to reach 5G latency targets.

# Acknowledgements

This master thesis report marks the culmination of my studies at Politecnico di Torino for obtaining a Master of Science (MSc) degree in Communications and Computer Networks Engineering. The experience for me while working on this thesis was both challenging and, more than that was intellectually stimulating. The successful completion of this work would not have been possible without the immense and constant support and guidance from several people. First,I would like to express my sincere gratitude to professor Rottondi, my supervisor at Politecnico di Torino, for her advices in shaping my research goals, measurement approach. I would also like to thank German Svridov, my daily supervisor, for devoting time out of his busy schedule to provide me with feedback on my work and helping me to find the fix for the various challenges that I faced during the development of the experimental setup.I would like to thank Vodafone Italia for providing me an opportunity to carry out this work at their facility. Vodafone ensured that I was provided with all the necessary resources to conduct my research with great quality. When I look back to the past six months, I spent on this work, and the time was filled with a great learning experience. I had the chance to put into practical use the knowledge I gained from my academic studies. Working alongside several experts in networks and communication has helped me develop my critical thinking and analytical skills.

*Omid Akbarzadeh*

# Table of Contents

**List of Tables** v

**List of Figures** vi

**Acronyms** ix

**1 introduction** 1

**2 Latency definition in a mobile network** 5
- 2.1 Various latency definitions . . . 5
- 2.2 Limitations of user plane latency definitions . . . 8
- 2.3 Latency targets . . . 9
- 2.4 Latency definition considered for this research . . . 10

**3 Objectives and related concepts definition** 12
- 3.1 Research objectives . . . 12
- 3.2 5G low latency services . . . 13
  - 3.2.1 Industrial automation . . . 13
  - 3.2.2 Intelligent transportation systems (ITS) . . . 15
  - 3.2.3 Robotic industry . . . 16
  - 3.2.4 Virtual reality (VR) . . . 16
  - 3.2.5 Augmented reality (AR) . . . 16
  - 3.2.6 Health care . . . 16
  - 3.2.7 Gaming . . . 17
  - 3.2.8 Smart Grid . . . 17
  - 3.2.9 Education and art . . . 17
- 3.3 Source of latency in a mobile network . . . 18
- 3.4 Low latency barriers . . . 22
- 3.5 RAN solutions . . . 23
  - 3.5.1 Backhaul network . . . 23
- 3.6 Core network solutions . . . 26

3.6.1 5G entities of core network . . . 26
3.6.2 Latency improvement methods: related studies . . . 28
3.7 Factors affecting packet latency . . . 30
3.7.1 5G and LTE Frame structure . . . 32
3.8 Latency measurement: related studies . . . 33

**4 4G and 5G Mobile networks and their components** 36
4.1 4G network architecture: E-UTRAN . . . 36
4.2 4G network architecture: EPC . . . 36
4.3 Delay of the mobile network . . . 39
4.4 Scheduling latency . . . 39
4.5 5G network architecture . . . 41
4.5.1 Full 5G system Architecture with Reference points . . . 45
4.6 5G: mm-Wave . . . 46
4.6.1 mm-Wave Challenges: Free Space path-loss . . . 46
4.6.2 mm-Wave Challenges: Blockage . . . 46
4.6.3 mm-Wave opportunities: Reduced latency . . . 48

**5 Measurement setup, scenarios, and results discussion** 50
5.1 Realistic wireless network . . . 50
5.2 Network virtualization . . . 51
5.3 Measurement setup components and networking . . . 52
5.4 Results discussion . . . 54
5.4.1 Measurement tools and defintions . . . 54
5.5 Packet loss . . . 56
5.5.1 Received packets delays histogram . . . 58
5.5.2 Delays average . . . 58
5.5.3 Delays mean values variation . . . 64
5.5.4 PDF and ECDF of delays . . . 66
5.5.5 RTT . . . 68
5.5.6 RTT – Jitter . . . 68
5.5.7 Bandwidth . . . 71
5.5.8 Packet loss . . . 72
5.6 CPEs location effect evaluation: first round of measurements . . . 74
5.7 CPEs location effect evaluation: second round of measurements . . . 77
5.7.1 Ping-jitter . . . 77
5.7.2 Bandwidth . . . 78
5.8 Results discussion . . . 80
5.8.1 Position 5 . . . 80
5.8.2 Position 4 . . . 80
5.8.3 Position 3 . . . 81

5.8.4 Position 2 . . . . . 81
5.8.5 Position 1 . . . . . 81
5.8.6 Observations . . . . . 82
5.9 Conclusion and future extensions . . . . . 82

**Bibliography** 84

# List of Tables


2.1 The summary of latency targets in different mobile generations. . . 10

3.1 The summary of latency targets in different mobile applications. . . 15
3.2 The latency value of each stage has contributed to control plane latency. . . . . 20
3.3 A summary of low-latency techniques in RAN [**3**]. . . . . 26
3.4 Overview of NFV and SDN technique for low latency [**3**]. . . . . 27

4.1 Quality indicator comparison of sub-6 GHz and millimetre-wave. . . 49

5.1 Data rates related to different measurement rounds. . . . . 55
5.2 Overall measurements statistics, all averages are calculated over a 1 minute window. . . . . 73

# List of Figures


1.1 A packet's end-to-end latency time interval in a 5G network.. . . . . 2

2.1 The control plane latency stages. . . . . 6
2.2 Figures (a) and (b) depict FDD and TDD LTE user plane latency frame. . . . . 7
2.3 Latency visualization . . . . . 8

3.1 A typical smart grid architecture using 5G mobile network. . . . . 17
3.2 The figure illustrates the stages that contribute to control plane latency . . . . . 19
3.3 End-to-end packet transmission latency contributions . . . . . 19
3.4 Network caching architecture. . . . . 23
3.5 Solutions for 5G RAN and core networks to achieve low latency. . . 24
3.6 Figures (a) and (b) depict the architectures of core networks for SDN and NFV. . . . . 27
3.7 Figures (a) and (b) illustrate 4G and 5G physical resource block. . . 32
3.8 End to End scenario in this research provided by Vodafone Italia. . 35

4.1 E-UTRAN architecture. . . . . 37
4.2 EPC Architecture. . . . . 38
4.3 The architecture of the NR radio interface protocol regarding physical layer. . . . . 40
4.4 5G overall network architecture. . . . . 42
4.5 RAN-level interworking architecture. . . . . 43
4.6 Core network-level interworking architecture. . . . . 44
4.7 5G system architecture with reference points. . . . . 45
4.8 Left and right Figures illustrate received power and path loss based on Friis equation. . . . . 47
4.9 Uplink scheduling procedure. . . . . 47

5.1 Tested networking including 5G CPEs. . . . . 53
5.2 MEVO internal architecture . . . . . 53

5.3 Packets dropping figures of 1st to 4th measurements. . . . . 56
5.4 Packets dropping figure of 5th to 8th measurements. . . . . 57
5.5 Received packets figures of 1st to 8th measurements. . . . . 57
5.6 Received packets distribution over first measurement duration. . . . 58
5.7 Received packets distribution over second measurement duration. . 59
5.8 Received packets delays distribution over third measurement duration. 59
5.9 Received packets delays distribution over fourth measurement duration. 60
5.10 Received packets delays distribution. . . . . 61
5.11 Received packets delays distribution. . . . . 62
5.12 Delays average of each measurement round. . . . . 63
5.13 Delays average of each measurement round. . . . . 63
5.14 Mean values variation over time, mean values are calculated over each 1-hour time slot for remote hosts during measurements. . . . . 64
5.15 Mean values variation over time, mean values are calculated over each 5-minutes time slot for remote hosts during measurements. . . 65
5.16 Mean values variation over time, mean values are calculated over each 1-minute time slot for remote hosts during measurements. . . . 65
5.17 Delays ECDF of 10.149.0.2. . . . . 66
5.18 Delays ECDF of 10.149.0.1. . . . . 67
5.19 Delays PDF. . . . . 67
5.20 RTT. . . . . 68
5.21 RTT – Day by day. . . . . 69
5.22 RTT – Jitter. . . . . 70
5.23 1 packet is generated every 100 ms (6 hours), average bandwidth = 11.24 Mb/s. . . . . 71
5.24 Packet loss. . . . . 72
5.25 Packet loss Day by day. . . . . 73
5.26 The graph on the left shows the uplink and downlink bandwidth obtained for CPE 1 (Serial No.7JK7N19614002044), previously called 10.149.0.1, and the one on the right shows the results of CPE 2 (Serial No.7JK7N19614002045), previously called 10.149.0.2. . . . . 75
5.27 The graph on the left shows the ping-jitter obtained for CPE 1 (Serial No.7JK7N19614002044), previously called 10.149.0.1, and the one on the right shows the results of CPE 2 (Serial No.7JK7N19614002045), previously called 10.149.0.2. . . . . 75
5.28 The graph on the left shows the RSSI, RSRP, RSRQ, SINR obtained for CPE 1 (Serial No.7JK7N19614002044), previously called 10.149.0.1, and the one on the right shows the results of CPE 2 (Serial No.7JK7N19614002045), previously called 10.149.0.2. . . . . 76
5.29 The figure illustrates CPEs five different locations on the Politecnico di Torino map during measurements. . . . . 76

5.30 The upside graph shows the ping-jitter values obtained for CPE 1 (Serial No.7JK7N19614002044) (IP:192.168.8.1), previously called 10.149.0.1, and the one on the downside shows the results of CPE 2 (Serial No.7JK7N19614002045), (IP:192.168.9.1), previously called 10.149.0.2. . . . . 77
5.31 The upside graph shows the bandwidth values obtained for CPE 1 (Serial No.7JK7N19614002044) (IP:192.168.8.1), previously called 10.149.0.1, and the one on the downside shows the results of CPE 2 (Serial No.7JK7N19614002045), (IP:192.168.9.1), previously called 10.149.0.2. . . . . 78
5.32 The upside graph shows the RSRQ, RSRP, RSSI, SINR values obtained for CPE 1 (Serial No.7JK7N19614002044) (IP:192.168.8.1), previously called 10.149.0.1, and the one on the downside shows the results of CPE 2 (Serial No.7JK7N19614002045), (IP:192.168.9.1), previously called 10.149.0.2. . . . . 79
5.33 RSSI range definition . . . . . 80
5.34 SINR range definition . . . . . 81
5.35 RSRQ range definition . . . . . 81
5.36 RSRP range definition . . . . . 82

# Acronyms

**AS**
Access Stratum

**AR**
Augmented Reality

**AMF**
Access and Mobility Management

**BLER**
Block Error Rate

**D2D**
Device to Device

**DRX**
Discontinuous Reception

**eMBB**
Enhanced Mobile Broadband

**eNB**
Evolved NodeB

**EPC**
Evolved Packet Core

**FFT**
Fast Fourier Transform

**GP**

Guard Period

**GGSN**

Gateway GPRS (General Packet Radio Service) Service Node

**HARQ**

Hybrid Automatic-Repeat-Request

**QPSK**

Quadrature Phase Shift Keying

**IoT**

Internet of Things

**IP**

Internet Protocol

**ITS**

Intelligent Transportation System

**IFFT**

Inverse Fast Fourier Transform

**ITU**

International Telecommunications Union

**MEC**

Mobile Edge Computing

**MAC**

Medium Access Control

**MCS**

Modulation and Coding Scheme

**MME**

Mobility Management Entity

**MIMO**

Multiple Input Multiple Output

**NFV**

Network Function Virtualization

**NAS**

Non-Access Stratum

**NR**

New Radio

**SDN**

Software-Defined Networking

**SINR**

Signal to Interference and Noise Ratio

**OFDM**

Orthogonal Frequency Division Multiplexing

**OFDMA**

Orthogonal Frequency Division Multiple Access

**OWD**

One-way delay

**PRB**

Physical Resource Block

**PUCCH**

Physical Uplink Control Channel

**RAN**

Radio Access Network

**RSSI**

Received Signal Strength Indicator

**RSRP**

Reference Signal Received Power

**RSRQ**

Reference Signal Received Quality

**SC-FDMA**

Single Carrier Frequency Division Multiple Access

**SGW**

Serving GPRS Gateway

**TCP**

Transmission Control Protocol

**TDD**

Time Division Duplex

**TTI**

Transmission Time Interval

**UTMS**

Universal Mobile Telecommunications System

**URLLC**

ultra Reliable Low Latency Communication

**UDP**

User Datagram Protocol

**UPF**

User Plane Function

**5G**

Fifth Generation Mobile Network

**4G**

Fourth Generation Mobile Network

**3GPP**

3rd Generation Partnership Project

**VR**

Virtual Reality

**VM**

Virtual Machine

# Chapter 1

# introduction

The introduction of new services and applications over mobile networks has drastically increased the request for ultra-low latency services with higher reliability and capability for massive connection and enhanced energy efficiency. To address this unprecedented demand, 5G has emerged promising features to meet all of the requirements. A primary case in this regard would be ultra-reliable low latency communication (URLLC). To meet all the needs of mission-critical applications such as smart grid, remote surgery, intelligent transportation, and industrial Internet, a set of new features and components have to be designed. Furthermore, several modifications have to apply in mobile architecture concerning previous mobile generations [**1**],[**3**],[**8**]. Regarding 4G LTE under 3GPP release 14, latency is currently at about four milliseconds. URLLC has been included in release 15, and the latency target has been considered at one millisecond. URLLC satisfies the requirements for application requested end-to-end security. Moreover, in URLLC, packet delivery time has been bounded strictly. These unique features demand novel approaches towards designing, operating, and mobile wireless technology [**3**],[**25**]. The physical layer would be an integral part of the system since low latency and ultra-high reliability are contrary. Therefore, a vast array of different quality of service (QoS) has to be satisfied to meet both requirements. Studies have illustrated that the number of applications with high communication performance and reliability requirements has grown substantially. For instance, high-speed trains, autonomous vehicles, robots, and drones are multiple examples where wireless should satisfy the need for high reliability. Based on studies and actual measurements over high communication performance and reliability, it has been rolled out that packet drop rate and latency should be at around $10^{-5}$ and one milliseconds, respectively [**1**], [**14**]. 5G leverages the combination of URLLC and enhanced mobile broadband (eMBB) services to meet all of the mentioned requirements. The primary issue in this regard is the one millisecond target of the end-to-end network latency, which has to be satisfied. In the following figure, a general definition of end-to-end latency

has been introduced. In most previous studies in this field, the end-to-end latency time interval, including the time that a packet traverses from application processing at the device modem to the application processing at the base station modem, Fig. 1.1 [**1**], [**3**].

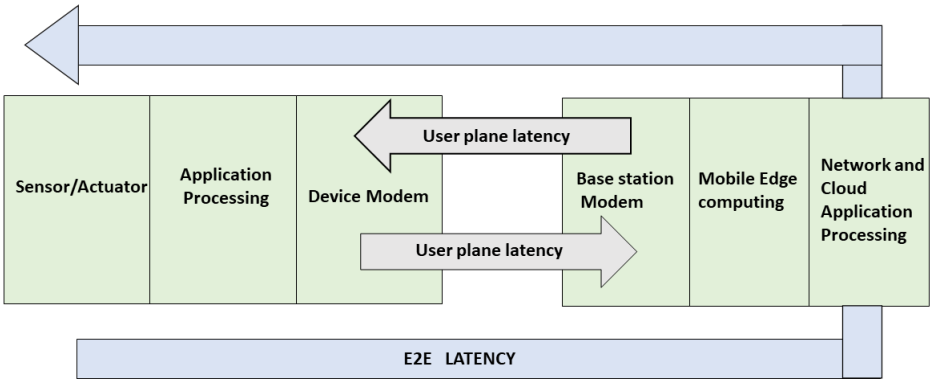


**Figure 1.1:** A packet's end-to-end latency time interval in a 5G network..

5G provides a wireless access solution for broadband communication, ensuring low latency for mission-critical communication (MCC). In general, the imposed latency in the network is generated by the radio access network (RAN), core network, and the backhaul between the core network and RAN. 5G mobile network consist of new entities such as software-defined network (SDN), mobile edge computing (MEC)/caching, and network virtualized function (NFV), which are the main ingredients used to reduce latency [**13**],[**14**],[**28**].

The mentioned entities lead to the capability of operation under desired latency boundaries and independence from hardware functionality. Moreover, new radio access with shorter interval transmission time, smaller packet size, new waveform, and new modulation and coding schemes are areas where low latency can be investigated. Additionally, optimization of radio resource allocation, massive MIMO, the priority scheduling of data transmission, and carrier aggregation in millimeter-wave are the subjects that have to be considered in this context. In summary, the efficient and fast deployment of 5G can be realized by integrating available LTE with new components. Although the prospect of 5G is ambitious according to the 4G viewpoint, a wide variety of research has been conducted to realize the 5G key performance indicator (KPI) goals, including low latency [**1**],[**3**],[**31**].

To concentrate on latency improvement and evaluation in the 5G and following generations, it would be fundamental to have a comprehensive perception of the latency concept in the currently existing mobile networks as the first step of the study. Furthermore, determining the types and the proportion of each component's effect on the overall latency of the mobile architecture is of capital importance. The

network parameters such as packet rate and network load are other performance figures that have to be concentrated on to evaluate how they affect the total latency of the network. Investigating the approaches that reduce latency in the currently existing mobile network is another crucial factor that enables us to move towards ultra-reliable low latency communication [**9**].

Most of the previous studies that have been performed on the latency evaluation in the 5G networks are based on analytical estimation. It is worth noting that no consolidated study conducted measurements over an actual 5G mobile network considering different factors affecting network latency. Most of the previous studies used the ping command to measure packet latency; in this study, the ping command has been used as a packet latency measurement tool. Through this command, we would be able to measure the RTT packet latency. The downside regarding the ping command is that we assume the uplink and downlink direction are symmetric concerning the latency. However, results have been illustrated that this assumption does not affect the measurement performance considerably. Chapter 3 will present a complete review of related studies on this subject. In the following, the set of targets that in the thesis will be followed is introduced [**12**],[**13**],[**44**].

In this study, we have attempted to measure the latency in an emulated network in conjunction with an existing mobile network. The measurement and evaluation of collected results give us an insight into the contribution proportion of each component over the network architecture. in this study, to conduct a wholesale research over this subject, we have introduced several scenarios for measurements. Various network parameters and application conditions have been included. First, The effect of the network traffic volume in the mobile network. Second,the effect of packet sizes and packet rates. In this thesis, the measurement setup is a network prototype implemented over a virtual machine with standard 5G devices (Huawei CPEs provided by Vodafone italia) equipped with tools and features for measuring and evaluating the actual network forming the emulated mobile network. Then, using the mentioned measurement setup and collecting the results, we will explain the reasons for the detected delay in the various scenarios. The rest of this thesis has been organized as follows. In chapter 2, the different definitions regarding the delay in the mobile network have been presented. Furthermore, in this chapter, the proposed latency targets of other releases of LTE and 5G have been studied. Then, we provide the advantages and disadvantages of various latency definitions mentioned earlier. Chapter 3 will provide this research objectives and the definitions of related concepts. Moreover, this section offers the descriptions of various source of latency in a mobile network and a complete review of related studies on delay evaluation in mobile networks. Then, the limitations of previous attempts will be investigated, and this thesis's methods to compensate for the earlier studies are presented. In chapter 4, 4G and 5G mobile networks architecture and their components are presented. In chapter 5, the measurement setup components are

introduced. This chapter also explains the different functions and their integration to shape the measurement setup. Furthermore, results discussion, research conclusion, and future extensions of this research are presented.

# Chapter 2

# Latency definition in a mobile network

In this study, our objective is to provide detailed research on latency in a 5G network. Therefore it is essential to have an insight into the different aspects of latency. The outline of this section is organized as follows. Section 2.1 provides the various definitions of latency in other sources such as standards, literature, and whitepapers. In the next section, the disadvantages of these definitions will be presented. In Section 2.3, latency targets are presented. The last section offers the definition of latency that satisfies the proposed demands over this thesis [**8**],[**20**],[**28**].

## 2.1 Various latency definitions

This section discusses latency definitions found in 3GPP documents. According to 3GPP, latency in the network can be divided into a control plane and user plane latency. Based on the definition, control plane latency is referred to as idle mode. The idle mode is the next stage after the inactivity period to reduce users' power consumption. In this stage, RRC disconnected users will listen to paging signals one time each paging round. This technique might reduce the power consumption of users' equipment. To enable users to receive and transmit packets, their mode has to be changed from idle to active if their current mode is idle. In the active mode, RRC connected users listen to the paging signal to receive data on the downlink more frequently than in idle mode at around every millisecond. Therefore, the delay has been generated by transmission from idle mode to active mode, referred to as control plane latency. Moreover, Because the above transition procedure relies on both radio access and core network contributions, the control plane latency comprises both radio and core network delays, Fig. 2.1,[**4**],[**26**].

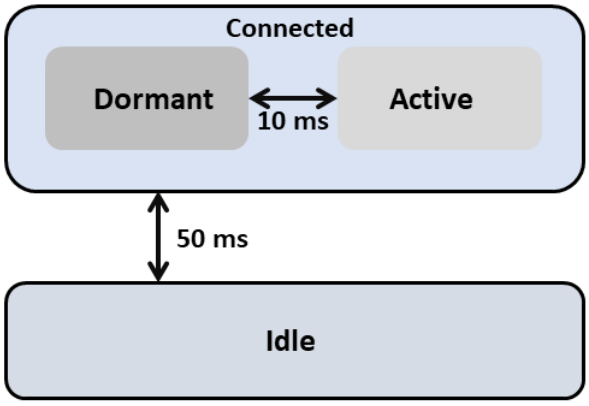


**Figure 2.1:** The control plane latency stages.

Based on 3GPP TR 38.913 version 14.3.0 Release 14 user 5G plane latency is defined as when an application layer packet is successfully delivered from the radio protocol layer 2/3 SDU input position to the radio protocol layer 2/3 SDU departure position through the radio interface, in both uplink and downlink directions, when neither device nor base station reception is limited by DRX. Alternatively, user plane delay is defined as the radio interface delay between when transmitter PDCP transmits an IP packet and when receiver PDCP successfully receives the IP packet. However, this does not provide information about the source and destination, which is the one-way delay between a source and its destination. Moreover, the end-to-end latency is not limited to the mobile network and includes other mobile or external networks that are involved in the communication path. In the following, we provide separate definitions for other latency metrics such as one-way latency, round-trip latency. FDD and TDD frame structures are used to analyze LTE user plane delay when analyzing 5G user plane latency, the same model may be re-used because it's sufficiently general, Fig. 2.2.

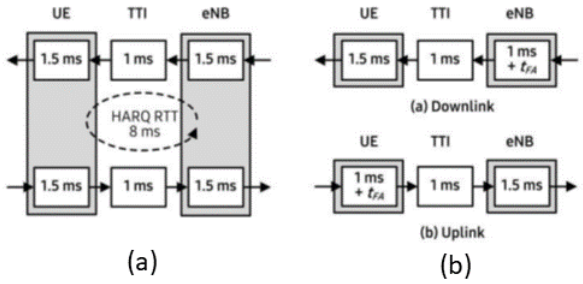


**Figure 2.2:** Figures (a) and (b) depict FDD and TDD LTE user plane latency frame.

- **One-way delay (OWD)**: the time it takes for a packet to travel across a network from source to destination, and the packet direction has to be specified beside its related delay value [**24**].

- **Round trip time (RTT)**: the time it takes to send a packet and receive an acknowledgment of that packet. RTT is also known as ping time and can be determined with the ping command [**26**].

- **Average end-to-end latency**: is defined as the average value of the end-to-end latencies over all the received packets [**26**].

- **Jitter**: this is defined as the variance of the end-to-end latencies over all the received packets.

- **Reliability**: is defined as the probability that the end-to-end latency is under the maximum allowable latency level. Here, it is derived by the Cumulative Distribution Function (CDF) of the end-to-end latency [**45**].

Regarding the definition of OWD, the destination node is usually located outside the mobile network, and RTT follows the same condition. One of the main issues concerning latency is that although many resources and definitions exist in this context, latency definitions provided by different sources contradict each other. Therefore, to avoid confusing the readers, we have tried to propose a unified definition of latency that covers all exiting definitions. However, the lack of general latency definition agreed by organizations and industries as a barrier for researchers working on this subject. The following figure depicts the various definitions of latency over the network, Fig. 2.3.

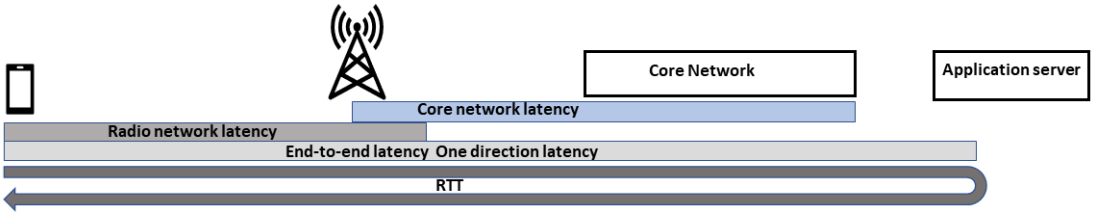


**Figure 2.3:** Latency visualization

## 2.2 Limitations of user plane latency definitions

Some other significant issues found in the latency definition provided by different sources are that they do not specify and consider the network parameters such as packet size and network configuration. the common assumption among all of the definitions is that all of the transmitted packets have no payload. For instance, based on the 3GPP definition of user plane latency, they assume an unloaded situation that a user with a single data stream and small packet size can achieve to user plane latency target. Moreover, it assumed that the user is synchronized with the network and can update network information. However, different data packets have varying payload sizes in the actual network, and network condition is not constant. It can be viewed that different types of network applications can significantly affect the data rate. To address this issue, the network configuration has to be modified continuously, which is not possible in most cases and leads to degradation of performance metrics of the network. In the case of synchronization, we should notice that, contrary to the assumption of latency definition provided by 3GPP, users may lose their synchronization with the network over time. Therefore,

the presented latency targets consider the lowest possible packet latency under predefined network conditions. Furthermore, 3GPP does not specify the type of latency used in its documents, such as average, minimum, or maximum. 3GPP papers do not consider the latency for large packets or networks with high load conditions, etc. Comparing these definitions with the results obtained from measurements in the actual network verifies the definitions of latency since latency varies packet by packet [**48**]. Consequently, based on all issues and reasons provided earlier in this text, for an accurate latency evaluation, we require knowledge about the latency experienced by the packets over the network considering larger packets, loaded network, higher packet arrival rate, and various user positions, etc. The latency definitions can be classified into the per-packet basis and network level. At the network level, the effect of several packets aggregated to form a KPI such as average network latency. It is worth noting that in both packet level and network-level scenarios, latency depends on network load condition, user location, etc.

## 2.3 Latency targets

If you're using LTE, the goal latency is fewer than 100 milliseconds for the switch from idle mode to active mode. It takes 50 milliseconds to go from inactive to active mode; these two latencies are considered to control plane latency targets. According to the specification of RRC connected status, the user can be put into the inactive mode to decrease the user's power consumption. In this state, the user listens to the paging channel more frequently than idle mode, and since the RRC connection was established, changing to active mode will be done faster. The user plane latency is specified at 5 ms. Regarding the LTE-Advanced, the latency target has improved drastically compared to LTE, which from 100 ms of LTE reduced to 50 ms of LTE-Advanced. In the case of transition from the inactive state into the active state, the same trend has been followed by decreasing from 50 ms in LTE to 10 ms in LTE-Advanced. However, user plane latency illustrates a stabilized figure at 5ms for both LTE and LTE-Advanced. Regarding the next mobile network improvement before 5G, so-called 4.5G or LTE-Advanced Pro, the user plane latency is less than 2ms. This considerable improvement is achieved by reducing the duration transmission time intervals (TTI) by using a smaller frame length. TTI for both LTE and LTE advanced is a fixed amount at 1ms while, depending on the application, LTE-Advanced Pro's latency might range from 0.14 to 0.50 milliseconds. According to 3GPP, the control plane latency of the fifth generation of the mobile network is less than 10 ms. The user plane latency is 0.50 ms and 4ms for both downlink and uplink for ultra-reliable low latency communications (URLLC) and enhanced mobile broadband (eMBB). In the following table, the summary of latency targets in different mobile generations

has been presented [**46**].

| **Mobile generation** | **User plane latency** | **Control plane latency** |
|---|---|---|
| LTE | 5 ms | Less than 100 ms |
| LTE-Advanced | 5 ms | Less than 50 ms |
| LTE-Advanced Pro | Less than 2ms | 10 ms |

**Table 2.1:** The summary of latency targets in different mobile generations.

## 2.4 Latency definition considered for this research

In this thesis, the round trip time of packets is considered as latency definition; this time interval includes the time that a packet reaches a destination and responds to the corresponding source. This thesis defines the source and destination nodes within the mobile network, implying no external network exists within the communication path between source and destination. As a result, this definition of latency conforms to RTT definitions available in literature and 3GPP's concept of end-to-end delay. The results of latency measurements should be understandable by readers, which gives them an insight into the meaning of latency and enable them to perceive the scale of measured latency and the network condition. Therefore we have considered four latency KPIs and our measurements are represented based on those. The considered KPIs consist of the average latency, per-packet latency, packet drop percentage, and packet jitter. The main reason for presenting packet drop rate as one of the KPIs is that it leads to latency, so it is required to evaluate this parameter to offer a realistic latency evaluation. Using these four KPIs of packet latency gives the readers complete insights into the packet latency distribution for each pre-defined scenario. In each pre-defined scenario, the network parameters such as the network load, user position, and traffic conditions are modified to show their effects in measurements results. The following three formulas have been suggested for the calculation of packet loss, packet loss ratio, and packet jitter where $N_T$ is the number of sent packets from source in the unit of time, $N_R$ is the number of received packets in the destination in the unit of time and $D_V$ is delay variations, Eq. 2.1, Eq. 2.2, and Eq. 2.3 [**24**],[**29**].

$$PacketLoss(PL) = N_T - N_R \tag{2.1}$$

$$PacketLossRatio(PLR) = \frac{100PL}{N_T} \tag{2.2}$$

$$Jitter = \frac{\sum D_V}{\sum N_R} \tag{2.3}$$

# Chapter 3

# Objectives and related concepts definition

This chapter defines our research objectives and provides a complete review of related work on this subject, including latency evaluation in 4G and 5G mobile networks. The rest of this chapter has been organized as follows. In Section 3.1, the major objective of this research has been discussed. The following section specifies a review of services presented over 5G mobile network and their requirements. Section 3.3 provides readers the primary latency resources in a mobile network, including process and components. In Section 3.4, a concise explanation of the restrictions and methods to attain reduced latency in a mobile network is provided. Sections 3.5 and 3.6 concentrate on RAN solutions for low latency and core network solutions for low latency. In Section 3.7, we study the factors affecting packet latency. The last section has been allocated to review the related studies on latency evaluation in 4G and 5G mobile networks and their advantages and disadvantages. At the end of this section, we present our evaluation method to describe the novelty and supremacy of this research compared to previous studies [**24**]

## 3.1 Research objectives

A mobile network architecture includes several components with various functionalities and procedures to transmit data from source to destination wirelessly. When a packet traverses such kind of network, experiences various delays in different components of this system. As mentioned previously, studying the delay of components is essential to have enough knowledge of end-to-end packet latency in a 5G network. Although this thesis uses a realistic 5G mobile network and 5G CPEs, the results are different from the actual latency figures of a 5G network since this research and previous studies perform their measurements under an idealistic

network condition. Generally, the reported network latency in such measures is lower than the actual latency experienced. The main reason is that multiple factors which degrade latency figures in a network are usually neglected in the lab environment. This knowledge will help us to enhance the network to achieve the desired latency targets. Therefore, we have categorized our objective into three classes. Realistic network's delay factors and causes are discussed in subsequent sections [**9**],[**11**].

- How and to what extent each network component contributes to overall latency.
- Causes of delay in the different components.
- Definition of several conditions that might impact network delay.

## 3.2 5G low latency services

Transportation, automatic manufacturing, robots, healthcare, entertainment, virtual reality, and education are examples of latency-critical applications. Intelligent home appliances such as smart fridges, thermostats, televisions, autonomous cars, sensors, drones, robots, intelligent wearable devices such as smartwatches, bracelets, and glasses, etc., are examples that require ultra-low latency networks to enhance our lifestyle [**15**],[**16**]. The existing mobile network such as 3G and 4G can not satisfy the requirements of IoT applications such as security, low latency, and high reliability. Therefore, implementing IoT applications over the next mobile generation is necessary to support the requirements for latency-critical services. In the following table 3.3, a summary of 5G applications and their needs have been provided.

### 3.2.1 Industrial automation

The main objective of industrial automation is to provide a synchronized and constant control over the processes, operations, and equipment, leading to an increase in production quality and speed with the lowest human involvement. The production procedure is a continuous process which is typically error-intolerant, and any error may leads to damages. Therefore this application demands a latency value of less than 10 ms and packet loss in the order of $10^{-9}$. Typically the automation in the industry includes data collection and then transmission of data for further processing in the programmable logic controller (PLC). It is worth noting that the proposed latency and loss rate is obtained through an empirical study, including conducting multiple surveys over a vast array of factories.

| **Mobile generation** | **Latency** | **Data rate** | **Note** |
|---|---|---|---|
| Industrial Automation | Less than 10 ms | Up to 1 $Mbps$ | Industrial Automation requires a low latency for its operations which in some latency-critical processes, a latency as less as 0.2 ms is desired. |
| Intelligent Transport Systems (ITS) | Less than 100 ms | Up to700 $Mbps$ | ITS requires a latency of 10 ms to provide safe roads; the other applications such as intelligent traffic control or virtual mirrors require a data rate of about 700 $Mbps$. |
| Robotics Industry | 1 ms | Up to 100 $Mbps$ | The robotic industry may require low latency and high bandwidth to perform sensitive and precise tasks and receive virtual haptic feedback. |
| Virtual Reality (VR) | 1 ms | Up to 1000 $Mbps$ | VR is one application that requests an extremely low latency and high data rate to provide High-resolution 360 VR. |
| Healthcare | Less than 10 ms | Up to 100 $Mbps$ | Remote diagnosis and remote surgery are applications in that latency plays a critical role in their performance quality. |

| **Mobile generation** | **Latency** | **Data rate** | **Note** |
|---|---|---|---|
| Education | Less than 10 ms | Up to 1000 *Mbps* | Education application includes several applications such as high-resolution 360 and haptic VR, tactile Internet-capable for human interface interactive which require extra data rate and lower latency. |
| Gaming | 1 ms | Up to 1 *Gbps* | New 3D and human interactive games with high visualization quality demand extremely low latency and high data rate compared to other applications. |
| Smart Grid | Less than 20 ms | Up to 1.5 *Mbps* | A delay of the range of one ms is required for adaptive activation and deactivation in the smart grid. Spatial awareness across a vast territory requires data speeds in the range of 1.5 *Mbps*. |

**Table 3.1:** The summary of latency targets in different mobile applications.

### 3.2.2 Intelligent transportation systems (ITS)

Intelligent transportation systems such as optimized traffic controlling and autonomous driving require ultra-reliable low latency communication. However, each application specifies its customized latency, packet loss, and data rate. To clarify, autonomous vehicles may require communication to perform actions such as platooning and overtaking. Therefore such types of operations need an end-to-end latency for data exchanging in the order of ten ms. Furthermore, assistive integrated video applications for autonomous driving purposes, (the so-called see-through-vehicle systems) require transmitting data with maximum delay of 50 ms. Optimized traffic

controlling including an intelligent warning system and a collision avoidance system to operate seamlessly may require the integration of local traffic information to be processed; the collection of local traffic data and transmission of those demands a latency up to 100 ms and packet loss rate of almost $10^{-5}$ [**20**].

### 3.2.3 Robotic industry

Remote-controlled robots will become more common in the future, with applications in a variety of fields, including construction and system maintenance in hazardous environments. Remote control with real-time visual-haptic input is essential for robotics and remote monitoring applications. In such kind of system, the feedback latency should be in the order of a few milliseconds. 5G network can provide the desired latency, loss rate, and data rate for such kinds of usages.

### 3.2.4 Virtual reality (VR)

Object manipulation is critical in micro-assembly and remote surgery applications, demanding extreme sensitivity and precision. VR technology includes services that enable users to communicate through a physical shared haptic environment—coupled with virtual reality. However, the existing mobile network can not meet the latency and data rate of VR technology for seamless and stable connection among users. For instance, typically, in VR technology, a display with refresh rates of 1000 $Hz$ is required to visualize information and physical simulation. Therefore a latency of 1ms for the round trip is demanded to provide a stable and seamless visualization for all users in shared communication [**3**], [**31**].

### 3.2.5 Augmented reality (AR)

Improved maintenance, museum guides, healthcare, remote education, and assistive technologies for security and emergencies can be supplied through the enrichment of information into the user's field of view. In this case, latency in the order of milliseconds is acceptable. However, a lack of computing power on mobile devices and the latency of typical mobile networks impede the applications' performance [**3**],[**31**].

### 3.2.6 Health care

Regarding the healthcare applications of ultra-low latency networks, remote-surgery, remote-rehabilitation, and remote-diagnosis can be mentioned. An ultra-low latency network with an adequate data rate that meets the requirements of healthcare applications can offer the health care services such as remote physical examination.

Moreover, a primary use case of ultra-low latency is remote surgery using robots and checking patients' status remotely. An average round-trip latency of fewer than ten milliseconds and reliable data transfer is needed [**31**].

### 3.2.7 Gaming

These days the gaming application is not restricted only to entertainment. Many innovative game paradigms, including critical-thinking challenges and hard-driving motivation, have use cases in education, athletic training, and process simulation such as flight simulators and healthcare. If the maximum latency exceeds 50 ms, gaming cannot provide a stable quality without degradation over time. Therefore, to satisfy the latency requirement of gaming applications, a round trip latency of about 1ms is desirable [**31**].

### 3.2.8 Smart Grid

It is an electrical grid that comprises sophisticated metering infrastructure, smart distribution boards, circuit breakers, load control switches, and renewable energy resources. Utility-grade fiber broadband is needed for smart grids to connect and monitor, with wireless as a backup. This means that the specified delay from start to finish is less than 20 milliseconds. A smaller latency is necessary, however, for synchronous co-phasing power supplies, Fig. 3.1 [**41**].

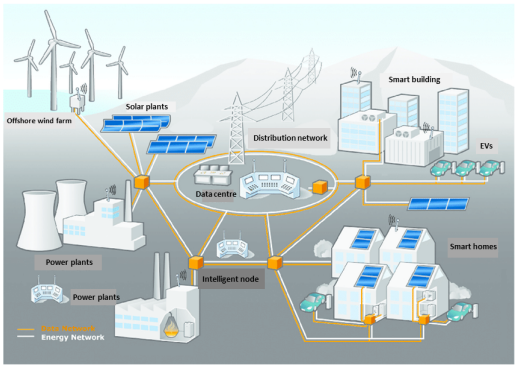


**Figure 3.1:** A typical smart grid architecture using 5G mobile network.

### 3.2.9 Education and art

One of the main applications of ultra-low low latency tactile Internet is remote education using a haptic overlay of teachers and students. Latency of less than ten ms is required to have a perceivable visual, auditory, and haptic interaction of

human-machine interfaces. Another notable application of ultra-low latency in this field provides musicians an opportunity to play musical instruments remotely and coordinate together. In this case, the maximum allowable round trip latency is less than a few tens of milliseconds [**31**].

Based on the several significant applications of latency-critical services in 5G networks provided above, it can be viewed that the end-to-end required latency is less than 100 ms in all cases. To summarize, ultra-low latency applications in 5G networks, such as gaming, healthcare, and VR, require lower round trip latency than other applications at around 1ms and a data rate of about 1 $Gbps$. Furthermore, the applications such as industrial automation and smart grid can tolerate a higher amount of round trip latency and lower data rate in the order of one $Mbp$s. To achieve the mentioned data rates, in the case of 1 $Gbps$, a frequency band of at least 40 $MHz$ should be shared among nodes in each kilometer, while for data rates in the range of $Mbps$, the bandwidth of at most 20 $MHz$ would be sufficient. The supported spectral efficiency by 5G and LTE-Advanced are supposed to be up to 30 $bps/Hz$ and 15 $bps/Hz$. mm-Wave could be one of the primary choices for high bandwidth requirements. Next, we'll cover the significant causes of latency in a mobile network, starting with a brief overview [**16**].

## 3.3 Source of latency in a mobile network

The control plane (C-plane) latency and the user plane (U-plane) latency are the two significant sources of delay in an LTE system. In the control plane, signaling and control functions are carried out, whereas actual user data is sent in the user plane. User plane latency is defined as the required time interval for a packet to become accessible from the IP layer in an edge node of E-UTRAN to the IP layer in the node of the user and vice-versa. Control plane latency is defined as the required time for the user equipment to transit from idle to active state, (see Fig. 3.2 and table 4.1). Since the latency of communication and consequently the performance of many applications depends mainly on the user plane latency, it is considered as the main focus in the study of ultra-low latency communication. As the calculation approach of control plane and user plane latency was mentioned extendedly in the text once, we recall it again in brief. In the following table, we have summarized the latency value of each stage that has contributed to control plane latency. User plane latency includes RAN, core network, backhaul, and data center latency in a mobile network, (see Fig. 3.3) [**3**],[**17**],[**27**].

We want to present a formulation for one-direction latency in user plane based on descriptions provided earlier in the text. First, we show the equation parameter and their definitions. A packet transmission to an eNB is delayed by $T_{Radio}$ , which is caused by the physical layer transmission and the contribution of eNBs, users,

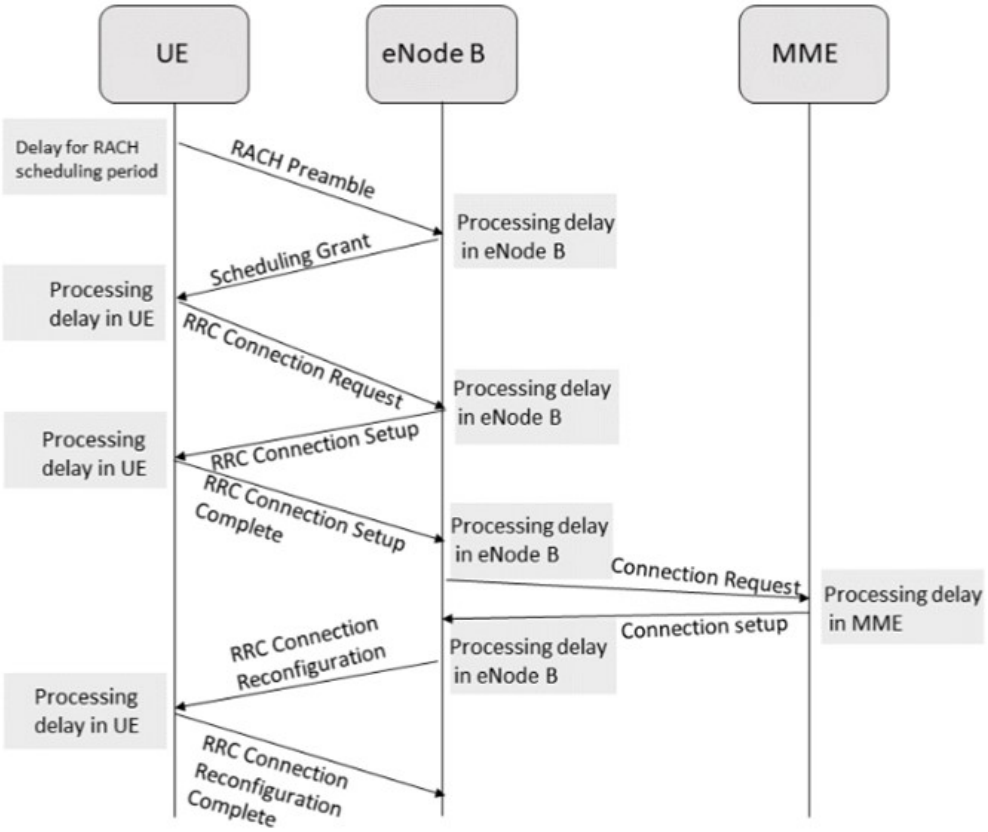


**Figure 3.2:** The figure illustrates the stages that contribute to control plane latency

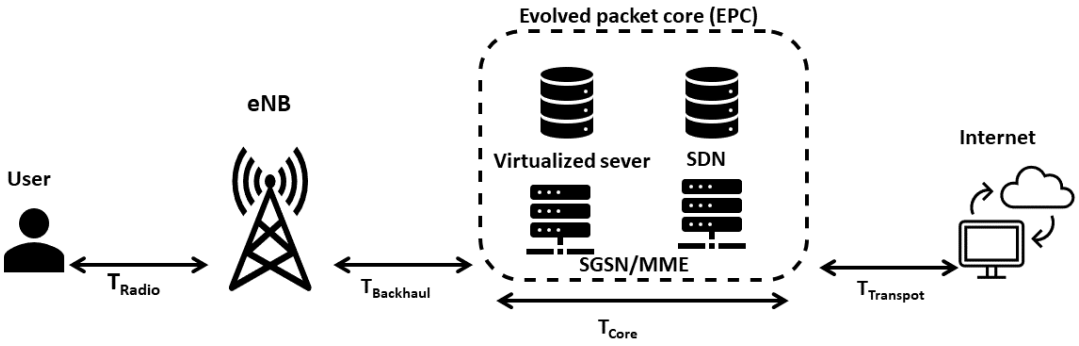


**Figure 3.3:** End-to-end packet transmission latency contributions .

| **Description** | **The Minimum Latency (ms)** | **The Averaged Latency (ms)** |
|---|---|---|
| RACH Scheduling Period Delay | 0.50 | 0.50 |
| RACH Preamble | 1 | 1 |
| RACH response | 3 | 5 |
| User Processing Delay | 5 | 5 |
| RRC Connection Request | 1 | 1 |
| eNB Processing Delay | 4 | 4 |
| RRC Connection Set-up | 1 | 1 |
| User Processing Delay | 15 | 15 |
| RRC Connection Set-up Complete | 1 | 1 |
| eNB Processing Delay | 4 | 4 |
| MME Processing Delay | 15 | 15 |
| eNB Processing Delay | 4 | 4 |
| Connection Reconfiguration | 1.5 | 1.5 |
| User Processing Delay | 20 | 20 |

**Table 3.2:** The latency value of each stage has contributed to control plane latency.

and the environment. $T_{Radio}$ consists of transmission time, processing time at use equipment and eNB, retransmissions, and propagation delay. Let's focus more in detail on $T_{Radio}$ components; processing delay at downlink includes channel coding, rate matching, scrambling, cyclic redundancy check (CRC), precoding, modulation, channel mapping, resource element distribution, and OFDM. The uplink processing delay contains rate matching, CRC, channel coding, data and control channel multiplexing, channel interleaver, code block segmentation, and concatenation [**3**]. Generally, $T_{Radio}$ is the sum of transmission and propagation delay and processing time, including channel estimation and first encoding and decoding. Propagation

delay is the parameter that depends on environment layout and the distance between user and eNB. Another parameter contributing to user plane total latency is $T_{Backhaul}$, i,e., the time interval for packet transmission between eNB and the core network. To link eNBs to the core network, copper cables, microwaves, or fiber-optic usually used as a means of communication. Among all options for connecting the core network to eNB, microwave allows for the lowest latency; however, there are some spectrum limitations regarding microwaves. The third parameter in the user plane's total latency formula is $T_{Core}$ defined as the processing delay imposed by the core network on total latency. $T_{Core}$ comprises many entities such as mobility management entity (MME), serving GPRS support node (SGSN), and SDN/NFV. Network-attached storage (NAS) security, evolved packet core (EPC) bearer control, inactive mode mobility handling and anchoring, users' IP address allocation, and packet filtering are part of the core network's processing stages. The fourth factor in the formula is $T_{Transport}$, the time interval for communication across core network and Internet which is strongly dependent on the distance between core network and server and bandwidth and communication protocol. As the user plane's total latency is considered one-directional, the end-to-end latency $T_{E2E}$ is obtained by 2T, Eq. 3.1 and Eq. 3.2 [**3**].

$$T = T_{Radio} + T_{Backhaul} + T_{Core} + T_{Transport} \tag{3.1}$$

$$T_{Radio} = t_Q + t_{FA} + t_{mpt} + t_{bsp} + t_{tx} \tag{3.2}$$

After formalization of user plane end-to-end latency, we want to formalize $T_{Radio}$ . As the first step, the parameters defined to cover all the mentioned latencies contribute to $T_{Radio}$ . First, we present $t_Q$, the buffering delay that relies on the number of users using the same resource. As the second parameter, we define $t_{FA}$, which is frame alignment leads to the delay that depends on the duplexing modes, including FDD, TDD, and frame structure. The third parameter is defined as the time for payload transmission and processing. The minimum required time-frequency resource is one TTI regarding $t_{tx}$ depending on available resources, radio channel condition, retransmission, and transmission errors [**3**], [**25**]. The last components of $T_{Radio}$ formula are $t_{mpt}$ and $t_{bsp}$, defined as user terminal and base station processing delays that rely on the user terminal and base station capabilities. Based on ITU specifications for ultra-low latency communication, $T_{Radio}$ is restricted to 0.5 ms. However, the configuration of 4G does not allow to achieve it as it is confined to 1 ms. To satisfy ultra-low latency communication latency requirements, radio transmission time $T_{Radio}$ should be in the order of microseconds. To address it, different aspects of RAN have to be improved. Modulation, frame structure, coding schemes, transmission techniques, new waveform designs, and symbol detection are the areas that can be focused on for RAN enhancement. $T_{Backhaul}$ is another

parameter that should be reduced to decrease the end-to-end latency of the user plane. This reduction can happen through the techniques such as fog-enabled networks and intelligent integration of access stratum (AS) and non-access stratum (NAS). Furthermore, $T_{Core}$ is another parameter that has to be optimized to reduce total latency, the approaches such as SDN, NFV can be used in this regard. The last parameter is $T_{Transport}$ that has to be considered as the optimization case. This dealy can be reduced by utilizing fog or MEC-enabled Internet, cloud, and caching. The following section discusses the limitations of the above approaches for latency reduction [**47**],[**48**].

## 3.4 Low latency barriers

As long as there is a trade-off among network quality indicators such as coverage, capacity, latency, and spectral efficiency, optimizing an indicator might affect the other indicators negatively. In the LTE, the transmission is organized into a 10 ms frame and TTI of 1 ms. The frame structure is one area that affects latency and depending on modulation and coding schemes for transmission rate adaptation along with constant control overhead. Overhead, including pilot symbols, cyclic prefixes, and transmission mode, has a considerable effect on latency. A large portion of transmission time is allocated to overhead at about 30% of the overall transmission time per packet. Therefore, frame/packet structure designing has become an essential field in which to have a practical frame/packet structure; the radio transmission time should not be lower than 1 ms as we know 30% of overall transmission time is allocated to overhead. The shorter transmission time leads to lower occupancy for user data. Moreover, the extra latency for retransmission per packet transmission has to be considered since retransmission positively affects the packet error rate. Consequently, based on the descriptions and reasons mentioned above regarding the packet structure and transmission time effect on latency, enhancing those would be an excellent choice. In the following, five approaches have been represented for this purpose [**38**],[**39**],[**40**].

- In the new frame structure design, the control overhead has to be smaller to ensure shorter transmission time. Control overhead might be reduced through integration or removal of different control overhead sections such as resource allocation, procedures for user scheduling, and channel training.
- New transmission techniques and new waveforms such as filtered OFDM can reduce transmission delay by increasing spectrum utilization and decreasing packet error at first transmission and the need for retransmission.
- Data have to be scheduled by prioritizing latency-critical data over standard data for immediate dispatching.

- OFDM main features such as orthogonality and synchronization lead to the high side lobe of the LTE system spectrum. Although these features are desirable in multiple access, they need extra spectrum and power resources and degrade latency performance.

- Network caching implies storing regularly reached information in a location near to the requester by decreasing the mass of traffic on WAN links and overloaded Web servers, caching benefits to ISPs, enterprise networks, and end-users, Fig. 3.4.

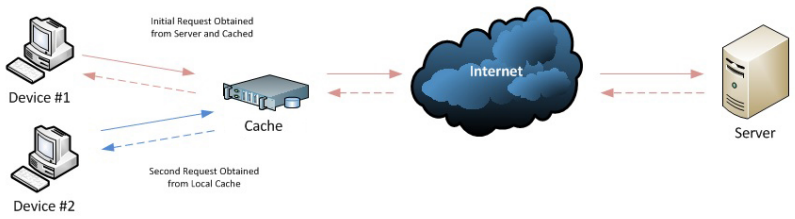


**Figure 3.4:** Network caching architecture.

In the following, we have proposed a chart including approaches for achieving low latency in a 5G network extracted from previous studies. We have divided the solutions into two significant sections in the proposed chart, RAN and core network solutions, (see Fig. 3.5, [**3**], [**31**]).

## 3.5 RAN solutions

RAN improvement is one of the critical solutions regarding ultra-low latency. The following table summarizes the proposed solutions for RAN improvement, such as advanced multiple access techniques designs, frame structure, diversity, antenna gain, etc, [**3**].

### 3.5.1 Backhaul network

Backhaul between the core network and base stations handle the data and signaling from the Internet and core network. The significant number of micro and macro cells support massive connectivity and capacity and latency-critical services in 5G. As a result, attaining low latency is constrained by the backhaul's capacity. According to the quality of the backhaul connection, copper, microwave, and fiber-optic lines are used. Compared to previous generations of mobile networks, there are several requirements for the backhaul of 5G networks, including increased

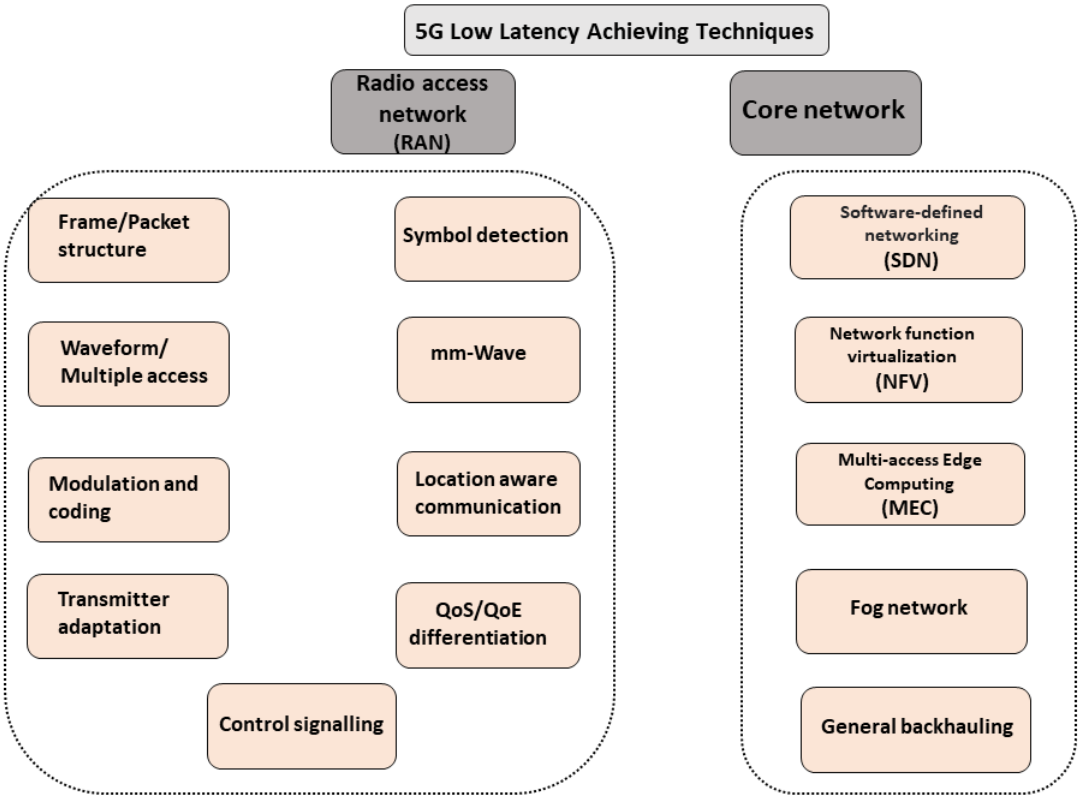


**Figure 3.5:** Solutions for 5G RAN and core networks to achieve low latency.

| **Technique** | **Description** | **Technique** | **Description** |
|---|---|---|---|
| Advanced multiple access/Waveform | Filtered CP-OFDM; UFMC; FBMC | Symbol detection | SM-MIMO detection scheme with ZF and MRC-ZF Linear MMSE Compressed sensing; Low complexity receiver design |
| Frame/Packet structure | Small packets/short TTI; Subcarrier spacing; TDD based OFDMA subframe; Physical subframe modification; Flexible subframe and resource allocation | mm-Wave | Multi-user massive MIMO; Beamforming gain |
| Modulation and coding | Polar coding; Turbo decoding with combined sliding window algorithm and cross parallel window (CPW) algorithm; Latency-critical data prioritizing; Balanced truncation | Location aware communication | Location information |
| Transmitter adaptation | Asymmetric window; Transmission power optimization; Path-switching method; Packet recovery method | QoS/QoE differentiation | Network parameters manipulation |

| Technique | Description |
|---|---|
| Control signaling | Control channel sparse encoding (CCSE); Scaled control channel; Radio bearer management; Outer-loop link adaptation (OLLA); |

**Table 3.3:** A summary of low-latency techniques in RAN [**3**].

capacity, greater security, effective coordination, greater flexibility, and reduced latency [**3**],[**15**],[**31**].

## 3.6 Core network solutions

Improved RAN and enhanced core networks are needed to ensure ultra-low latency in 5G networks. The new core network leverages a new network architecture equipped with new entities such as SDN, MEC, NFV, and new backhaul techniques. These improvements reduce the processing time of the core network and bypass several protocol layers [**3**],[**31**].

### 3.6.1 5G entities of core network

Assumedly, SDN and NFV will play a significant role in the 5G core network. Therefore, the 5G core network's latency can be reduced by using SDN and NFV technology. In Fig. 3.6, the architectures of NFV and SDN have been depicted where ONOS is open network operating system, APP is application, OSS is operations support systems forming, engineering, VNFM is virtual network function manager, EMS is element management system, VIM is virtualized infrastructure manager, and BSS is base station subsystem. Moreover, we have provided the list of general concepts underlying the 5G core network. The following table summarizes NFV and SDN techniques for low latency [**18**],[**28**].

- Virtualization and Network Function (NF) modularization.
- Unified Service Based Architecture and Interfaces.
- Control plane and user plane separation (CUPS).
- Mobility management and session management function decoupling.
- New Quality of Service (QoS) architecture for introducing the new services.
- Network slicing for supporting the new business domains.

- Minimize dependencies between the Access Network and the Core Network.
- Support a unified authentication framework.
- Support enhanced capability exposure.

| Technique | Note |
|---|---|
| SDN-based architecture | SDN-based architecture satisfies massive connectivity requirements, large throughput, and low latency in a 5G network |
| NFV-based architecture | NFV decouples physical network equipment from the NFs running on them, which leads to the deployment of EPC functions and the sharing of resources in the RAN. These features reduce end-to-end latency and enhance throughput performance. |
| Multi-access Edge Computing (MEC)/Fog Computing-based networ | Using MEC/fog, computation and storage are located close to the end-user, separating the data plane and control plane. |

**Table 3.4:** Overview of NFV and SDN technique for low latency [**3**].

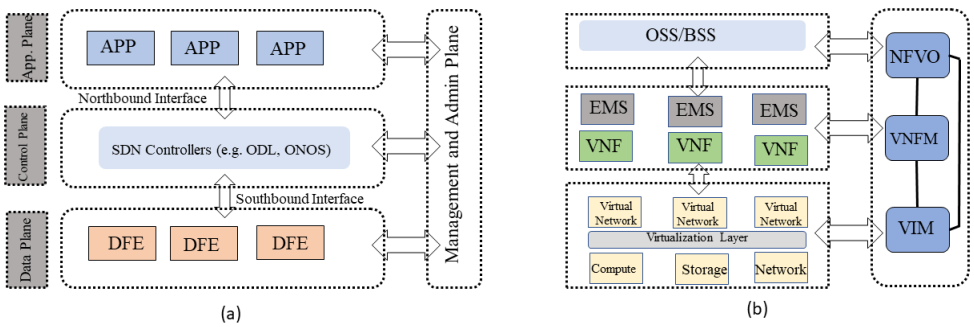


**Figure 3.6:** Figures (a) and (b) depict the architectures of core networks for SDN and NFV.

The EPC of LTE has several limitations that affect the total mobile network latency. It's important to note that with EPC, there is no complete separation

between the control plane and the data plane. For instance, packet data network gateway (PGW) and serving gateway (SGW) are dependent when they require a different network QoS. Therefore, It would be essential to separate the control plane and data plane. Remarkably, While the control plane requires low latency to process signals, the data plane demands high throughput to process the data. Consequently, SDN and NFV must be utilized in EPC to decouple the data and control planes [**3**],[**36**],[**47**].

The whole network parts are built using software running on Virtual Machines (VM), control plane, and user plane may be separated by utilizing SDN in EPC once the NFV based EPC has been modified. Between the separated planes, an SDN controller can function as an interface. Users' mobility and flow distribution flexibility are only a few of the benefits of SDN/NFV-based detachment of the user plane from the control plane; it may also minimize latency. Due to the reduced latency, this planes decoupling can enable mobile edge computing technologies to function more efficiently. It's also possible that the addition of an SDN controller will increase latency. Deploying several controllers can, on the other side, alleviate the issue of scalability. As a result, there is a cost-benefit analysis involved in the process. Consideration should be given to the scalability of controllers and the rise of latency while designing applications [**3**],[**25**],[**33**].

### 3.6.2 Latency improvement methods: related studies

LTE EPC's data plane is centralized, which adds to the list of limitations. End to end latency is increased even for users that just need to interact locally since their traffic is routed up the hierarchy to a small number of central PGWs. While the network's centralization facilitates operator administration and monitoring, it increases end-to-end latency, which is incompatible with applications including self-driving cars, smart grids, and factory automation. It's important to note that this sort of approaches result in poor system operation and extra delay, which are incompatible with 5G's goals and objectives. A more decentralized implementation of the network is now possible with the advent of new technologies such as cloud computing, fog networks, and mobile edge computing,NFV, and SDN. The network's CAPEX and OPEX can be significantly lowered by implementing such technology. The end-to-end latency can be considerably reduced by putting the main network components closer to the users. Because SDN/NFV allows the data plane to establish a decentralized MEC as the authors recommended SDN/NFV-based MEC networks in [**48**]. The SDN-based core network's mobility management may cause some delays in the core network. Authors describe in [**35**] how SDN-based mobility management systems might suffer from processing delays. A study suggested, that a carrier cloud architecture could be introduced with decentralized virtual machines (VMs) deployed in different locations [**25**].

The use of Follow-Me-Cloud has been proposed as a means of reducing end-to-end latency for consumers. The basic idea behind this notion is that all network components can maintain a record of the user's mobility, which translates into seamless connectivity and decreased end-to-end latency. "Soft Mow" was a proposal in the field for a hierarchical configurable network-wide control plane [**54**]. To achieve this, highly dispersed controllers are charged for supporting the network in their respective locations. The applicable delay limitation may determine the number of layers in this hierarchical architecture. According to the 5G core network, NFV is a crucial component of the core network. NFV reduces hardware platform reliance and allows for scalable resource sharing in RAN and EPC services. Throughput can be improved while the end-to-end latency is reduced through the 5G design based on NFV and SDN.

In [**33**],[**36**],[**37**], researchers present a network controller with distinct control and data planes, separated network functionality from hardware, and centralized network intelligence. To merge wireless network virtualization with network systems, a new information-centric method is suggested. Low-latency services are supported by critical parts, namely the wireless spectrum resource, the mobile network infrastructure, virtual resources, including content slicing, network-level slicing, flow-level slicing, and the information-centric wireless virtualization controller which found in the proposed architecture.

A related study introduces an EPC based on NFV, an EPC as a service to facilitate mobile core network management [**51**]. A virtual machine (VM) is used to represent the EPC's parts in this system. Another attempt presents EPC deployment as a service with fewer disadvantages, one of which is a delay increase across virtualized network functions and the EPC components [**52**]. According to [**53**], it is essential to separate the virtualized network operations into various groups/subsets based on interactions and responsibilities to decrease network latency. This distributed control plane can serve applications requiring high mobility and reduced latency because it could be located closer to users at the network edge. Network designers should consider regulations and billing enforcements while designing networks depending on user requirements. The authors suggested the optimization issue of computation, composing, and connecting virtualized functions as a part of the strategy for reducing overall delays, including network and computation delays. After an initial assessment, the optimization issue is a resource-constrained optimization problem on a supplementary multi-layered network [**55**].

In [**50**], comprehensive strategy for implementing a self-organizing network (SON) powered by large data have been proposed. As a result, a relatively more optimal SON may be implemented using machine learning and data analytics where eNBs' uplink transmissions are scheduled using the smart gateway (SM-GW). Simulated results show that the SM-GW scheduling may assign the balanced

bit rate in uplink transmission to the eNBs while minimizing packet delays [**50**]. Moreover, since a significant number of eNBs are linked to a single SM-GW, the traffic of fully loaded eNBs can saturate the SM-GW's buffer, leading to additional delay. SM-GW connections may be spread among eNBs, and preserving QoS with appropriate scheduling.

## 3.7 Factors affecting packet latency

In [**3**] , the authors stress the need to minimize control overhead for shorter packet transmissions. Then they presents criteria to assess the functionality of minimizing control overhead method and a detailed explanation of the fundamental principles underlying the transmission of small-scale packets with high reliability and low delay. [**31**] has been suggested a configurable 5G radio frame layout, in which the TTI size may be customized to fulfill the demand of certain services requirements. The TTI of 0.25 ms is a practical option for reduced delay at light loading due to the minimal control overhead. However, control overhead rises with higher traffic and impacts packet solutions, reliability, and latency. An efficient hybrid automatic repeat request (HARQ) solution with reduced transmission time interval and round-trip time improves URLLC's failure capacity and fulfills the low latency requirements of 5G mobile networks. For ultra-reliable low latency communications, it appears that a lot of simulations are being conducted to understand better the fundamental relation between failure capacity, bandwidth, and latency requirements [**26**],[**31**]. Regarding low latency 5G networks, the subframe structure and numerology are specified depending on a variety of frequency spectrum and bandwidths where the carrier frequency, sampling frequency, FFT size, cyclic prefix, and subcarrier spacing were determined. An SDR-based 5G framework with stringent latency requirements is described. The suggested subcarrier spacing has been widened to reduce the OFDM symbol runtime according to the author's suggestion in [**31**]. The number of OFDM symbols within every subframe has been constant in the redesigned frame structure for TDD downlink transmission. The subcarrier gap is adjusted to 30 kHz, yielding a symbol length T = 33.33 $\mu s$ for OFDM transmission. Assuming the sampling rate $f_s$ remains at 30.72 $MHz$ and frame duration Ts remains at ten milliseconds, the FFT size N is set at 1024, and $T_x$, and $R_x$ control resource blocks are distinguished from each other. The data resource block by guard periods allows highly flexible assignments of various control and data resource blocks in successive subframes. This yields the overall number of guard periods for each subframe, distinguishing the two different subframes. To estimate the subframe length, we may apply Eq. 3.3, using the same subcarrier spacing for the data plane and control plane, assuming identical $T_x$ and $R_x$ control sections with N control symbols each, Where $T_{symbol}$ is the length of an OFDM symbol, $N_c$ and

$N_d$ are the numbers of data symbols and control symbols, $T_{cp}$ and $T_{GP}$ are cyclic period and gaurd period [**56**], [**31**].

$$T_s = (2N_c + N_d(T_{symbol} + T_{cp})) + 2T_{GP} \tag{3.3}$$

As a result, radio resources are required to transmit the packet from the buffer. This means that each packet's latency may vary depending on how many resources a user has available at any particular time. Packet delay can be affected by factors that demand varying resource needs or parameters that influence resources allocation within every subframe interval ($T_s$) . Three characteristics might impact packet delay in a mobile network that has been investigated in this research. Each of these aspects will be described thoroughly in the subsequent section. As a result of this investigation. [**3**].

- Network Load
- Rate of packet
- Size of packet

When it comes to network load, both the radio and core networks are considered. Resources per user are reduced due to heavy load on the radio network. Data packets sent by a user may be delayed until the radio network allocates adequate resources. There is also a difference between the amounts of inter-cell interference in the network because of the decreased SINR. Lower modulation and coding scheme (MCS) might result in a reduced data rate and more significant packet delay. As a result of shared resources, packets are queued in the core network owing to differing load levels, resulting in packet delay [**21**]. In contrast with an ordinary core network, a core network equipped with edge computing would allocate network resources flexibly based on instant network load conditions. Therefore, the effect of network load on latency will reduce. The other benefit of using edge computing in the core network is that using this technology allows placing the core network near the readio network, leading to substantial reduction of transport latency and backhauling latency [**21**],[**23**].

Quantitatively, in terms of packet length and rate, various applications demand distinct traffic parameters. As a result, a more extended packet uses more network resources than a shorter packet. This means that the queueing delay between eNB and users increases for packets coming at a higher rate than those receiving at a lower rate. As a result, packets may face varying delays in accordance with the packet's length and speed. If the radio network employs a system of distinguished scheduling, a foremost priority is given to the marked user compared to other network users. The users may have multiple levels of priority ascribed to them. Packet latencies are decreased because the user has access to more radio resources

than if there was no prioritizing. Using a system of customized scheduling, the chosen user receives a greater priority in the radio network. The user may be assigned various levels of priority; as a result, latencies are reduced to a minimum value [**22**].

### 3.7.1 5G and LTE Frame structure

Downlink and uplink transmissions are organized into frames with duration $T_f$=10 ms, consisting of ten subframes of $T_{sf}$=1 ms duration. Each subframe is divided into a certain number of slots, and each slot is composed of 14 consecutive OFDM symbols. The slot represents the basic scheduling unit for slot-based scheduling and slot length in time scale with the subcarrier spacing [**11**], [**17**]. The number of the slot in a subframe depends on $\mu$, Eq. 3.4.Time in LTE is divided into 10 ms frames and each frame is divided into ten sub-frames or Time Transmission Interval (TTI) of 1 ms. Each TTI is composed of 14 OFDM symbols. A frequency-time grid of 12 sub-carriers (180 $kHz = 12x15KHz$) and 1 TTI (i.e., 14 OFDM symbol) is referred to as Resource Block (RB), and one sub-carrier in one OFDM symbol is named Resource Element (RE), Fig. 3.7 [**33**],[**34**].

$$Slotlength = 1ms/2\mu \tag{3.4}$$

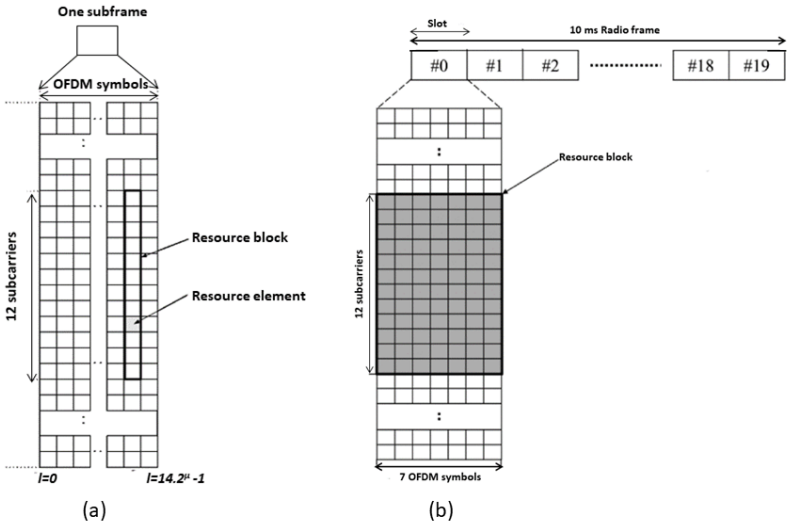


**Figure 3.7:** Figures (a) and (b) illustrate 4G and 5G physical resource block.

## 3.8 Latency measurement: related studies

Several publications have looked at the latency of 4G/LTE networks. For example, multiple studies examine user and control planes latency in LTE networks. However, control plane latency receives more significant attention [**38**]. As a result, the findings of these latency evaluations are based on estimates, not actual measurements. According to these studies, the user plane latency results are based on estimations and reflect the user's radio interface processing delay induced by the user plane protocol stack . As a result of these measurements, we can determine the minimum delay produced by the different sub-layers. There are not significant differences in processing delays between uplink and downlink. Therefore, these investigations indicate that total latencies between uplink and downlink are comparable. It's not reasonable to consider the same latency values for LTE's upper and lower layers since there is a difference across their signaling in LTE.

In [**58**] authors have suggested a framework for latency evaluation of a 5G network; in this study, they have tried to propose a method to define and evaluate latency as KPI of 5G communication. To obtain this goal, several scenarios and technical solutions have been presented by the authors. This paper has reviewed the prior studies, including their ecosystem, challenges, components, and testbeds. Moreover, in this study, a complete review of latency-critical services has been provided. However, this study is limited to a qualitative assessment of the previous research without real measurements or simulations and numerical results for further comparisons.

An actual network may have various results because these delays aren't standard but rather varying by operators. Additionally, the findings do not include the delay produced in either the delay in the core network or the radio network because of the various signaling methods in uplink and downlink. Apart from that, the researchers do not consider the multiple scenarios and factors affecting latencies that have been mentioned above. [**39**] illustrates that the ping tool is the most used measuring system as a latency analysis method in actual networks. For example, data packets are delivered to a target with the ping program, and the target responds with an acknowledgment. The RTT is calculated depending on the packet's transmission time and the acknowledgment from the receiver. For example, in [**39**], the authors describe the RTT measured on the network using ping for a user situated near and distant from the radio base. While working on it, they considered two different packet sizes, including 32 and 1400 byte for ping and RTT is measured for only the minimum, maximum, and average values.

[**3**] offered a study over the latency evaluation of 4G networks. This study has used a measurement setup emulating the effects of a realistic mobile network where the traffic is generated using Ostinato traffic generator. This study investigates packet traversing through various network components based on packet payload

content as a latency reduction solution. Furthermore, this study has used an emulator and degrader in order to simulate the actual network by reducing the performance of the simulated network. The effect of several latency reduction techniques has been investigated. However, the lack of using an existing 4G mobile network and direct focusing on a 5G mobile network is observed.

There are still many questions about how latency changes with packet size and user distance. As a result, it is impossible to determine what percentage of the whole packets examined had an inferior or greater latency than the mean value or what percentage had a delay near to maximum or minimum value. It's also important to note that the results don't consider latency independently in the uplink and downlink. Furthermore, the studies don't investigate how packet rate affects latency. In [**41**] RTT measurement using ping for packets of different sizes is discussed. Even though this research covers a wide range of packet sizes, this study does not provide a comprehensive picture of the relationship between packet size and delay. Moreover, this research does not consider the other aspects that might impact latency, as mentioned earlier. Furthermore, this study's outcomes do not consider the impact of latency specifically in the uplink and downlink in the radio and core networks. Transmission and reception of packets must be recorded simultaneously to measure packets' one-way latency in a network accurately. In addition, the capturing locations must be synchronized in accordance with time. In [**42**] the authors examine packets of different sizes to determine the one-way latency. In this case, the findings do not reflect the packet delay contribution made by each network component since there is no access to them, which is a disadvantage. It is also worth noting that results of this study consider potential delay contribution from an external network, affecting packet latency. This study lacks information on packet latency behavior; it only focuses on the maximum and minimum packet latency values. With complete network access, there are very few studies on packet delay evaluation. [**43**] offers a helpful scope of the user plane latency, aligning with the thesis's goals. Uplink, downlink, radio network, and core network one-way and round-trip delay are discussed for packets of varying sizes. Determining latency behavior throughout the whole traversed path is also included in the study results. Latencies are examined in this thesis to allow the readers to understand better how latency behaves by showing how latency varies over time—network load, packet size, and packet rate which impact round trip delay in the network. There are also extensive explanations of why the observed latency behavior occurred for each factor that affects latency. In the following figure, we have illustrated the end-to-end path that we have measured latency, including the latencies imposed by backhauling network, radio access node, and core network, Fig. 3.8.

[**60**] focused on designing 5G transport networks with ultra-low latency communications, including mobile backhaul networks, and supporting latency-critical

applications. To address it, two research areas have been investigated titled "PON-based mobile backhaul networks" and "service migration in fog computing enabled cellular networks."

In [**59**] a study over one-way Packet delay in a radio network has been performed using a simulator named CSIM-FsUE, which is a verification tool used to test and verify the performance of a base station. This study has formulated problems over parameters such as simulation parameters affecting quality of service (QoS), system behavior in terms of packet delay, and simulator at high data load effect on performance. This study has several restrictions, such as simulator software and hardware resources, that affect the overall performance of the simulation process.

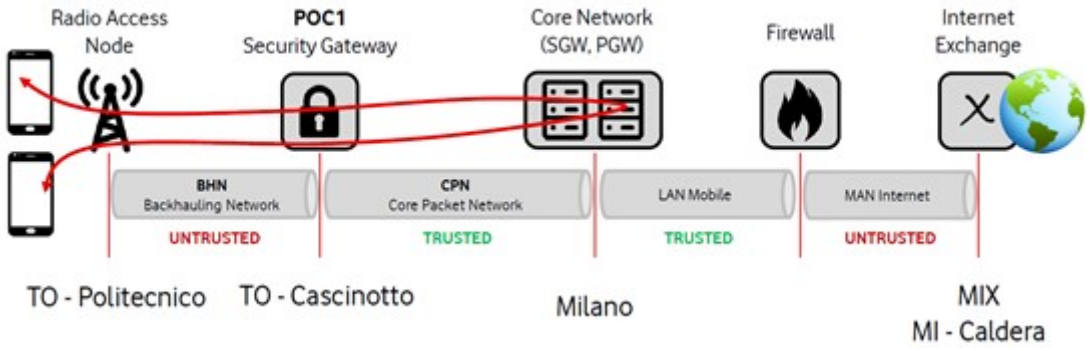


**Figure 3.8:** End to End scenario in this research provided by Vodafone Italia.

# Chapter 4

# 4G and 5G Mobile networks and their components

Mobile networks include the radio and the core networks with different processes in the two parts. A wireless connection sends packets from the user equipment to the radio base station, also known as eNodeB in LTE and gNB in the 5G. It is necessary to employ a wired media for packet exchange in the core network [**8**].

## 4.1 4G network architecture: E-UTRAN

The Evolved Packet System's access component is the Evolved Universal Terrestrial Access Network (E-UTRAN). Orthogonal frequency division multiple access (OFDMA) is used as the multiple access scheme in the downlink, and Single Carrier - Frequency Division Multiple Access (SC-FDMA) is used in the uplink. For the downlink, it uses MIMO with spatial multiplexing. Fig. 4.1 illustrates the architecture of E-UTRAN [**5**],[**6**],[**7**].

## 4.2 4G network architecture: EPC

Throughout this study, the evolved packet core is employed as the core network (EPC). As 3GPP's 4th generation, the EPC is selected as the core network. 4G systems were developed by the 3GPP using an IP packet switching network and EPC successfully evolves a packet-switched architecture such as that seen in GPRS/UMTS networks. Since LTE is a packet-switched network, there is no requirement for protocol transformation in the core network; contrasting GPRS and UMTS, fewer nodes in the network deal with users' traffic [**7**]. In comparison to the GPRS and UMTS core network architecture, the EPC is considered a flat layout.

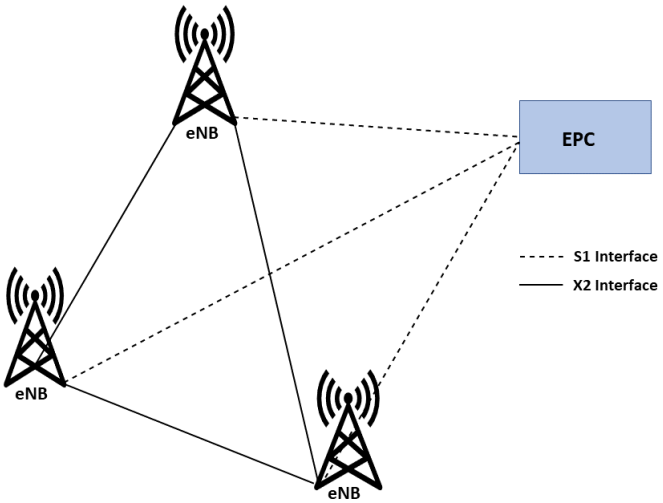


**Figure 4.1:** E-UTRAN architecture.

The core network might be scaled separately based on user plane or control plane requirements. The EPC's fundamental architecture is depicted in the diagram below, with the user linked through E-UTRAN. The EPC has consisted of four main network components, Fig. 4.2 [**12**],[**13**].

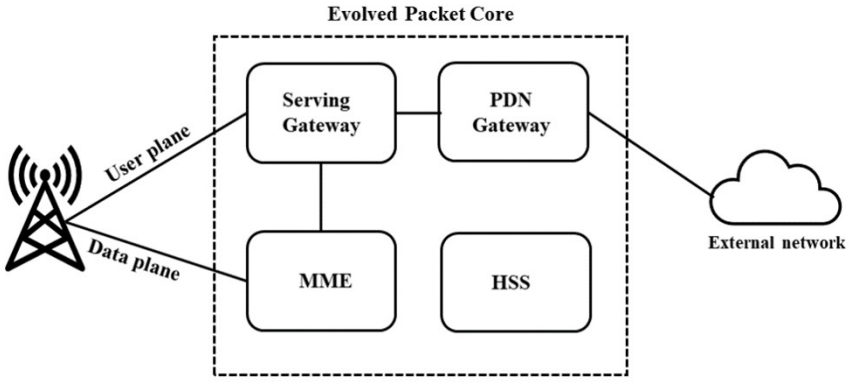


**Figure 4.2:** EPC Architecture.

- **Mobility Management Entity (MME)**: is the LTE core network's primary control node in charge of the control plane. It is responsible for signaling for the E-UTRAN's security and mobility management. It is also the node in charge of tracking user devices in the network and paging them for updated data, which is similar to the control plane part of SGSN in 3G networks [**13**].

- **Serving Gateway (S-GW)**: connects the network's radio side to the EPC and routes user plane IP packets to and from the user. Anchor point for inter 3GPP RAN mobility similar to the user plane part of SGSN in 3G network [**13**].

- **Packet Data Network Gateway (PDN-GW)**: interconnects the EPC with external IP networks. To and from external networks, it routes user plane IP packets. It also assigns IP addresses and prefixes to the users, as well as policy control, which is gateway to the Packet Data Network similar to GGSN in 3G.

- **Home Subscriber Server (HSS)**: save information about users and subscriptions in a database. The HSS carries out mobile management and user authentication in cooperation with the MME.

## 4.3 Delay of the mobile network

As has already been described, both the uplink and downlink packets are processed by the E-UTRAN user plane protocol stack. These numbers cannot be determined in a mobile network without measuring the processing delays, since they are operator-specific. These delays occur regardless of the direction in which packets are sent or any other parameters in the network. As seen below, the user plane protocol stack consists of a number of levels, (see Fig. 4.3).

- **Packet data convergence protocol layer (PDCP)**: Over the RLC layer, in the Radio Protocol Stack, there is PDCP. These services are provided to the RRC and user at higher levels, such as IP at the user and relay in base stations. The services including transfer of user plane data, transfer of control plane data, header compression, ciphering, integrity protection have been offered by PDCP.
- **Radio link control layer (RLC)**: To guarantee that higher-layer data is the right size for transmission over the air interface, RLC includes a segmentation and reassembly mechanism. When needed, the protocol can additionally offer concatenation and error correction.
- **Medium access control layer (MAC)**: Controls hardware responsible for interacting with cable, optical and wireless communication mediums; the data link layer comprises the MAC sublayer and the logical link control (LLC) sublayer. The LLC offers flow control and multiplexing for the logical connection for the data link layer. The MAC provides flow control and multiplexing for the transport medium [**12**], [**49**].
- **Physical layer (PHY)**: Interface between Medium Access Control (MAC), RRC, and the Physical Layer. The physical layer provides transport channels. In the radio interface, a transport channel determines how information is sent and received [**12**].

## 4.4 Scheduling latency

Due to a scheduling delays, packet latency may be further impacted. Latencies in the network might vary depending on the current cell load, the user's position in the cell, and the uplink and downlink scheduling method. For the limited network resources to be allocated and shared among the network users, a scheduler is employed in the user and eNB buffer. Therefore, it does not allow a packet to be sent unless adequate radio resources are provided. Latencies are consequently

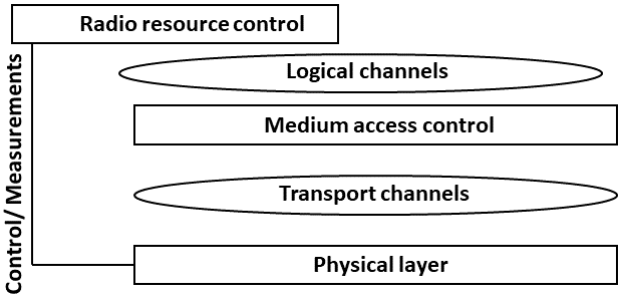


**Figure 4.3:** The architecture of the NR radio interface protocol regarding physical layer.

affected by the rate of allocated radio resources. Scheduling is influenced by the current network load, the user's location inside a cell, and the scheduling method. In the resource grid, the smallest data-carrying unit is termed a resource element, which consist of a subcarrier and one OFDM symbol. Because the quantity of data transported by a resource element depends on SINR, as a result of a low SINR, which indicates a poor radio channel, less data is sent by a resource element.

QPSK, a lower-order modulation method than 64 QAM, can transport fewer data per resource element because it is less error-prone than higher-order modulations. A Block error rate (BLER) of about ten percent is required to receive LTE signals. To achieve this goal, it is necessary to select acceptable transmission parameters. Due to a worse signal-to-noise ratio (SINR), a more robust, lower-order modulation will be used to guarantee that the desired BLER is maintained. The modulation and coding system (MCS) of LTE is changed to accommodate link adaptability capability. Because of this, a particular MCS is allocated to a user based on channel quality estimation. Additionally, the spatial multiplexing, transmission modes, and transmit diversity can also be modified to fulfill the BLER objective. During a transmission time interval, a user is assigned to radio resources and users may be assigned to different packet resource blocks and TTIs, based on the scheduling method used to assign them. Consequently, the amounts of data that may be transmitted in a TTI are determined by the number of packet resource blocks allocated at that moment, the MCS, and the transmission method.

## 4.5 5G network architecture

The 3GPP new radio access network (NG-RAN) access can be provided to capable users through different access technologies. An NG-RAN node is a gNB, providing new radio (NR). User plane (UP) and control plane (CP) protocol terminations towards the UE, or an ng-eNB, providing E-UTRA (i.e., LTE-like). User plane and control plane protocol terminations towards the user gNBs and ng-eNBs are interconnected via the $X_n$ interface. NG-RAN nodes connect to 5G Core network (5GC) nodes known as access and mobility management function (AMF) and User plane function (UPF) via NG-Cand NG-U interfaces, respectively. AMF is for control signaling to/from UE, similar to MME in EPC, but is only responsible for UE registration, reachability, connection, and mobility. It performs access authentication and authorization, Fig. 4.4 [**14**],[**35**].

- **AMF**: Handling UE data sessions (i.e., session establishment and management) in charge of the Session Management Function (SMF), formerly part of MME. It supports the allocation of UE IP addresses, UPF selection, and control [**31**].
- **UPF**: responsible for data plane handling, including providing an anchor

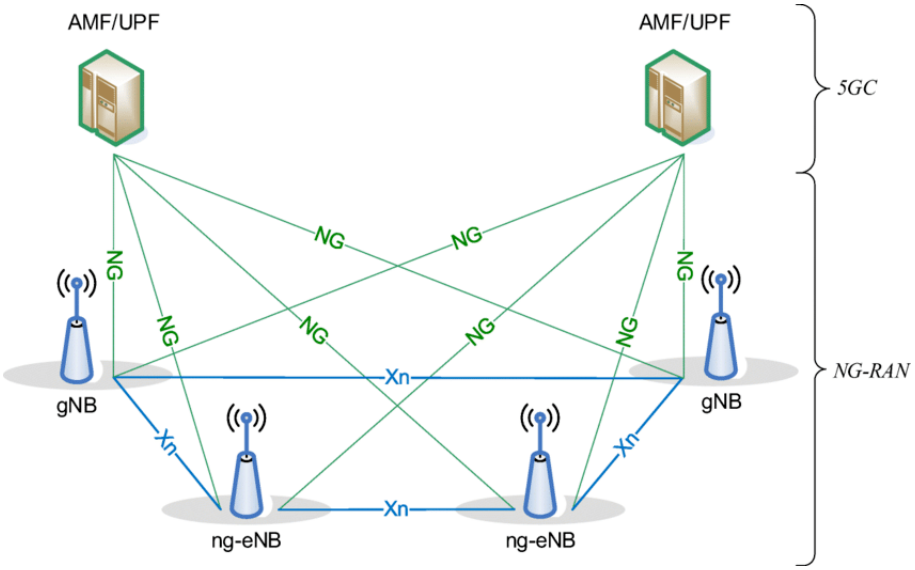


**Figure 4.4:** 5G overall network architecture.

point for mobility such as packet routing and forwarding functions, previously performed by S-GW and P-GW in EPC [**31**].

A flawless service may be distributed to the customers even if 5G cells are not fully deployed. This can be accomplished by interacting with the existing LTE network, which is already fully deployed. the transition from 3G to LTE, LTE cells were installed with limited coverage when LTE first became commercially available. 3GPP is now debating two types of solutions for 4G-5G interworking: RAN-level interworking and CN-level interworking.

- **Interworking, (RAN-level)**: a LTE eNB and 5G base station interface is used to enable interworking services between LTE and 5G at the RAN level. In NSA (Non-Standalone Architecture), when 5G Radio (NR) cannot be used without LTE Radio, RAN-level interworking is required, Fig. 4.5 [**10**], [**14**].

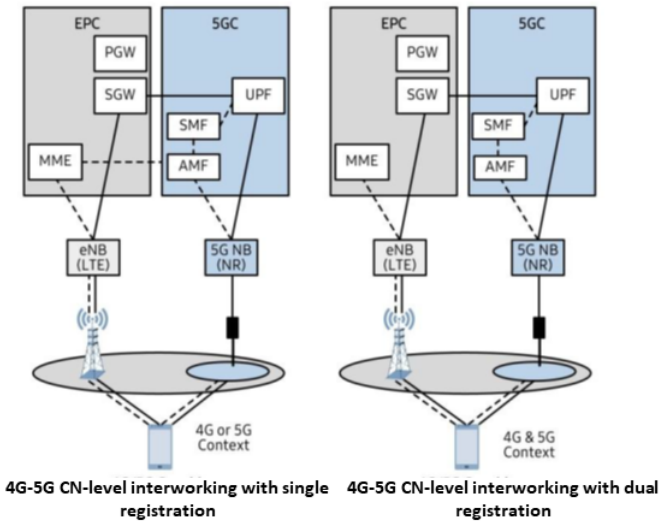


**Figure 4.5:** RAN-level interworking architecture.

- **Interworking, (core network-level)**: a direct communication between the LTE eNB and the 5G NB isn't required for CN-level interworking, but the EPC entity is connected to the 5GC entity. When using 5G Radio (NR) without LTE Radio, CN-level connectivity is required in SA (Standalone Architecture). Dual and Single Registration are both options for CN-level connectivity. Fig. 4.6 [**10**], [**14**].

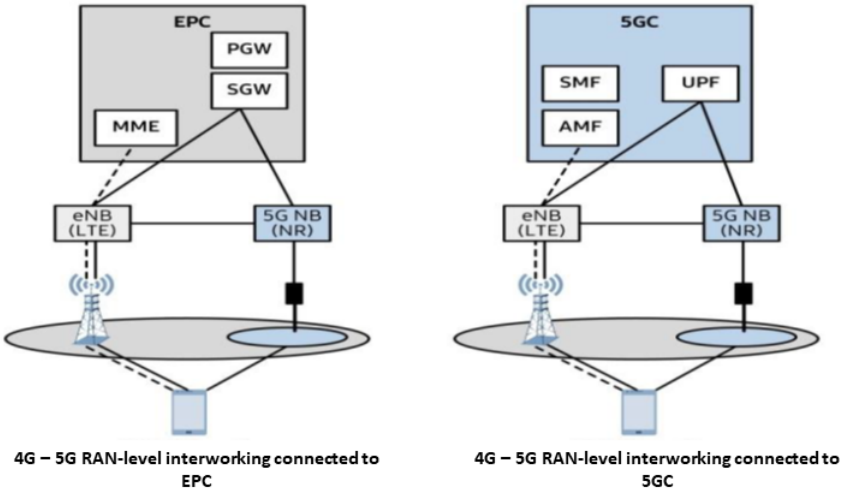


**Figure 4.6:** Core network-level interworking architecture.

### 4.5.1 Full 5G system Architecture with Reference points

Besides the RAN (gNB/ng-eNB) and AMF/SMF/UPF, the 5G architecture also includes the following network functions (NF), Fig. 4.7.

- Authentication Server Function (AUSF)
- Unified Data Management (UDM.)
- Network Slice Selection Function (NSSF)
- Policy Control Function (PCF)
- Application Function (AF.)
- Data Network (DN), e.g., operator services, Internet access, or 3rd party services

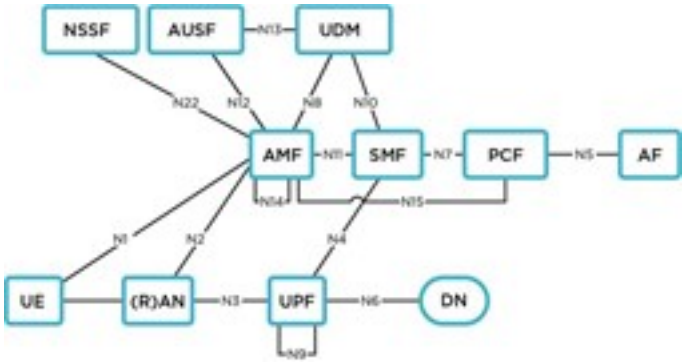


**Figure 4.7:** 5G system architecture with reference points.

Both 5G NR Control Plane (CP) and User Plane (UP) protocol stacks have many similarities with LTE protocol stacks. LTE protocol stacks are the baseline for the standardization of 5G NR.

- Non-Access Stratum (NAS): protocols for non-radio signaling between 5GC (protocol terminated in AMF) and UE Performing functions described in 3GPP TS 23.501 e.g., authentication, mobility management, security control.
- Access Stratum (AS): protocols between gNB and UE including RRC, only for CP, handles radio resource configuration and L2 protocols

## 4.6 5G: mm-Wave

Because of the rising demand for higher data rates, mm-Wave technology has been a significant focus of 5G due to its broad bandwidth. As a result, it is crucial to understand mm-Wave wireless systems' propagation behaviour to optimize the design process. There are numerous ways in which mm-Waves operate differently from the sub 6 GHz frequencies that are more often employed by cellular networks. When it comes to a given use-case scenario, these differences can either cause problems or opportunities. In the following, we study the challenges and opportunities regarding mm-Wave. Furthermore, a list of mm-Wave challenges has been provided.

- Free Space Pathloss
- Blockage
- Penetration Loss
- Foliage Loss
- Body and Hand Losses
- Scattering
- Atmospheric Loss

### 4.6.1 mm-Wave Challenges: Free Space path-loss

Increasing the carrier frequency $f_c$ by order of magnitude adds 20 $dB$ of power loss for a given distance $d$ in the case of free space propagation. As an electromagnetic wave propagates, it is attenuated or reduced in power density. Because it is essential to the analysis and design of the wireless communication system's link budget, free Space path-loss effect must be considered. mm-Waves seem to have a more significant path loss than lower frequencies [**57**]. Therefore, it's vital to understand the origins of this frequency-dependent loss. When transmit and receive power are measured in line-of-sight conditions, the Friis equation gives the essential connection between transmit and received power, Fig. 4.8. To preserve path loss at a constant level or even lower it as the frequency increases, it's necessary to keep the same physical size of the antennas.

### 4.6.2 mm-Wave Challenges: Blockage

When objects in the physical environment between the mm-Wave transmitter and receiver create large-scale variations in signal intensity, this phenomenon is known

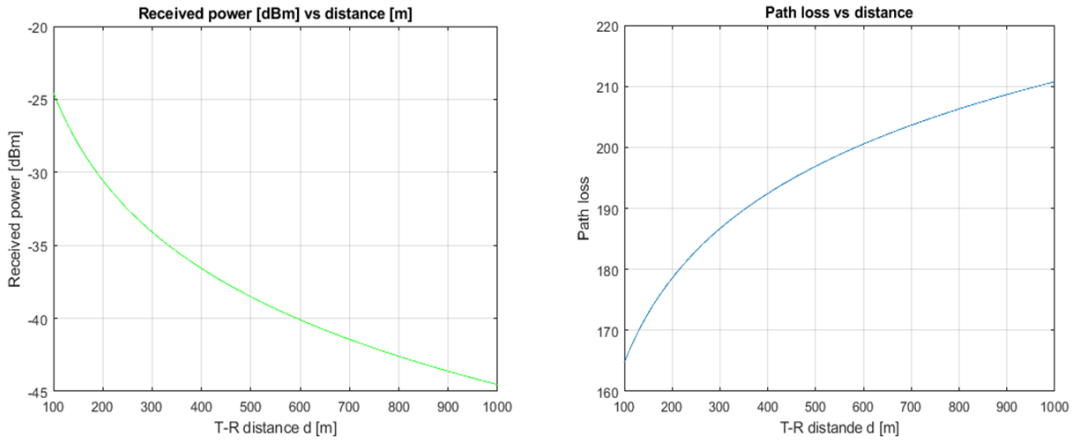


**Figure 4.8:** Left and right Figures illustrate received power and path loss based on Friis equation.

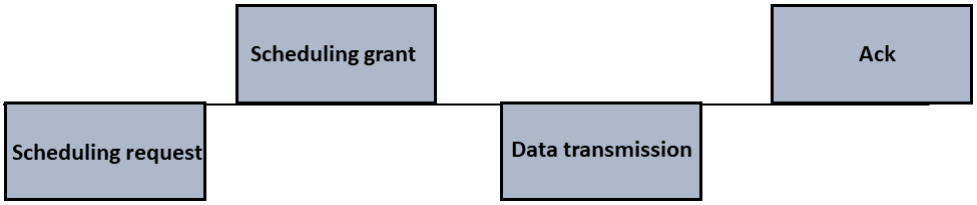


**Figure 4.9:** Uplink scheduling procedure.

as blockage or shadowing, depending on the context. Due to the shorter wavelength of mm-Wave signals, they are more sensitive to environmental barriers than sub-6 GHz transmissions. This is because objects in the environment seem to be larger at mm-Wave frequencies point of view. mm-Wave signals may be absorbed, reflected, scattered, or diffracted when they come into touch with these objects. Multi-site connectivity or dual connectivity is one possible way to counteract blocking.

### 4.6.3 mm-Wave opportunities: Reduced latency

Sub-6 $GHz$ spectrum traditionally has been utilized for many uses other than mobile broadband communications, such as radio, broadcast television, and radar. The orthogonality of neighboring subcarriers in OFDM/OFDMA technology is one of its most significant technological features. Due to the fact that subcarrier spacing $(SCS) = m/Ts$, where m is a positive integer, and Ts is symbol duration, this orthogonality is achieved. Of course, additional factors such as multi-path, Doppler, and delay are also taken into account regarding systems based on OFDM. Because 5G NR uses a flexible subcarrier spacing rather than LTE's fixed 15 $kHz$ subcarrier spacing, it's compatible with a wide variety of frequencies, ranging from 800 MHz to 39 $GHz$ and beyond. If n is a number Between 0 and 4, then the SCS range is specified as $15x2n$. Most commercial installations seem to be utilizing SCSs of 120 $kHz$ in all of the mm-Wave bands that have been specified so far. For mm-Wave, these bigger SCSs are required to counteract inter-symbol interference and phase noise. For mm-Wave at higher SCS, this shorter slot corresponds to a considerably shorter TTI for mm-Wave on 5G NR. With this TTI, 5G NR at a particular subcarrier spacing is subjected to a fundamental system delay. However, the user can't accomplish this independence and must ask the network to deliver data to the user. After receiving the scheduling permit, the network informs the device when and how to provide data. Following transmission, the network must confirm whether the data was received correctly, (Fig. 4.9). In any case, a bigger SCS will reduce latency by lowering the time between each of these processes, independent of the band. This would take four milliseconds to finish the operation on 15 $kHz$ SCS, but just one millisecond on 120 $kHz$ SCS. TTI and SCS decoupling in 5G NR significantly reduce latency. When transmitting, a mini slot permits the transmission of a few symbols across a whole slot, rather than needing to use the full slot. As a result, URLLC applications can benefit from this. In the following, a list of mm-Wave benefits has been provided, table 4.1.

- Higher Densification
- Channel Reciprocity
- More extensive Antenna Array in Small Form Factor

| **Quality indicator** | **Sub − 6** $GHz$ | **Millimetre-wave** |
|---|---|---|
| Pathloss | Low | High |
| Radio channel | Multipath | Line of sight (LOS) |
| Indoor penetration | Average | Poor |
| Channel size | Medium | Wide |
| Cell size | Small to Average | Small |
| Challenge | Capacity | Coverage |

**Table 4.1:** Quality indicator comparison of sub-6 GHz and millimetre-wave.

# Chapter 5

# Measurement setup, scenarios, and results discussion

It is required to develop a measuring setup that emulates the impacts of an actual mobile network. This chapter aims to provide a comprehensive explanation of the designed measuring setup. The radio network is described in Section 5.1. It is explained in detail in section 5.2 which components and technologies are being utilized to turn the previously existing radio network into a mobile network with the required specifications. For further information on how to assess actual traffic characteristics, section 5.3 would present a suitable description. Moreover, we have suggested a measuring setup with the specification, as illustrated in Fig. 5.1.We will then detail each component's purpose and how it fits into the overall configuration.

## 5.1 Realistic wireless network

The wireless network utilized in this thesis is described in the following sections. Regarding channel fluctuations and wireless network load conditions, the radio network configuration designed for this research to enable us the study of selected parameters such as packet size effect on a real-world wireless network. When this thesis was started, early measurements were conducted under totally unloaded network conditions. CPEs have been situated in Politecnico di Torino in the proximity of gNB. Additional equipment and software are used to shape the actual behaviour of a wireless network. To affect the performance of an existing network, these mentioned components decrease the performance of an ideal or high-performing network. Therefore, the measurements over such a network are

reliable and represent an acceptable estimation of the network behaviour under realistic conditions. This means that latency will be affected when packets in the uplink direction are routed through a gNB and leave towards the core network. On the other hand, effects will be applied to the downlink packets that enter from the core network and depart the core network through the gNB on their way to the user.

## 5.2 Network virtualization

On a Linux machine, different scenarios have been applied by using the command-line interface. More information on virtual machines will be provided later. We would be able to apply extra latencies to degrade the packet's overall latency through a Linux command. There are specified parameters for matching packets, including source IP address and transport protocol. As a result, each UDP packet in the uplink and downlink might have a particular delay based on the IP address of the distant hosts. It is possible to run an operating system or other applications on an actual device that already has its operating system and applications loaded on it using a virtual machine. The operating system that runs on the actual computer is the guest, while the physical computer it runs on is a host. A single physical host can host many virtual machines or guests. As a result, each virtual machine may run on its dedicated hardware. In other words, each virtual machine has the same resources as the host computer's CPU, storage, and other hardware resources. According to this thesis' kind of virtualization implementations, four types of virtual network options can be selected based on desired scenarios.

- **Private networking**: as with bridged networking, the VM can connect directly with the outside world. As a result, only VMs on the same host that connects to the same internal network is accessible to the outside world. As requested, internal networks are immediately formed by the system and a centralized configuration does not exist. It is possible to identify each internal network by its name alone. One or more virtual network cards with the same internal ID will be automatically wired together with more than one active virtual network card. The support driver implements an Ethernet switch for Oracle VM VirtualBox which supports both broadcast and multicast frames [**2**].

- **Host-only Networking**: this mode is a mix of bridged and internal networking. Bridged networking allows virtual computers to communicate as if they were linked to a physical Ethernet switch. Virtual machines are not connected to an actual networking interface in virtual networking, so they can't communicate with anyone outside the host network. Host-only networking

seems to be very effective in pre-configured virtual applications, where several virtual machines are provided together and configured to cooperate [**2**].

- **Bridging**: such capability allows virtual machines like Oracle's VirtualBox to collect data from the actual network and insert data into it, establishing an entirely new networking interface within the software. Because of the new software interface seems to the host system that guests are linked to the interface utilizing a wired connection when they are not. Data can be sent and received using this interface to establish routing between the guest and the rest of network [**2**].
- **Network address translation (NAT)**: an IPv4 and IPv6-based TCP/UDP network address translation (NAT) service combines the machines that use it into a single system and restricts the external networks from directly contacting the internal network while allowing the inside systems to interact both with each other and with the external network. To take advantage of it, the VMs must be connected to the internal network. This is done when the NAT service is created, and if it does not currently exist, the internal network is established [**2**].

## 5.3 Measurement setup components and networking

Each of the three interfaces on the virtual machine is named ensp3s0, enp5s0f0, and enp5s0f1. SSH is enabled on enp3s0, the virtual machine's public interface. Vlan0 is linked to MEVO through enp5s0f0 and enp5s0f1 interfaces. vlan0 interface is connected to the wlan0 interface on the physical hosts where The port 8080 is used for NAT traversal. A delay of one millisecond is imposed on packets that traverse the path from the public interface to the interfaces of enp5s0f0 and enp5s0f1.moreover, uplink, and downlink delays (one-direction latency) can be measured using this setup, Fig. 5.1. Wlan0 is an interface including 2 Huawei 5G CPEs provided by Vodafone Italia [**3**]. Fig. 5.2 illustrates the internal component of mentioned virtual machine (MEVO). However, the description of interior architecture and components of MEVO is out of this thesis scope.

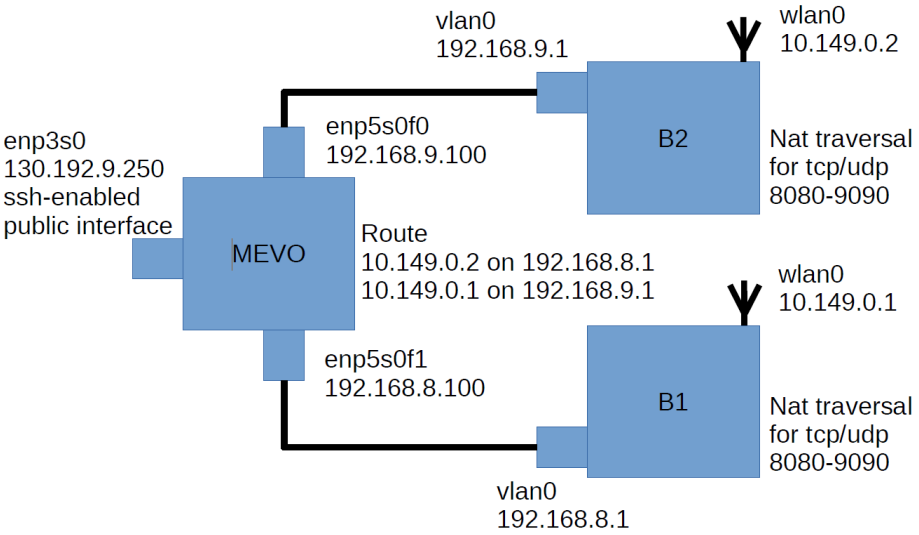


**Figure 5.1:** Tested networking including 5G CPEs.

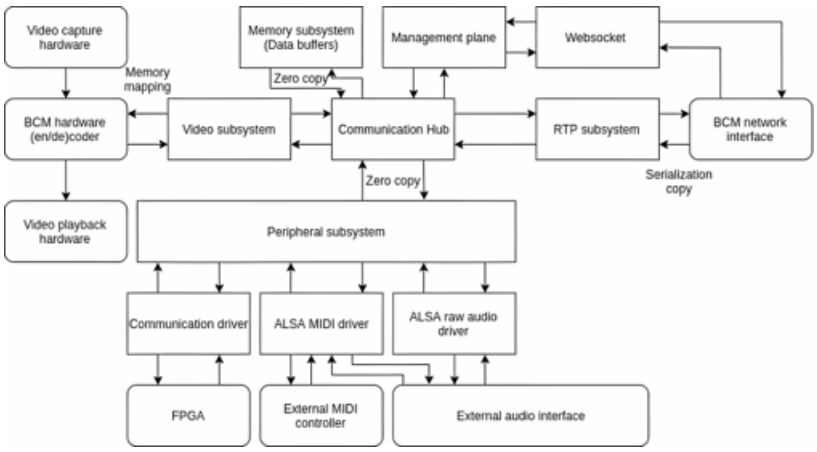


**Figure 5.2:** MEVO internal architecture .

## 5.4 Results discussion

The measurements were performed within six consecutive days to reduce the disconnections probability. Due to COVID-19 restrictions, there was no permission for physical presence in the laboratory, and all measurements had to be done remotely. Subsequently, the connection between my system and the virtual machine (MEVO) acting as a router with two static routes 10.149.0.2 -> 192.168.8.1, and 10.149.0.1 -> 192.168.9.1 has been made through secure shell protocol (SSH). Furthermore, the measurements, including delay and bandwidth related to each day, have been divided into two consecutive 12 hours measurements. Each day measurement is performed with different data rates and all data rates related to each measurement are provided in Table 5.1. In Table 5.2, the details related to the number of transmitted packets, dropped packets and receiving rate are provided. To perform this process, after collecting the results and importing them to the Matlab algorithm, then change it to a matrix of data, the index of rows representing the dropped packet is extracted, and its related time instant is marked. The time designated on the y-axes is measuered in the seconds, and it starts from 1, which is the start time of measurement. To clarify, if the measurement begins at 9:30:45, this time is mapped on $10^0 = 1$ on the y axes and increases second by second as the measurement progress, and x-axes show the dropped packets. For instance, one packet is dropped at the 4th second (4th represents a specific time concerning the start time of measurements) next packet is dropped at the 6th second, and so on. In the case of received packets, the same logic is followed with one difference that at the first step, instead of dropped packets, the index of rows in the data matrix related to received packets has been extracted.

### 5.4.1 Measurement tools and defintions

- **My trace route (MTR)**: MTR sends ICMP packets with incrementally increasing TTLs to view the route or series of hops that the packet makes between the origin and its destination. The TTL, or time to live, controls how many hops a packet will make before "dying" and returning to the host. By sending a series of packets and causing them to return after one hop, then two, then three, MTR can assemble the route that traffic takes between hosts on the Internet [**61**].

- **Timeouts**: can happen for various reasons. Some routers will discard ICMP, and absent replies will be shown on the output as timeouts. Alternatively, there may be a problem with the return route.

- **Iperf**: is an open source networking tool used to measure throughput or performance of a network. It can be used to test TCP and UDP. Iperf can be

used in Windows, Linux, and MAC etc operation systems [**63**].

- **Ping**: Ping is a computer network administration software utility used to test the reachability of a host on an Internet Protocol (IP) network.Ping measures the round-trip time for messages sent from the originating host to a destination computer that are echoed back to the source.Ping operates by sending Internet Control Message Protocol (ICMP) echo request packets to the target host and waiting for an ICMP echo reply. The program reports errors, packet loss, and a statistical summary of the results, typically including the minimum, maximum, the mean round-trip times, and standard deviation of the mean [**62**].
- **Adaptive ping**: Typically, ping requests are sent across the network at a set interval, usually 1 second. This is configurable with the -i flag. Generally, this means a complete ping request and response occurs, then the tool does nothing for the rest of that second. Adaptive ping flag -A tries to adjust the interval to the RTT of the network link. This way, on average, a new ping request goes out as soon as the last reply is received [**62**].

| **Measurement round** | **Average data rate** ($Mbit/sec$) |
|---|---|
| 1st | 11.9 |
| 2nd | 11.9 |
| 3rd | 8.35 |
| 4th | 8.35 |
| 5th | 5.96 |
| 6th | 5.96 |
| 7th | 4.77 |
| 8th | 4.77 |
| 9th | 4.77 |
| 10th | 23.8 |

**Table 5.1:** Data rates related to different measurement rounds.

## 5.5 Packet loss

Packet loss occurs when packets do not reach the destination, mainly due to transmission errors. Packet loss is measured concerning sent packets. Packet loss is an indicator that shows the network's reliability. In a wireless network, unstable channel characteristics, high bit error rate (BER), and user mobility are significant reasons for packet loss. Generally, network congestion can cause packet loss in all types of networks, and bottlenecks are usually the main reason for congestion in a network regardless of its architecture. Fig. 5.3 and Fig. 5.4 illustrate by increasing the time during each 12-hours measurement duration, the dropped packets increase by almost the same degree four all measurement rounds. It is worth noting since all figures nearly have the same trends in all measurement rounds; only four figures have been provided. A complete comparison of received packet distribution over the measurement time interval have been provided in Fig. 5.5.

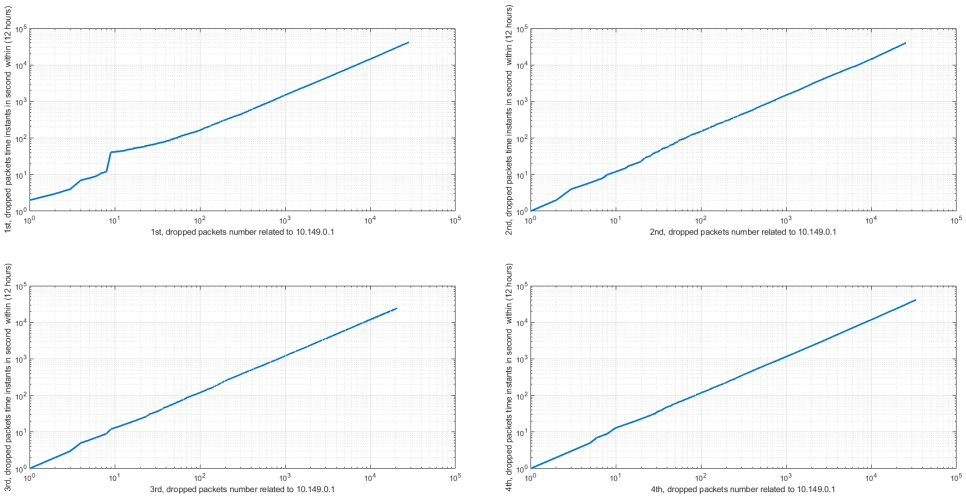


**Figure 5.3:** Packets dropping figures of 1st to 4th measurements.

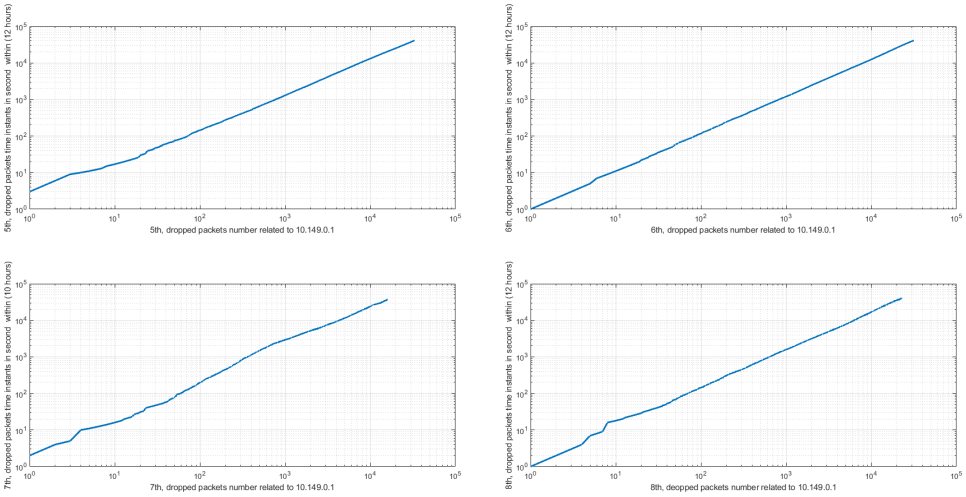


**Figure 5.4:** Packets dropping figure of 5th to 8th measurements.

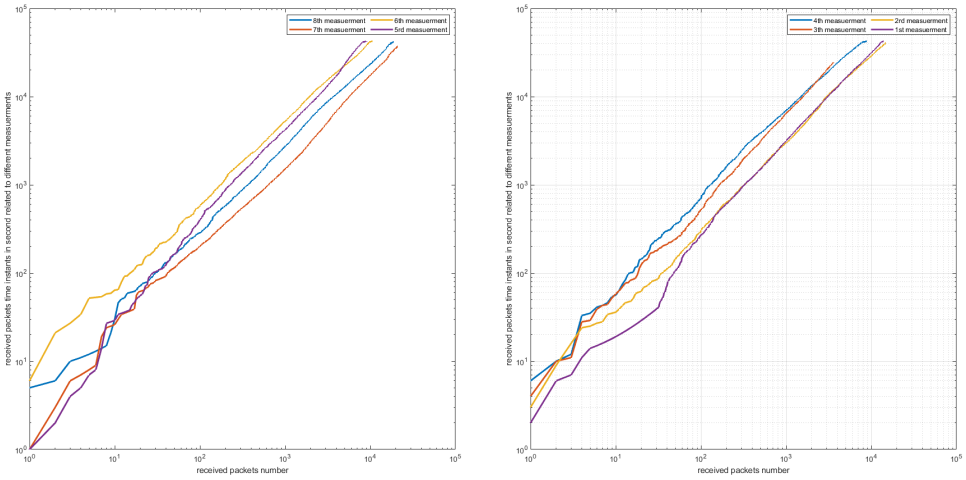


**Figure 5.5:** Received packets figures of 1st to 8th measurements.

### 5.5.1 Received packets delays histogram

The following histograms demonstrate the approximation of the probability distribution of a given delay by depicting the frequencies of observations occurring in specific ranges of delay values. Distribution of delays which shows a nearly symmetric distribution over the delay interval of 300 ms to 400 ms, (see Fig. 5.6, Fig. 5.7, Fig. 5.8, and Fig. 5.9). Moreover, these figures illustrate that the frequency of packets with an approximate value between 300 ms and 400 ms is high. These outcomes are more elevated than expectations since most latency-critical applications demand a delay lower than 50 ms. To have a wholesale overview of the network's latency behaviour, other rounds of latency measurements have been repeated. Although the results show latency improvement compared to the previous set of results in which the frequency of packets in the range of 50 ms and 200 ms are high, the latency performance of the network is not appropriate for most of the latency-critical services. More importantly, this comparison shows network instability in terms of latency as round by round; the latency figure behaves differently, Fig. 5.10, Fig. 5.11.

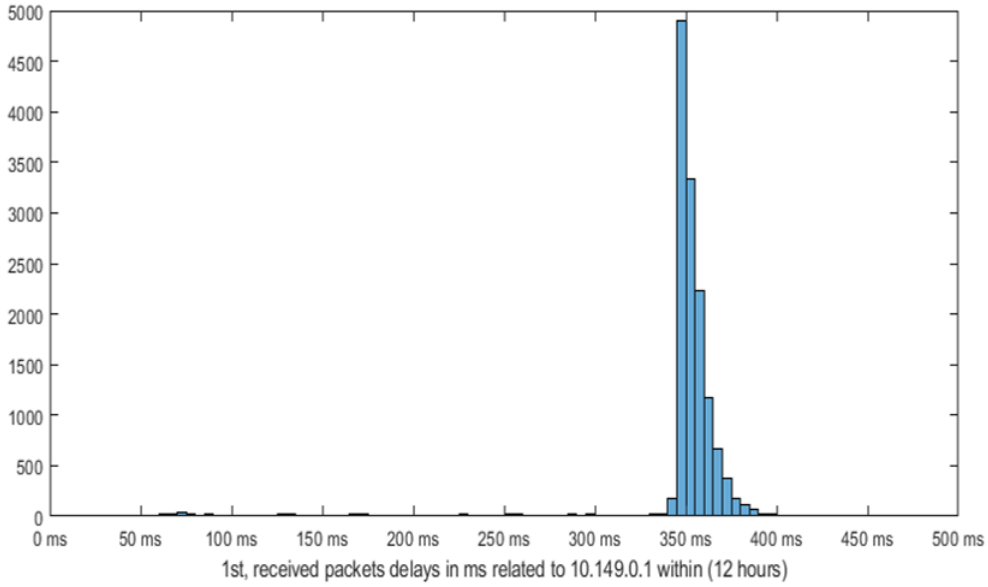


**Figure 5.6:** Received packets distribution over first measurement duration.

### 5.5.2 Delays average

Fig. 5.12 and Fig. 5.13 illustrate the variation of delays average based on measurement rounds. As it could be seen, there is a considerable variation in average

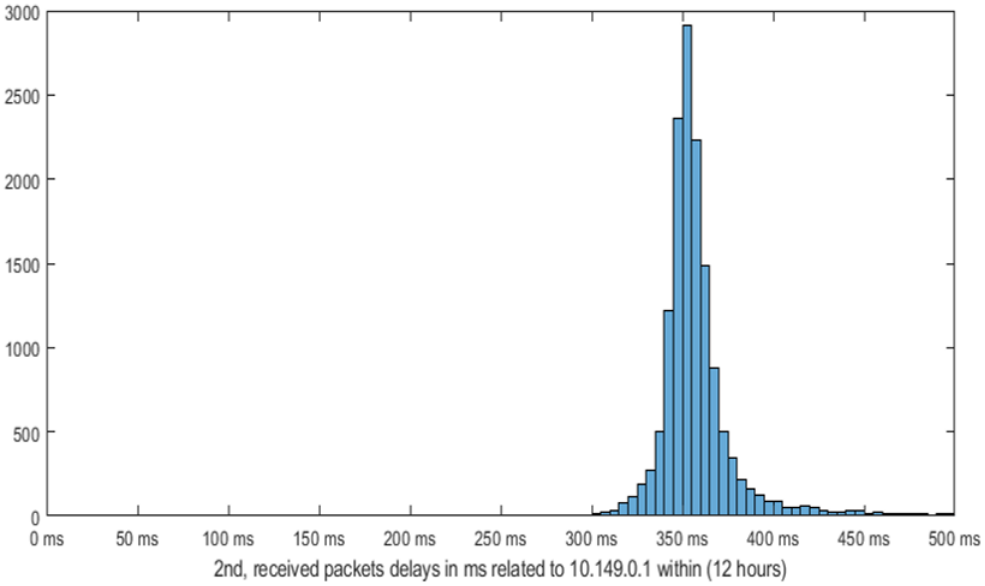


**Figure 5.7:** Received packets distribution over second measurement duration.

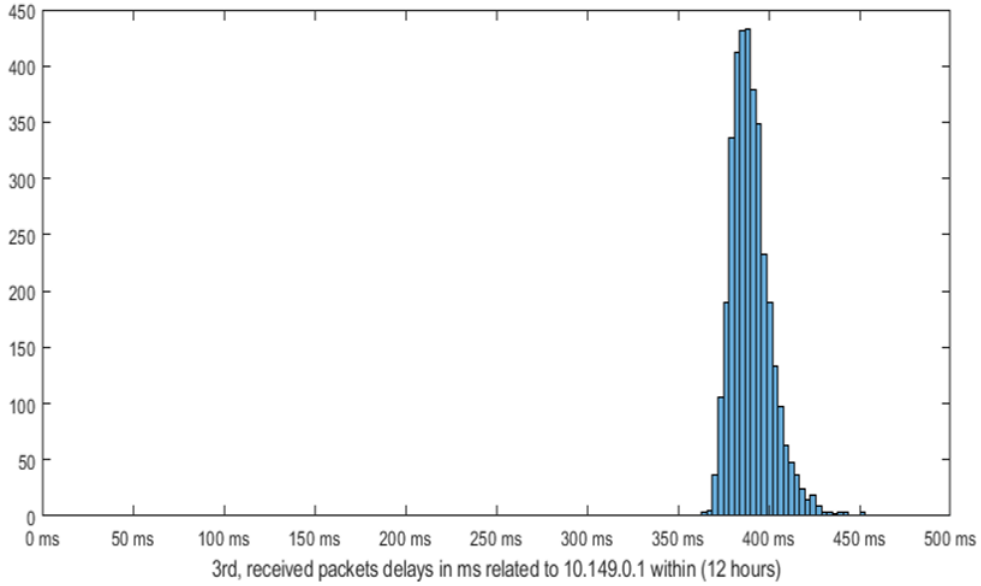


**Figure 5.8:** Received packets delays distribution over third measurement duration.

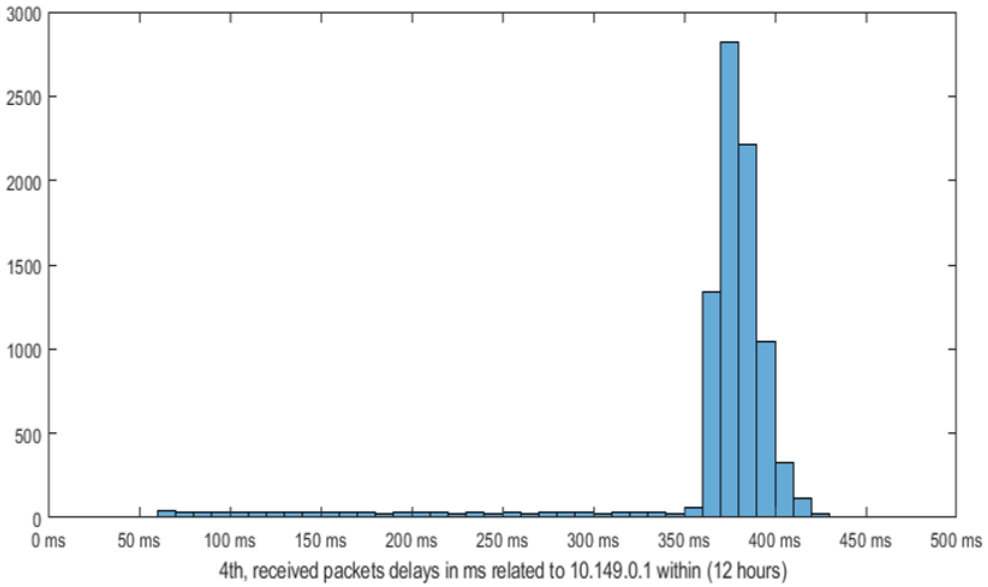


**Figure 5.9:** Received packets delays distribution over fourth measurement duration.

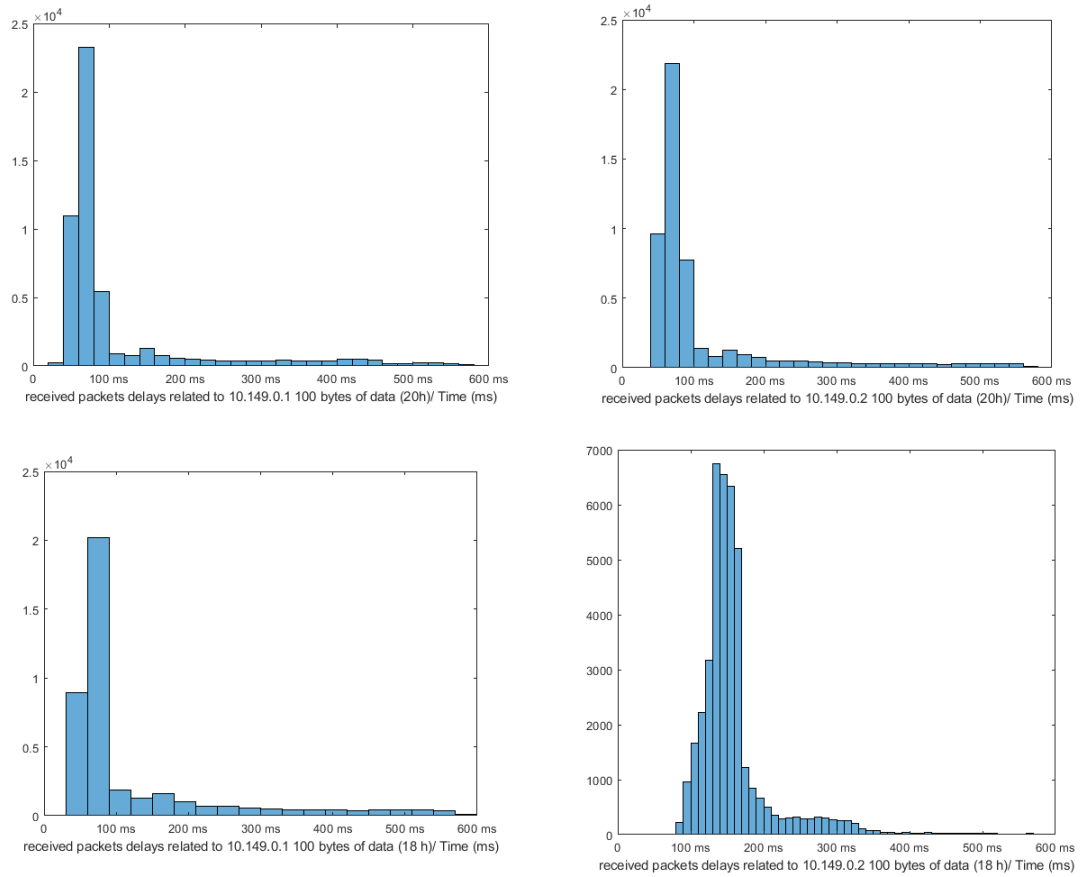


**Figure 5.10:** Received packets delays distribution.

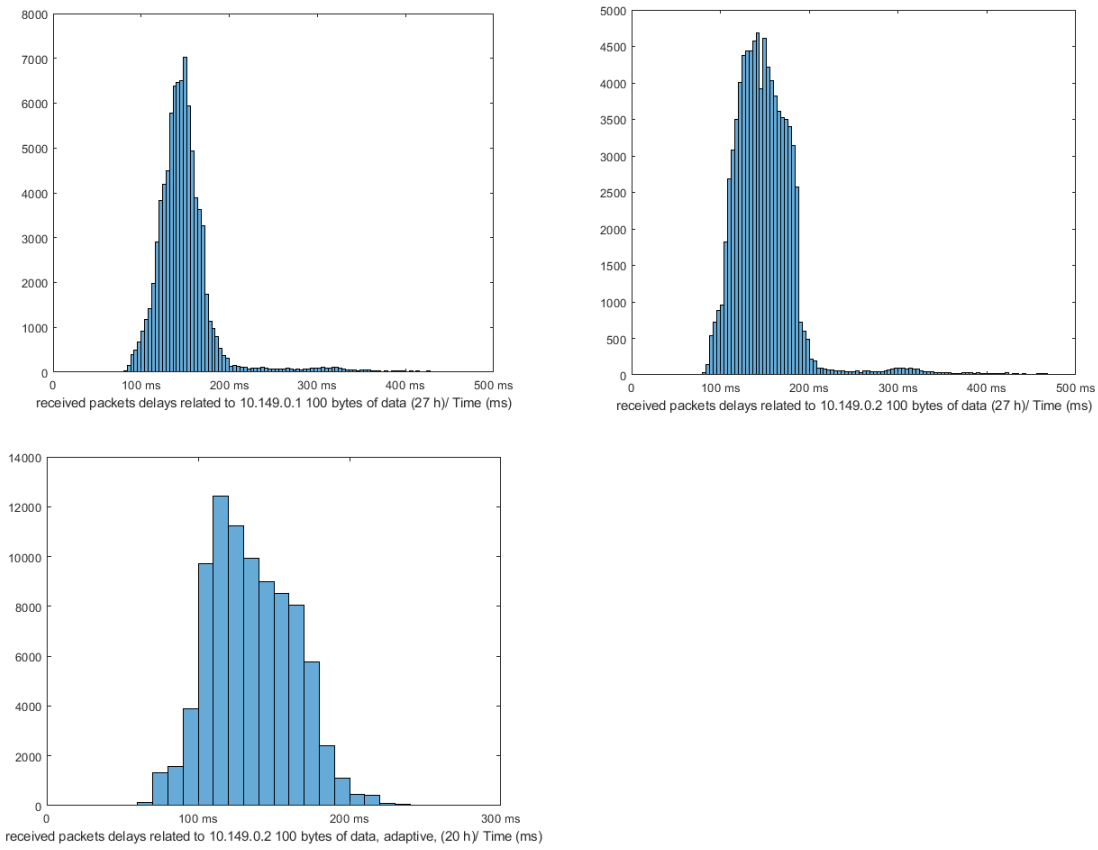


**Figure 5.11:** Received packets delays distribution.

values over rounds that if average delay values are considered as KPI, It will not satisfy the desired requirement for latency-critical applications science it drastically fluctuate round by round of measurements.

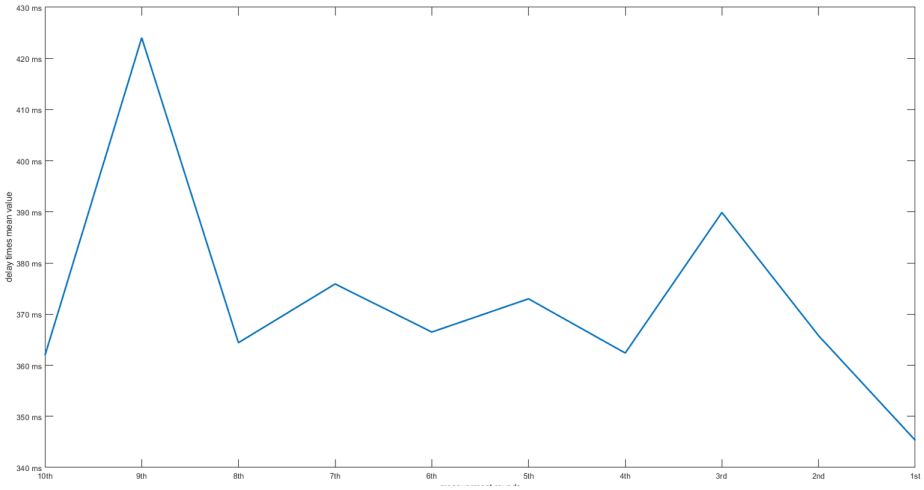


**Figure 5.12:** Delays average of each measurement round.

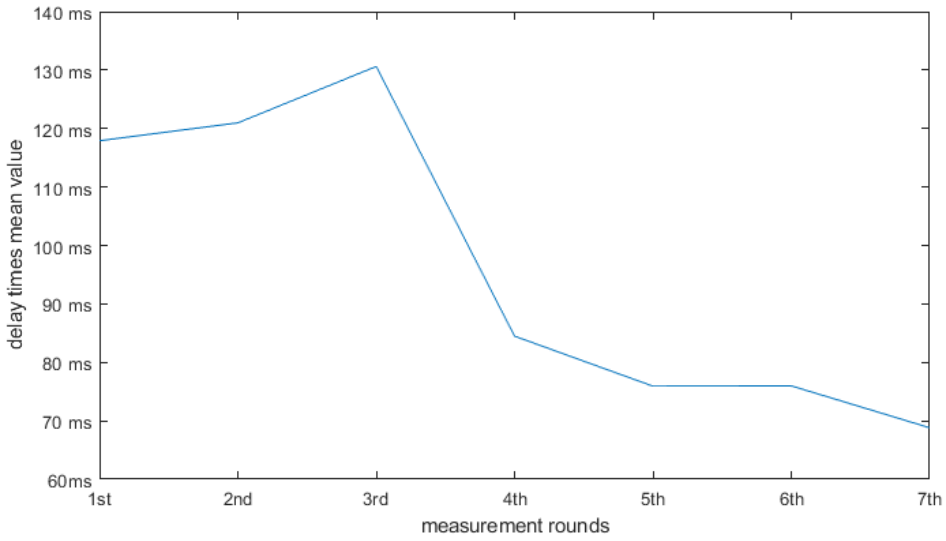


**Figure 5.13:** Delays average of each measurement round.

### 5.5.3 Delays mean values variation

Fig. 5.14, Fig. 5.15 and Fig. 5.16 indicate a more accurate and understandable analysis of latency behavior over time. Due to disconnection, some figures are shorter and have not covered all duration of measurements. The average values have been taken over 1-minute, 5-minutes and 1-hour time windows for a comprehensive overview of latency behaviour. Expectedly, this KPI shows unacceptable latency performance over time as it fluctuates substantially, which leads to high jitter.

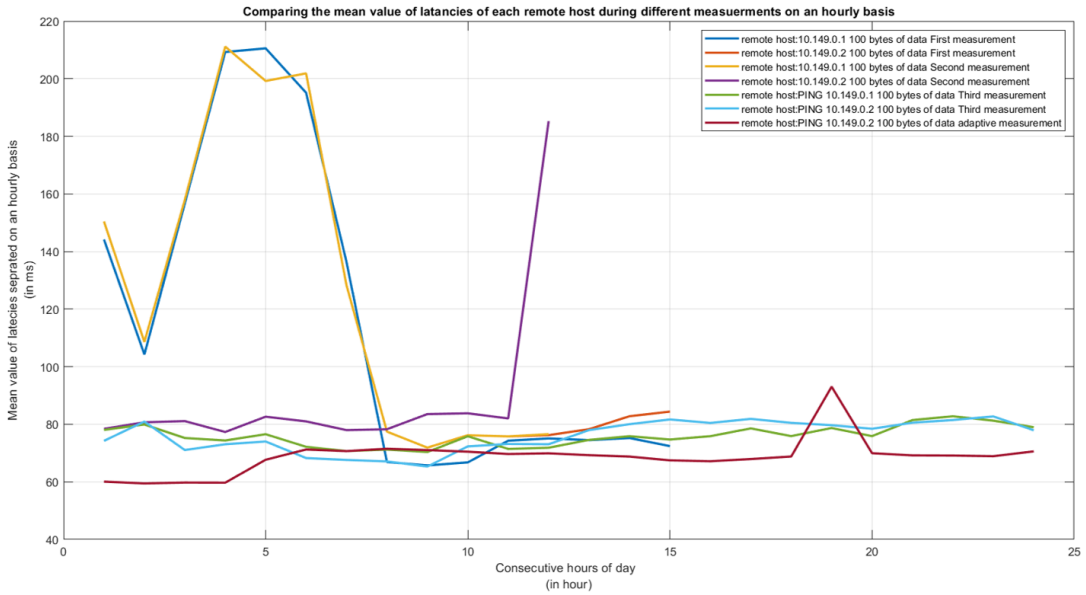


**Figure 5.14:** Mean values variation over time, mean values are calculated over each 1-hour time slot for remote hosts during measurements.

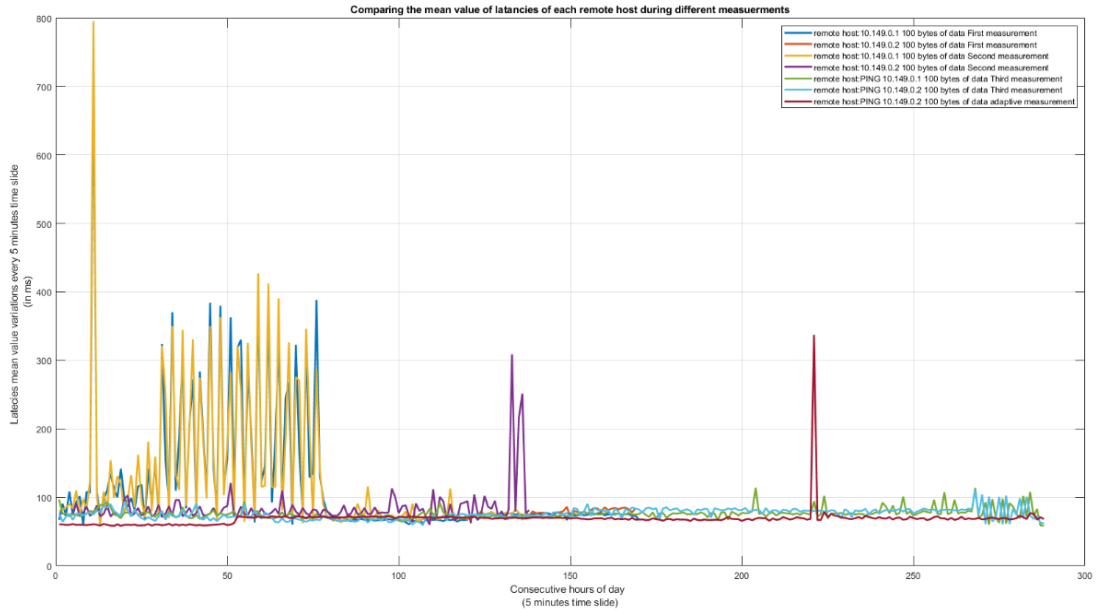


**Figure 5.15:** Mean values variation over time, mean values are calculated over each 5-minutes time slot for remote hosts during measurements.

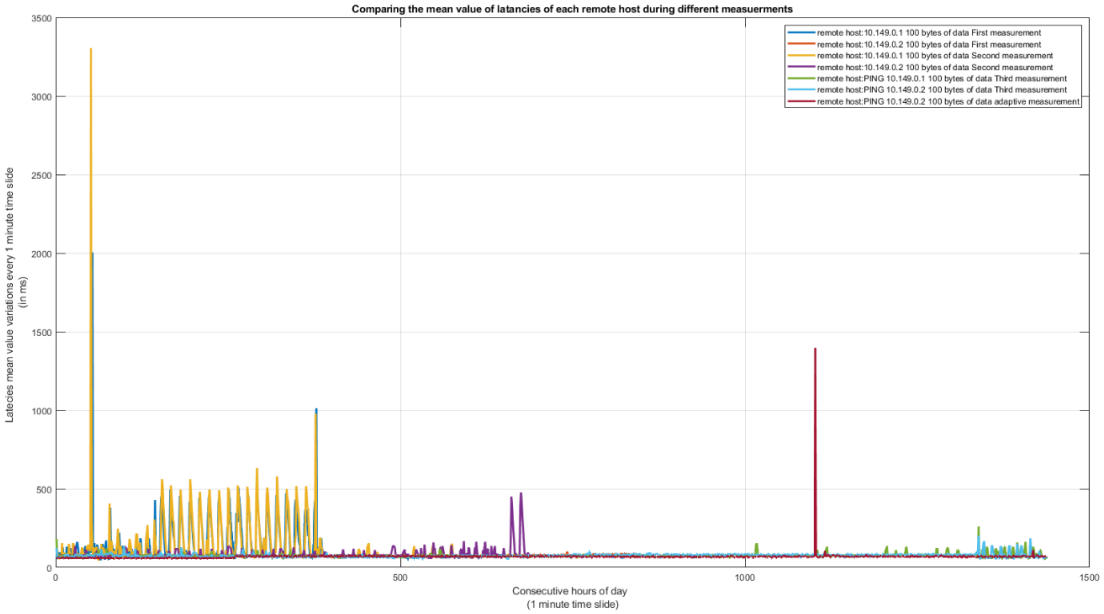


**Figure 5.16:** Mean values variation over time, mean values are calculated over each 1-minute time slot for remote hosts during measurements.

### 5.5.4 PDF and ECDF of delays

Fig. 5.19 illustrates probability density function of delays distribution, as it can be viewed the in one of measurements sets including 3 rounds of measurements the probability of delays values between 50 ms and 100 ms is high while the values between 100 ms and 200 ms show lower probabilities. Fig. 5.17 and Fig. 5.18 depict the ECDF figures of delays related to both 5G interfaces during all measurements rounds.

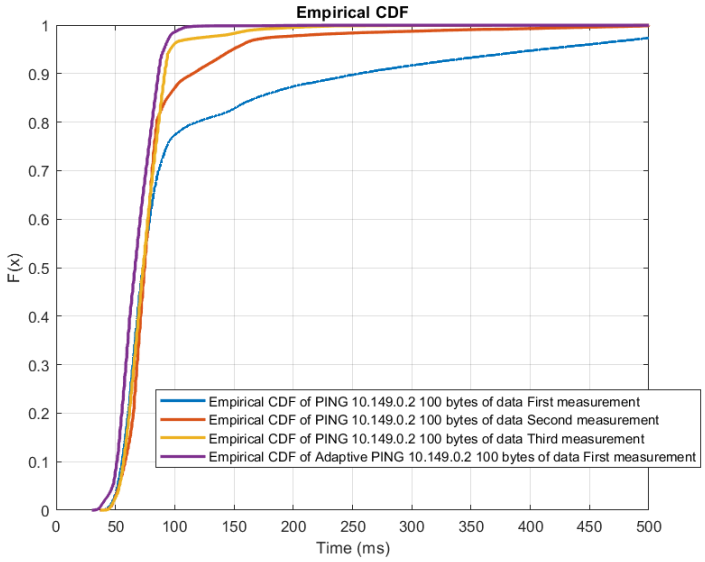


**Figure 5.17:** Delays ECDF of 10.149.0.2.

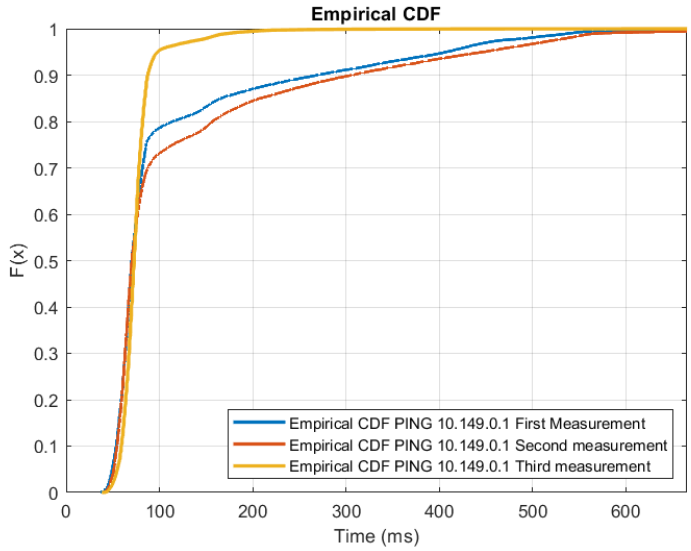


**Figure 5.18:** Delays ECDF of 10.149.0.1.

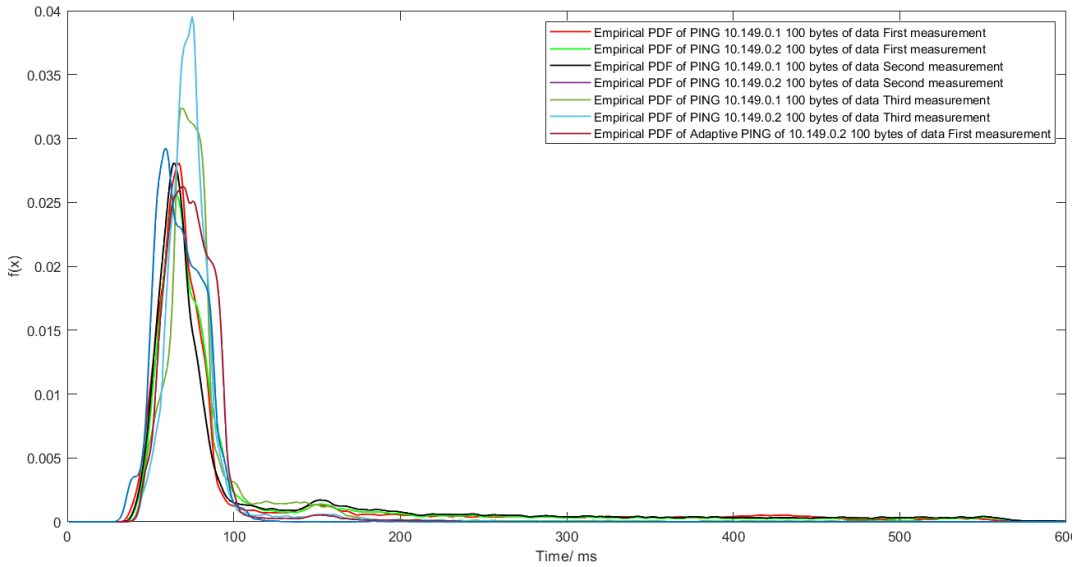


**Figure 5.19:** Delays PDF.

### 5.5.5 RTT

As mentioned in the last section, round trip time (RTT) is usually measured by a command-line tool called ping and presented in milliseconds as a critical parameter for network latency evaluation. However, the actual RTT is generally higher than one measured by ping duo to some network issues such as network congestion and server throttling. RTT is affected by parameters such as distance, traffic levels, and server response time. Fig. 5.20 and Fig. 5.21 illustrate a substantial variation over the RTT figure and an average value more than expected for 5G networks. This result shows that the implementation of latency-critical applications over 5G networks would be challengeable. For typical network usages such as real-time audio/video streaming or loading pages over the network, the network's latency performance is also lower than expectations regarding the previous mobile networks generations.

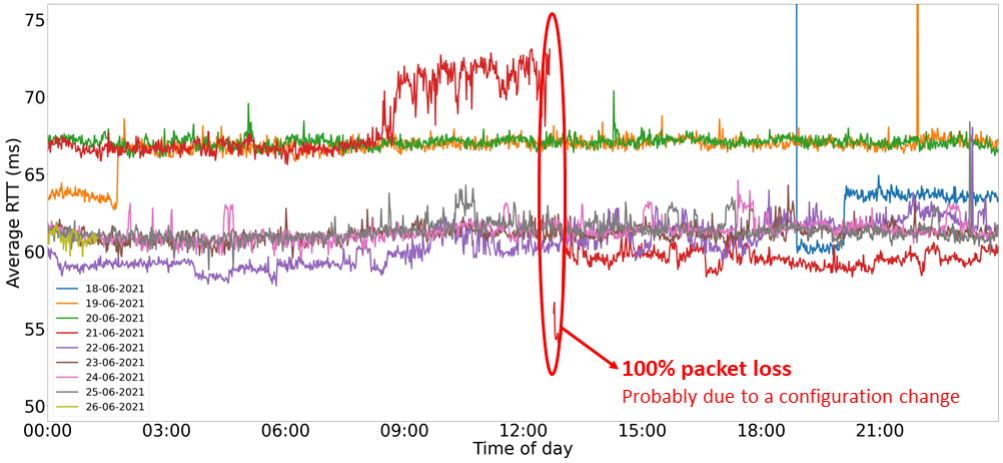


**Figure 5.20:** RTT.

### 5.5.6 RTT – Jitter

Based on the definition, jitter is variation over the packets' delays arriving at the destination. Jitter is mainly due to network congestion or route changing. The high jitter of receiving packets increases network latency and packet loss, drastically affecting real-time audio streaming quality. Choppy and static audio, delayed or dropped calls result from the network's high jitter effect on audio services. Fig. 5.22 shows a high jitter average value and considerable fluctuations of jitter figure, which negatively affects the network latency performance. Jitter is one of the network KPIs presented in this research, affecting real-time sound and video quality. Delay offset and buffering are the ways to address the jitter effect on the receiver side.

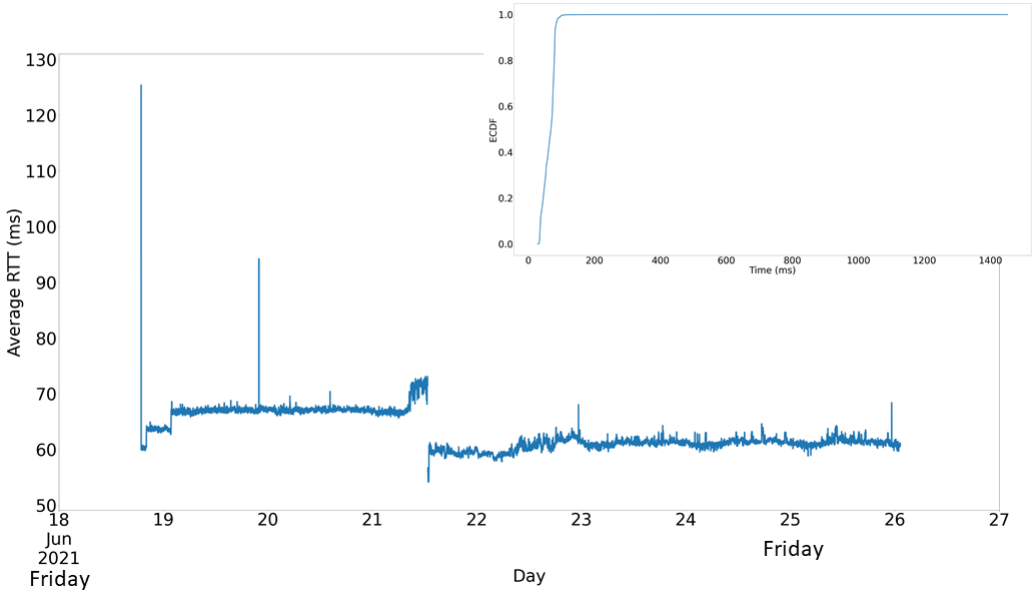


**Figure 5.21:** RTT – Day by day.

The procedure of storing received packets in the receiver system buffer enables the system to play the desired sound blocks or video frames without lags is named delay compensation. Therefore, the higher jitter means more delay offset for packet buffering leads to a trade-off between flawless real-time data streaming and the imposed delay on the packets. As it can be viewed in Fig. 5.22 the jitter figure fluctuates drastically over time, but its average value is higher than our expectation concerning the requirement of latency-critical applications. The main reason behind this issue is the considerable variation of RTTs shown in Fig. 5.20 and Fig. 5.21.

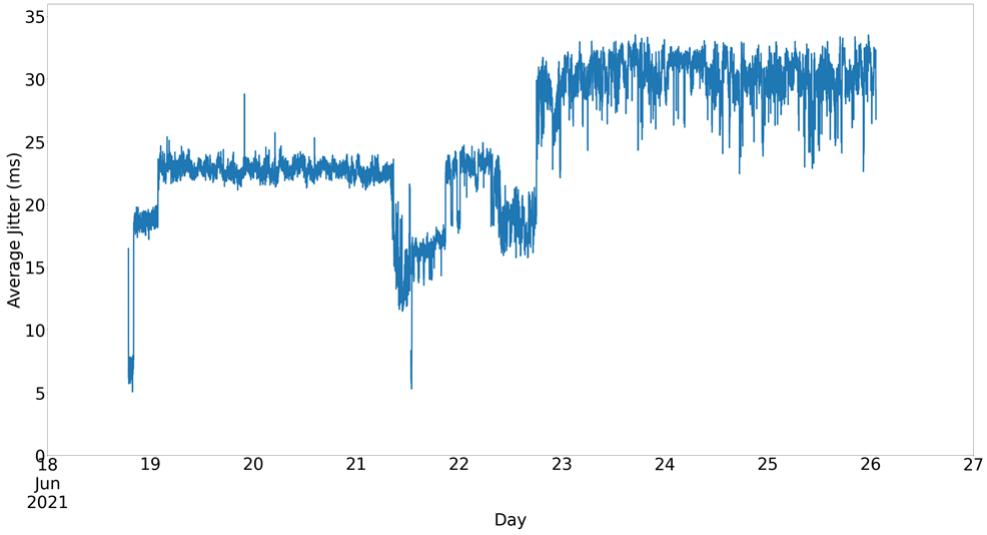


**Figure 5.22:** RTT – Jitter.

### 5.5.7 Bandwidth

Based on the definition, network bandwidth indicates network throughput that is measured in the particular time and network conditions when a data block with a specific size transmit. The bandwidth is calculated by dividing the summation of sent data (bit) into data delivery time (second). Bandwidth fluctuation can affect the quality of experience (QoE) regarding real-time video and audio streaming. Although the volatile network bandwidth is one of the characteristics of wireless networks, the bandwidth figure obtained from collected results shows substantial fluctuations and an average bandwidth value considerably lower than expectations for a 5G mobile network, Fig. 5.23.

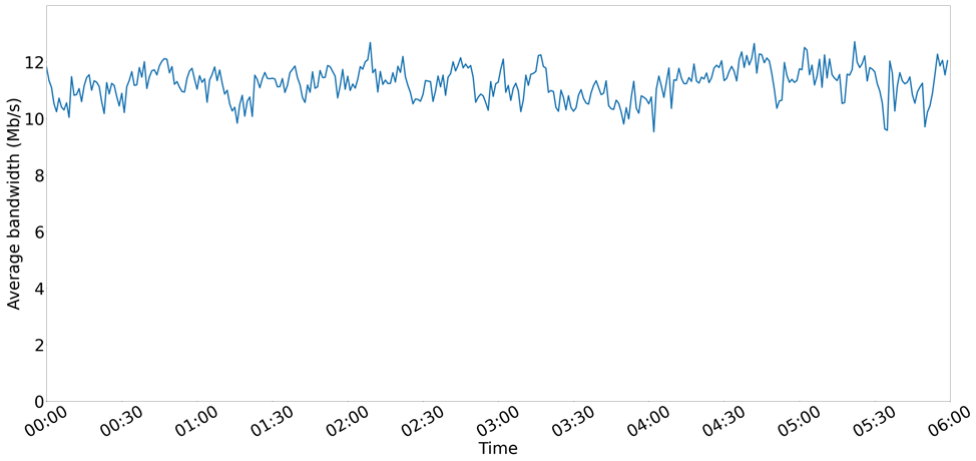


**Figure 5.23:** 1 packet is generated every 100 ms (6 hours), average bandwidth = 11.24 Mb/s.

### 5.5.8 Packet loss

Packet loss happens when a packet fails to reach the destination or is damaged mainly due to congestion, exceeding the buffer's capacity, and network control policy. Packet loss is one of the metrics that guarantee QoS of user applications such as VoIP. As described in the previous seasons regarding the demanded packet loss for mission-critical applications, and since most of these applications, including real-time video/audio streaming, a packet loss of less than 0.08% for seamless streaming quality is required. Moreover, Packet loss between 0.08% and 0.5% lead to the ghosting effect, packet loss between 0.5% and 1% causes frames to drop. Therefore, the loss packet figure illustrates this KPI cannot satisfy the requirement for a seamless video/audio streaming which is an integral part of mission-critical applications, Fig. 5.24 and Fig. 5.25.

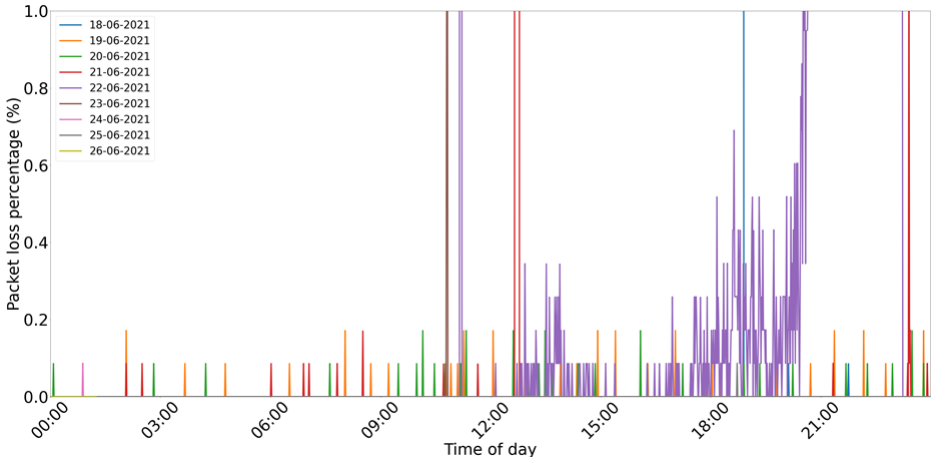


**Figure 5.24:** Packet loss.

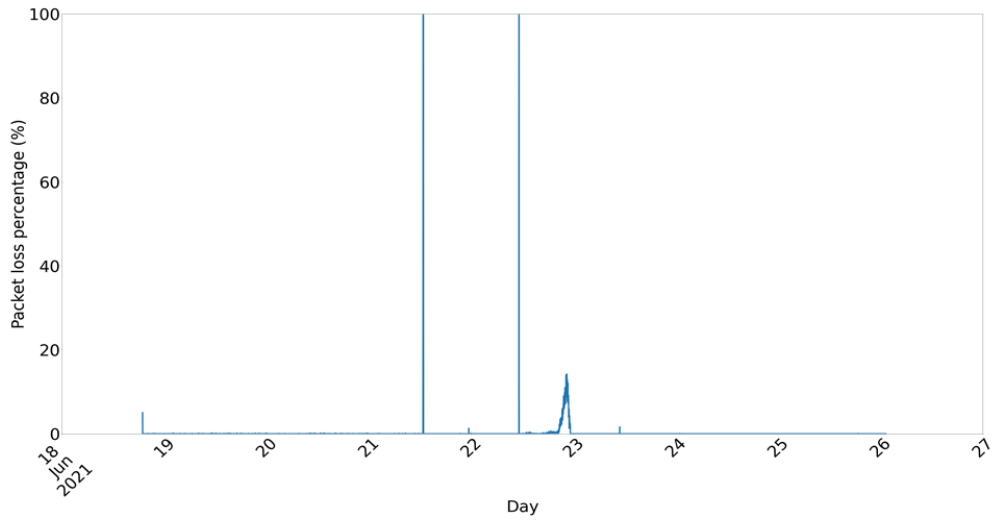


**Figure 5.25:** Packet loss Day by day.

| **Transmitted packets** | **Received packets** | **Destination unreachable** | **Lost packets (%)** | **Average RTT** | **Average jitter** |
|---|---|---|---|---|---|
| 12168000 | 12148190 | 2 | 0.16 % | 63.14ms | 25.14ms |

**Table 5.2:** Overall measurements statistics, all averages are calculated over a 1 minute window.

Table 5.2 shows the overall statistics of measurements such as total transmitted packets, total received packets, and loss rate. Fig. 5.23 shows the bandwidth figure over six hours of measurement; the high fluctuation of bandwidth values along with the packet loss figure over time decreases the network reliability as the bandwidth and packet loss are the indicators of network reliability, Fig. 5.24 and Fig. 5.25.

## 5.6 CPEs location effect evaluation: first round of measurements

This section investigates the reasons for different bandwidths observed in the measurements obtained from the two CPEs. To address this question, a scenario has been defined in which five different locations have been selected in the vicinity of the Politecnico di Torino LTE site (1-TOL0019921) to deploy the CPEs. Fig. 5.29 indicates these locations on the Politecnico di Torino building map. For each position, six parameters, including received signal strength indicator (RSSI), reference signal received power (RSRP), reference signal radio quality (RSRQ), signal to interference and noise ratio (SINR), bandwidth, and delay, have been measured. The bandwidth and delay values are collected using Speedtest-CLI by Ookla. The values are measured against the Vodafone IT Milano server, and RSSI, RSRP, RSRQ, and SINR values are collected through the CPEs profile. In the following, several figures have been provided indicating the average values of the six parameters mentioned above, (see Fig. 5.26, Fig. 5.27, and Fig. 5.28). During measurements, it has been observed that although both CPEs are located in the same physical location, in positions 1 and 3, they connected to the same cell ($CELL - ID$ : 5099827), while in positions 2, 4, and 5, they connected to two different cells where CPE1 connected to ($CELL - ID$ : 5099829) and CPE2 connected to ($CELL - ID$ : 5099809). Fig. 5.28 illustrates that even though each CPE uses a different frequency band in the positions of 2, 4 and 5, the RSRP, RSRQ, and SINR figures seem to be quite similar for both CPEs. Therefore, it can be interpreted that the difference in bandwidths is due to the different configurations of the cells.

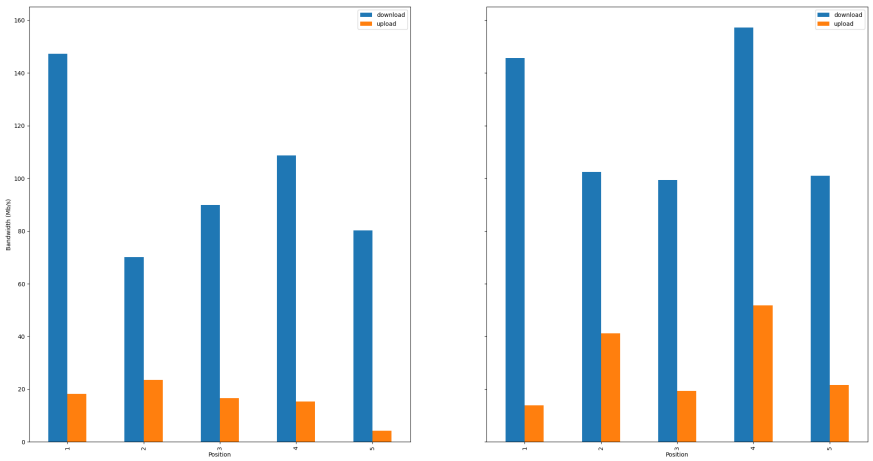


**Figure 5.26:** The graph on the left shows the uplink and downlink bandwidth obtained for CPE 1 (Serial No.7JK7N19614002044), previously called 10.149.0.1, and the one on the right shows the results of CPE 2 (Serial No.7JK7N19614002045), previously called 10.149.0.2.

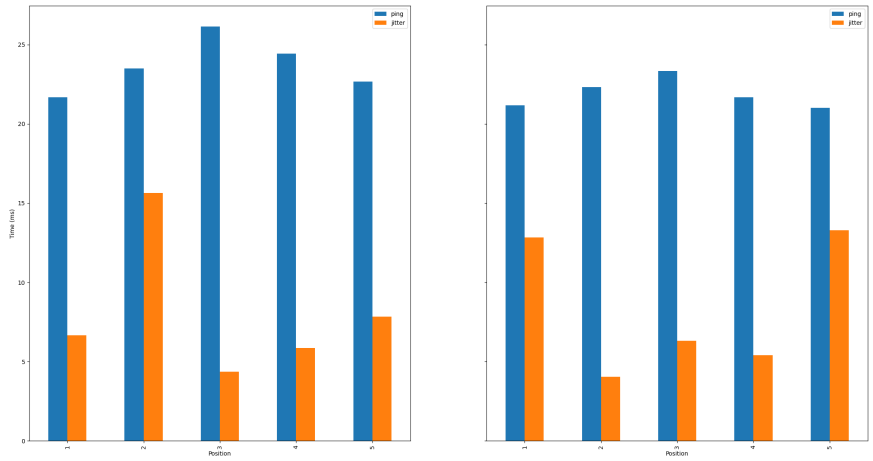


**Figure 5.27:** The graph on the left shows the ping-jitter obtained for CPE 1 (Serial No.7JK7N19614002044), previously called 10.149.0.1, and the one on the right shows the results of CPE 2 (Serial No.7JK7N19614002045), previously called 10.149.0.2.

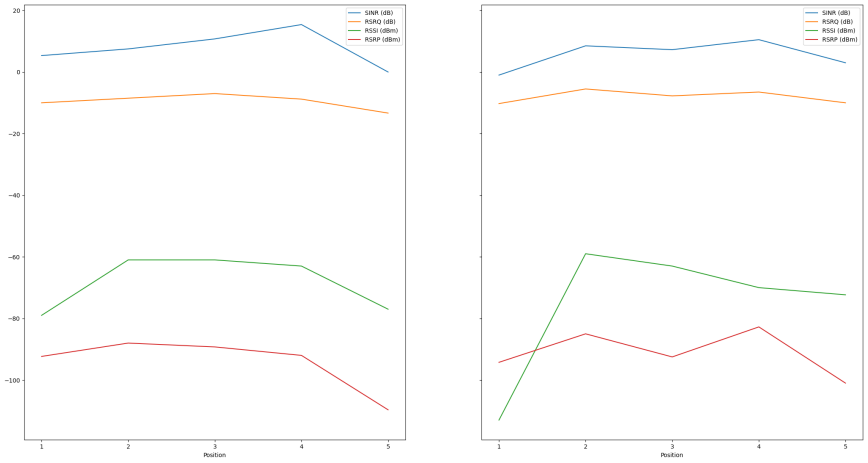


**Figure 5.28:** The graph on the left shows the RSSI, RSRP, RSRQ, SINR obtained for CPE 1 (Serial No.7JK7N19614002044), previously called 10.149.0.1, and the one on the right shows the results of CPE 2 (Serial No.7JK7N19614002045), previously called 10.149.0.2.


**Figure 5.29:** The figure illustrates CPEs five different locations on the Politecnico di Torino map during measurements.

## 5.7 CPEs location effect evaluation: second round of measurements

### 5.7.1 Ping-jitter

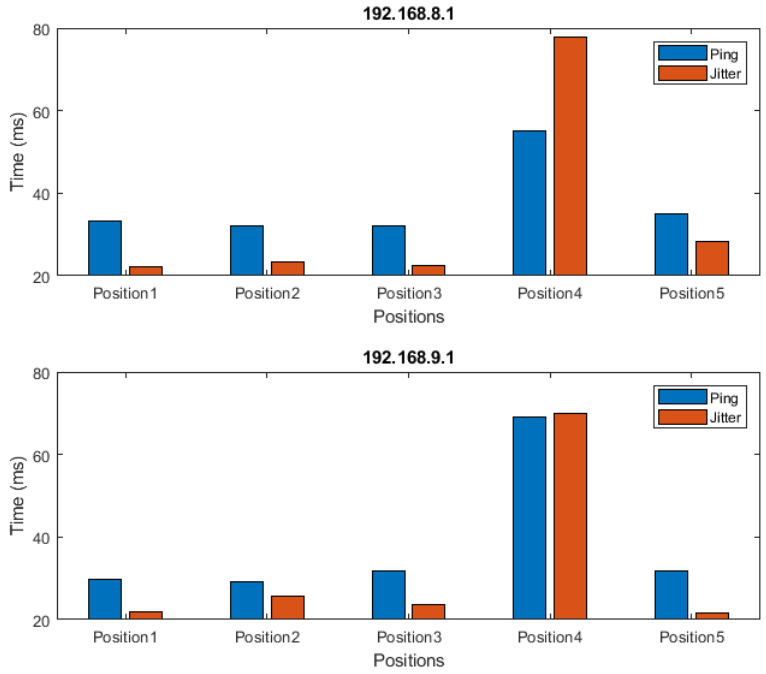


**Figure 5.30:** The upside graph shows the ping-jitter values obtained for CPE 1 (Serial No.7JK7N19614002044) (IP:192.168.8.1), previously called 10.149.0.1, and the one on the downside shows the results of CPE 2 (Serial No.7JK7N19614002045), (IP:192.168.9.1), previously called 10.149.0.2.

### 5.7.2 Bandwidth

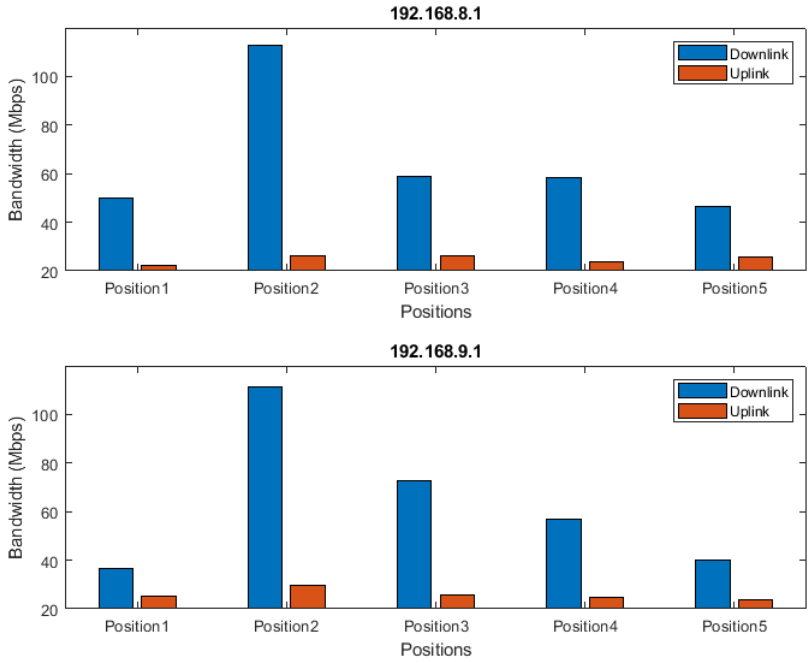


**Figure 5.31:** The upside graph shows the bandwidth values obtained for CPE 1 (Serial No.7JK7N19614002044) (IP:192.168.8.1), previously called 10.149.0.1, and the one on the downside shows the results of CPE 2 (Serial No.7JK7N19614002045), (IP:192.168.9.1), previously called 10.149.0.2.

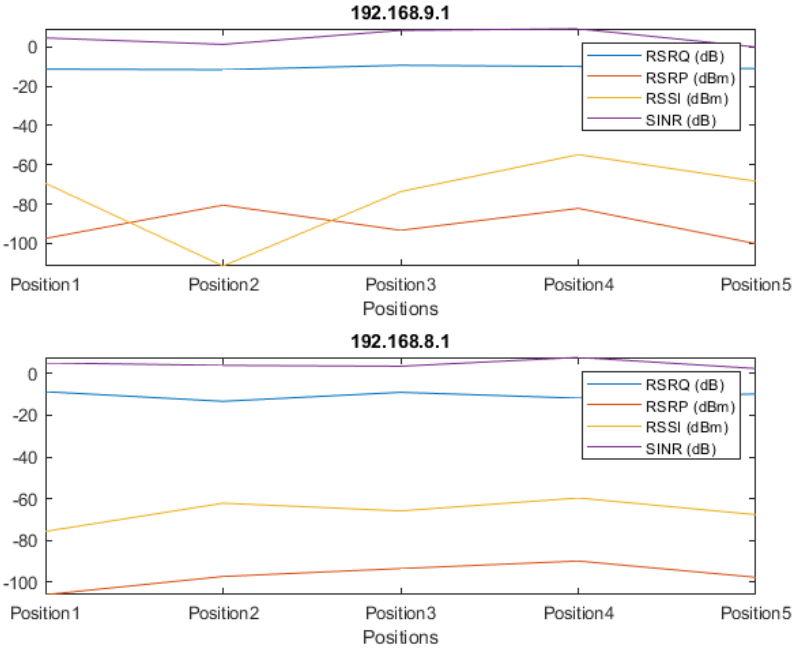


**Figure 5.32:** The upside graph shows the RSRQ, RSRP, RSSI, SINR values obtained for CPE 1 (Serial No.7JK7N19614002044) (IP:192.168.8.1), previously called 10.149.0.1, and the one on the downside shows the results of CPE 2 (Serial No.7JK7N19614002045), (IP:192.168.9.1), previously called 10.149.0.2.

## 5.8 Results discussion

### 5.8.1 Position 5

In the fifth position, where the CPE 1 ($IP$ : 192.168.8.1) is connected to cell ($ID$ : 5099829), and CPE 2 ($IP$ : 192.168.9.1) is connected to cell ($ID$ : 5099809), the delay figure and bandwidth performance of both CPEs are nearly identical, but CPE 2 ($IP$ : 192.168.9.1) show lower jitter. The same performance roots in the same level of SINR at this position, and since the connection in the fifth position was only over 4G, and -65 dBm to -75 dBm would be considered as the appropriate level for 4G connection considering the value of RSSI, at the fifth position this performance would be justifiable, Fig. 5.33 and Fig. 5.30 [**65**].

| RSSI | Signal strength | Description |
|---|---|---|
| > -65 dBm | Excellent | Strong signal with maximum data speeds |
| -65 dBm to -75 dBm | Good | Strong signal with good data speeds |
| -75 dBm to -85 dBm | Fair | Fair but useful, fast and reliable data speeds may be attained, but marginal data with drop-outs is possible |
| -85 dBm to -95 dBm | Poor | Performance will drop drastically |
| <= -95 dBm | No signal | Disconnection |

**Figure 5.33:** RSSI range definition

### 5.8.2 Position 4

In the fourth position where the CPE 1 ($IP$ : 192.168.8.1) is connected to cell ($ID$ : 5099829) and ($ID$ : 5099809) and CPE 2 ($IP$ : 192.168.9.1) is connected to cell ($ID$ : 5099809), the delay figure, including ping and jitter, is higher than other locations for both CPES; this behavior would be justifiable by considering the values of SINR for this location, Fig. 5.34 and Fig. 5.30 [**65**].

| SINR | Signal strength | Description |
|---|---|---|
| >= 20 dB | Excellent | Strong signal with maximum data speeds |
| 13 dB to 20 dB | Good | Strong signal with good data speeds |
| 0 dB to 13 dB | Fair to poor | Reliable data speeds may be attained, but marginal data with drop-outs is possible. When this value gets close to 0, performance will drop drastically |
| <= 0 dB | No signal | Disconnection |

**Figure 5.34:** SINR range definition

### 5.8.3 Position 3

Regarding position three, both CPEs indicate acceptable performance concerning the bandwidth and delay figure. At this position, the CPE 1 ($IP$ : 192.168.8.1) is connected to the cell ($ID$ : 5099829), CPE 2 ($IP$ : 192.168.9.1) is connected to the cell ($ID$ : 5099829) and ($ID$ : 5099809). At this position, RSRQ has good characteristics for both CPEs based on Fig. 5.35, Fig. 5.30 and Fig. 5.31 [**65**].

| RSRQ | Signal quality | Description |
|---|---|---|
| >= -10 dB | Excellent | Strong signal with maximum data speeds |
| -10 dB to -15 dB | Good | Strong signal with good data speeds |
| -15 dB to -20 dB | Fair to poor | Reliable data speeds may be attained, but marginal data with drop-outs is possible. When this value gets close to -20, performance will drop drastically |
| <= -20 dB | No signal | Disconnection |

**Figure 5.35:** RSRQ range definition

### 5.8.4 Position 2

In the case of position 2, it can be viewed that the highest Downlink bandwidth and acceptable values for the delay have been obtained. At this position, the CPE 1 ($IP$ : 192.168.8.1) is connected to the cell ($ID$ : 5099829), CPE 2 ($IP$ : 192.168.9.1) is connected to the cell ($ID$ : 5099829). At this position, both RSRQ and RSRP are situated at reasonable amounts based on Fig. 5.35 and Fig. 5.36.

### 5.8.5 Position 1

At the first position where the CPE 1 ($IP$ : 192.168.8.1) is connected to the cell ($ID$ : 5099827) and CPE 2 ($IP$ : 192.168.9.1) is connected to the cell ($ID$ : 5099817), the CEP1 has higher latency and bandwidth compared to CPE2.

| RSRP | Signal strength | Description |
|---|---|---|
| >= -80 dBm | Excellent | Strong signal with maximum data speeds |
| -80 dBm to -90 dBm | Good | Strong signal with good data speeds |
| -90 dBm to -100 dBm | Fair to poor | Reliable data speeds may be attained, but marginal data with drop-outs is possible. When this value gets close to -100, performance will drop drastically |
| <= -100 dBm | No signal | Disconnection |

**Figure 5.36:** RSRP range definition

### 5.8.6 Observations

It is worth noting that in the positions of 2 and 3, 5G network was available, and CPEs have experienced a hybrid connection which can justify the higher Downlink bandwidth in these locations. Moreover, based on results that have been obtained, the probable reasons for various behavior of network within five selected positions can be categorized into four parts.

- The distance from base stations leads to path loss, according to Friis equation.
- The thickness of building walls and roofs leads to attenuation depending on the refraction coefficient of the used material. Moreover, the obstacles between CPEs and base stations cause diffraction leading to extra attenuation. In positions 2, 3, 4, and 5 walls are made of glass while in the first position walls are considerably thick and made of concrete.
- The number of active mobile terminals existing in the vicinity of CPEs affects SINR which can be observed in position 1 where it is located near to a crowded area with the high number of active mobile terminals.
- Cells configuration effect can be observed in the case of connection to different cells.

## 5.9 Conclusion and future extensions

Several questions mentioned below have been shaped the initial idea for this study

- What is and with which extent the effect of networks parameters such as packet size and network load on latency?
- What would be the behavior of the network in the case of latency during the measurements?
- Can 5G mobile network satisfy the latency requirement of latency-critical applications, in other words, the outcomes are higher or lower than our expectations concerning previous mobile network generations such as LTE?

Addressing the above questions can ensure the required quality of service (QoS) for different applications. One feature that distinguishes this study from previous related studies is that instead of using network simulators, an actual 5G mobile was investigated. All the numerical results are obtained through actual measurement considering the actual network conditions. To have accurate and comprehensive outcomes, several measurements rounds have been performed to observe the effect of network configuration and parameters on the latency figures hourly and daily. Surprisingly, the obtained KPIs have illustrated that the indicators, such as jitter, and RTT, are higher than our expectations and far from the required values for almost all latency-critical applications, especially for use cases involving real-time audio-video playing. Loss packet heavily affects the network reliability and is considered an imperative parameter for applications such as remote surgery and automation is higher than the demanded values. Therefore, the outcomes show the desired KPIs have not been obtained yet, and latency improvement methods such as core network and radio access solutions have to be performed to approach the desired KPIs. This study provides a unique opportunity for future extensions, and the open related areas have not been studied in this research, such as latency improvement methods such as edge computing, backhaul network topology effect on latency, and measurements considering real audio/video traffic.

# Bibliography


[1] https://www.5gamericas.org/

[2] https://www.virtualbox.org/manual/ch06.html

[3] http://resolver.tudelft.nl/uuid:e1badd8d-a384-49a1-b958-a0c1e499c539

[4] https://images.samsung.com/is/content/samsung/p5/global/business/networks/insights/white-paper/4g-5g-interworking/global-networks-insight-4g-5g-interworking-0.pdf

[5] hhttps://www.etsi.org/deliver/etsi_TS/122200_122299/122261/15.07.00_60/ts_122261v150700p.pdf

[6] https://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.02.00_60/tr_138913v140200p.pdf

[7] Sasha Sirotkin, "5G Radio Access Network Architecture: The Dark Side of 5G," *Wiley, 2020.*

[8] Jyrki T. J. Penttinen, "5G Explained," *Wiley, 2019.*

[9] Devaki Chandramouli, Rainer Liebhart, Juho Pirskanen, "5G for the Connected World," *Wiley, 2019.*

[10] http://www.diva-portal.org/smash/get/diva2:1346021/FULLTEXT01.pdf

[11] "5G Second Phase Explained," *Wiley, 2021.*

[12] Martin Sauter, "From GSM to LTE-Advanced Pro and 5G," *Wiley, 2021.*

[13] Martin Sauter, "From GSM to LTE," *Wiley, 2010.*

[14] "5G and Beyond," *Springer Science and Business Media LLC, 2021.*

[15] Patrick Marsch, Navid Nikaein, Mark Doll, Tao Chen, Emmanouil Pateromichelakis, "RAN Architecture," *Wiley, 2018.*

[16] "Enhanced Radio Access Technologies for Next Generation Mobile Communication," *Springer Science and Business. Media LLC, 2007.*

[17] Wan Lei, Anthony C.K. Soong, Liu Jianghua, Wu Yong, Brian Classon, Weimin Xiao, David Mazzarese, Zhao Yang, Tony Saboorian, "5G System Design," *Springer Science and Business Media LLC, 2021.*

[18] Strahil Panev, Pero Latkoski, "Study of 5G Services Standardization: Specifications and Requirements," *Transactions on Emerging Telecommunications Technologies, 2019.*

[19] A. A. Ateya, A. Muthanna, M. Makolkina and A. Koucheryavy, "Study of 5G Services Standardization: Specifications and Requirements," *2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2018*, pp. 1-6, doi: 10.1109/ICUMT.2018.8631201.

[20] Sassan Ahmadi, "Chapter 12 - Performance of IEEE 802.16m and 3GPP LTE-Advanced, Editor(s): Sassan Ahmadi," *Mobile WiMAX, Academic Press, 2011*, Pages 657-721, ISBN 9780123749642, https://doi.org/10.1016/B978-0-12-374964-2.10012-8.

[21] L. Li and X. Tan, "Big-Data-Driven Intelligent Wireless Network and Use Cases," *2021 IEEE International Conference on Communications Workshops (ICC Workshops), 2021*, pp. 1-6, doi: 10.1109/ICCWorkshops50388.2021.9473653.

[22] S. Abdelwahab, S. Zhang, A. Greenacre, K. Ovesen, K. Bergman and B. Hamdaoui, "Big-Data-Driven Intelligent Wireless Network and Use Cases," *When Clones Flock Near the Fog," in IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1914-1923, June 2018, doi: 10.1109/JIOT.2018.2817392.

[23] F. J. Dian and R. Vahidnia, "A Simplistic View on Latency of Random Access in Cellular Internet of Things," *2021 IEEE International Conference on Communications Workshops (ICC Workshops), 2020*, 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2020, pp. 0391-0395, doi: 10.1109/IEMCON51383.2020.9284948.

[24] Q. Yang, A. Lim, X. Ruan and X. Qin, "Location Privacy Protection in Contention Based Forwarding for VANETs," *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pp. 1-5, doi: 10.1109/GLOCOM.2010.5684166.

[25] S. Choi, K. Shin and H. Kim, "End-to-End Latency Prediction for General-Topology Cut-Through Switching Networks," *in IEEE Access*, vol. 8, pp. 13806-13820, 2020, doi: 10.1109/ACCESS.2020.2966139.

[26] J. Li et al., "Service Migration in Fog Computing Enabled Cellular Networks to Support Real-Time Vehicular Communications," *in IEEE Access*, vol. 7, pp. 13704-13714, 2019, doi: 10.1109/ACCESS.2019.2893571.

[27] R. Jin, X. Zhong and S. Zhou, "The Access Procedure Design for Low Latency in 5G Cellular Network," *2016 IEEE Globecom Workshops (GC Wkshps)*, 2016, pp. 1-6, doi: 10.1109/GLOCOMW.2016.7849058.

[28] Masashi IWABUCHI, Anass BENJEBBOUR, Yoshihisa KISHIYAMA, Guangmei REN, Chen TANG, Tingjian TIAN, Liang GU, Yang CUI, Terufumi TAKADA, "5G Experimental Trials for Ultra-Reliable and Low Latency Communications Using New Frame Structure, IEICE Transactions on Communications, 2019," Volume E102.B, Issue 2.

[29] Bo Yan and H. Gharavi, "Multi-Path Multi-Channel Routing Protocol," *Fifth IEEE International Symposium on Network Computing and Applications (NCA'06)*, 2006, pp. 27-31, doi: 10.1109/NCA.2006.41.

[30] H. Chen et al., "Ultra-Reliable Low Latency Cellular Networks: Use Cases,

Challenges and Approaches," *in IEEE Communications Magazine*, vol. 56, no. 12, pp. 119-125, December 2018, doi: 10.1109/MCOM.2018.1701178.

[31] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *in IEEE Communications Surveys and Tutorials*, vol. 20, no. 4, pp. 3098-3130, Fourthquarter 2018, doi: 10.1109/COMST.2018.2841349.

[32] Khan, S., Akram, A., Alsaif, H. et al., "Emulating Software Defined Network using Mininet-ns3-WIFI Integration for Wireless Networks," *Wireless Pers Commun* ,118, 75–92 (2021). https://doi.org/10.1007/s11277-020-08002-w.

[33] L. Zanzi and V. Sciancalepore, "On Guaranteeing End-to-End Network Slice Latency Constraints in 5G Networks," *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*,2018, pp. 1-6, doi: 10.1109/ISWCS.2018.8491249.

[34] M. Tayyab, X. Gelabert and R. Jäntti, "A Survey on Handover Management: From LTE to NR," *in IEEE Access*, vol. 7, pp. 118907-118930, 2019, doi: 10.1109/ACCESS.2019.2937405.

[35] C. I, Q. Sun, Z. Liu, S. Zhang and S. Han, "The Big-Data-Driven Intelligent Wireless Network: Architecture, Use Cases, Solutions, and Future Trends," *in IEEE Vehicular Technology Magazine*, vol. 12, no. 4, pp. 20-29, Dec. 2017, doi: 10.1109/MVT.2017.2752758.

[36] S. A. Gbadamosi, G. P. Hancke and A. M. Abu-Mahfouz, "Building Upon NB-IoT Networks: A Roadmap Towards 5G New Radio Networks," *in IEEE Access*, vol. 8, pp. 188641-188672, 2020, doi: 10.1109/ACCESS.2020.3030653.

[37] F. J. Dian and R. Vahidnia, "A Simplistic View on Latency of Random Access in Cellular Internet of Things," *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2020, pp. 0391-0395, doi: 10.1109/IEMCON51383.2020.9284948.

[38] T. Blajic, D. Nogulic, and M. Druzijanic, "Latency Improvements in 3G Long Term Evolution," *in MIPRO'07*, Opatija, 2007.

[39] D. Singhal;M. Kunapareddy;V. Chetlapalli, "Latency Analysis for IMT-A Evaluation," *Tech Mahindra Limited*, 2010.

[40] M. P. Wylie-Green and T. Svensson, "Throughput, Capacity, Handover and Latency Performance in a 3GPP LTE FDD Field Trial," *in 2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, Miami, 2010.

[41] . Xu and C. Fischione, "Real-time scheduling in LTE for smart grids," *in 2012 5th International Symposium on Communications, Control and Signal Processing*, Rome, 2012.

[42] P. Arlos and M. Fiedle, "Influence of the Packet Size on the One-Way Delay on the Down-link in 3G Networks," *in IEEE 5th International Symposium on Wireless Pervasive Computing 2010*, Modena, 2010.

[43] M. Laner, P. Svoboda, P. Romirer-Maierhofer, N. Nikaein, F. Ricciato and M.

Rupp, "A Comparison Between One-way Delays in Operating HSPA and LTE Networks," in 2012 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), Paderborn, 2012.

[44] O. O. Erunkulu, A. M. Zungeru, C. K. Lebekwe, M. Mosalaosi and J. M. Chuma, "5G Mobile Communication Applications: A Survey and Comparison of Use Cases," *in IEEE Access* , vol. 9, pp. 97251-97295, 2021, doi: 10.1109/ACCESS.2021.3093213.

[45] S. Abdelwahab, S. Zhang, A. Greenacre, K. Ovesen, K. Bergman and B. Hamdaoui, "When Clones Flock Near the Fog," *in IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1914-1923, June 2018, doi: 10.1109/JIOT.2018.2817392.

[46] G. J. Sutton et al., "Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives," *in IEEE Communications Surveys and Tutorials*, vol. 21, no. 3, pp. 2488-2524, thirdquarter 2019, doi: 10.1109/COMST.2019.2897800.

[47] S. K. Sharma, I. Woungang, A. Anpalagan, and S. Chatzinotas, "Toward Tactile Internet in Beyond 5G Era: Recent Advances, Current Issues, and Future Directions," *in IEEE Access*, vol. 8, pp. 56948-56991, 2020, doi: 10.1109/ACCESS.2020.2980369.

[48] Strahil Panev and Pero Latkoski. 2020. SDN-based failure detection and recovery mechanism for 5G core networks. Trans. Emerg. Telecommun. Technol. 31, 2 (February 2020). DOI:https://doi.org/10.1002/ett.3721.

[49] "Transactions on Computational Science XXXIII", *Springer Science and Business Media LLC, 2018.*

[50] A. S. Thyagaturu, Y. Dashti, and M. Reisslein, "SDN-based smart gateways (Sm-GWs) for multi-operator small cell network management," *IEEE Trans. on Netw. and Serv. Manag.*, vol. 13, no. 4, pp. 740–753, 2016.

[51] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "EASE: EPC as a service to ease mobile core network deployment over cloud," *IEEE Network*, vol. 29, no. 2, pp. 78–88, 2015.

[52] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, Nov 2014.

[53] A. Jain, N. Sadagopan, S. K. Lohani, and M. Vutukuru, "A comparison of SDN and NFV for re-designing the LTE Packet Core," *Proc. IEEE conf. Netw.Function Virtualization and Software Defined Networks (NFV-SDN)*, 2016, pp. 74–80.

[54] M. Moradi, L. E. Li, and Z. M. Mao, "SoftMoW: A dynamic and scalable software defined architecture for cellular WANs," *in Proc. 3rd workshop on Hot topics in software defined networking. ACM*, 2014, pp. 201–202.

[55] R. Casellas, R. Munoz, R. Vilalta, and R. Martinez, "Orchestration of IT/cloud and networks: From Inter-DC interconnection to SDN/NFV 5G services," *in*

*Proc. Intern. Conf. on Optical Netw. Design and Model. (ONDM)*, May 2016, pp. 1–6.

[56] P. Guan, X. Zhang, G. Ren, T. Tian, A. Benjebbour, Y. Saito, and Y. Kishiyama, "Ultra-Low Latency for 5G - A Lab Trial," CoRR, vol. abs/1610.04362, 2016. [Online]. Available: `http://arxiv.org/abs/1610`. 04362

[57] S. Yoshioka, Y. Inoue, S. Suyama, Y. Kishiyama, Y. Okumura, J. Kepler, and M. Cudak, "Field experimental evaluation of beamtracking and latency performance for 5G mmWave radio access in outdoor mobile environment," *in Proc. IEEE Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.

[58] K. X. Du et al., "Definition and Evaluation of Latency in 5G: A Framework Approach," *2019 IEEE 2nd 5G World Forum (5GWF)*, 2019, pp. 135-140, doi: 10.1109/5GWF.2019.8911629.

[59] D. Fahlborg, 'Measuring one-way Packet Delay in a Radio Network', Dissertation, 2018.

[60] J. Li, 'Ultra-low latency communication for 5G transport networks', PhD dissertation, KTH Royal Institute of Technology, 2019.

[61] `https://askubuntu.com/questions/688495/what-is-the-exact-meaning-of-ping-a`

[62] `https://en.wikipedia.org/wiki/Ping_(networking_utility)`

[63] `https://linuxhint.com/iperf_command_usage/`

[64] `https://en.wikipedia.org/wiki/Empirical_distribution_function`

[65] `https://www.https://wiki.teltonika-networks.com/view/Mobile_Signal_Strength_Recommendations/`