Latency Analysis and Reduction in a 4G Network

by

Ashish Kurian

in partial fulfilment of the requirements for the degree of

Master of Science
in Electrical Engineering
Telecommunications and Sensing Systems

at the Delft University of Technology,
to be defended publicly on Monday, February 19th, 2018 at 2:30 PM

Thesis Committee	Dr. Remco Litjens MSc	TU Delft, TNO
	Dr. Ir. Fernando Kuipers	TU Delft
	Dr. Iko Keesmaat	TNO

Preface

This master thesis report marks the culmination of my studies at the Delft University of Technology for obtaining a Master of Science (MSc) degree in Electrical Engineering (Telecommunications & Sensing Systems). 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 thank the Networks department of TNO for providing me an opportunity to carry out this work at their facility. TNO ensured that I was provided with all the necessary resources to conduct my research with great quality. When I look back to the past one year and five months I spent at TNO for this work, 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 the field of networks and communication at TNO has helped me to develop my critical thinking and analytical skills.

I would like to thank Iko Keesmaat, my daily supervisor, for devoting time out of his busy schedule to provide me with feedback on my work (especially writing this report) and helping me to find the fix for the various challenges that I faced during the development of the experimental setup. His experience and expertise in writing and framing of technical reports have helped me to develop this report. I would also like to express my sincere gratitude to Remco Litjens, my supervisor at TU Delft, for his advices in shaping my research goals, measurement approach and critical feedback of the intermediate results obtained during this work. Although, his critical feedbacks were quite challenging, working based on those feedback has helped me to immensely improve the quality of this work. I would like to also thank Fernando Kuipers for agreeing to be part of my thesis committee. I cannot miss to acknowledge the help and support that I received from Daan Ravesteijn, Wieger IJntema, Niels van Adrichem, Jeffrey Panneman, Ille-Daniel Gheorghe-Pop and José Luis Almodóvar Chico for my research work.

Finally, I would like to thank my parents who have supported me financially and mentally during my study in the Netherlands. Without their support, I would not have completed this work, which took an extended duration of time. I would also like to thank my friends who have provided me their support and encouragement during the ups and downs that I went through.

Ashish Kurian

Abstract

5G, the next generation of mobile network, is expected to be launched commercially around 2020. Compared to the present generation – 4G mobile network, a significant improvement in terms of performance and reliability is considered for 5G. One of the important factor in the design of 5G is – about 10 times lower packet latency than 4G. Some of the use cases identified for 5G require packet latency as low as 1 ms. Such stringent latency targets are essential to enable new services like virtual reality streaming of live content over mobile network, automated vehicle platooning over mobile network and tactile internet where machines and tools can be controlled remotely with extreme responsiveness over the mobile network.

The main goal of this thesis is to understand how packet latency is affected by the various factors observed in a realistic environment. In contrast to lab environments, where the packet latency reported would be very low, a consolidated study on the various factors affecting packet latency in a 4G (LTE) network in a realistic environment is missing. To this extent, the results of this work have enabled to identify the various factors affecting packet latency in a realistic 4G network. This further led to identifying the latency contribution of the various components to the overall packet latency. Later on, two different latency reduction techniques were evaluated to verify the possible latency reduction achievable on a 4G network, using those two techniques.

To reduce packet latency to achieve the latency targets for 5G, first it was necessary to identify how packet latency is caused and affected in a 4G network. This work was aimed at achieving this goal. As the latency reduction techniques were evaluated at their best configuration in terms of latency, results from the latency reduction techniques also identifies the lower limit of latency improvement achievable in a 4G network. The inference from the results suggests that in order to achieve the latency targets specified for 5G networks, a redesigned radio access technology of 4G is essential.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

The future, so-called 5th generation (5G) mobile network, strives to achieve significant performance and reliability improvements compared to the present (4G) network. In addition to high throughput, extreme reliability, and support for a massive amount of (machine-type) devices, very low latency is an important aspect of 5G [1].

Latency simply speaking is the delay between source and destination caused by the (mobile) network. Latency requirements put on 5G networks sometimes are as low as 1 ms. Example applications requiring such low latencies are autonomous driving (e.g. in platoons), tactile internet (e.g. remote surgery, or remote drone control), and virtual reality (VR). For instance, for safe autonomous platooning of vehicles, the braking distance of the cars should be within 0.025 meter at a speed of 100 km/h and to achieve such low braking distances, the latency for the information that instructs the car to brake should be approximately 1 ms (or less). For tactile internet where real and virtual objects can be controlled remotely, the response time of the system should be in the millisecond range. For VR to handle head movement without causing nausea it requires latency to be approximately 1 ms or less [2] [3] [4].

In order to be able to improve latency (i.e. to reduce it), it is necessary to understand latency in (current) mobile networks first. That is, to understand what is the contribution of the various (network) components to the total network latency, what are the main reasons latency occurs in the various components, what is the effect on latency of various parameters (e.g. packet rate, network load, distance to a base station), and how latency can be improved in current networks (e.g. by using differentiated scheduling, or by using Edge Computing). Answering these questions will enable to identify the most crucial elements in the search for low latency.

Although some works on latency analysis on 4G network are available, the majority of those are based on analytical estimations. In the measurement aided packet latency experiment results available in the literature, none of them provide a consolidated study on the various factors affecting the packet latency. Moreover, works done to measure one-way packet latency in the network are mostly done using ping programming with the underlying assumption that the links in the uplink and downlink direction are symmetric with respect to latency. A detailed review in the already available literature will be presented in Chapter 3.

Considering the state of the art of the available literature on latency in this thesis the following goals are set:

- Determine the latency breakdown, separately for UL and DL - by measuring - in an emulated mobile network under realistic network and application conditions. The latency breakdown will provide insight in the contribution to the total latency of the various network components. Realistic network and application conditions include
    - the effect of the network load (both in radio network and in core network);
    - the effect of distance (the mobile device to base station distance);
    - the effect of backhaul topology (effects of multiple hops and distance of the links between the radio and core network);
    - the effect of application-level aspects (e.g. packet sizes and packet rates).

  The main tool for the measurements will be a measurement setup consisting of a prototype mobile network with standard (4G) mobile devices enhanced with tools for

simulating realistic network and application conditions forming the emulated mobile network.

- Explain the causes for the observed latency behaviour in the various scenarios.
- Assess the potential of latency reduction techniques by applying them in the developed measurement setup.

The outline of this thesis is as follows.

In Chapter 2, various definitions of latency in a (mobile) network are presented. The latency targets defined for the various releases of LTE (Long term evolution) until and including 5G are also presented in this chapter as well as the need for low latency. Then a discussion on the drawbacks of these definitions in the context of this thesis is presented. Following the drawbacks, the definition of latency used in this thesis is given and it is shown how this definition is suitable for the goals set for this thesis.

In Chapter 3, the research questions that are addressed by this thesis are presented. Along with the research questions, a brief description of the basic and relevant concepts, factors causing latency in a mobile network and a review of the available literature on latency analysis in mobile networks are presented. The limitations of the studies found are discussed and to what extent the approach of this thesis differs from the existing work on latency.

In Chapter 4, the requirements to the measurements and the measurement setup are listed as they follow from the research questions given above. Consequences following from the basic requirements are also presented.

In Chapter 5, the measurement setup developed for this thesis work is discussed. In this chapter, the various components and tools constituting the measurement setup are listed and the reasons for their selection are discussed. This chapter also presents a detailed explanation of the functioning of the different tools used and their integration to form the complete measurement setup.

In Chapter 6, a description on the measurement strategy and the default scenario which is used as the base for comparing the results are presented.

In Chapter 7, the results of the various scenarios considered and the detailed analysis of the results for the measurements in both uplink (UL) and downlink (DL) are presented. The results of the processing delay in the UL and DL in the radio network will also be presented in this chapter.

Finally, in Chapter 8 the key conclusions from the results of the measurements, answering the various research questions presented in Chapter 3 are presented. Along with the conclusion, suggested directions for future research are presented.

Chapter 2. Latency in a mobile network

Since the aim of this thesis is to analyse the latency in a 4G network, it is important to familiarise ourselves with the various aspects of latency. In Section 2.1, the definitions of latency found in standards, in literature and other industry whitepapers are presented. In Section 2.2, the need for low latency and the latency targets for the various 'versions' of LTE and 5G are presented. In Section 2.3, the limitations and uncertainties of those definitions in the context of this thesis are presented. In Section 2.4, a description of latency, which is in alignment with the requirements of this thesis is presented.

2.1 Latency definitions found in the literature, white papers and standards

In this section, latency definitions as found in the literature are discussed with a special focus on latency definitions as used/defined in 3GPP documents, since the main reason for the study into latency is the requirement for low latency in 5G networks.

According to 3rd Generation Partnership Project (3GPP) [5], latency in the network can be classified into *control plane latency and user plane latency*.

Control plane latency is related to the so-called idle mode. In LTE, in order to save the power of the user device/user equipment (UE), the UE enters into an idle mode after a defined period of inactivity. In idle mode where the UE is in RRC disconnected state, the UE listens to paging signals, only once in every paging cycle [6]. Before the UE can transmit or receive any packets, the UE must be in active mode and hence transit from idle mode to active mode, if the UE is in idle mode. In active mode, the UE is in RRC connected state and listens to detect any DL data in every millisecond. The transition delay for this process is defined as the control plane latency. For the UE to transit from idle to active state, some operations must be performed in the radio and the core network. Thus, control plane latency is the combined delay for this process in both the radio and core network.

Within 3GPP there are essentially two definitions for user plane latency for 5G. The first and most important definition of latency is defined and discussed in 3GPP TS 22.261 [7], where end-to-end latency is defined as "*the time that takes to transfer a given piece of information from a source to a destination, measured at the communication interface, from the moment it is transmitted by the source to the moment it is successfully received at the destination*". This definition does not specify what the source and destination are, but it clearly defines latency as the one-way transit time between source and destination (as contrasting to round trip time). A note in the document states, however, "The end-to-end latency is not completely allocated to the 5G system in case other networks are in the communication path." which seems to imply that end-to-end latency also includes systems outside the 5G networks. As TS 22.261 is putting requirements on the whole 3GPP network, it seems logical that end-to-end latency is not only restricted to the radio network only. Note that in the same document other latency terminology is used, such as "one-way latency", "round-trip latency", "two-way end-to-end-latency", and "motion-to-photon latency".

In contrast to the above, in 3GPP documents concerned with radio network requirements [5], user plane latency is defined as "*the one-way transit time between a packet being available at the IP layer of the UE/RAN edge node and the same packet being available at the RAN edge/UE. The RAN (Radio Access Network) edge node is the node providing the radio access*

network interface towards the core network". That is, in this definition only latency in the radio network is considered. A similar definition of latency in a mobile network is provided by ITU [8].

According to the latency definitions considered in the next generation mobile networks (NGMN) whitepaper [9], two definitions are considered:

- The *E2E (end-to-end) latency*: is the time taken for the transport of a small data packet from the application layer at the source node until its reception at the application layer at the destination node, plus the equivalent time required to carry the response back;
- The *user plane latency*: is the time taken to transfer a small data packet from the user device to the layer 2/3 interface at the 5G system destination node plus the equivalent time to carry the response back.

For the E2E latency definition, the placement of the destination node is not explicitly specified. It may be placed outside the 5G system, on an external network. Thus, this definition seems similar to the 3GPP end-to-end definition found in [7] in terms of the placement of the destination node, but the definition of 3GPP considers only the delay in one direction. However, for the user plane latency definition by NGMN, the destination node must also be in the same 5G network. This definition is different from the user plane latency according to 3GPP in [5], where it is restricted to one-way transit time only on the radio network.

According to latency definitions found in the literature, two definitions are considered:

- *One-way delay (OWD)*: The delay for a packet to reach from the source to the destination. This delay must be specified along with the direction of the packet travel.
- *Round trip time (RTT)*: The delay for a packet to reach from the source to the destination and a response back to the source. This metric is usually specified with latency estimation based on ping. This definition of latency is coinciding with the E2E latency definition specified by the NGMN whitepaper.

The placement of the destination node considered for the OWD is usually outside the mobile network and in an external network. For RTT, the placement of the server that receives the packet and sends the response, is also usually located in an external network.

Figure 2-1 visualises the various latency definitions discussed in the above section.

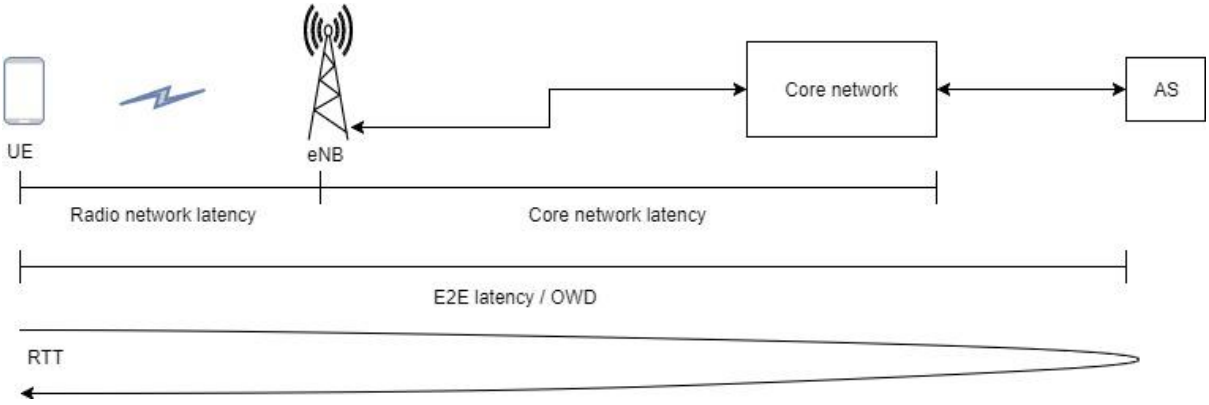

Figure 2-1: Latency visualisation

Although many industrial reports and white papers are available, which defines latency and latency targets for the various use cases in 5G, they are sometimes contradicting with each other. For example: the user plane latency definition by 3GPP and NGMN white paper are not

the same. Thus, it is not worthwhile to discuss the latency definition in all the sources, as it can confuse the reader. After the search for latency definition, a general conclusion is that: a single definition of latency agreed by all the organisations is missing.

2.2 Need for low latency and common latency targets

As already mentioned briefly in Chapter 1, the necessity for the ultra-low latency targets for 5G network originates from the new use cases to be facilitated by 5G. The already mentioned use cases are not the only use cases where low latency is required. There are many other uses cases requiring lower latency than 4G, but the mentioned ones are some of the use cases requiring the lowest latency targets.

For the case of automated vehicle platooning [2], vehicles are moving close to each other. Overall fuel consumption can be reduced and more efficient road usage can be achieved with a coordinated close together movement of trucks. To achieve this level of fuel saving, the vehicles must be around 1 meter apart. For such close coordinated movement of vehicles, a lot of information needs to be exchanged between all the vehicles within a short duration. These messages ensure that all the vehicles are informed about the status of other vehicles and the required distance is kept between each of them. Very small distance between each vehicle also means the requirement for a very short braking distance. In case of some emergency situation to one of the vehicles, the network must be responsive enough to ensure that the rest of the vehicles brake immediately and avoid any collision. For ensuring safety for vehicle platooning with an inter-vehicle gap in the range of a meter, the braking distance should be around 0.025 meter. With the current 4G network, with an end-to-end latency of around 50 ms, the vehicle would have moved a distance of 1.4 meter from the instant the sensors on the vehicle under collision course detect an obstacle and the braking command is executed on the rest of the vehicles. But with a one-way delay latency target less than 10 ms for 5G, the vehicles would be moving about. 0.025 meter. Thus, latency as low as the 5G latency targets are essential for automated vehicle platooning.

Another use case requiring very low latency is the tactile internet [3]. The idea behind tactile internet is to control machines remotely in real time. Tactile internet can enable users to wirelessly control real and virtual objects. Such a technology can have great practical advantages. For example, in case a car breaks down in the middle of nowhere, tactile internet can enable a mechanic at a remote location to fix the car by enabling the mechanic to control the necessary tools to repair the car. Another example is remote surgery. A surgeon can remotely perform surgery on a patient by remotely controlling the tools for the surgery. These use cases require the network to be responsive enough to enable very precise and responsive manoeuvring of the tools. Such a network should have a one-way delay in the range of 1 ms.

According to an Ericsson white paper [10], the 5G use cases can be categorised into three: massive machine type communication (MMTC), critical machine type communication (CMTC) and enhanced mobile broadband (eMBB). The already mentioned use cases of 5G fall into the category of CMTC (termed as URLCC – Ultra-Reliable and Low Latency Communication in 3GPP). For these use cases, it is very important that the packet delivery happens within the specified latency targets or else it can lead to a catastrophe. Yet another use case that require a similar low latency target is VR. However, as a failure to meet the latency requirement in VR does not cause any catastrophe, it is not falling into the category of CMTC. One important aspect in VR is to ensure a low motion to photon latency (MTP). The MTP is the delay between the head movement of the user and the view port content being updated in response to the

head movement. The view port of a user is the field of view of the user, from the available 360^0-video. For non-interactive content, buffering is not an issue, such as for streamed 360^0 videos. But for interactive content like a first-person shooter online game using VR, latency of the network can be the deciding factor between shooting or being shot by your opponent. For such applications, the content delivery should be immediate, requiring very low latency.

Moreover, the view port of the user on the VR headset should move immediately in response to the head movement for such applications. If there is high MTP latency, it can cause disorientation and dizziness to the user. Therefore, ensuring low latency for such application is important to ensure that there is complete immersive experience from the application while maintaining the comfort of the users [4]. The use case of VR requires a round trip time of 10 ms.

Based on considerations such as this for 5G and previous consideration for LTE networks, latency targets have been formulated for the various 'versions' of LTE, i.e. LTE, LTE Advanced, and LTE Advanced Pro.

For LTE, the latency target for the transition from idle mode to active mode is 100 ms. The latency target for transition from the dormant mode to the active mode is 50 ms. The dormant mode is also an energy saving mode for the UE where the UE is in RRC connected state. The UE context is available in the cell and UE listens to paging signals more frequent than the idle mode. As there is already RRC connection established, transition to active mode is faster than from the idle mode. These are the two control plane latency targets. For the user plane latency, the target value is 5 ms [5]. The transitions with their targets are shown in Figure 2-2.

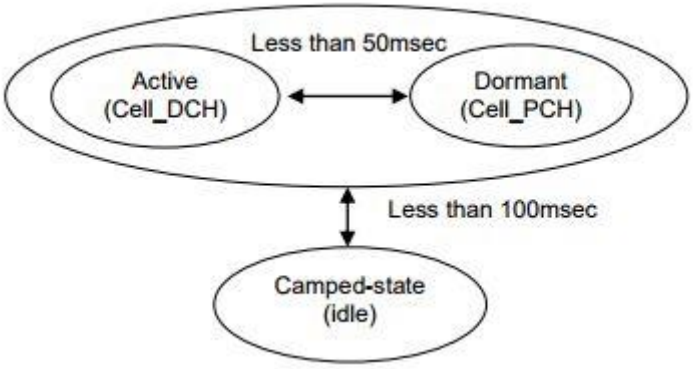

Figure 2-2: LTE control plane latency [5]

For LTE-Advanced, the latency targets have been further reduced in comparison to LTE. The latency target for the transition from idle mode to connected mode is 50 ms. The latency target for transition from dormant state to active state is 10 ms. However, for the user plane latency, the target value of 5 ms is kept the same as in LTE [11]. The transitions with their targets are shown in Figure 2-3.

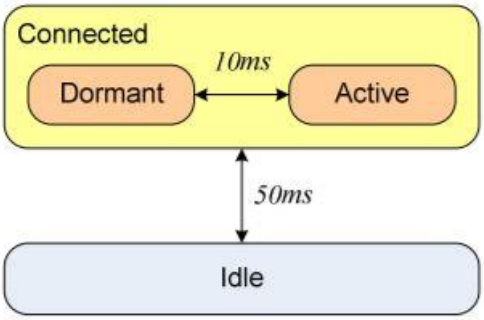

Figure 2-3: LTE-Advanced control plane latency [11]

The LTE-Advanced Pro is also referred as 4.5G or pre-5G. 3GPP has not yet published an official latency target specification for LTE-Advanced Pro. However, the proposal to reduce user plane latency to 1 ms for one-way delay is provided. This reduction in user plane latency is achieved by using a shorter frame length [12]. A shorter frame length is achieved by reducing the duration of each of the ten transmission time intervals (TTIs), that constitute one frame. In the LTE and LTE-Advanced, the duration of a TTI is 1 ms. For LTE-Advanced Pro, the proposal is to reduce and make it into a flexible value between 0.14 ms and 0.5 ms depending on the requirement as shown in Figure 2-4.

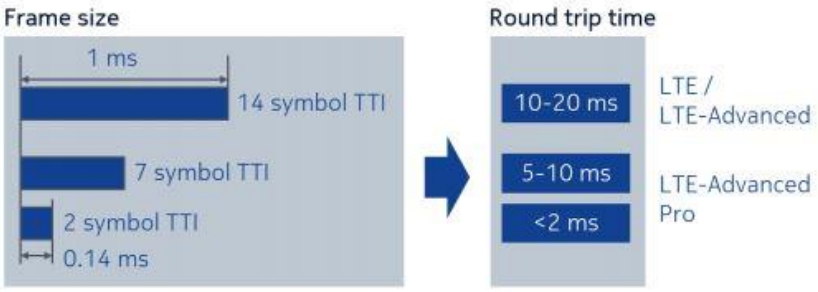

Figure 2-4: LTE-Advanced Pro flexible frame

According to 3GPP latency targets specified for 5G, the control plane latency should be 10 ms. The user plane latency should be 0.5ms for UL, and 0.5ms for DL for the use cases of Ultra-Reliable Low Latency Communications (URLLC) [13] which is the same service category of CMTC by Ericsson in [10]. For the use case of enhanced Mobile Broadband (eMBB) [13], the target for user plane latency should be 4ms for UL, and 4ms for DL [14].

The NGMN white paper specifies the latency targets for a 5G system. The latency target for E2E latency is 10 ms in general and 1 ms for use cases that require extremely low latency [9]. This target values agree with the target values specified by 3GPP (although a different definition of latency is used). Also, the targets values for the different use cases agrees with the targets specified by 3GPP.

2.3 Uncertainties and limitations of user plane latency definitions

In the latency definitions from different sources, a consideration of the actual packet size or the network configuration are not specified. It is assumed that the IP packets considered in these definitions have no payload and has only IP headers.

According to 3GPP, the definition for user plane latency is defined under the situation where the system specification shall enable the network to achieve the user plane latency target (unloaded condition i.e. single user with single data stream and small IP packets). It is

assumed that the UE is synchronised to the network so that it can extract information from the network. Also, it is assumed that the UE is already scheduled by the radio network.

But in practice, data packets seen in a network will have varying payload size and the network conditions would vary over time. The packet arrival rate can also vary depending on the type of application used. The network may not always be optimised for the best packet latency performance, as such a configuration could affect other performance-related metrics of the network. Moreover, the UE can lose synchronisation with the network over time. A user may not always be scheduled in an actual network. Thus, considering all these aspects, it can be assumed that these latency targets are considering the lowest possible packet latency under the assumed network conditions.

Latency targets specified in e.g. 3GPP or NGMN documents do not mention whether the latency is the minimum or the average or the maximum of all packet latency. One could wonder, what would be the latency of the network for a larger packet or what would be the latency of the network when the loads are high or when the user is positioned very far from the radio base station etc.

Therefore, such a definition of network latency, for measurements done on an actual network may lead to an ambiguity as all the packets will not report the same value of packet latency. As the results of latency measurement might possibly be different from these defined values, the reader might wonder what is the merit of such a definition. A more elaborate insight into the packet latency of the network is required to fully understand the latency experienced by the packets in the network. The questions of what is the latency value for a larger packet, higher packet arrival rate, loaded network, various user position etc. need to be addressed to fully understand the latency in the network.

Latency can be well-defined on a per packet basis (one-way or roundtrip; RAN or core network or E2E). At the packet flow or network/system level, the latency incurred by multiple packets are aggregated into a KPI (e.g. a percentile or average). Latencies both at the packet, flow or system level may depend on load conditions, user locations and other scenario aspects.

2.4 Latency definition considered for this research

To avoid any confusion with the latency definition, a definition of latency is formulated for this thesis. The one-way packet latency considered for this thesis is the time taken for a packet to be available at the source node until the same packet is available at the destination node. The node acting as the destination node for the UL and the source node for the DL, is placed next to the considered mobile network and the packets are not traversing through any external network. The latency will also be specified by the direction of packet travel.

This definition of latency corresponds to the OWD definition found in the literature and end-to-end latency definition by 3GPP in [7], with the exception that the destination node is not placed on an external network, but next to the core network. In the results, the term end-to-end latency will be used in the plots, which is in accordance with the definition of 3GPP.

It is intended that the latency value reported on a network should provide the reader with insights into what is the range of the packet latency in the network. In this thesis therefore, three latency KPIs are considered: the 10th latency percentile, the average latency and the 90th latency percentile. Along with the three different levels of packet latency, the reader is also provided with complete packet latency distribution for each of the considered scenarios.

These three latency KPIs will depend on e.g. the network load, user position and traffic conditions considered in the assessment scenario.

Another KPI, that is not directly related to the latency is the packet drop percentage. If packets are dropped in the network due to some reason or other, it means an infinite latency for that packet. Therefore, it is important to evaluate this KPI also, while determining the packet latency in the network.

Chapter 3. Research objectives and relevant concepts

In this chapter, the reader is presented with the research objectives of this thesis along with already existing work on latency analysis in a 4G network. In Section 3.1, the research objectives of this thesis work are presented. In Section 3.2, a brief description of the components of a mobile network is presented. In Section 3.3, a brief description on the various processes in the network components that causes latency is presented. In Section 3.4, a brief description on the various scenario aspects that can affect latency in a mobile network is presented. In Section 3.5, review of already available works on latency analysis in a 4G network is presented, along with the contribution of this thesis, in addressing the gaps of the already available literature.

3.1 Research objectives

A mobile network consists of various components that functions cooperatively to transport data to and from a UE reliably. The functions carried out in the various components are different and therefore the packet latency experienced in each of these components could be different.

To achieve the latency targets of 5G, it is necessary to have a complete understanding of how the end-to-end packet latency is built up in a 4G network. To design new and improved technology for 5G, it is necessary to understand the reasons why the latency observed in a 4G network is as it is. This understanding involves the knowledge of the causes for latency in the various components.

A 4G mobile network in a lab environment, such as the one used in this thesis, might already deliver quite low latency for packets. However, the latency reported on such an idealistic network cannot be considered as a genuine latency figure of a 4G network, since there can be various factors that negatively affect the packet latency in a realistic environment.

After gaining a complete understanding of the various components of a 4G mobile network, the various factors affecting packet latency in a realistic environment and the causes for latency in the various components, it is possible to understand the major reason for packet latency in a 4G network. Only with this knowledge, it will be possible to design solutions for improving latency.

Thus, the main research objectives for this thesis work are to understand:

- Latency contribution of the various network components
- Reasons for latency in the various components
- Various scenario aspects that can affect latency in a realistic network
- Possible solutions to improve latency

In the following sections, an introduction into the various components, the factors affecting latency in a realistic network and the reasons for latency in the various components are presented.

3.2 Mobile network and its components

In a mobile network, there are two components: the radio network and the core network. In the radio network, the packets are transmitted wirelessly from the UE to the radio base station

- also referred as eNodeB (eNB) in LTE. In the core network, the packet exchange uses a wired medium. The various processes carried out in the two components of the network are different.

3.2.1 Evolved Universal Terrestrial Access Network

The Evolved Universal Terrestrial Access Network (E-UTRAN) is the access part of the Evolved Packet System (EPS). The E-UTRAN is based on the principle of Orthogonal Frequency Division Multiple Access (OFDMA) in the DL and Single Carrier - Frequency Division Multiple Access (SC-FDMA) in the UL. Its features include modulation up to 64QAM, a carrier bandwidth to a maximum of 20 MHz and the use of MIMO with spatial multiplexing (SM; downlink) [15]. Figure 3-1 shows the architecture of the E-UTRAN.

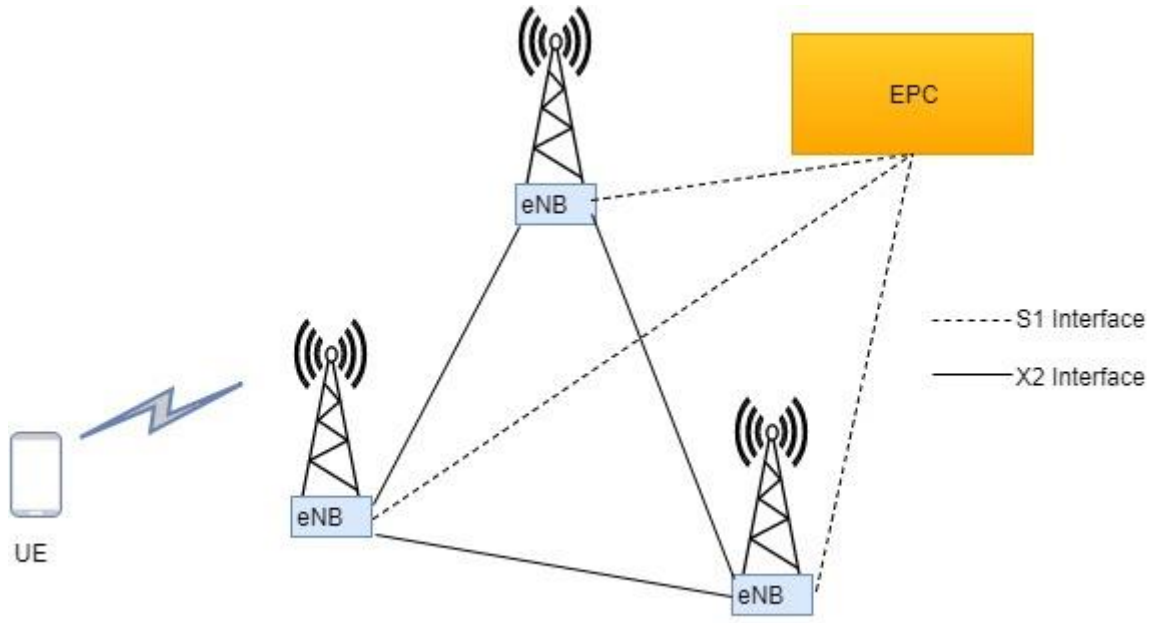

Figure 3-1: E-UTRAN architecture

The eNBs are connected to each other via the X2 interface and to the core network via the S1 interface. An advantage of E-UTRAN compared to UMTS, is the reduced Transmission Time Interval (TTI) of 1 ms. It is in this interval that a UE is assigned the network resources and is permitted to send or receive data. In every TTI, the scheduler chooses a particular link adaptation level to ensure the best overall performance. With a TTI of 1 ms, UEs are always assigned resources for a duration of 1 ms which leads to more frequent resource allocation in a given period of time. This can lead to reduced packet latency in the radio network.

3.2.2 Evolved Packet Core

In this thesis, the core network used is the Evolved Packet Core (EPC). The EPC is the 4th generation 3GPP core network architecture. When developing the 4G systems, the 3GPP community decided to have an all IP network with packet switching architecture. Thus, the EPC is effectively an evolution of the packet-switched architecture seen in the GPRS/UMTS network [16].

In the LTE core network, as there is only a packet-switched domain, all the services are designed based on packet switched domain and thus there is no protocol conversion required

unlike GPRS/UMTS. Also, fewer number of network nodes are used for the handling of user traffic. For these reasons, the EPC is considered as a flat architecture compared to that of GPRS/UMTS core network. It was also decided to have a split in the user plane and control plane so that the scaling is independent. This enabled the core network to be dimensioned independently according to the user plane or control plane demands.

The Figure 3-2 shows the basic architecture of the EPC where the UE is connected over E-UTRAN. The EPC is composed of four network elements [17]:

- *Mobility Management Entity (MME)*: is the key control node of the LTE core network that deals with the control plane. It takes care of the signalling related to security and mobility for the E-UTRAN. It is also the node responsible for tracking the user devices in the network and paging in case of any new data for the devices.
- *Serving Gateway (S-GW)*: is the interconnection point between the radio side of the network and the EPC. It routes incoming and outgoing user plane IP packets towards the UE.
- *Packet Data Network Gateway (PDN-GW)*: is the interconnection point between the EPC and the external IP networks. It routes the user plane IP packets to and from the external IP networks. It also performs the functions of IP address/prefix allocation to the UEs and policy control and charging.
- *Home Subscriber Server (HSS)*: is a database that stores the various user-related and subscription-related information. In coordination with the MME, the HSS performs the functions of mobility management and user authentication.

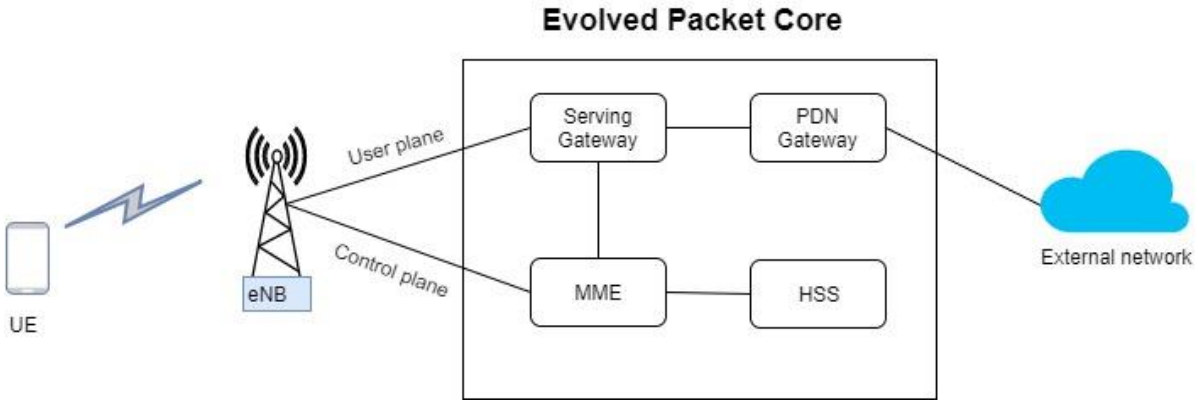

Figure 3-2: EPC Architecture

3.3 Reasons for latency in network components

In the *radio network*, the eNB takes care of assigning the radio resources, also called Physical Resource Blocks (PRBs) to the users in the network. While scheduling the PRBs, the eNB estimates the wireless channel quality for each user and adapts the transmission parameters (including the selected Modulation and Coding Scheme (MCS) and the transmit mode, e.g. transmit diversity of MIMO/SM) to meet the target Block Error Rate (BLER) which is typically 10% [15]. The eNB must ensure that the packets are arriving in the proper order and are received without any errors. These functions are performed by the various sub layers in the E-UTRAN user plane protocol stack and therefore the packets will experience some definite delays. A similar process is carried out in the LTE UE to reconstruct the packets received from the eNB. Thus, packet latencies are caused inside the UE and eNB due to the sub layers in the E-UTRAN user plane protocol stack.

The packet transfer procedures in the radio network is different for the UL and DL. Therefore, depending on the direction of the packet, different signalling procedures happens. The combined delay from the transfer procedure and the delays from the different sub layers in the E-UTRAN is called the *processing delay*.

Besides the processing delay, as the scheduling mechanism may vary in the UL and DL direction, there may be different *scheduling latency* in the UL and DL. Thus, the processing and scheduling latency will lead to different latencies in the UL and DL of the radio network. A detailed explanation of the processing and scheduling latency will be presented in the next section.

In the *core network*, processes which ensure proper routing of the packets towards the destination are performed. Compared to the radio network, there is no need for link adaptation or scheduling latency caused in the core network. However, there are some definite delays involved in the core network due to the packet processing inside the SGW and PDN GW. Moreover, the packets may be buffered in the core network causing some buffering latency to the packets. In addition to the latency caused due to processing and buffering, the backhaul topology can also cause latency due to the transport delay and other latencies due to the multiple hops in the backhaul topology.

The various processing and scheduling latencies are not standardised but are vendor-specific, and therefore latency depends on the hardware configuration of network considered.

3.3.1 Processing delay in the radio network

As already mentioned, the packets in both the UL and DL are processed by the various layers of the E-UTRAN user plane protocol stack. As the processing delays are vendor specific, there is no possibility to determine these values in a mobile network without actually measuring them. All the packets exchanged between the UE and the eNB experience this delay irrespective of the direction of packet transfer or any other conditions in the network.

The different protocol layers in the user plane protocol stack are [18]:

- *Packet data convergence protocol layer (PDCP)* – The PDCP layer is responsible for the header compression and decompression of all the user data, handover management: re-ordering and sequencing of packet data units (PDU) during handover and performs encryption and decryption for user and control plane data along with integrity protection and verification of the control plane data.
- *Radio link control layer (RLC)* – The RLC layer is tasked with segmentation and concatenation of PDCP PDUs to fit the size requirement for the MAC layer. RLC also reorders packets received out of order.
- *Medium access control layer (MAC)* – The MAC layer distributes the available frequency resources to the active UEs. MAC enables a UE to access and synchronise with the network through random access procedure. The MAC layer also performs the HARQ operation to retransmit and combine received data blocks in case an error occurred.
- *Physical layer (PHY)* – The PHY layer is the layer that receives all the data exchanged between the UE and the eNB. The PHY layer uses an OFDMA multiple access scheme in the DL and a Single Carrier FDMA multiple access scheme in the UL. It is in the PHY layer that the transmission parameters are adapted while exchanging data between the UE and eNB using frequency- and time-domain resources.

The PDCP, RLC and the MAC layer constitute layer 2 in the protocol stack and PHY layer constitute layer 1.

Apart from these delays, extra delay is caused by the different signalling procedures involved for the packet transfer, that are different for the UL and DL as discussed in Sections 3.3.1.1 and 3.3.1.2 respectively.

3.3.1.1 UL signalling

The Figure 3-3 indicates the UL signalling procedure in the radio network. The values for each of the latency component in the UL signalling procedure are specified by 3GPP in [19].

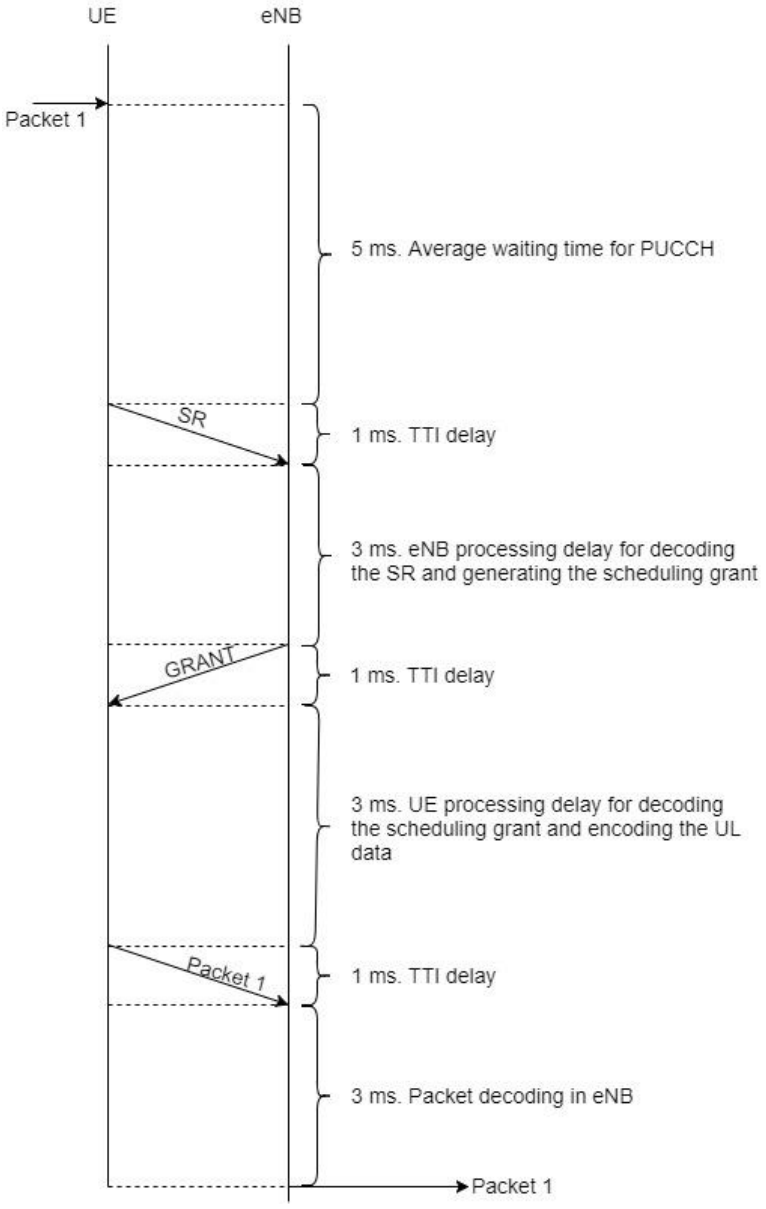

Figure 3-3: Uplink signalling

After the UE has detected that it has a packet to send in the UL, the UE must generate an uplink Scheduling Request (SR) so that eNB can allocate resources to the UE to transmit the packets. However, this SR will not happen immediately. A UE generates the SR request after a period of 5 ms. The SR transmitted on the Physical uplink control channel (PUCCH) format 1, will have an average waiting time of 5 ms with a SR periodicity set to 10 ms.

After the generation of the SR, the SR will be transported to the eNB within a TTI duration which is 1 ms.

The eNB takes approximately 3 ms to process the SR request and until an uplink grant is generated. The UL grant message is sent in the control channel in the DL and the generated uplink grant is then transported to the UE within a TTI duration.

After the UE receives the UL grant, it decodes the UL grant to identify the transmission parameters to be used in the UL. Due to the time required for the UE to process the UL grant and prepare the data, after 3 ms, the UE transmits the packet in the UL using the information obtained from the UL grant.

The UE transmits the UL packet within a TTI duration.

The eNB decodes the packet, involving all the processing delays in the different layers in the protocol stack. This can be approximated to 3 ms. The eNB then reconstructs the UL packets as an IP packet with proper headers to be forwarded to the core network.

All these delays add up to 17 ms.

3.3.1.2 DL signalling

The Figure 3-4 indicates the DL signalling procedure in the radio network. When a packet arrives into the eNB towards a UE and if the UE is in idle mode, the eNB pages the UE. The UE then enters the active mode and starts listening on the control channel for any data in the DL. As long as there is data coming to the eNB in the DL for a UE, the control channel broadcasts the information on how to extract those packets. The active UE, constantly listens to the control channel to detect any DL packets. From the information broadcast on the control channel, the UE then decodes the resource blocks and the MCS it should use to extract the data. If the UE does not find any DL data information destined to it for a specific period of time, it enters the idle mode.

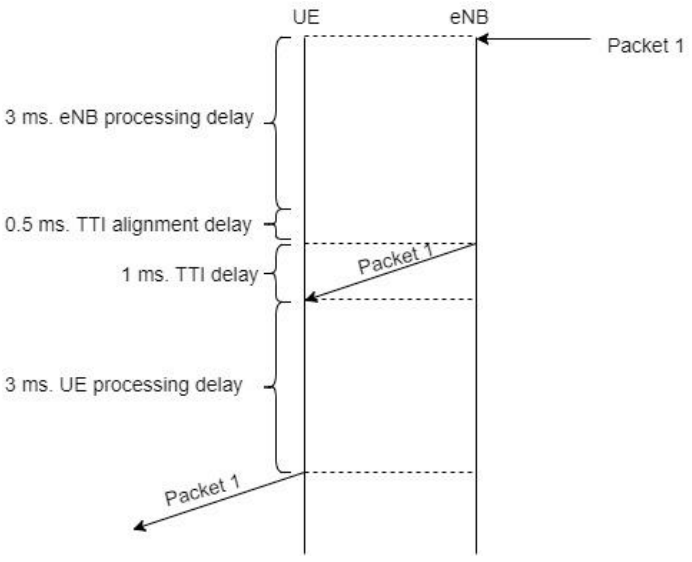

Figure 3-4 : Downlink signalling

After the packet enters the eNB in the DL direction, it takes approximately 3 ms for processing the incoming data packet in the eNB. It takes another 0.5 ms for the TTI alignment. The packet is then sent in the DL within a TTI duration.

The packet is then received at the UE and it takes approximately 3 ms for the UE processing until the packet is reconstructed as the original, before it is available for use in the UE. The total delay for the whole process is then 7.5 ms.

3.3.2 Scheduling latency

The packet latency is further affected also by the potential scheduling latency in the network. Depending on the current cell load, position of the user in the cell and the scheduling strategy used for UL and DL, packets experience different latencies in the network. A scheduler is used in LTE to assign and share the limited network resources among the users of the network. A packet is not completely transmitted from the UE/eNB buffer until enough radio resources are assigned. Therefore, depending on the rate at which the radio resources are assigned, which depends on the current network load, position of the user in the cell and the scheduling strategy itself, the latencies experienced by the packets varies.

In LTE, the minimum unit of radio resource allocation is called Physical Resource Block (PRB). A PRB is composed of a time-frequency plane. Thus, a PRB can also be considered as a resource grid as shown in Figure 3-5. A resource grid is consisting of seven OFDM symbols in the time domain and twelve adjacent subcarriers in the frequency domain. The subcarrier spacing of each subcarrier is 15 kHz resulting in a total of 180 kHz. The seven OFDM symbols form a slot with a total duration of 0.5 ms.

The smallest data carrying element in the resource grid is called the resource element – formed by one subcarrier and one OFDM symbol. Depending on the SINR of the received signal, the amount of data carried in a resource element is varied. If the SINR is very low, indicating a poor radio channel, lower amount of data is transmitted in a resource element. Such a lower amount of data transferred per resource element is enabled by lower order modulation schemes such as QPSK which is less error prone compared to higher order modulations like 64 QAM. In LTE, the requirement for the received signal is to achieve the target BLER of 10%. This target is ensured by choosing appropriate transmission parameters. Thus, as SINR decreases, a more robust, lower order modulation will ensure that the target BLER is maintained even in case of a poor radio channel. This link adaptation feature in LTE is carried out by varying the modulation and coding scheme (MCS). Therefore, based on the channel quality estimation, a particular MCS is used for the set of PRBs assigned to a user. Besides varying the MCS, other aspects that can be adapted to meet the BLER target are the two transmission modes: transmit diversity and spatial multiplexing. Therefore, based on the SINR of the received signal, the transmission parameters (MCS and transmission mode), are adapted and due to this the amount of data cleared from the buffer varies over time.

The Transmission Time Interval (TTI) is the interval during which a UE is assigned the radio resources by the scheduler. Depending on the scheduling strategy, a UE might be assigned a different number of PRBs and over different TTIs. Therefore, the amount of data that can be transported in a TTI depends on the number of PRBs assigned in that instant and also on the MCS and the transmission mode used for the assigned PRBs.

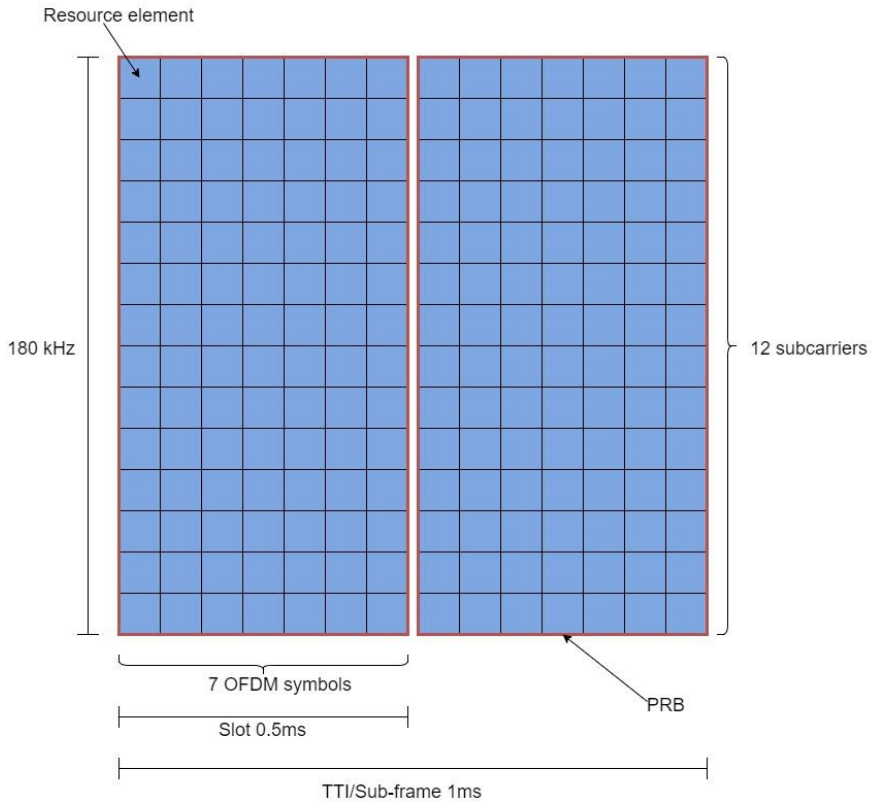

Figure 3-5 : Physical resource blocks constituting resource grid

3.4 Factors affecting packet latency

As explained in the section above, to clear the packet from the buffer, radio resources must be utilised to transport them. Therefore, depending on the amount of resources available to a given user at every time instant, the packet might require different duration to be completely cleared from the buffer, resulting in potentially different latencies for each packet. Therefore, parameters that demand different resource requirement or parameters that affect the resource availability in every TTI can in turn affect the packet latency. In this thesis, six parameters have been identified that could potentially affect the packet latency in a mobile network. In the following sections, a detailed description of each of these factors will be presented.

3.4.1 Parameters affecting packet latency

These six parameters will be referred as the six scenario aspects of the measurements performed in this thesis. The six scenario aspects identified are as follows:

- Network Load
- Positioning of the user with respect to the eNB
- Packet size
- Packet rate
- Differentiated scheduling in the radio network
- Edge Computing (EC) enabled core networking

The *network load* aspect considers the load on both the radio network and the core network. Due to the load in the radio network, the resource allocation per user is reduced. Thus, user's data packets may experience some latency while waiting for sufficient resource assignment

from the radio network. The different network load also causes different levels of inter-cell interference in the network. This can cause a reduced SINR and hence a lower MCS, which causes a lower bit rate and consequently higher packet latency. The mechanism on how SINR affect packet latency will be explained in Section 5.2.5. In the core network, due to different levels of load, the resource sharing among the different users causes the packets to be buffered in the core network causing latency to the packets.

The *positioning of the user* with respect to the eNB consider the variations in the SINR due to path loss and inter-cell interference. As distance increases, due to increased path loss, the received signal strength is reduced. Besides the path loss, an increased distance typically also means an increased inter-cell interference. Due to the increased path loss and inter-cell interference the SINR of the signal in reduced. A reduced SINR causes lower bit rates and hence higher latency.

Different applications require different traffic characteristics in terms of packet size and packet rate. A larger *packet size* requires more resource assignment from the network than a smaller packet. For packets arriving at a higher *packet rate*, the queuing latency in the eNB/UE buffer would be higher than packets with a lower packet rate. Therefore, packets may experience different latencies depending on the packet size and the packet rate.

Applying *differentiated scheduling* in the radio network, the tagged user is assigned the radio resources with a higher priority compared to the other users in the network. The user might be assigned with different levels of prioritisation. As the user gets more radio resources than in the case with no prioritisation, the packets experience reduced latencies.

In the Edge Computing enabled core networking, the considered user is served by a core network (edge computing node) or a processing unit, kept very close to the radio network. As the edge computing node is placed close to the radio network, the packets experience significantly lower transport delay and the delays caused by multiple hops in the backhaul topology, as it would have been in a normal core network. In a normal EPS architecture, the core network is placed usually hundreds of kilometres away from the radio network resulting in significant transport delay for packets.

Besides the advantage of low transport delay, the edge computing node can be dynamically provisioned with resources according to the instantaneous traffic demand, independent of the normal core network. Therefore, the packet processing in the edge computing node is not affected by load in the normal core network resulting in a lower latency for packets.

After identifying the latencies involved in the various network components and how the various parameters affect latency, a complete picture of how the packet latency, as in an actual network would be, is obtained.

With this knowledge, it is possible to identify the most crucial parameter or component causing the packet latency. The various latency reduction techniques can then be implemented to reduce the latency in the network. By evaluating the latency improvement achieved using differentiated scheduling, it will be possible to identify the maximum possible latency reduction achievable in an LTE radio network. By evaluating the latency improvement achieved using EC, it will be possible to identify the maximum possible latency reduction achievable on the core network.

3.5 State of the art in latency measurement in LTE

Latency in 4G/LTE network has been investigated in various papers. For instance, authors in [20] and [21] discuss control plane latency and user plane latency in LTE network, but more attention is given to control plane latency. These two works on latency assessments are based on estimations and the results are not obtained based on actual measurements. The user plane latency results in these works show only the processing delay caused in the radio interface due to the user plane protocol stack, based on estimation. What these measures give is the bottom limit of the latency caused by the various sub layers. As the processing delays of the various sub layers remain the same in the Uplink (UL) and in the Downlink (DL), these studies are concluding that the overall latency in the UL and DL are symmetric. However, as there is a difference in the UL and DL signalling in LTE, such an assumption of a symmetric latency for the UL and DL is invalid.

As these delays are not standard delays but are vendor-specific, in an actual network, the results may be different. Moreover, the results do not report the latency caused in the core network or the latency caused in the radio network due to the different signalling procedures in UL and DL. Besides that, the studies do not take into account the various scenario aspects influencing latency as discussed above.

In measurement-aided latency assessments in real networks, most of the works available are based on the simple technique of using ping program. In ping program, packets are sent towards a destination and the destination sends an acknowledgement back to the sender. Based on the time of the sent packet and the received acknowledgement on the sender, the RTT is assessed.

In [22], the authors discuss the RTT measured using ping on the network for a user placed close and far from the radio base station. The work also considers two packets sizes for the ping: 32 bytes and 1400 bytes. However, the results show only the minimum, maximum and the average of the RTT. Such a result is not sufficient to completely understand how latency varies with packet size or user distance. The results do not show what fraction of the total packets tested reported a latency lower or higher than the average, or what fraction reported a latency close to the lowest/highest value. Moreover, the results are not considering the latency separately in the UL or DL or separately in the radio network or core network. Besides that, the results are also not investigating on the how packet rate is affecting the latency.

In [23], the authors discuss RTT using ping for packets of various sizes. Although this work considers the whole range of packet size: 10 bytes to 1000 bytes, the results are reporting only the average RTT values. Therefore, this study is also not providing a complete understanding of latency dependence on packet size. Moreover, the study is not considering the other factors that can affect latency as discussed in the above section. Similar to the drawbacks of [22] as discussed above, the results of this study are not considering the latency separately in the UL or DL or separately in the radio network or core network.

To truly asses the one-way delay of packets in the network, it is required to capture the packets at both the sender and receiver. The capturing points must also be synchronised in time. In [24], authors asses the one-way delay in the DL for packets of varying size. The drawback of this work is that as there is no access to the various components of the network, the results does not indicate the packet latency contribution of each of the various components of the network. Moreover, the results include the latency contribution from the external network which may bias the packet latency reported on the network. The result also does not provide a

complete understanding of the packet latency behaviour of the network as the result just concentrates on the maximum and minimum values of the packet latency. The results on one-way delay in the UL is not considered in this work.

Works on packet latency assessment with direct and complete access to the network are few. In [25], a better analysis of the user plane latency is provided and is in close congruence with what is intended with this thesis. The results discuss the one-way delay separately in the UL and DL and also separately in the radio network and core network for packets of different sizes. The results are also providing the reader with a complete understanding of the latency behaviour covering the entire range of latency.

This work is used as a strong base for this thesis as it mentions the various aspects to be considered when measuring the user plane latency in a mobile network. However, the results are not discussing how latency is affected by packet rate, positioning of the user or the network load.

In this thesis, the results of latency analysis are presented, which can enable the reader to completely understand the latency behaviour by indicating the variation of latency within its complete range. The results are presented for the one-way delay separately in the UL and DL and also separately for the radio network and core network for each of the various scenario aspects that affect latency in the network: network load, user position, packet size and packet rate. Moreover, the results presented is also indicating the maximum achievable latency reduction using the mentioned techniques. Apart from presenting the results, a detailed analysis of the reasons for the observed latency behaviour is also presented for each of the aspects affecting latency.

Chapter 4. Requirements for measurement setups

In this chapter we discuss the requirements on the measurement setup in light of the research objectives mentioned in Chapter 3. The developed measurement setup should also be able to investigate the latency improvements possible by the two latency reduction techniques used. In Section 4.1, the requirement to have realistic effects on the developed measurement setups is discussed. Along with the requirement, the approaches taken in achieving them are also presented (italicised) in this section. In Section 4.2, the need for the ability to assess one-way latency, to clearly assess the latency separately in the UL and DL, on the measurement setup is presented. The consequences or requirements to carry out such measurements are presented as subsections of Section 4.2, along with the approaches taken (italicised) in fulfilling those requirements.

4.1 Need for realistic effects on measurement setups

The measurement setup developed to measure the packet latency of the network should enable to assess the packet latency in as realistic a scenario as possible. What this means is that, all the components of the measurement setup and other tools and configuration developed should also be as realistic as possible. Finally, the developed setup should also be able to enable to measure the latency reduction techniques considered for this thesis work. The three crucial components or aspects of the measurement setup, which must be realistically modelled are the radio network, the core network and the user traffic characteristics.

4.1.1 Realistic radio network

In order to measure the packet latency of the mobile network in a realistic scenario, the radio network should emulate the effect of an actual radio network in terms of channel variations and the radio network load conditions. Therefore, the radio network should be developed so as to imitate a realistic modelling of the propagation and interference environment. At the beginning of this thesis, the already available setup had an Ericsson small cell LTE base station which was directly connected to the core network via an ethernet cable.

As the already existing LTE radio network was unloaded and directly connected to the core network, the radio network was not a fair imitation of the radio network as observed in a real situation. The packets transmitted over this ideal radio network were not experiencing latency as in a realistic radio network conditions. So, for this thesis, steps had to be taken to develop the already existing ideal LTE radio network into a realistic radio network. The developed radio network must be able to apply latencies to the packets, based on a pre-defined pattern. If this pre-defined latency pattern can somehow be generated based on the realistic network conditions, it can enable to have the realistic effects of a radio network on the already available ideal radio network.

Extra component called degrader had to be integrated between the eNB and the core network which emulated the effects of a realistic radio network. Finally, the degrader should also enable to the apply the effects of differentiated scheduling in the radio network.

To provide input to the degrader, other tools and simulators are used. A radio network simulator developed by TNO, that can simulate the effects of a realistic radio channel condition provides the realistic channel conditions for the various scenarios considered for the

measurements. Then using a python script developed for this thesis, the radio network simulator trace is converted to individual packet latency trace. The developed packet latency trace is then finally provided as the input to the degrader. A detailed description of the radio network simulator and the degraders will be presented in later sections.

4.1.2 Realistic core network

Similar to the need for a realistic radio network, the already available LTE core network had to be developed into a realistic core network. The developed core network should emulate the effects of the transportation delay, delay due to multiple hops in the backhaul topology and the load conditions in the core network.

At the beginning of this thesis, the core network that was already available was the EPC implementation from Fraunhofer Fokus. As the EPC was directly connected to the radio network using an ethernet cable, there was no transport delay for packets that would have been otherwise observed in an actual mobile network operator's core network. In a typical mobile network operator's core network, the radio access network is usually hundreds of kilometres away from the core network and this causes some transport delay to the packets. Moreover, the packets reach the core network from the eNB after multiple hops, depending on the backhaul topology. Also, the already existing setup at the beginning of this thesis had only a limited number of UEs that could be attached to the network and transmit/receive data. Due to the lack of possibility to attach multiple users to the core network, it was not possible to have the effect of a loaded core network as it would have been in an actual core network.

Extra tools and components were used inside the core network to add the effects of transport delay to the packets and also the effects of loading on the core. Finally, the developed realistic core network should also enable to apply the latency reduction technique in the core network. A detailed explanation of how these were achieved will be presented in later sections.

4.1.3 Realistic user and service characteristics

After developing the two components in the measurement setup into a realistic one, the final aspect that also had to be realistic is the user traffic characteristic considered for the measurements. Different user applications used on a mobile network have different traffic characteristics. A traffic characteristic of an application is characterised by the packet size and generation rate of the packets. Besides the realistic traffic characteristic, it is also required to have the effects of different positions of the user in the radio network.

In order to ensure that the measurement results obtained for the packet latency in this thesis are realistic as it would have been in an actual network measurement, it is necessary that the traffic characteristics used are also realistic. *Therefore, a packet generator that can generate packets with different traffic characteristics as specified, is used in this thesis to generate the user packets.*

The effects of different user positions in the radio network is achieved using the radio network simulator mentioned above. The radio network simulator takes into account the SINR variation for the user due to the pathloss and inter-cell interference that are dependent on the user position in the radio network.

4.2 Ability to measure one-way latency

To identify the packet latency caused by the various network components and other identified factors affecting packet latency, separately in the UL and DL direction, one-way latency

measurement must be performed. Such a one-way latency measurement result should be ensured that it is very accurate, free from any external network's influence and free from any kind of network performance bias. To ensure such a measurement setup, following consequences or requirements have to be taken care.

4.2.1 Access to network

As already mentioned in the previous section, the intention of this thesis is to determine the one-way packet latency separately in the UL and DL in as much realistic scenario as possible. The Figure 4-1 shows a simple mobile network without showing the inside details and other involved complexities.

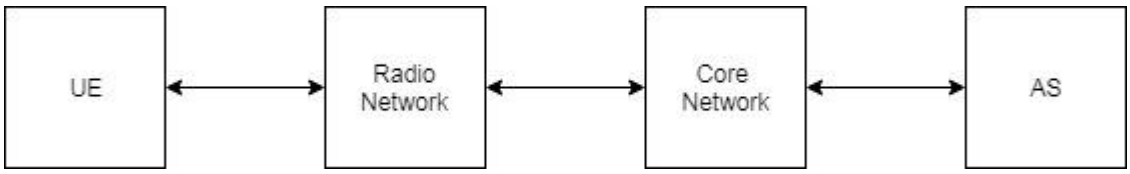

Figure 4-1: A simple mobile network

The Application Server (AS) is the component in the measurement setup that receives the packets sent by the UE in the UL. The AS also act as the packet source, where it generates packets towards the UE in the DL.

As a normal user (customer) of a mobile network, it is impossible to have explicit access to the various components in the network. A normal user of an LTE network can use the network service, using an LTE device, and all the network configuration and topology remains unknown to the user. A normal user's access is in most of the cases restricted to the component UE as shown in the Figure 4-1. To clearly estimate the latency contribution from the various components of the network, explicit access to the entire network is necessary. To identify the reasons for the observed packet latency from the network, it is also essential to have a complete knowledge of the system level aspects as well as other system configurations of the network.

In this thesis, explicit access to the network components is available with the knowledge of the network topology and other system configurations like the available bandwidth etc. With the explicit access to the network components, the next step is to decide on where and how to have the measurement points so as to ensure that it is possible to determine the latency contribution from the various network components individually.

4.2.2 Placement of measurement points

With explicit access to the network components and the knowledge of all the required network configurations and parameters to assess the packet latency, measurement points are to be placed in the appropriate locations. The placement of the measuring points has to be such that it enables to measure the packet latency separately in the radio network and the core network.

The measurement tool used in this thesis consists of a packet capture tool called Wireshark and python scripts developed to extract the latency of the individual packets in each components of the network. The Wireshark is a tool that can capture network packet entering or exiting a network interface of a machine [26]. While capturing the packets, the Wireshark time stamps each of the packets according to the time obtained from the system clock.

Based on the time stamp of the packets entering each of the components of the mobile network, the developed python script extracts the packet latency incurred for that packet.

4.2.3 Appropriate transport protocol

In IP networks, the most commonly used transport protocols are UDP and TCP. Transport protocols are used in IP networks to ensure that the packets sent from the source is received at the destination as intended.

TCP is a delivery guaranteed transport protocol with end-to-end flow control. In TCP, the receiver acknowledges all the correctly received packets. What this mean is that, in TCP, a traffic flow is generated also in the opposite direction of the intended traffic flow. Therefore, TCP based one-way latency measurement result on a mobile network can be affected by the flow of the acknowledgement packets in the opposite direction.

If an acknowledgement is dropped on the network, it can lead the transmitter to retransmit the same packet. In this thesis, the latency for each of the packets are calculated based on the time stamp of unique packets sent and received. Therefore, the capturing of duplicate packets can cause confusion as to which of the two copies should be regarded for determining the packet latency.

Moreover, the flow control mechanism of TCP optimises the network link performance so as to ensure that receiver nor the network link is congested.

These kinds of flow control and error control would bias the accurate one-way delay measurement on the network [27].

Therefore, an appropriate transport protocol must be used to determine the one-way latency on the mobile network without any underlying optimization and moderation. In this thesis, the transport protocol used is UDP. The traffic generator used in the measurement setups are set to generate UDP packets of specified size and at specified packet rate. Thus, ensuring a realistic user and service characteristic also means that an appropriate transport protocol is used to determine unbiased one-way latency.

4.2.4 Placement of the application server

While measuring the packet latency in a mobile network, it is important to avoid any kind of influence from sources that are external to the network. Otherwise, the latency reported will also include the latency contribution from the external sources. As the intention of this thesis is to determine the latency contribution purely from the network, any influence in the results from external sources must be avoided.

If the packet end points are placed on any external network, there should be the possibility to measure precisely how much latency was contributed to the overall packet latency from the external network. Usually this is not easily achievable as this will require to have explicit access to the external network also. Another approach to avoid such a situation would be to place the end points next to the mobile network and not on any external network.

In this thesis, the packet end point (destination in the UL and source in the DL) is placed next to the core network and not on any external network. Such a setup will ensure that there is no excess latency caused in the latency measurement of the mobile network from any external network.

4.2.5 Eliminating measurement error

In any kind of measurement experiments, it is important to avoid measurement errors. Especially in experiments, where accuracy is an important aspect, possible sources of time error must be avoided so as to achieve the required level of accuracy. In this thesis, the latency measurement on an LTE network is performed, for which the latency values will be presented in millisecond ranges. Therefore, the measurements in this thesis require an accuracy level under a millisecond.

In this thesis, it is ensured that there isn't any time mismatch across the various measurement points in the setup by having time synchronised across all the measurement points. As the time is synchronised across all the machines where measurement points are located, the system clock in each of the machines are having the same copy of time. The Wireshark instances uses this system clock to time stamp the captured packets and therefor the latency calculated for each of the packets are eliminated of any time error. The time synchronisation is achieved using NTP (Network Time Protocol) which will be discussed in detail in Section 5.5.

Chapter 5. Measurement setup

As mentioned in Chapter 4, it is required to develop a measurement setup which emulates the effects of a realistic mobile network. This chapter provides a detailed description of the developed measurement setup. In Section 5.1, a description of the UE with the necessary tools running inside to generate realistic traffic characteristics is presented. The various traffic generators available and the reason for choosing the particular traffic generator for this thesis is explained in this section. In Section 5.2, a detailed explanation of the components and tools used to develop the already available radio network into a realistic radio network is presented. In Section 5.3, a description of the components and tools used to develop the already available core network into a realistic core network is presented. In Section 5.4, a description of the Application Server (AS) is presented. In Section 5.5 a description of how time synchronisation is achieved across all the measurement points in this thesis is presented. In Section 5.6, an overview of the various latency reduction techniques used in this thesis is presented along with a detailed description of the EC setup developed for evaluating latency reduction in the core network.

The developed measurement setup with the requirements specified in the previous chapters is as shown in Figure 5-1. Following that, a detailed description of each of the components, their function and how they are integrated into the setup will be presented.

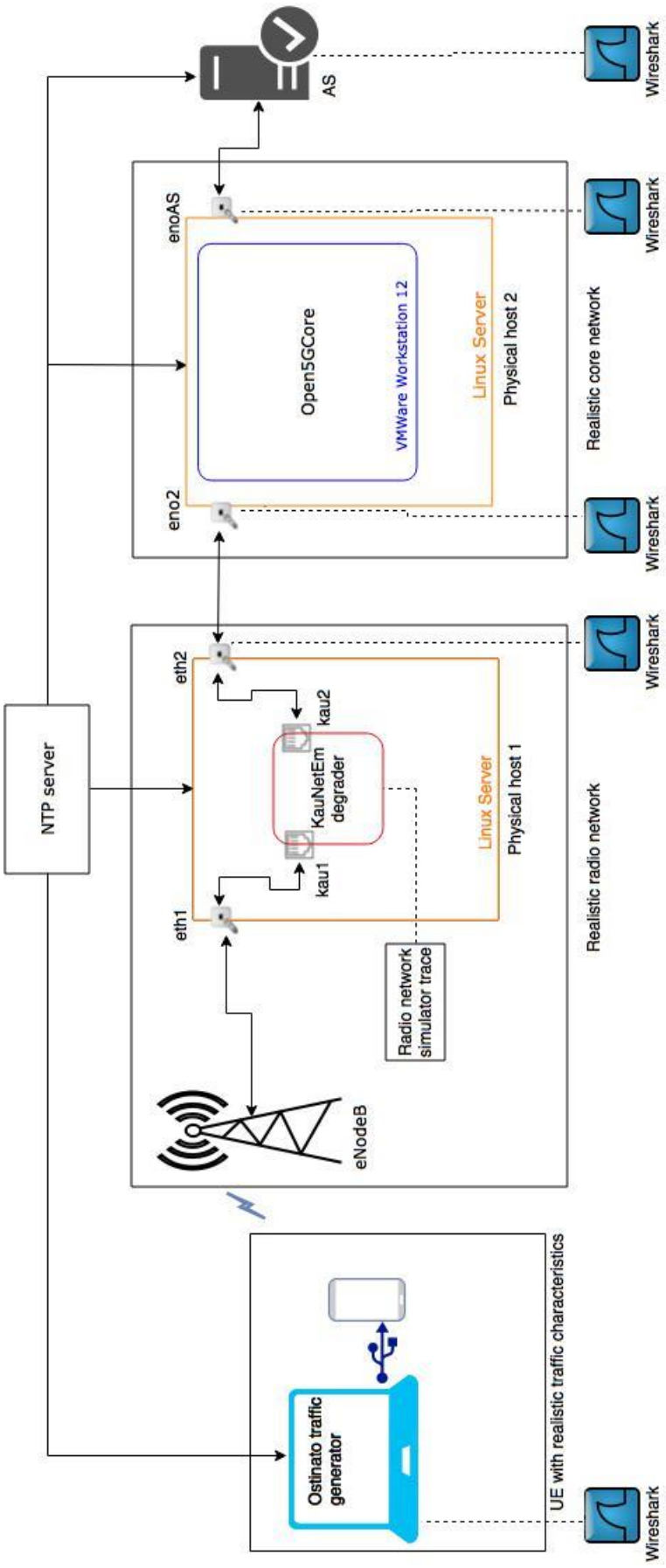

Figure 5-1: Measurement setup

5.1 The user equipment

In this section, a description of the UE component is presented. The most commonly used UE in a mobile network is an LTE mobile device like a smartphone. In this thesis, the UE is required to run a packet generator that can generate packets according to the various realistic traffic characteristics as observed in an actual network.

The packet generator used for this thesis is a Linux-based software tool and is not supported to run independently on an LTE mobile device. For this reason, the UE used for this thesis is not an LTE mobile device but an Ubuntu machine which is in turn connected to the mobile network via the option of USB tethering using an LTE mobile device. The LTE mobile device used for USB tethering is a Google Nexus 5X device. The LTE device has a SIM card that is configured to connect only to the mobile network that is considered for this thesis research.

The UE acts both as a packet sender and a packet receiver. In the UL direction, it acts as a packet sender where the traffic generator generates packets according to the specified traffic characteristics. The generated packets are then sent to the LTE device. From the LTE device, the packets are then sent wirelessly to the radio base station. In the DL direction, the UE receives the packets originating from the Application Server (AS). The LTE device receives the packets in the DL direction from the radio base station and delivers the packets to the Ubuntu machine.

The packet generator that is running on the UE is the Ostinato packet generator which can generate packets according to the traffic characteristics that have been specified in the traffic configuration options [28]. The traffic generator is set to use only UDP packet for reasons that have been specified in Section 4.2.3. In the UL direction, the traffic generator sends UDP packets towards the AS with the source port as 5555 and destination port as 8888. The choice of these port numbers is arbitrary. The AS is set to listen for any UDP packet on the port 8888 and when it receives any packet, those packets are captured and timestamped.

When the UE is receiving packets in the DL direction from the AS, it is set to listen for UDP packets on the port number 8888. An instance of Wireshark is used to capture the UDP packets received on the UE. The latency of the packets are determined from the time stamps of the captured packets.

5.1.1 Traffic generator

As mentioned in the previous section, the traffic generator used is the Ostinato traffic generator [28]. In an actual user application, packets are generated with different traffic characteristics to satisfy the needs of that particular application. In this thesis, the packet size is specified in bytes and the packet generation rate in the number of packets generated per second.

There are hardware and software-based traffic generators. In hardware-based traffic generators, a hardware device is used to generate packets of desired traffic characteristics, e.g. the Spirent Testcenter [29]. In software-based traffic generators, special software utility runs on top of a specific operating system (OS) to generate packets of desired traffic characteristics. The software-based traffic generators are also referred to as synthetic traffic generators. There are multiple options for software based packet generators that are supported on different OS.

Ostinato is a software-based traffic generator that is supported on the Linux platform. The reason for this choice is the simplicity of the packet generator and the support for it to create python scripts automatically from pre-specified traffic characteristics. This python script can

then later be executed from a Linux terminal to generate packets according to the saved traffic characteristics in the python script. Another important reason for using Ostinato as a packet generator for this thesis is that it has the ability to set the payload content of each of the packets in a particular pattern so that each packet has a unique payload content. The advantage of this feature is that it makes it possible to uniquely identify each of the packets traversing the different network component by inspecting the packet payload content. This feature enabled the use of other python scripts to identify the latency of each packet in the different network components by inspecting the packet payload content.

5.1.2 Ostinato architecture

Ostinato has a controller – agent architecture [30]. The agent is the most important component of Ostinato. It does the actual work of sending the packets towards the destination or receiving any incoming packets. The controller is the component that acts to control the agent and fetches the data from the agent to report the statistics as shown in Figure 5-2.

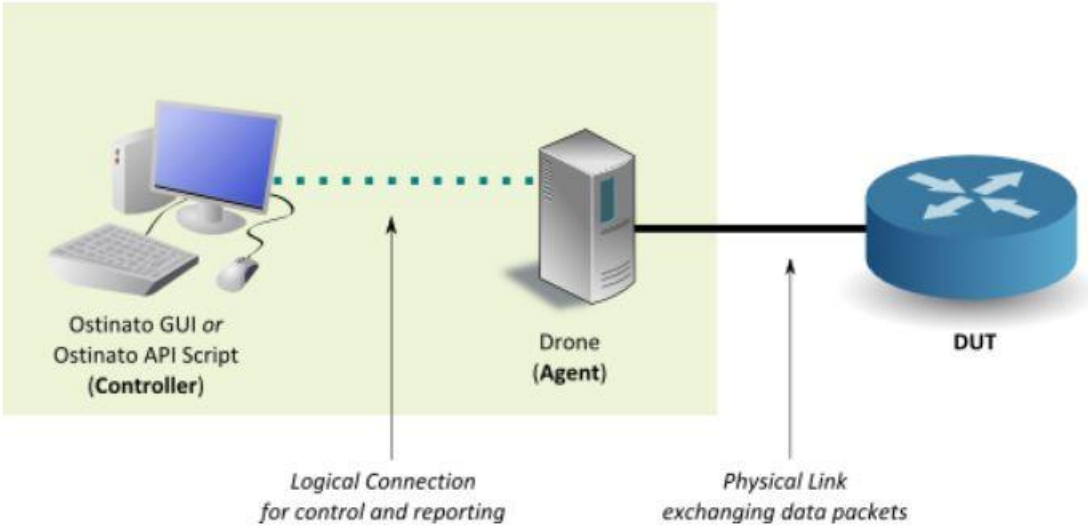

Figure 5-2: Ostinato architecture [30]

This controller-agent architecture can also be considered as a frontend-backend architecture. The frontend part consists of the Ostinato GUI or the Python API which is used to configure the traffic characteristics and the backend is the Agent (Drone) which sends the actual data packets based on the traffic characteristics from the Controller. The packets generated by the Drone is then transmitted over the link to the Device Under Tests (DUT).

Ostinato can operate in three different modes of operation. They are:

- Default mode
- One controller many agent mode
- Many controller one agent mode

In the Default mode as shown in Figure 5-3, the controller and the agent run on the same computer to generate the traffic. Whenever the Ostinato GUI is started, it will spawn both the controller and the agent on the same computer.

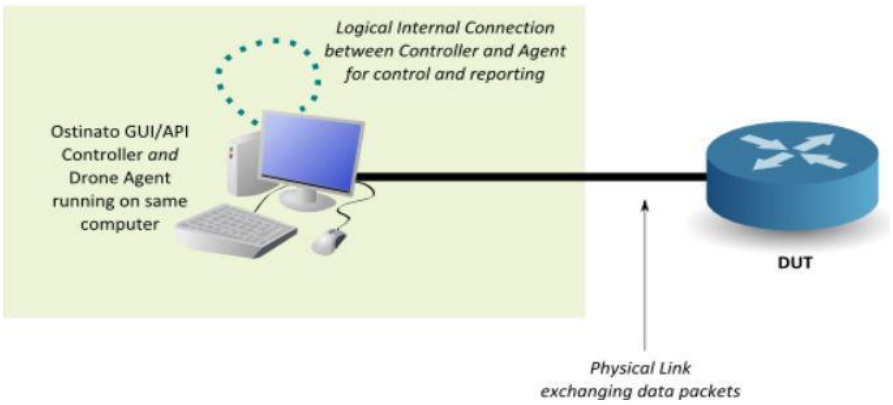

Figure 5-3: Ostinato default mode [30]

For this thesis, the default mode of Ostinato operation is used. The packets are generated from a single device and the traffic parameters of the generated packets are controlled by the local agent which is started using the Ostinato GUI.

5.2 The radio network

In this section, a description of the radio network developed for this thesis is presented. The tools and extra components integrated into the radio network to emulate a realistic radio network is presented as different subsections under this section.

5.2.1 Realistic radio network

The radio network setup developed for this thesis emulates the effect of an actual radio network in terms of channel variations, the effect of the user location with respect to the radio base station and the radio network load conditions.

As the already available radio network at the beginning of this thesis was in ideal conditions i.e. completely unloaded and the users are placed very close to the eNB, it was necessary to develop the radio network into a realistic radio network. For this purpose, some extra components and software tools are used to emulate the realistic behaviour of a radio network. These additional components and software tools used are referred to as ‘network emulators’. These components are also named ‘degraders’ as they actually degrade the performance of an ideal or well performing network or systems to emulate the performance of a realistic network or system.

So, when the packets in the UL direction enters the eNB and exits towards the core network via the degrader, the packets will have effects of latency behaviour as observed in a real network. Similarly, the effects will be added to the packets in the DL direction which are entering the degrader from the core network and exiting towards the UE via the eNB.

Moreover, as the Ericsson eNB is a proprietary device and cannot be accessed with full rights, it is impossible to capture packets by running a Wireshark instance or any other kind of packet capturing tool inside the eNB. With the new setup of degrader placed between the eNB and the core network, it is possible to capture the packets exiting and entering the interface of the degrader. These captured packets will have the effects in terms of latency as observed in a real network. The Figure 5-4 shows the overview of the realistic radio network developed. Detailed explanation of the various components is presented in the following sections.

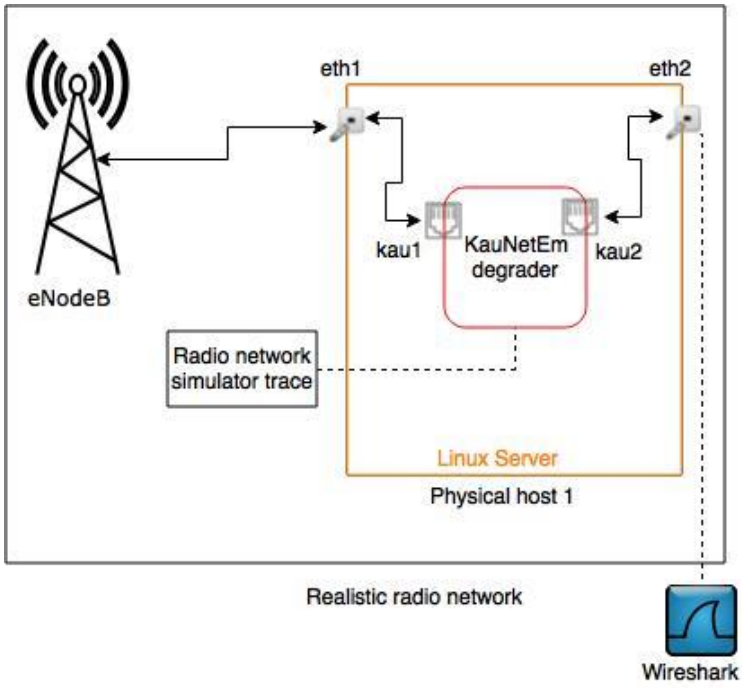

Figure 5-4: Overview of realistic radio network

5.2.2 KauNetEm overview

As software-based network emulators/degraders are much easier to be incorporated into the measurement setup for this thesis, it was the choice of network emulator/degrader used in this thesis work. As the aim of this thesis is to emulate a realistic network by degrading the performance of the already available network, the network emulator/degrader will be referred as the degrader from now on.

KauNetEm is developed by Prof. Johan Garcia and Prof. Per Hurtig of Karlstad University [31]. Comparing to the NetEm network emulation utility available natively on Linux [32], KauNetEm differs by offering the support for deterministic network emulation. The NetEm utility is lacking the ability to introduce network emulation based on a pre-defined pattern, which is a requirement for this thesis as mentioned in the Section 4.1.1. With KauNetEm, the ability to add dynamic changes to network parameters such as the packet latency, packet loss or bandwidth limit, on a Linux operating system is possible. This feature of network emulation based on a pre-defined pattern is required as the latency applied to the packets in the realistic radio network is predetermined using the radio network simulator and the python script.

With the feature of pattern-based emulation, KauNetEm allows to introduce specific behaviours on the network that can be specified as a pattern input. The pattern file is a file that instructs the network emulation utility of Linux (NetEm), what effect in terms of e.g. packet delay, packet loss or bandwidth limit, should be applied on a per packet basis or per millisecond basis. An effect can be applied on each packet, irrespective of their arrival time into the KauNetEm, using the *data-driven* mode of operation. For example, it is possible to specify how much delay should be applied to the n^{th} packet. Similarly, an effect can be applied on the link at each time interval irrespective of the number of packets flowing over the link in that interval, using the *time-driven* mode of operation. For example, it is possible to specify how much bandwidth in bits per second should a link have during the n^{th} millisecond.

The pattern creator utility of KauNetEm is a user-space program in KauNetEm called the ‘pattgen’, which takes the network traces as input files and creates the pattern file, which is structured similar to other NetEm emulation patterns that are natively supported by the Linux kernel.

The trace files provided as input to the pattern generator must be a text file which specifies the packet on which the effect to be applied separated by a comma and followed by how much of that effect to be added in data-driven mode. In time-driven mode, the trace file must be a text file which specifies the time instant at which the effect must be applied separated by comma and followed by how much of that effect to be added.

In this thesis, the packet latency trace generated using the radio network simulator and the developed python script to convert the simulator trace output into a packet latency trace, is applied as the pre-defined packet delay change pattern in a data-driven mode of operation.

The Table 5-1 explains the various effects that can be emulated using KauNetEm and how each of them is operated in the two modes.

Pattern Name	**Time-driven Mode**	**Data-driven mode**
Packet loss	Packets are dropped during specific interval of time	Packets are dropped on per packet basis
Bandwidth	Bandwidth changes can be implied during a certain interval of time	Bandwidth restrictions can be placed after a certain number of packets are passed
Delay change	Delays to the packets are applied during a certain interval of time	Delays are applied on specified packets
Packet reordering	Packets are reordered in specified time duration	Packets are reordered according to packet numbers
Bit-error	Bits can be flipped during certain interval of time	Bits can be flipped at defined positions of data

Table 5-1: KauNetEm modes of operation

The Figure 5-5 shows an overview of the data-driven packet delay operation of KauNetEm. As shown, each of the packet can be applied a specific delay as specified to the delay pattern file.

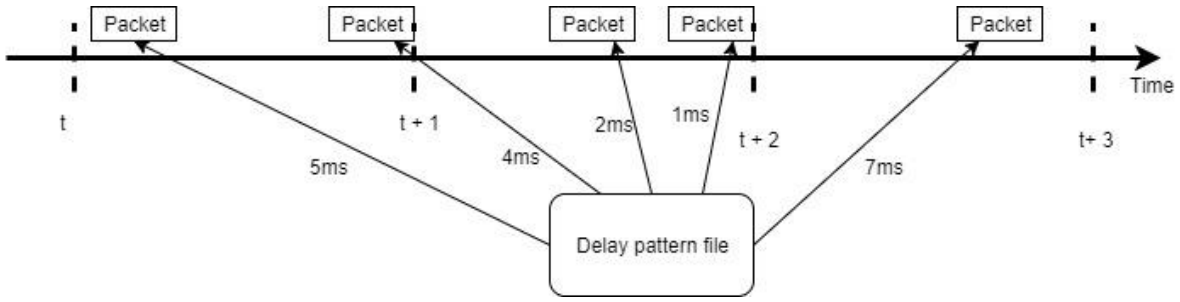

Figure 5-5: KauNetEm data-driven packet delay operation

5.2.3 Radio network simulator

For this thesis, the latency trace used in the degrader is created using a throughput-based trace generated with a system-level radio network simulator, using python code that was developed specifically to convert the simulator trace into a latency trace fed into the degrader.

When a wireless signal propagates, the signal attenuates. Such a radio channel, where the received strength gets attenuated is referred as a fading channel. There are two kinds of fading: large-scale fading and small-scale fading.

Large-scale fading considers the effect of fading in the signal strength due to path loss as a function of the distance between the transmitter and receiver. Small-scale fading considers the effect of multipath fading. The Figure 5-6 shows the various causes for degradation of the transmitted signal in a wireless mobile channel.

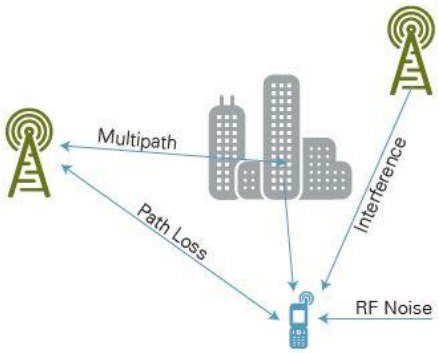

Figure 5-6: Fading radio channel [33]

In multipath fading, the transmitted signal gets reflected from various obstacles in the signal path. Thus, multiple copies of the same signal may interfere with each other constructively or destructively at the receiver [34]. Depending on the whether the signal interfered constructively or destructively at a particular instant, the signal strength might be boosted or degraded. This will result in signal strength that varies over time as shown in Figure 5-7.

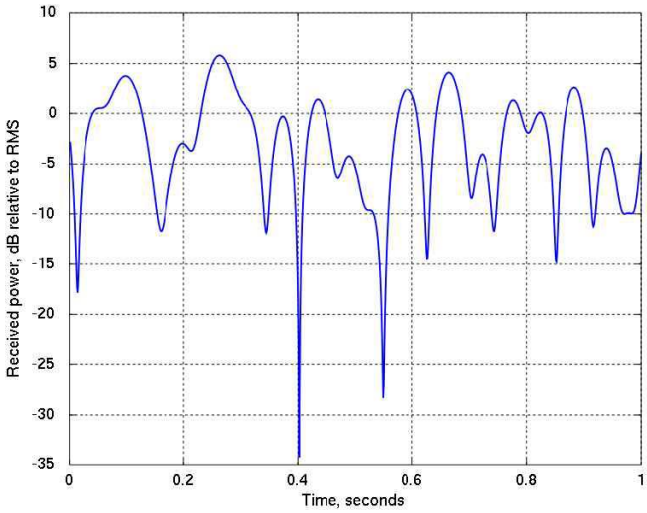

Figure 5-7: Multipath fading [35]

Another aspect affecting the quality of the received signal is the interference in the network. There are two kinds of interference: inter-cell interference and intra-cell interference [36].

Inter-cell interference is caused due to the same frequency resource assignment to UEs in the neighbouring cells. The effect of inter-cell interference in the DL depends on the position of the UE and the power at which the neighbouring cells are transmitting. For example: a UE close to the eNB is far from the neighbouring eNBs and therefore experience lower inter-cell interference than a cell edge user. Suppose the case for a cell edge user: a cell edge user experiences much higher inter-cell interference if the neighbouring cells are transmitting at a

higher power. In the UL, interference is experienced at the eNB. Therefore, depending on different levels of uplink transmission activity in the neighbouring cells, the aggregated interference (I) from all the neighbouring cells adds up together to create a noise rise above the noise level (thermal noise plus the received noise figure). The ratio of the interference (I) plus noise (N) and the noise level (N) is represented as the noise rise ((I+N) / N).

In LTE, due to the orthogonality in frequency domain in the UL and due to orthogonality in time domain in the DL, there is no *intra-cell interference*.

The quality of the received signal is indicated by the metric Signal to Interference plus Noise Ratio (SINR). The higher the SINR, the better the signal received and thus lower the probability of error in data transmission, assuming a given MCS. The pathloss, multipath fading and interference affect the SINR of the received signal. An important requirement for the received signal in LTE is to maintain a network specific BLER. When a lower SINR is reported, to maintain the target BLER, the LTE uses the feature of link adaptation. When a lower SINR is reported on the network indicating a poor radio link, to maintain the target BLER, a more robust modulation and coding scheme must be used. For example, if a more robust modulation technique such as QPSK is used instead of 64QAM, it would be possible to maintain the target BLER even in cases of low SINRs. The robustness can also be increased by reducing the code rate. The code rate determines what percentage of the maximum bits transferrable in a particular transport block is actually transmitted as effective data bits. The general trade off in adaptive modulation and coding is one of robustness versus spectral efficiency: the higher the degree of robustness (achieved by applying lower order modulation and/or a lower code rate), the lower the BLER but also the lower the effective throughput.

Thus, when a higher SINR is reported due to lower path loss or constructive interference by multipath fading or lower interference in the network, a higher MCS is selected and consequently a higher number of bits can be transferred in that TTI. Similarly, when a lower SINR is reported due to higher path loss or destructive interference by multipath fading or higher interference in the system, lower number of bits will be transferred in that TTI.

The radio network simulator is a system-level LTE simulator developed by TNO. Two separate simulators are used: an UL simulator and a DL simulator. Both the simulators, consider a network with twelve sites where each of the sites are divided into three sectors as shown in Figure 5-8. A hexagonal site layout is considered and a wraparound feature is applied to avoid boundary effects. The radio environment considered is a suburban environment with a typical inter-site distance of 1.5 km.

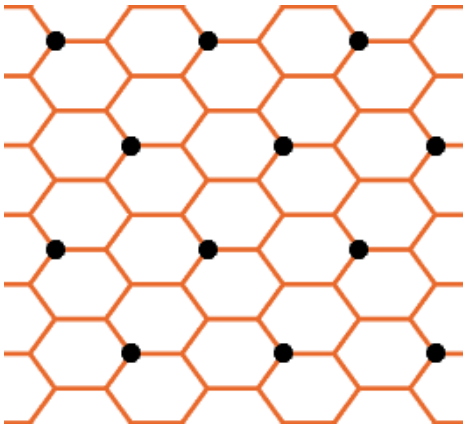

Figure 5-8: 12 site network layout

Both the simulators consider the distance-based Okumura-Hata path loss model [37] for a suburban environment to model the distance based path loss due to different user position in the radio network. In order to model the effects of signal reflections or obstructions that could happen in a suburban environment due to e.g. moving vehicles and buildings, the IMT Advanced multipath model for macro-cells in a suburban environment is considered with an assumed UE speed of 0.8 m/s [38]. To account for the inter-cell interference in the considered cell (cell which is serving the UE under test), in the DL, it is modelled by assuming that all the neighboring cells of the considered cell are transmitting at an a priori fixed power, expressed as a given percentage of their maximum transmit power. In the UL, it is modelled by different levels of noise rise, as more explicitly defined above.

The output of the radio network simulator is a trace file which specifies the number of bits that can be cleared from the buffer in each millisecond or TTI depending on the direction of packet transfer. The number of bits that can be cleared in each TTI may vary depending on the instantaneous radio channel quality and the network load. In the UL, the data to be transmitted is temporarily stored in the UE buffer. Based on the amount of radio resource assignment to the UE in each TTI, data from the buffer gets emptied in the UL direction. In the DL, the data for each of the UE gets buffered in the corresponding PDCP buffer for that particular radio bearer [39]. Then based on the instants of radio resource assignments and the amount of resource assigned in those instants, the data from the PDCP buffer gets emptied in the DL direction. From now on a reference to buffer in the context of UL refers to the UE buffer and in the context of DL refers to the PDCP buffer in the eNB.

The Figure 5-9 shows an example of the radio network simulator trace. The trace indicates the amount of data that can be cleared from the buffer in every TTI. The observed fluctuations in the attainable bit rates are due to multipath fading. Periods of more favourably multipath fading are observed as ‘spikes’ in the bit rate trace where a relatively higher number of bits are transferred during those TTIs (e.g. in the period between 1201 ms and 1350 ms). This period corresponds to periods where the signals are interfering constructively, resulting in high SINR and consequently high MCS. Similarly, there are periods of multipath fading dips where fewer number of bits are transferred (between 900 ms and 1000 ms). This corresponds to periods where the signals are interfering destructively, resulting in low SINR and consequently low MCS.

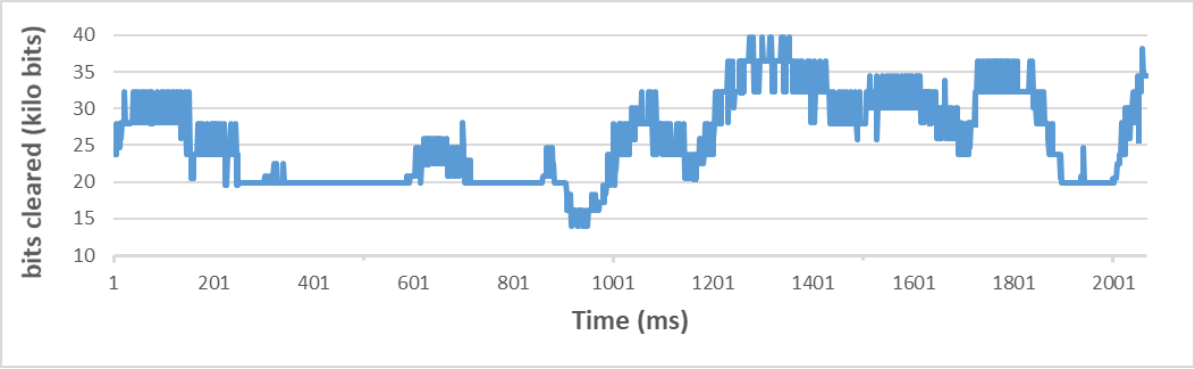

Figure 5-9: Radio network simulator trace

From the trace output of the radio network simulator, the number of bits that can be cleared in each TTI is obtained for the various scenarios considered for the measurements. The developed python script then uses this trace along with other factors like the size of the packet,

arrival time of the packet, number of users considered in the system and prioritisation level of the considered user and the scheduling strategy used to determine how much latency each of the packets experience.

For the DL, the developed python script considers a round robin scheduling as will be explained in more detail in Section 6.2.1.1. Based on the packet size and the arrival time of the packets, the script checks if there is enough bit transfer possible in that particular arrival instant of the simulator trace. If not, the script continues to check for the next scheduling instant for that user to see if the number of bits in the two instants combined, can clear the packet completely. The script determines the latency for the packet as the duration between the packet arrival and the instant at which packet transfer was complete. Figure 5-10 shows an example of how the latency is determined using the script in the DL for a total of 3 users.

For the UL, the developed python script can attempt to clear the packet from the buffer in every TTI as the user is assigned consecutive TTIs due to a frequency-domain only based scheduling as will be explained in more detail in Section 6.2.1.2. Based on the packet size and the arrival time of the packets, the script checks whether there is enough bit transfer possible in that particular arrival instant of the simulator trace. If not, the script checks for the next TTI to see whether the number of bits in the two instants combined, can clear the packet completely. The script determines the latency for the packet as the duration between the packet arrival and the instant at which packet transfer was complete. Figure 5-11 shows an example of how the latency is determined using the script in the UL.

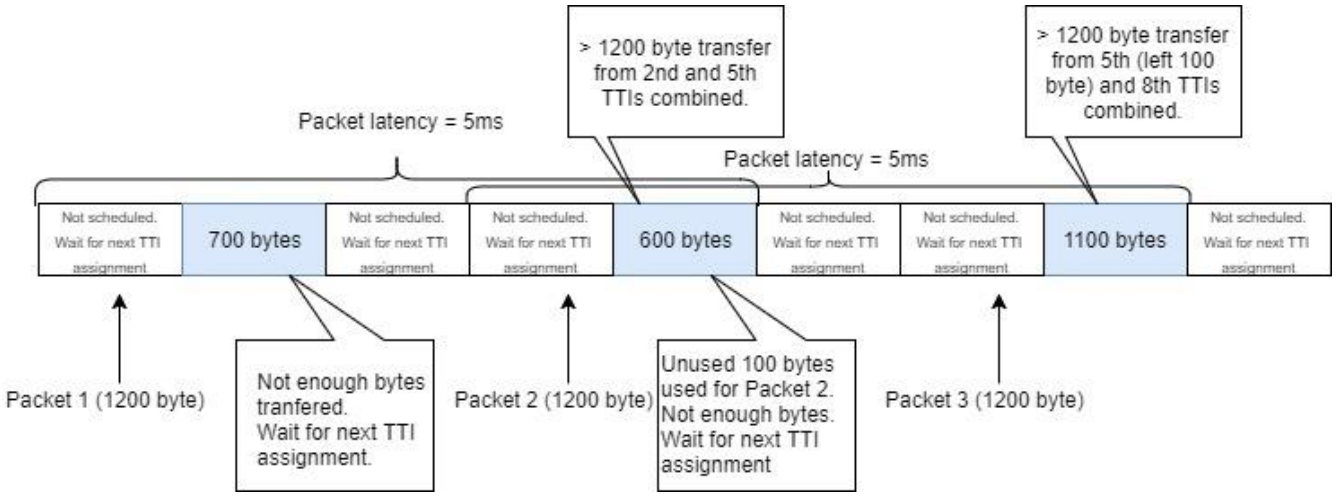

Figure 5-10: DL latency calculation

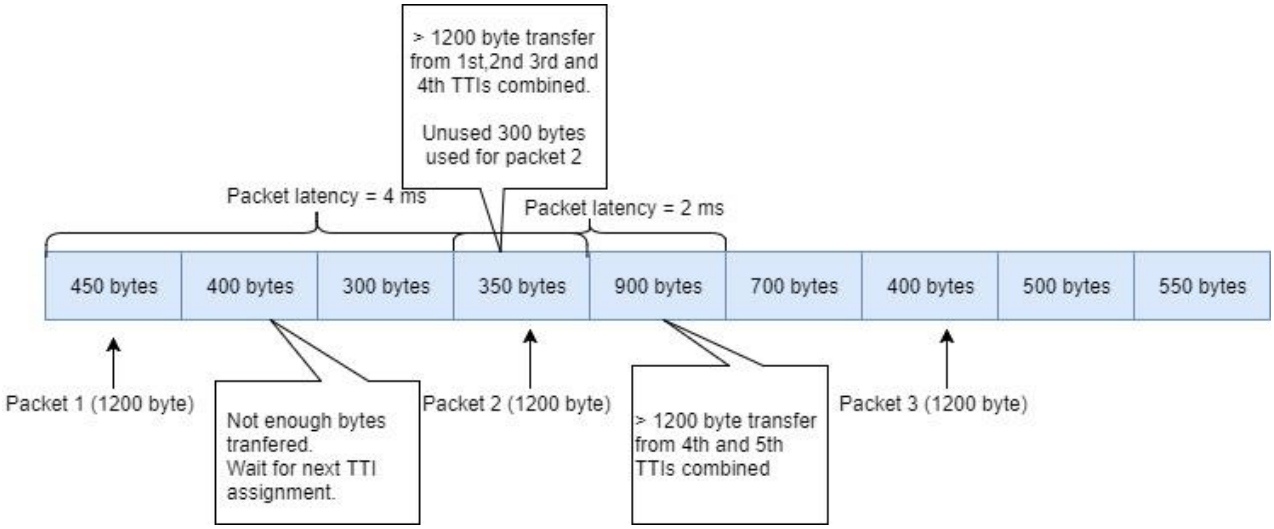

Figure 5-11: UL latency calculation

The python script finally generates a text file which specifies the latency that needs to be applied to each of the packets by KauNetEm. The Figure 5-12 shows an example of the trace file input for the KauNetEm generated by the python script. The trace file specifies that the 1st packet should be applied a latency of 5 ms, the 2nd packet with a latency of 7 ms, 3rd with 8 ms and so on.

```
1,5,2,7,3,8,4,9,5,2,6,3,7,3,8,9,9,4,
```

Figure 5-12: KauNetEm trace file

It is this generated packet latency trace that is eventually used inside the KauNetEm degrader to be applied on the actual user packet sent over the developed measurement setup.

5.2.4 Integrating radio network and KauNetEm

The degrader is run inside a Linux machine which is named physical host 1 as shown in Figure 5-1. The degrader is run as a virtual machine inside the physical host 1. More detailed description of virtual machines will be presented in later sections.

The degrader running on the physical host 1, along with the radio network simulator trace and the developed python script, enables a realistic behaviour on the radio network. An instance of Wireshark running on one of the interface of the physical host 1 is used to capture and time stamp packets in the radio network.

The physical host 1 running the degrader has two physical interfaces eth1 and eth2. The interface eth1 is connected directly to the eNB and eth2 is connected directly to the core network. The degrader virtual machine has two interfaces kau1 and kau2. The interface kau1 on the degrader is bridged to the eth1 and the interface kau2 is bridged to eth2.

The degrader applies latency specified by the latency trace to the packets matching specific criteria. The packet matching criteria used is the source IP address and the transport protocol. Thus, using the IP address of the eNB, it is possible to apply specific latency to each of the UDP packets in the UL. Similarly, using the IP address of the PGW-U, it is possible to apply specific latency to each of the UDP packets in the DL.

The packets in the UL direction, will be applied the effects of a realistic radio network condition inside the degrader and exits towards the realistic core network. Therefore, the packets

captured on the interface eth2 will have the latency effect of a realistic radio network. Similarly, the packets in DL direction, will be applied the effects of a realistic radio network condition inside the degrader and exits towards the UE. Thus, the packets captured on the UE, will have the latency effect of a realistic radio network.

5.3 The Core network

The core network setup developed for this thesis emulates the effect of a realistic core network as observed in an actual core network in terms of transportation delay for packets, delay due to multiple hops in backhaul topology and the load conditions in the core network. In the Open5GCore EPC implementation from Fraunhofer Fokus, various functionalities of an EPC are implemented as software functions that are running inside multiple virtual machines [40].

As the Open5GCore was directly connected to the eNB using an ethernet cable, there was no transport delay or delay due to multiple hops for packets that would have been otherwise observed as in a typical mobile network operator's core network where the radio access network may be hundreds of kilometres away from the core network. Thus, to emulate this effect in the Open5GCore, a deterministic latency of 7ms was added to the packets as a value of around 7 ms is typically observed in very large networks [41] [42]. To emulate the effects of a loaded core network, the virtual machine named 'bt' is used, which emulates the effects of multiple LTE devices attached to multiple eNodeBs and send data into the core network. As the Open5GCore is not designed as a high-performance EPC but rather as a virtualised EPC with a focus on enabling new features, the loading conditions considered for core network loading is far less than what would be observed in a typical mobile network operator's core network. Due to the performance limitations of the Open5GCore, the buffer sizes used on the testbed are very limited. This means that, even at the limited loading levels considered for the experiments, packets may be dropped in the core network due to buffer overflow. However, as this limitation is specific to this testbed, in reality packet dropping rate should be significantly lower in a performance oriented platform.

5.3.1 Virtualisation techniques used

As the Open5GCore used for this thesis is virtualised and the various functions of an EPC are running as different virtual machines (VM), it is important to discuss in detail the different types of virtualisation techniques used in the setup and what a virtual machine means.

In simple terms, virtual machine means to run an operating system or other software, inside a physical machine that is already running its own operating system and software. The new instance of a virtual OS running as a new independent system on top of the physical machine is referred to as guest and the machine on which the VM run is called the physical host or simply host. Thus, virtualisation enables multiple VMs or guest to be run simultaneously on a single physical host. This technique has the advantage that there is no need to have standalone hardware for each of the VMs that need to be run. For each of the VMs, it is as if the host hardware's processor, memory and other resources are available for its functioning. The component that enables the various VMs running on the host to access the host's hardware resources co-operatively between each other is a software called the Hypervisor that is running on the host. It is the responsibility of the hypervisor to allocate the correct CPU resources, memory, bandwidth and disk storage space for each virtual machine.

There are two types of hypervisors: type 1 hypervisors also known as bare metal hypervisors and the type 2 hypervisors also known as hosted hypervisors. For this thesis, both the kinds of hypervisors are used to host the VMs. Some examples of type 1 hypervisors are: VMware

ESX/ESXi, Oracle VM Server for x86, KVM, or Citrix XenServer. Some examples for type 2 hypervisors are: Virtage hypervisor, VirtualBox and VMWare Workstation. Thus, KVM hypervisor is a type 1 hypervisor and VMWare Workstation is a type 2 hypervisor. For this thesis work, both KVM and VMWare Workstation based VMs are used [43].

In type 1 hypervisors, the hypervisor is run as a software directly on the host's hardware like how an OS runs on a machine's hardware. This type of hypervisors has direct access to the host's hardware and other resources. In type 2 hypervisors, the hypervisor is run as an extra software that is installed on top of the native OS that is running on the host machine. Thus, for type 2 hypervisors, there is no direct access to the host's hardware and other resources. For type 2 hypervisors to provide access of the hardware or other resources of the host machine to the VMs, it has to request for the access through the native OS running on the host. Thus type 2 hypervisors make the performance of the VMs lower due to this bottleneck compared to the type 1 hypervisors. The hypervisor is also called virtualisation machine manager or virtual machine manager [44].

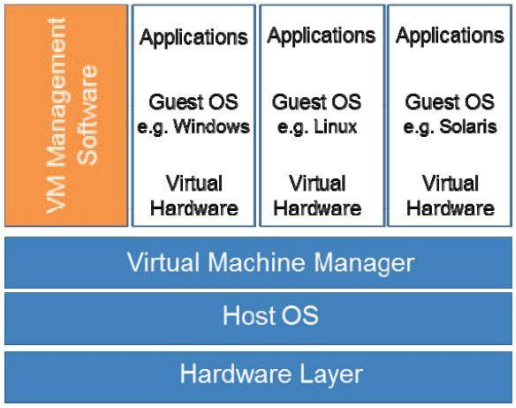

Figure 5-13: VMWare Workstation architecture [44]

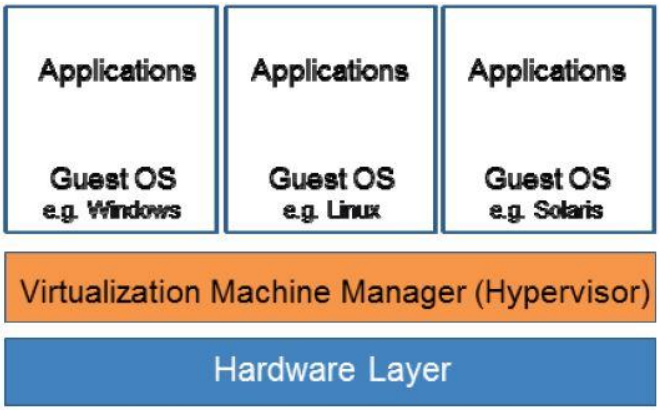

Figure 5-14:KVM architecture [44]

In type 2 hypervisors, there is an extra software component called the VM management software that is running on top of the hypervisor software. This software is used to provide easy management and creation of new VMs. With this software, it is very easy to create the various networking and other hardware devices to be allotted for the VMs. Another advantage of type 2 hypervisors is the ease of installation as the native OS running on the host takes care of the installation of the hypervisor and OS takes care of the resource allocation via the hypervisor.

In the type of virtualisation implementations used for this thesis i.e. KVM [45]and VMWare Workstation [46], there are (at least) four kinds of virtual network options that are common. They are:

- Internal network
- Host only network
- Bridged network
- NAT network

In an internal network, called custom network using LAN segments in VMWare and called network config with no gateway address in KVM, the VMs can only communicate with other VMs connected to the same internal network. The host cannot communicate with the VMs and neither external systems (i.e. external to the host) can communicate with the VMs.

In a host only network, called isolated network config in KVM, the VMs can only communicate with the host and with other VMs connected to the same host-only network. External systems, however, cannot communicate with the VMs.

In a bridged network, the VM virtual network interface are bridged to a physical interface on the host and the VMs communicate to all other VMs connected to this bridged network and they can communicate to all systems connected to the physical host interface (and conversely: the external systems can communicate to the VMs: each VM gets its own IP address which is reachable from the systems to which the host is attached).

In a NAT network, the VMs can communicate with other VMs connected to the same NAT network and they can access external systems, but only through NAT. The VMs get local addresses inside the host that are not directly accessible from outside the host. The internal addresses are mapped through NAT onto a single externally visible address, and through this mapping the VMs can access external systems, but conversely external systems cannot access the VMs directly.

5.3.2 Open5GCore architecture

The realistic core network developed for the measurement setup is based on the Open5GCore [40]. Figure 5-15 depicts the architecture of the Open5Gcore. An explanation of the functions of the various VMs of the Open5GCore will be explained in this section.

The architecture of the Open5Gcore consists of the radio access network and the Open5Gcore core network. Although Open5Gcore is a virtualised implementation of the EPC, it has the option to have an emulated radio network. As shown in Figure 5-15, the access network can be in the form of a software with emulated eNodeB and emulated UE or an actual eNodeB and an actual UE or even other access technologies. Moreover, there is also another functionality called the benchmarking tool (bt), which is not shown, that is used to test the Open5Gcore for its performance in terms of throughput and other packet handling and signalling message handling performance. In this thesis, for the mobile network setup, the configuration option with an actual LTE UE and an actual LTE eNB is used as we have an actual radio network setup with the Ericsson small cell eNB and the degrader.

Inside the core network architecture of the Open5Gcore, there are various VMs that carry out the standard LTE functions of the core network. The VM named ctrl, performs the functions of the MME, SGW-C and the PGW-C. All these three functions are collocated inside this single VM. The SGW-C and the PGW-C are the control plane counterpart of the standard SGW and

PGW LTE functionalities. These components are required to properly update and modify the SGW-U and PGW-U functionalities. The SGW-U and the PGW-U are the user plane counterpart of the standard SGW and PGW LTE functionalities. Thus, this single VM can take care of the complete signalling procedure required for a UE access. As this VM takes care of all the control plane data handling, hence the name of the VM as ctrl. The ctrl VM is connected to the VM named HSS. This VM performs the function of the HSS in an LTE network. The HSS works in coordination with the MME functionality inside the ctrl, to perform the authentication of UEs trying to attach to the LTE network. The ctrl and the HSS communicates between each other using the diameter protocol.

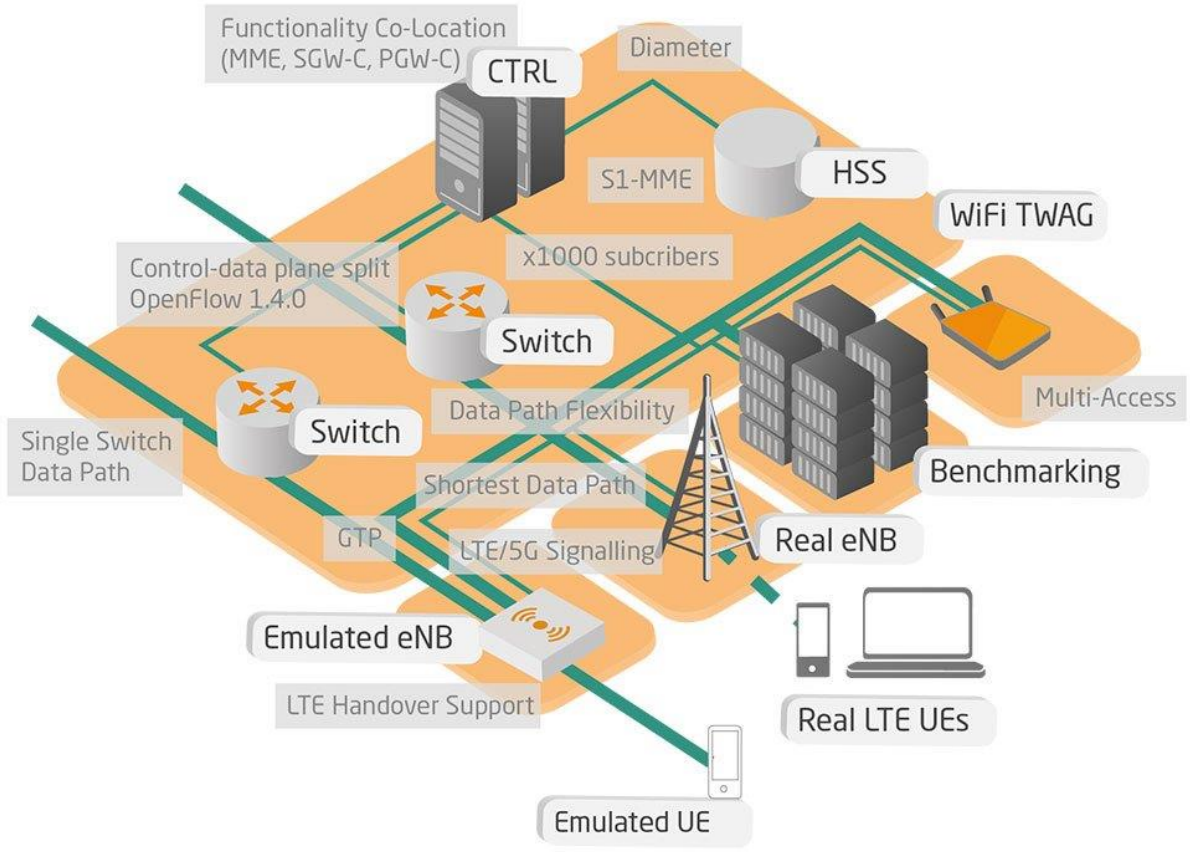

Figure 5-15: Open5GCore Architecture [40]

The VM named switch takes care of the function of SGW-U and PGW-U. These two functions are co-located on this single VM. This feature of the Open5Gcore to split the control and the user plane separately allows the measurement of user plane packet latency independent of any control plane data flow influence. A huge advantage of this feature is that as the control and user plane is separate, the control plane loading of the network is not affecting the user plane performance as they are occurring on separate virtual machines. The VM named switch performs all the packet processing and switching between the eNB and the VM named inet-gw. From the switch VM the packets are forwarded to the destination outside the core network via the VM called the inet-gw.

5.3.3 Realistic core network

The realistic core network developed for the measurement setup is based on the Open5GCore which is placed inside the physical host 2 as shown in Figure 5-1. The physical host 2 has three physical interfaces that are used for the measurement setup. They are eno1, eno2 and enoAS. The interface eno2 is connected directly to the realistic radio network. The eno1 is the physical interface through which all the VMs of the Open5GCore can communicate to the outside world. The enoAS interface was setup additionally to directly connect the AS to the core network via an ethernet cable.

The various VMs running in the Open5Gcore to carry out the functionalities of a realistic core network for this thesis are: backhaul delay emulation, ctrl1, sw, hss, inet-gw and bt.

The VM 'backhaul delay emulation' is the one that applies the deterministic transport delay to the packets in both UL and DL. As explained in previous section the VM 'bt' applies the effect of a loaded core network. Inside the bt, the number of users attached to the core network and their traffic characteristics can be specified. By varying the number of users attached or their traffic characteristics, the load applied on the core network can be altered. The VMs bt and backhaul delay emulation combined, emulates the effect of a realistic core network on the Open5Gcore.

An instance of Wireshark running on the interfaces eno2 and enoAS are used to capture and time stamp the packets to determine the packet latency in the core network. In the following section, a detailed explanation of all the VMs running inside the Open5GCore and their networking are presented to have an in depth understanding of the setup.

5.3.4 Open5GCore components and networking

The Figure 5-16 shows how the various VMs running on the Open5GCore are interconnected to each other, the host and to the outside networks.

On the VMWare Workstation, there are three virtual switches: vmnet0, vmnet8 and vmnet3. The virtual switches on the VMWare Workstation is also referred as virtual network. The vmnet0 virtual switch is bridged to the physical interface eno2 on the physical host 2. It is via this virtual switch that the packets from the realistic radio network enters the Open5GCore. The virtual switch vmnet3 is bridged to the physical interface enoAS. It is via this virtual switch that the packets are forwarded to the AS. The virtual switch vmnet8 is configured as the Network Address Translation (NAT) interface. As the default interface on the physical host 2 for outside network access is eno1, the packets from the NAT virtual switch vmnet8 are forwarded to the outside network via the eno1 physical interface.

The VM backhaul delay emulation, is connected to the virtual switch vmnet0. The VM has three interfaces: ens33, ens38 and ens39. The interface ens38 is connected to the virtual switch vmnet0 and the interface ens39 is connected to the internal network net_core. The interface ens33 interface is connected to the NAT switch vmnet8 for outside network access. The VM is configured so that any packets entering on the interface ens38 and ens39 is applied a latency of 7ms. This configuration allows to emulate the transport delay effects for packets in the UL and DL direction. As the interfaces ens38 and ens39 are attached to a Linux bridge running on the VM, the packet destined to other VMs in the Open5GCore from the realistic radio network is forwarded appropriately.

The VM ctrl1 has two interfaces. One interface is connected to the internal network net_core and the other to the virtual switch vmnet8. The interface connected to the net_core internal

network receives the packet from the realistic radio network via the backhaul delay emulation VM. Thus, the VM ctrl1 receives all the signalling messages destined to the MME from the realistic radio network. The interface connected to the vmnet8 gives access to the outside network for the VM ctrl1. This interface is also used to forward the control plane signalling towards the VMs switch and hss and also the DNS request to the inet-gw.

The VM sw has three interfaces. One interface is connected to the internal network net_core via which the VM receives the data packets from the realistic radio network via the backhaul delay emulation VM. The second interface is connected to the virtual switch vmnet8 via which the VM has access to the outside network. This interface also receives the signalling messages from the VM ctrl1. The third interface is connected to the internal network net_a via which the VM switch forwards the data packets after processing towards the inet-gw.

The VM bt which emulates the loading effect on the core has one interface. The interface is connected to the internal network net_core via which it sends both the control plane and user plane packets towards the VMs ctrl1 and switch respectively.

The VM hss, has only one interface which is connected to the virtual network vmnet8. Through this interface the VM has access to outside network and also receives the control plane signalling messages from the VM ctrl1.

The VM inet-gw has three interfaces. The first interface is connected to the internal network net_a via which it receives the user data packets from the VM sw. The second interface is connected to the virtual switch vmnet8. The VM inet-gw is the component that resolves the DNS request for all other virtual machines in the Open5GCore. All other VMs in the Open5GCore forwards the DNS request towards the inet-gw via the interface that is connected to the virtual switch vmnet8. This interface also serves as the access for the VM inet-gw to the outside network. The inet-gw also acts as the gateway to the PGW functionality running inside the VM sw. The third interface is connected to the virtual switch vmnet3 via which the packets destined towards the AS are forwarded towards the physical interface enoAS on the physical host 2.

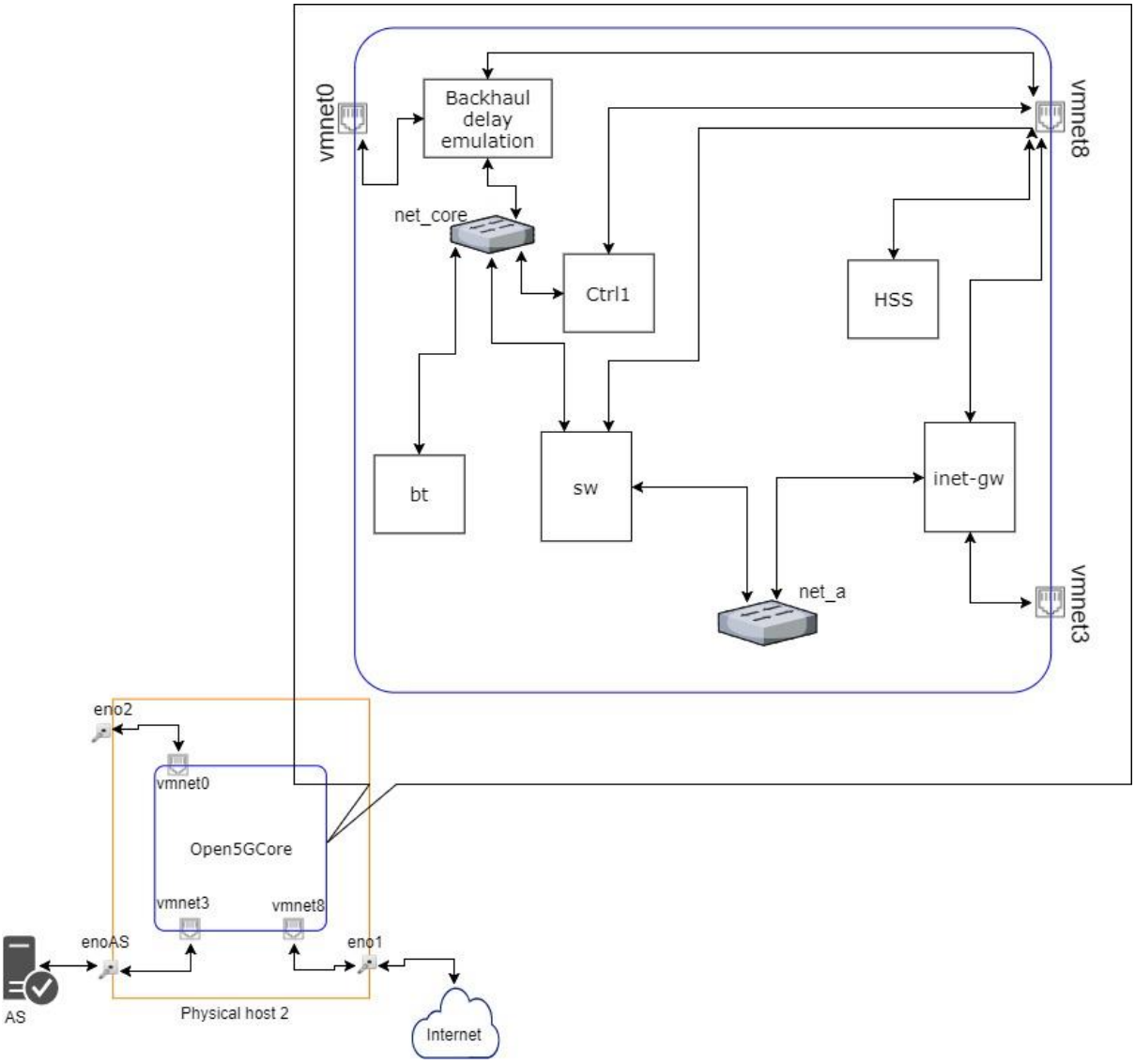

Figure 5-16: Open5GCore internal networking

5.4 Application Server

Immediately after the Open5GCore, the AS is placed in the measurement setup. The AS is directly connected to the Open5GCore via the interface ens15f2 using an ethernet cable. Such a direct connection serves the requirement for avoiding any excess latency caused from an external network. The AS is an Ubuntu machine similar to the UE. Similar to the functioning of the UE, the AS operates both as a packet sender and packet generator. In the UL direction, the AS functions as a *packet receiver* where it captures the packets received from the UE over the mobile network and time stamps them. An instance of Wireshark running on the AS is used to capture and time stamp the packets in the UL direction.

In the DL, the AS functions as a *packet sender*, where the same packet sender application as in the UE, Ostinato is used. The AS is setup to send packets in the DL direction according to the specified traffic characteristics. While functioning as a packet sender in the DL direction, the AS has to listen for any UDP packet received on the UDP port number 8888 so as to recognize that there is already a default bearer established inside the core network for the proper delivery of the DL packets towards the UE. A simple C program is run to listen for this

initial UDP packet from the UE on the UDP port 8888 and upon reception of this packet, the code is designed to initiate the DL packet transfer using Ostinato.

5.5 Time synchronisation

When measuring packet latency in a mobile network, it is very important that every measurement point in the developed measurement setup is synchronised to a single source of time. In this thesis, the time synchronisation among the various measuring points are achieved using Network Time Protocol (NTP).

In a NTP configured system, the NTP daemon running on the system request for time from a remote server at a regular interval. Based on the reported time from the NTP server, the system calculates its time drift with respect to the remote NTP server and adjusts the system clock value of the system. As the drift of the system gets gradually reduced and the system time converges towards the remote NTP server time, the frequency of time request from the system reduces [47].

One important aspect in NTP time synchronisation is the delay in reaching the remote NTP server. If there is large delay in reaching the remote NTP server, the jitter to the remote server will also vary and thus the time drift on the system will be very large compared to the remote NTP server. Therefore, it is very important that the delay for the time request packets to reach the remote NTP server is kept very low. For this thesis, all the measurement points are synchronised to the NTP server and the routing to the server is ensured to have a very low delay. Also, it is ensured that the time synchronisation request packets are not affecting the mobile network in any way. The machines which are connected to the NTP server are configured such that the packets towards the NTP server are exchanged through interfaces that are not connected to the mobile network.

Wireshark instances running on each of the measurement points obtains the time for time stamping the captured packets from the system clock of the machine. As NTP ensures that the system clock across all the measurement points are synchronised, time stamps of the captured packets across all the measurement points are also synchronised.

5.6 Latency reduction techniques

In this thesis, two latency reduction techniques are investigated. The first one is the *differentiated scheduling* in the radio network. With different levels of differentiation, the value by which the considered user is prioritised over other users is varied to investigate the latency improvements. In differentiated scheduling, the latency experienced from the core network is not affected.

In the second latency reduction technique, *Edge Computing (EC)* is used to reduce the core network latency. In EC, the data service for the devices attached on the mobile network is provided within the radio access network without having to be passed through the whole core network. It can enable the services and applications to be accelerated, thus increasing the responsiveness of the services [48].

Thus, to have an EC setup and to verify the benefits of EC, the core network processing and the application server is shifted closer to the realistic radio network. The advantage of the developed EC setup compared to the normal setup is that the EC setup does not suffer from the latencies due to the transport delay and the latency caused due to multiple hops in the backhaul topology. Another advantage is that if the core network of the operator is under heavy load due to some reason, a newly spawned processing node can be setup quickly in a

virtualised core network using the Network Function Virtualisation/Software Defined Networking (NFV/SDN) technology, close to the radio network. As the newly spawned processing node is independent of the core network, it not affected by the load on the core network. This is a win-win situation for both the operator and the user. The user is getting excellent service even though the core network is heavily loaded. The operator's advantage is that as the user's service is processed outside the core network, the service is not causing extra loading on the already loaded core network.

In the EC setup as shown in Figure 5-17, the VM sw that was running inside the Open5GCore as a VMWare Workstation VM on the physical host 2 is shifted into the physical host 1 as a KVM virtual machine. As the new VM 'edge sw' is inside the physical host 1, it is not experiencing the latency added by the VM backhaul delay emulation and the loading of the core network by the VM bt. Also, the AS that was placed after the physical host 2 is now placed inside the physical host 1 as VM 'edge AS' which is connected directly to the edge sw. The VM edge sw have two interfaces: esw1 and esw2. A new interface kau3 is added to the degrader to connect to the VM edge sw. kau3 and esw1 are connected using the option of isolated network in KVM. In a similar way, the interfaces esw2 and eas1 are connected. Wireshark instance are run inside the VMs the edge AS and edge sw to determine the one-way latency of packets separately in the UL and the DL.

An important aspect to be brought to the attention of the reader here is the discussion on the reasons for better performance of a type 1 hypervisor in Section 5.3.1. As KVM is a type 1 hypervisor, it is having a better performance than a type 2 hypervisor. However, the reader must not conclude that this advantage of a type 1 hypervisor is also a factor for lower packet latency on this KVM based EC setup.

As already mentioned, Open5GCore is not designed as a high-performance core network. Therefore, the tests performed for this thesis is far less than the load levels observed on an actual core network. The tests performed for this thesis is well under the performance limits of both KVM and VMWare based VMs. Thus, any better results reported on the EC setup cannot be reasoned due to the performance advantage of a type 1 hypervisor. The choice for a KVM based EC setup was due to the simplicity in setting up a new VM using KVM.

To enable such a setup, it was required to change the path of the OpenFlow commands from the VM ctrl1 towards the VM edge sw. In the normal setup, the OpenFlow commands were exchanged through the vmnet8 virtual switch. As the VM edge sw is not having connection to the vmnet8 virtual switch, the OpenFlow messages in the EC setup is now exchanged via the vmnet0 virtual switch though the physical interface eno2 on the physical host 2 and eventually into the VM edge sw via the KauNetEm degrader through the physical interface eht2 on the physical host 1. Thus, the VM edge sw receives all the OpenFlow messages to properly configure the flows from the core network running inside the physical host 2 while the user data packet processing remains inside the physical host 1.

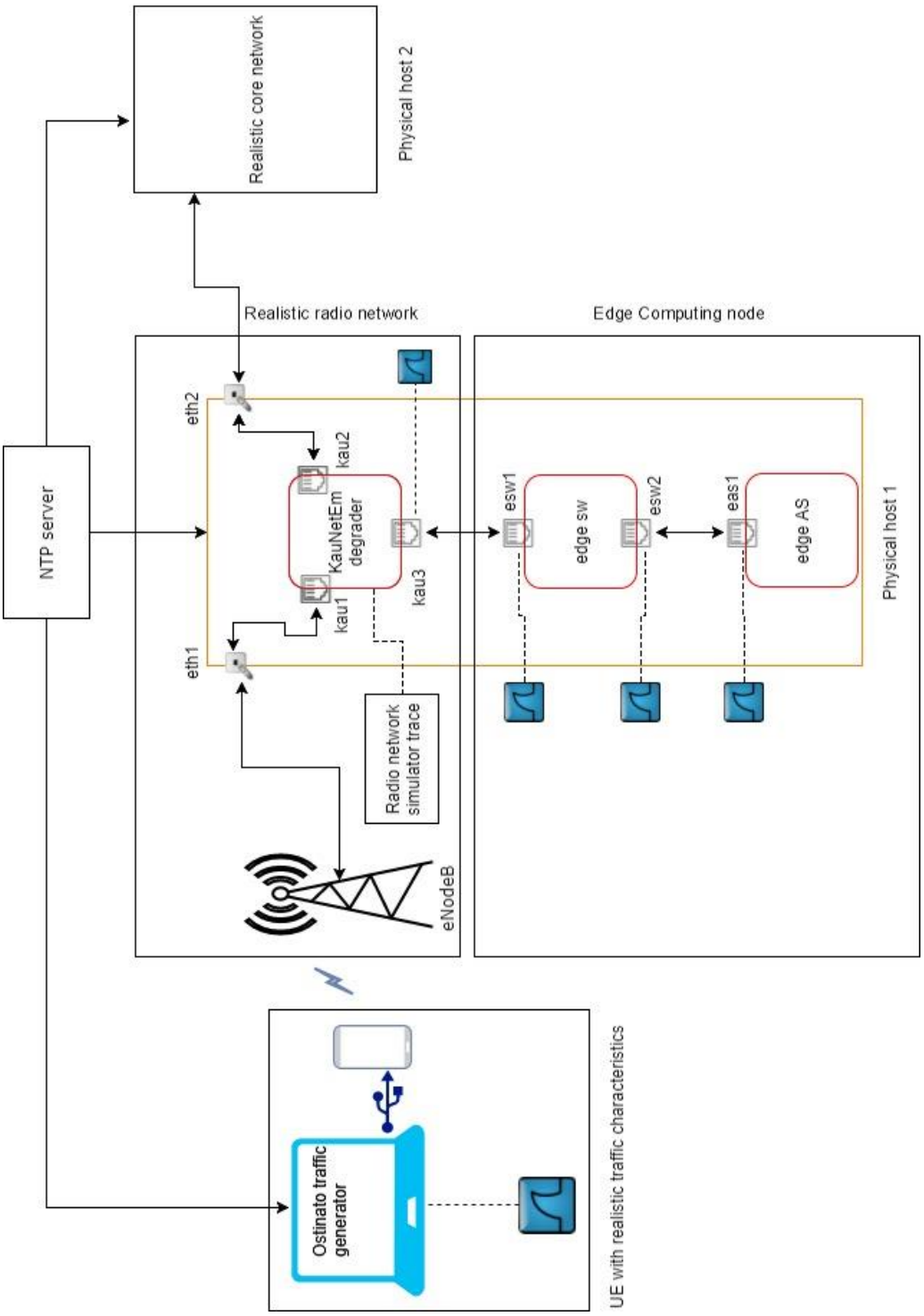

Figure 5-17: Edge Computing setup

Chapter 6. Scenarios and scheduler modelling

In this chapter, the latency assessment strategy is presented in the form of the different scenarios that are evaluated with the developed measurement setup. More specifically, the default and the alternative settings are specified for the six distinct scenario aspects that are considered, i.e. the network load, user distance from the eNB, packet size, packet rate, differentiated scheduling and the edge computing. The reasoning for choosing a particular value as the default setting of the six distinct scenario aspects will also be presented along with the description. This chapter also presents the scheduling models considered for the UL and DL measurements.

6.1 Measurement strategy

In the measurements performed to understand the effects of each of the six scenario aspects on the packet latency, a unilateral variation of the different values of each of it is considered. To study the effect of the variation of a scenario aspect on packet latency, the different values of that aspect are evaluated using separate measurement experiments in a ceteris paribus manner.

For example, to study the effect of variation of network load on packet latency, different network load levels are considered in separate experiments. While a different network load level is considered in each experiment, the values of the other five scenario aspects will be set to their default setting.

Again, for a particular value of the considered scenario aspect, the measurements are repeated five times to ensure statistical accuracy. When the result for that particular value of the considered aspect is generated, the results from the five experiments are combined together. The number of repetitions was restricted to five after an initial analysis of the results. A comparison of results was performed for the default setting, with measurements repeated for five times and ten times. The results for both the cases were similar, indicating a repetition of five times was sufficient to eliminate any statistical errors from the measurement setup.

In the results, a comparison between the packet latency statistics for all the distinct values of the considered scenario aspect are presented.

Such a scenario enables to understand the effect of each of the six-scenario aspect on packet latency without the influence from any other scenario aspect.

The Table 6-1 shows the distinct values considered in the experiments for the six scenario aspects. The default value of each of the aspect is indicated as **red** in the table.

The network load aspect considers the load on the radio network and the core network. Four different load levels are considered for this thesis: low load, medium load, high load and very high load, represented in their respective order in the table. For a particular load level, a similar load level is considered in the radio and core network.

In the *core network*, a particular level of loading is achieved using the VM bt in the Open5GCore. To achieve a particular level of loading in the core, a particular number of emulated UEs are attached to the Open5GCore and traffic is generated in the core in both UL and DL directions.

Scenario aspect	Direction			Values considered				
Network load	Radio network	Inter-cell (Interference from neighbouring cells)	UL (noise floor rise in dB)	0	1.5	**3**	6	-
			DL (transmit power of neighbouring cells in % of max transmit power)	0	25	**50**	100	-
		Intra-cell (number of other users in the cell)	UL / DL	0	1	**2**	6	-
	Core network (total no of other users in the core network)	UL / DL		0	12	**25**	50	-
User distance from eNB (km)	UL / DL			0.10	0.50	0.70	**0.90**	-
Packet size (bytes)	UL / DL			100	500	1000	1100	**1200**
Packet rate (pps)	UL / DL			100	110	120	**130**	180
Differentiation weights in the radio network	UL			**1**	2	5	10	-
	DL			**1**	2	3	4	-
Edge Computing	UL / DL			Yes	**No**	-	-	-

Table 6-1: Default scenario

In the *radio network*, a particular level of load is consisting of two factors: inter-cell and intra-cell loading. The inter-cell loading effects in the DL is modelled as the inter-cell interference due to the various levels of transmit power of the neighbouring cells expressed as percentage of their maximum transmit power. In the UL, the inter-cell loading is modelled as the inter-cell interference, causing a noise floor rise due to different levels of UL activity in the neighbouring cells.

In intra-cell loading, the loading effect on the considered cell due to other users sharing the radio resources between themselves is considered.

To study the effects of user distance on latency, different distances as shown in the table are considered. As the network layout considered is hexagonal with an inter-site distance of 1.5 km, the cell range corresponds to 1 km. Therefore, a user at a distance of 0.9 km is considered as a cell edge user. Note: In the results for latency variation for various user distance in the DL, the curves for the distances 0.1 km and 0.5 km are overlapping, suggesting a similar latency variation for both. Experiments were done also for the distance of 0.3 km and it also resulted in an overlapping curve. Therefore, the curve for the distance of 0.3 km is omitted as the reason for a similar curve for 0.3 km as 0.5 km is the same as that for 0.1 km. In the UL

measurement results, it can be observed that for the cases of distance 0.1 km and 0.5 km, the curves are close together. Measurements were also performed for the distance of 0.3 km and the curve was between the 0.1 km and 0.5 km case. As, the curve for 0.3 km is omitted for DL, it is also omitted for the UL.

In the experiments for a particular scenario aspect, where different values of the aspect are considered for the tagged user, the other users in the system are set to the default setting of the various scenario aspects. Results presented in this thesis are the results for this tagged user. Depending on the scenario aspect, the variation of its value may or may not affect the other user. For example, considering network load, a variation of load in the network affects all the users in the network. However, for other aspects, a variation in its value for the considered user is not affecting the default setting for the other users. For example, when different weights of differentiation are considered for the tagged user, the other users in the system are given a weight of 1.

Therefore, except for the scenario aspect of network load, different values of the scenario aspect can affect the latency of the other users in the system. For example, when the tagged user is placed close to the eNB, the user experiences relatively better channel quality than the default settings. Due to a better radio channel, the resource utilisation by the tagged user is different than the default scenario. Due to the different resource utilisation, the resource availability for the other users in the system also varies resulting in a different latency statistic than the default setting. This aspect is of great importance, especially for the case of differentiation. As the tagged user is explicitly prioritised in the network, it is important to assess how the other users in the system are affected. Although, the other users are affected in other scenario aspects also, as this effect does not occur due to any explicit change in the system aspect, it is not considered in this thesis. Only the effects on other users with differentiation is considered in this thesis.

The specific value for each of the scenario aspect in the default scenario is not chosen arbitrarily.

In network planning by any mobile network operator, the different KPI targets are specified for a cell edge user. It is under the assumption that, if the cell edge users – worst case users in terms of performance, are ensured that the KPI targets are achieved, all other users in the network are naturally assumed to achieve them as well. Therefore, in this thesis, the distance of a cell edge user (0.90 km) is set as the default value for distance from the eNB.

The default packet size should be the most commonly observed packet size in an IP network – maximum possible packet size of 1500 bytes. But due to the limitation of the Open5GCore for a maximum packet size of 1200 bytes, the default packet size is set to 1200 bytes.

For DL inter-cell loading, the default value is set to 50% transmit power by the neighbouring cells as it is a middle value. The default level of inter-cell loading in the UL is set to a noise rise of 3 dB.

The radio network simulator was used to generate channel bit rate trace for different user distance from the eNB and at each inter-cell load levels in the DL. From the bit rate trace for a cell edge user experiencing default inter-cell loading in the DL, the average bit rate was found to be 5.2 Mbps. As a bit rate above 1.5 Mbps would be considered acceptable for a cell edge user, a total of three users, providing 1.7 Mbps per user, is chosen as the default value for intra-cell loading in the DL. The same number of user is set as the default value for intra-cell loading in the UL.

As the default value of intra-cell loading is three users and as there are twelve cells considered in the network layout, there are a total of thirty-six users in the network for the default scenario. It is a fair assumption that out of the thirty-six users, only twenty-six users are actively engaged in the network, sending traffic to the core network. Thus, for the default load level, a total of twenty-five other users are assumed in the core network.

With an average bit rate of 1.7 Mbps for the default scenario and the default packet size of 1200 bytes, a packet rate of 180 pps seems valid. However, initial results from the python script indicated that with 180 pps, the latencies determined were extremely high.

It was identified that due to the fading dips in the default channel, with a packet rate of 180 pps, packets were experiencing high latency due to queuing in the eNB buffer and latency statistics kept rising exponentially. After multiple tests a packet rate of 130 pps was identified to report a stable latency statistic. Thus 130 pps is set as the default packet rate in DL. The same packet rate is set as the default value for UL. A detailed mechanism on how packet rate is affecting latency is presented in the results section.

With the default values for packet size, packet rate, user position, inter-cell loading in DL and intra-cell loading in the DL with a total of three users, it was observed that even a slight increase in the intra-cell loading to four users, produced unstable latency statistics. Thus, this level of intra-cell loading in the DL is considered a high value and anything above this is considered a very high value, producing unstable latency statistics. Thus, all other default values of load, corresponding to the high intra-cell loading in the DL is also referred as high load value.

For the default scenario, there is no differentiation applied to the considered user. All the users in the radio network are assigned radio resources with equal priority. However, it is to be noted that the prioritisation weights considered for the UL and DL are different. For UL, higher weights are considered (2,5,10) than DL (2,3,4). This is because, the results of the experiments for DL with a weight of even three indicates that the maximum latency reduction is already achieved and higher weights like in the UL of five or ten would results in a similar curve.

For the default scenario, there is no EC applied and the considered user is served by the realistic core network. Note: In the scenario with EC, the edge computing core network is considered to have a low load level (only the tagged user's traffic) while the radio network is considered to have the default load level. This is because, in Edge Computing, the network is flexible in the sense that it is able to assign enough resources to the edge computing node dynamically based on the instantaneous resource requirement. Thus, with proper provisioning of resources dynamically to the edge computing node, it can be ensured that the node is always having sufficient resources resembling an unloaded node.

6.2 Scheduler modelling

In this section, the modelling of the scheduling schemes used to calculate the latency trace from the radio network simulator's channel bit rate trace is discussed. A scheduler is used in LTE to assign and share the limited network resources among the users of the network.

In this thesis, the developed radio network for the measurement setup has a channel bandwidth of 5 MHz which corresponds to a total of 25 PRBs.

6.2.1.1 DL Scheduling

In the DL, the assumed scheduler considers a simple round robin scheduling purely in the time domain. What this means is that in every TTI, all the 25 PRBs are assigned to a particular UE while the rest of the UEs are assigned no PRBs and are waiting for their turn to come.

More specifically, if the network has M users and the tagged user is assigned all the PRBs in a particular TTI, the tagged user then has to wait for another M-1 TTIs before it gets the next PRB assignment. In differentiated scheduling, the tagged user is assigned a modified round robin scheduling where that user is prioritised over the rest of the users. For example, with prioritisation weight of 2, the tagged user is assigned all the 25 PRBs in two consecutive TTIs instead of a single TTI compared to case of non-differentiated scheduling. The tagged user then has to wait for M-1 TTIs from the instant of the 2nd TTI assignment until the next resource assignment is done. Thus, with a weight of 2, the user is assigned 50 PRBs over two consecutive TTIs (2 x 25 PRBs). This will enable the user to clear/receive more data compared to the non-differentiated case. A similar trend is followed for higher weights. However, this kind of resource allocation to the tagged user in differentiated scheduling is under the assumption that only the tagged user is prioritised.

6.2.1.2 UL Scheduling

In the UL, the scheduling scheme is based purely on frequency domain. What this means is that every user in the UL is scheduled with at least one PRB in every TTI. So, unlike DL, the user does not have to wait for PRB assignment and can clear or receive some bits in every TTI.

In UL, out of the assumed spectral availability of 25 PRBs, only 24 PRBs are used to assign resources to the users as one PRB is reserved for signalling. In every TTI, the number of PRBs assigned to each user depend on the total number of users served by the eNB and SINR per PRBs received at the eNB. For example, if there are three users, the 24 PRBs get shared among the three users resulting in eight PRBs per user in every TTI.

In the differentiated scheduling in the UL, the tagged user is assigned frequency-domain resources with a higher priority compared to the other users in the network. Therefore, a prioritisation weight of x means that out of the 24 PRBs available for data transmission, it will be shared among the users in such a way that the tagged user is assigned with resources x times as that of the other users.

In the LTE UL, the maximum transmit power limit for a UE is p_{MAX}. The transmit power control (TPC) is essentially driven by a target *receive* power spectral density (i.e. per PRB), from which the target *transmit* power spectral density is derived. Given a by default equal distribution of PRBs over the different users, a cell edge UE will typically require more aggregate transmit power to utilise the assigned PRBs than a UE closer to the cell site, due to the differences in path loss. Consequently, a cell edge UE is less likely to have enough power headroom to be able to utilise any additionally assigned PRBs with the needed transmit power. Hence any extra PRBs that are available for assignment to a cell edge user by means of differentiated scheduling, may in fact not be assigned in practice. In contrast, a UE close to the cell site is more likely to have enough power headroom to utilise a more generous PRB assignment and thus achieve higher bit rates. Back to cell edge UE, even if it would be assigned additional PRBs and reduce the per-PRB transmit power in order to indeed utilise them, then the eventual aggregate bit rate may not necessarily improve by much, as the potential gain from adding PRB may be largely cancelled out by the reduced per-PRB SINR.

In conclusion, for a cell edge UE it is a priori not obvious that a more generous PRB assignment would lead to an increase in the experienced bit rate. For a UE close to the cell site, such an increase is more probable given the power headroom is may likely have.

Chapter 7. Results and analysis

In this section, the results and analysis of the measurements performed in DL and UL for packet latency are presented. In Section 7.1, the results of measurements performed in the DL for the six scenario aspects are presented along with their analysis. In Section 7.2 the results of measurements performed in the UL for the six scenario aspects are presented along with their analysis. In Section 7.3, the results for radio network processing delays in the UL and DL are presented. As explained in the Section 3.3.1, the processing delay is experienced by all the packets, due to the different signalling procedure in the radio network, irrespective of the what scenario aspects are considered. Therefore, results that are presented in the Sections 7.1 and 7.2, the processing delay are already included in the determined packet latencies.

In the results for each of the six scenario aspects, Empirical Cumulative Distribution Function (ECDF) plots for all the considered values of a particular scenario aspect and a bar plot representing the 10^{th} percentile, 90^{th} percentile and the average of the latency are presented.

An ECDF plot represents the fraction of the total samples that are having a value less than or equal to a value represented on the horizontal axis. In the ECDF plots of the results, the end-to-end latency for the packets is represented on the horizontal axis and the fraction of the total number of packets that are having an end-to-end latency less than or equal to a particular value on the horizontal axis is represented on the vertical axis. In the ECDF plots, a curve towards the left indicates that the packet latency is lower across all percentile levels. Therefore, in the analysis, such a curve would be considered as a better curve in terms of packet latency distribution.

In the ECDF plots, separate plots are presented for the latency statistics in the core network, radio network and the end-to-end latency. The bar graphs only show the different percentiles and average for the end-to-end latency.

If a particular trend is observed across the results for different scenario aspects, the reasoning for that will be discussed only at the first occurrence of the trend and subsequent occurrences will be referred to the first.

7.1 Results of measurements in the Downlink

In this section, the results and analysis of the packet latency measurements in the DL are presented. The measurement strategy considered for the measurements are as explained in the Section 6.1, where, to study the effect of a particular scenario aspect, a default scenario is considered and a unilateral deviation of the scenario aspect is considered. The results order are as follows:

- Network load variation
- User distance variation
- Packet size variation
- Packet rate variation
- Differentiation variation
- Edge Computing

7.1.1 Impact of load on latency in DL

Figure 7-1 shows the results for the impact of load on latency and packet drop percentage in DL.

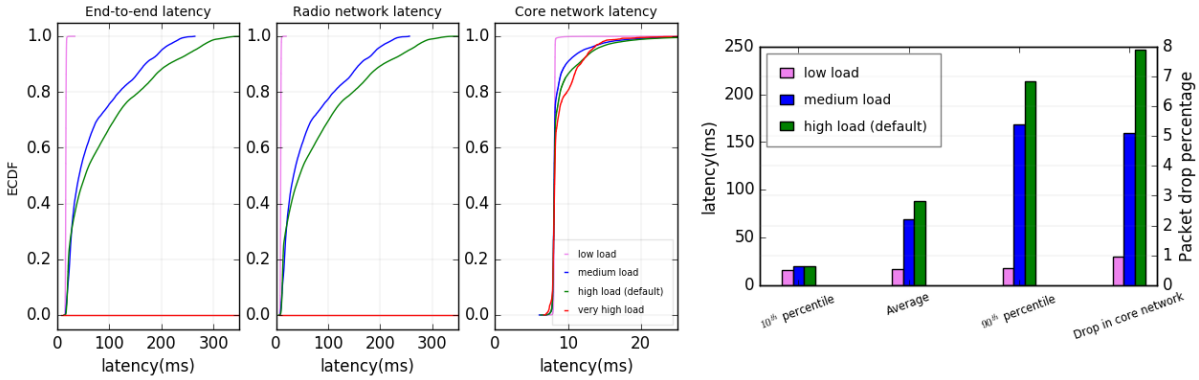

Figure 7-1: Impact of load on latency and packet drop percentage in DL

In the *core network*, results indicate that packet latency is insensitive to the load (for the considered load range) in the core network for percentiles up to the 70th percentile.

At low load level, the core network does not have any other emulated users besides the tagged user whose latency is explicitly measured. There are sufficient resources available in the core network to process all the user data packets and almost all the packets are experiencing very low packet latency. As the load level is increased, the available resources are shared with an increased number of other emulated users and therefore the tagged user's packets are not always getting enough resources to be processed immediately. As a result, some of the packets, that happened to enter the core network when the resources are fully utilised for processing the other user's packets, are buffered in the core network until resources are available again to process the tagged user's packets. Therefore, such packets exit the core network with higher latency. This is reflected in the higher latency percentiles. The instantaneous resource availability in the core network for the tagged user decreases with an increase in load level resulting in slightly higher latency.

Although the variation in the packet latency in the *core network* is not as prominent as in the radio network, another important aspect is the packet drop percentage. As explained above, the tagged user's packets arriving into the core network are buffered in the core network. As the load level varies, the buffer occupancy also varies. Thus, some packets that arrive into the core network are dropped as the buffer is already full. The number of packets getting dropped increases with increase in the load level. The percentage of packets dropped in the core network at different load level are represented in the bar plot.

In the *radio network*, the packet latency for the low load level is very low compared to all other load levels. As the load level is increased, the packet latency also increases.

At low load, as there is only one user in the radio network, the DL scheduler assigns all resources to the user in consecutive TTIs. Therefore, the packets do not have to wait, and an attempt to clear the packet from the eNB buffer can be made in every consecutive TTIs. Moreover, the neighbouring cells are not transmitting causing no interference and hence better radio channel quality.

At higher load levels, the latency is higher than the low load level. The following are the reasons for this:

- Transmission latency – As the load level is increased, the DL scheduler does not assign all the resources to the tagged user in consecutive TTIs. If there are M users in the radio network, the tagged user is not assigned any resources for M – 1 TTIs, after every TTI assignment. With an increase in load, this interval, with no resource assignment for the tagged user increases as the number of users are increasing. Therefore, a packet of 1200-byte will obtain sufficient resources to completely clear the packet from an extended duration, resulting in a higher transmission latency. Moreover, an increased load level means an increased interference and therefore lower SINRs hence lower bit rates per TTI, in the channel bit rate trace. Due to this lower bit rates, a packet of same size will need more number of TTI assignments to clear than compared to a lower load level. Thus, an increase in load level also contribute to a higher transmission latency.
- Scheduling latency – Due to a higher transmission latency as explained above with an increase in the load level, a packet remains in the eNB buffer for an extended duration before it is completely transmitted. Therefore, the subsequent packets arriving to the eNB buffer, is buffered until all the previous packets are cleared. Therefore, as transmission latency increases due to an increase in load level, it will consequently cause a higher scheduling latency.

For the extreme case of a very high load level, as a result of a very poor channel condition due to very high interference and more number of users in the network, the latency experienced by the packets are extremely high. For the given range in the plots, this results in a horizontal line as shown.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for end-to-end latency as the radio network latency.

In the bar plots, a similar 10^{th} latency percentile is noticed across the different load levels. The 10^{th} latency percentile is likely determined by the relatively low number packets experiencing favourable channel conditions (multipath fading spikes). Such periods of favourable channel conditions occur equally often at all load levels. Even at a high load level, the few packets arriving during these periods, experience favourable channel conditions such that a 1200-byte packet will experience a lower transmission latency due to the higher bit rates during that period. Consequently, the subsequent packets experience a lower scheduling latency. As a result, a few packets experience a similar lower latency as compared to the case of a low load level.

At higher load level, consider for instance the 90^{th} latency percentile, we note that such high percentiles are likely determined by the major fraction of the total packets, with relatively unfavourable channel conditions (multipath fading dips), in which case packet transmissions take longer and, consequently, the scheduling latency also increases for buffer packets, particularly so under higher packet rates. However, a few packets arriving during a period of more favourable channel conditions, experience lower latencies as explained above, determining the 10^{th} latency percentile. Therefore, a significant difference in the levels between 10^{th} and 90^{th} percentile is observed for higher load levels. This will consequently lead to a significant difference in the average latency.

A similar level for the 10^{th} and 90^{th} latency percentile for a low load level indicate that even for the major fraction of the total number of TTIs, which experience a relatively lower channel

quality compared to the favourable channel conditions, due to the absence of interference, the bit rate is high enough to cause very low latency. Thus, majority of the packets experience a similar low latency as compared to the few packets arriving during the favourable channel conditions, resulting in a similar level for the 10th and 90th percentile. This will consequently cause the average value also to be similar.

7.1.2 Impact of user distance on latency in DL

Figure 7-2 shows the results for the impact of user distance on latency and packet drop percentage in DL

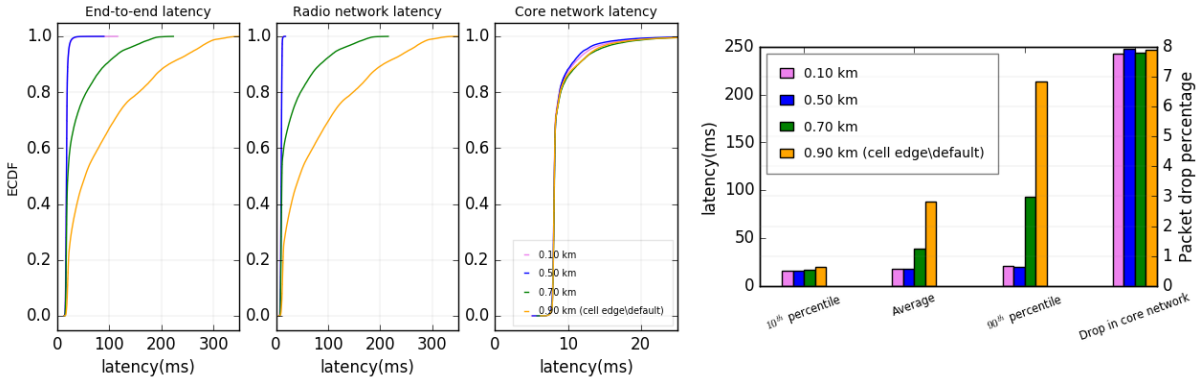

Figure 7-2: Impact of user distance on latency and packet drop percentage in DL

In the *core network*, the packet latency distribution curves for the different tagged user's positions in the cell are almost overlapping each other suggesting that packet latency is hardly affected by user distance, in the core network. In this experiment, only the position of the tagged user is varied while the other users remain in their default location and the load level considered is the default level of high load. So, this change in position, only of the tagged user is not significant enough to cause any noticeable effect in the core network, where traffic from a large number of cells and user are aggregated. However, an insignificant variation in the packet latency is noticed in the core network and this is due to the sporadic variation in the resource availability in the core network. As a high load level is considered in the core network for all the cases, the packet drop percentage remains almost the same in all cases, as shown in the bar plots.

Due to the lack of ability to strictly control the load level on the core network in the experiments, the actual load level varies around the target default load level of 25 Mbps. This variation is realistically due to the fact that emulated users attached to the core network are sending packets without any kind of coordination between them and consequently the number of packets coming into the core network at any time instant is not fixed. Thus, at every instant the buffer utilisation in the core network is different and hence the packets of the tagged user entering the core network will have different waiting times. This explains the small variation in the packet latency distribution in the core network for the same load level.

In the *radio network*, the packet latency for the user increases as the distance of the user from the eNB increases. This is expected as the channel quality degrades as distance increases due to an increase in both the path loss and inter-cell interference level. As the load level and traffic characteristics are kept fixed in the analysis, the observed variation in packet latency is only caused by channel quality variations.

The packet latency curve for a distance of 0.1 km and 0.5 km are overlapping. This suggests that the channel qualities in both cases are similar in terms of the rate at which the packets can be cleared from the eNB buffer. If we look at the specifics of the applied traces, we see that for user at 0.1 km, the experienced bit rate under the worst fading conditions is 11.8 Mbps suggesting that even then a packet of size 1200 byte can be cleared from the eNB buffer within a single TTI. For user at 0.5 km from its serving cell, only about 0.6% of the total TTIs in the trace experience bit rates too low to clear a packet within that single TTI. Thus, the curves for the tagged user at distances 0.1 km and 0.5 km are overlapping.

For the channel quality at 0.7 km, there are many more TTIs in which the experienced bit rate implies a packet transfer time of multiple assigned TTIs. Due to the need to wait in the eNB buffer for more TTI assignments, the subsequent packets arriving to the eNB will experience higher scheduling latency.

For the distance of 0.9 km, these effects are further amplified causing even higher latencies.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for end-to-end latency as the radio network latency.

In the bar plots, favourable channel conditions (multipath fading spikes) that occur equally often across all the distances, are favourable enough to causes a similar low latency to a few packets arriving during those periods as compared to a user close to the eNB, resulting in a similar 10th latency percentile across the different user distances.

At long distance, higher latency percentile (e.g. 90th latency percentile) that is likely determined by periods of relatively unfavourable channel conditions (multipath fading dips) causes a major fraction of the total packets, a higher transmission latency and consequently higher scheduling latency for the subsequent packets, resulting in a high 90th latency percentile. Due to reasons explained above, few packets experience a lower latency resulting in a lower 10th latency percentile. Therefore, at long distances, a significant difference is observed for the 10th and 90th percentiles. This will consequently lead to a significant difference in the average latency. Due to the reasons explained above, the levels for 0.1 km and 0.5 km are similar.

7.1.3 Impact of packet size on latency in DL

Figure 7-3 shows the results for the impact of packet size on latency and packet drop percentage in DL.

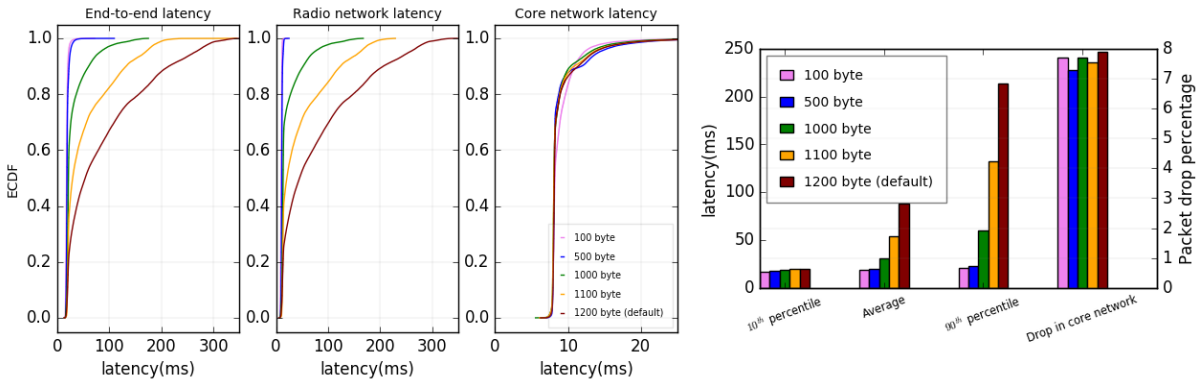

Figure 7-3: Impact of packet size on latency and packet drop percentage in DL

In the *core network*, the packet latency distribution for packets with different size are overlapping, suggesting that core network latency is hardly impacted by packet size. As the

load level in the network is considered as the default load level of high load for all the cases, the small variation in the packet latency observed in the results is due to the sporadic variation in the resource availability in the core network. Due to this reason, the core network packet drop percentage remains almost the same in all cases, as shown in the bar plots.

The results further show that the packet size affects the *radio network* packet latency. As the packet size is increased, the packet latency increases. However, latency for packets of size 100 and 500 bytes are very low and similar, compared to packets of larger size.

This is due to the fact that in the default scenario, the channel quality is such that in majority of the TTIs in the considered trace, the bit rate experienced is high enough, to clear packets up to a size slightly higher than 500 byte, within that TTI while all the TTIs have sufficiently high bit rate to clear a 100 byte packet within that TTI. Hence a packet of up to 500 bytes experience a similar low transmission latency, resulting in an overlapping curve for 100 and 500-byte packets.

On the other hand, for packets of 1000 bytes, given that the average bit rate per TTI is 5.2 Mbps in the trace, a lot of packets need more than two TTI assignments resulting in a high transmission latency. Moreover, a packet of 1000 bytes arriving during a period of unfavourable channel condition (multipath fading dips) will experience higher transmission latency. Consequently, this can lead subsequent packets to have higher scheduling latency. These effects are further amplified for larger packets and therefore their packet latencies are higher.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for end-to-end latency as the radio network latency.

In the bar plots, the favourable channel conditions (multipath fading spikes) in the considered channel trace, are favourable enough to causes a similar low latency to a few packets of 1200 bytes, arriving during those periods as compared to packets of 100 bytes, resulting in a similar 10^{th} latency percentile across the different packet sizes.

For large packet size, higher latency percentile (e.g. 90^{th} latency percentile) that is likely determined by periods of relatively unfavourable channel conditions (multipath fading dips) causes a major fraction of the total packets, a higher transmission latency and consequently higher scheduling latency for the subsequent packets, resulting in a high 90^{th} latency percentile. Due to reasons explained above, few packets experience a lower latency resulting in a lower 10^{th} latency percentile. Therefore, for large packets, a significant difference is observed for the 10^{th} and 90^{th} percentiles. This will consequently lead to a significant difference in the average latency. Due to the reasons explained above, the levels for 100 and 500-byte packets are similar.

7.1.4 Impact of packet rate on latency in DL

Figure 7-4 shows the results for the impact of packet rate on latency and packet drop percentage in DL.

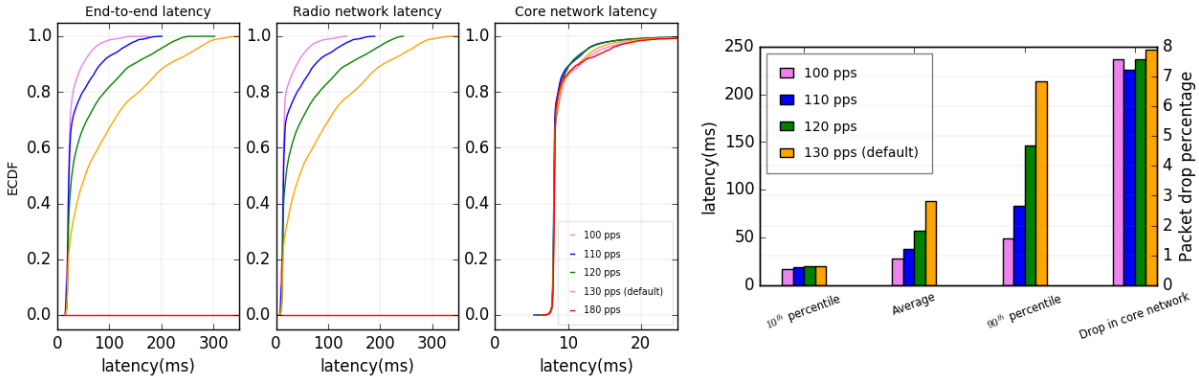

Figure 7-4: Impact of packet rate on latency and packet drop percentage in DL

In the *core network*, considering that the same load level is assumed in all scenarios which is hardly affected by the variation of the packet rate of one individual (the tagged) user, the different packet latency distributions are overlapping, suggesting that core network latency is hardly affected by the packet rate variation of the tagged user. However, a small variation in the packet latency is again observed at the higher percentiles due to the sporadic variation in the resource availability in the core network. Due to this reason, the core network packet drop percentage remains almost the same in all cases, as shown in the bar plots.

In the *radio network*, the packet latency increases in the packet rate. As the channel quality is not affected by the variation in packet rate, this variation in packet latency is due to the fact that when packet rate is increased, there is an increased number of packets already in the buffer to be cleared. Therefore, an arriving packet experience higher scheduling latency due to this increased number of packets already in the buffer.

In the considered channel bit rate trace, the bit rate experienced by a vast majority of TTIs is such that a 1200-byte packet need multiple TTI assignments to obtain enough resources, resulting in a high transmission time. Moreover, during a period of unfavourable channel conditions (multipath fading dips), the packet need even more transmission time due to lower bit rates during those TTIs. In such a channel quality, where a packet of 1200-byte experience high transmission latency, an increase in the packet rate means that a packet coming to buffer will experience an increased number of packets, already in the buffer. Therefore, the arriving packet will experience higher latency due to the increased scheduling latency. Thus, as packet rate is increased, the packet latency also increases as observed in the results.

For the extreme case of a packet rate of 180 pps, for similar reasons as for the very high load level, the latency curve results in a horizontal line.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

In the bar plots, the favourable channel conditions (multipath fading spikes) in the considered channel trace, are favourable enough to causes a similar low latency to a few packets with rate as high as 130 pps, arriving during those periods as compared to packets with a lower rate of 100 pps, resulting in a similar 10^{th} latency percentile across the different packet rates.

At higher packet rates, higher latency percentile (e.g. 90th latency percentile) that is likely determined by periods of relatively unfavourable channel conditions (multipath fading dips) causes a major fraction of the total packets, a higher scheduling latency due to an increased number of packets already in the buffer, resulting in a high 90th latency percentile. Due to reasons explained above, few packets experience a lower latency resulting in a lower 10th latency percentile. Therefore, for high packet rates, a significant difference is observed for the 10th and 90th percentiles. This will consequently lead to a significant difference in the average latency.

7.1.5 Impact of differentiated scheduling on latency in DL

Figure 7-5 shows the results for the impact of differentiated scheduling on latency and packet drop percentage in DL.

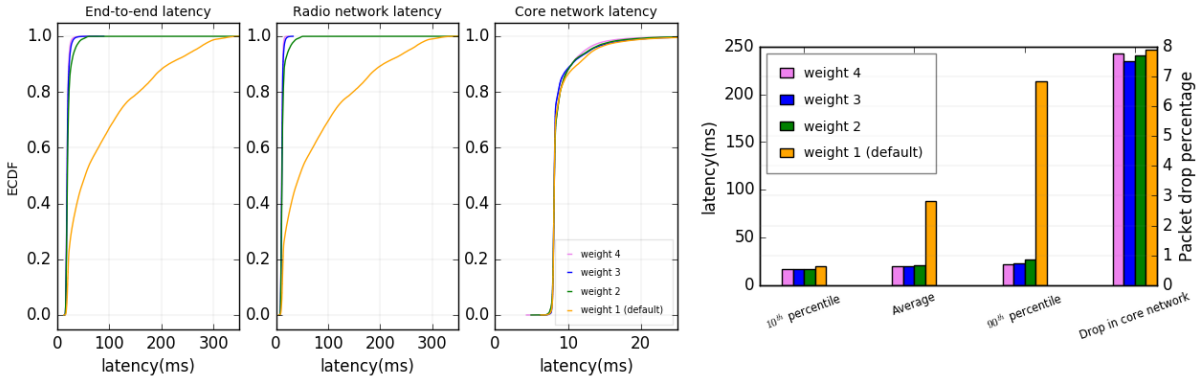

Figure 7-5: Impact of differentiated scheduling on latency and packet drop percentage in DL

In the differentiated scheduling in the radio network, the tagged user is prioritised over the rest of the users, for radio resource assignment as explained in Sections 6.2.1.1 and 3.4.1. In this experiment, only the tagged user is prioritised while the rest of the users are assigned resources with the default priority of weight 1 (meaning no prioritisation). However, in reality, a network operator can take an approach of prioritising x % of the total number of users. To determine the maximum possible latency reduction, it is necessary to consider the case where only the tagged user is prioritised. In all other cases where there are more high-priority users in the network, the gain in latency reduction for the tagged user (users) will be lesser than the case with only the tagged user prioritised. In the worst case, where all the users are a high-priority user, there is no gain at all.

In the *core network*, as expected, the differentiated scheduling, which takes place in the radio network only, is not affecting the packet latency in the core network. Hence the packet latency distribution is insensitive to the different levels of differentiation indicating a similar performance for all the cases. As the load level in the network is considered as the default load level of high load for all the cases, the small variation in the packet latency observed in the results is due to the sporadic variation in the resource availability in the core network. Due to this reason, the core network packet drop percentage remains almost the same in all cases, as shown in the bar plots.

In the *radio network*, a significant reduction in the packet latency is observed with an increase in the level of differentiation with respect to the default scenario. With a mean bit rate of 5.2 Mbps per TTI, as taken from the channel bit rate trace for the default scenario, a 1200-byte packet will experience a transmission latency of 2 TTIs. In the default prioritisation weight 1 with a total of three users in the radio network, a 1200-byte packet will experience a latency

of 4 ms, under the assumption that the packet arrived exactly at an instant of TTI assignment for the tagged user and there are no other packets already in the buffer. In the case for a prioritisation weight 2, the same packet will experience only 2 ms latency. Moreover, during the unfavourable channel conditions (multipath fading dips), this difference in latencies are further increased. The Figure 7-6 shows a worked-out example for such a situation, considering that the bit rate experienced per TTI is 3.5 Mbps. It can be observed that with a weight of 2, apart from a reduced transmission latency for packet 1, the packet 2 is also not experiencing any scheduling latency as it is not having to wait for any previous packet to be cleared from the buffer. Therefore, packet latency decreases with an increase in the level of prioritisation.

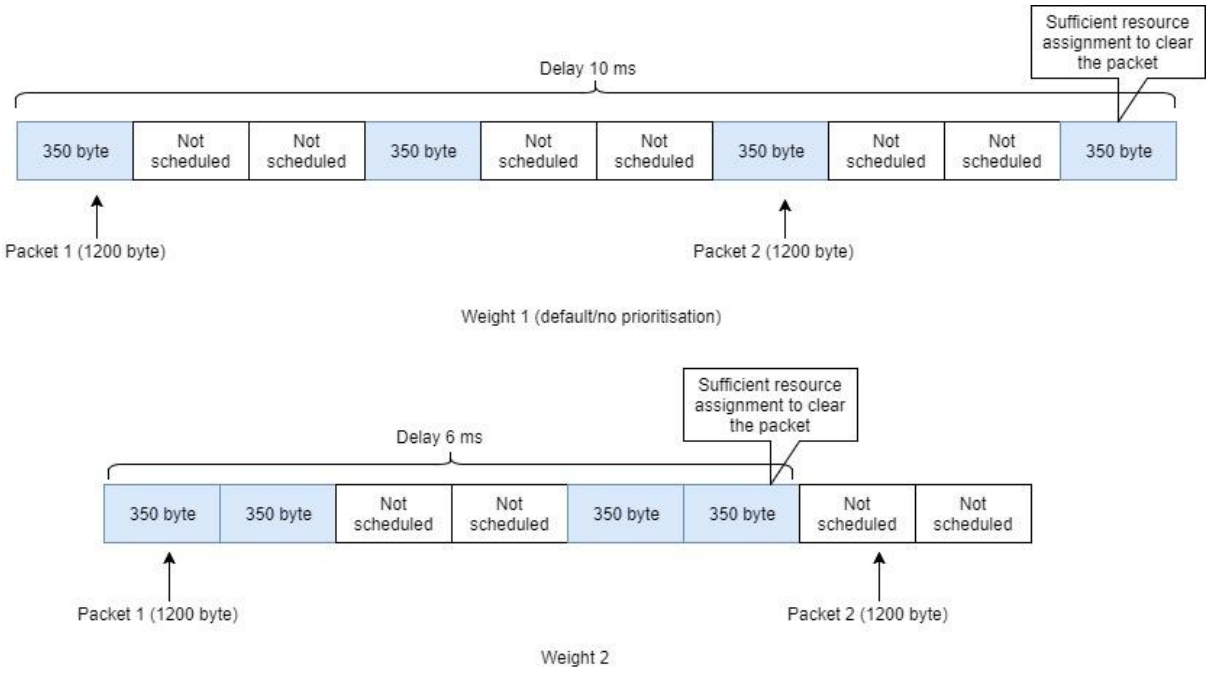

Figure 7-6: Packet latency for prioritisation weight 2

With further increase in the level of prioritisation, a point reaches, where there are sufficiently many TTIs assigned to the tagged user in a given scheduling instant to completely transfer the packet. Therefore, any further increase in the differentiation weight will not have any effect on the packet latency. This reason explains the similar packet latency for the cases of differentiated scheduling with weights 3 and 4.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

In the bar plots, a similar 10^{th} percentile latency for different prioritisation weights suggest that, even for the default weight of 1 with no prioritisation, the periods of favourable channel conditions are favourable enough to cause a low latency to the few packets arriving during those periods. With an increase in the prioritisation weight, the packet latency is significantly reduced for most of the packets resulting in very low 90^{th} latency percentile. As explained above, due to the saturation of maximum latency reduction, even with a weight of 2, higher weights have a similar and low level for 90^{th} latency percentile, consequently causing a similar average latency.

With differentiated scheduling, as the tagged user is prioritised at the expense of resource allocation to the rest of the users, it is important to assess how the other users are affected by differentiation. In a mobile network with differentiated scheduling enabled, care must be taken to ensure that the non-prioritised users are able to use the network service with some decent performance. In order to also quantify this downside of differentiated scheduling, the packet latency distribution for the other user in the network is as shown in Figure 7-7.

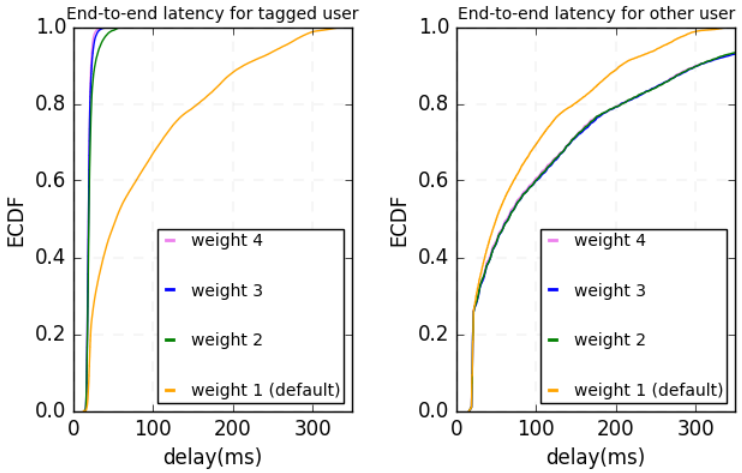

Figure 7-7 : Impact of differentiated scheduling of the tagged user for the packet latency for the other users

The ECDF plot on the left is the end-to-end latency distribution for the tagged user for different weights of differentiation. This is the same plot as shown in Figure 7-5. The plot on the right show the end-to-end packet latency distribution for the other users in the system. All users in the cell are assumed to be in the default scenario conditions with the exception that different prioritisation levels are considered for the tagged user.

For the default scenario with a weight of 1, where the tagged user is not prioritised, radio resources are equally distributed among all the users. Therefore, all users have the same latency experience and hence the same packet latency distribution as indicated by the yellow curve.

With a prioritisation weight of 2, the latency for the tagged user is reduced as explained in the above section, while the packet latency distribution for the other users increases. This is due to the fact that with a weight of 2, other users have to wait longer between subsequent TTI assignments.

With further increase in the weights, as the minimum latency possible for the tagged user is achieved, the resource utilisation by the tagged user is remaining the same as with a prioritisation weight of 2. As the packets are already transmitted from the buffer for the tagged user, there is no need for assigning the tagged user with excess resources, as it would have been for higher weights, if the packets were still not cleared from the buffer. Therefore, these non-assigned radio resources are distributed equally among the other users in the system. Therefore, the packet latency distribution for the other users in the system are identical for higher prioritisation weights for the tagged user.

7.1.6 Impact of Edge Computing on latency in DL

Figure 7-8 shows the impact of Edge Computing on latency and packet drop percentage in DL.

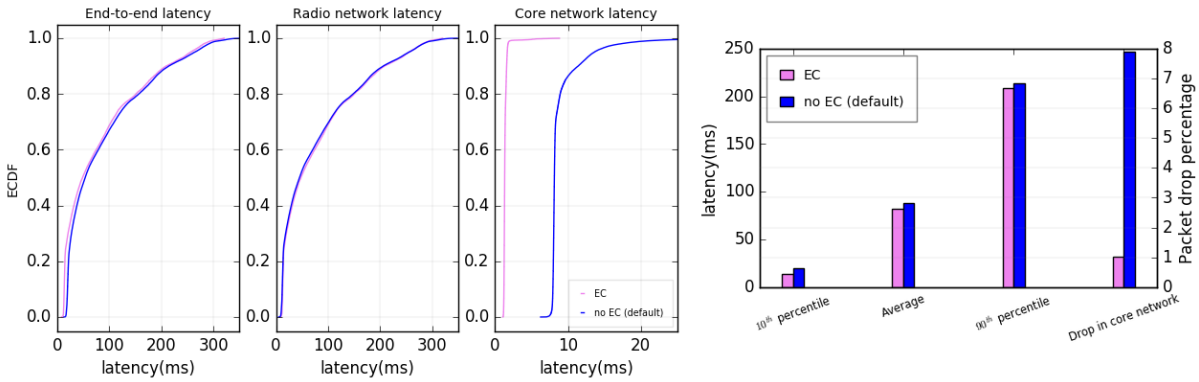

Figure 7-8: Impact of Edge Computing on latency and packet drop percentage in DL

Edge computing is considered as a crucial technology in 5G for enabling applications like Augmented Reality (AR) and Content delivery and caching [49]. In AR, which demand low latency, it is required that the packet delivery must be immediate. In content delivery and caching, frequently accessed contents by the users in the network are cached in the edge computing node, so as to deliver the contents to the users immediately. However, as the edge computing node is typically designed for general purpose processing, where a lot of application services can be run in parallel, it may not be suitable for applications requiring for e.g. Digital Signal Processing (DSP) [50]. So, the presented gains from the Edge Computing setup may not be applicable for all application scenarios.

In EC, the data service is provided within the radio access network without having to be passed through the whole core network as explained in the Section 5.6.

In the EC enabled core networking, there is a clear advantage for the packet latency distribution in the core network. In the EC setup, the user data packets are not experiencing any transport delay of the realistic core network due to the placement of the EC node next to the eNB. As discussed in Section 6.1, the edge computing node is considered congestion-free where there is only the traffic from the tagged user, corresponding to the low load level as in this thesis. Thus, the user packets are having enough resources always, to be processed with very low latency. Therefore, the ECDF curve for the EC setup is a steep curve compared to the default scenario. Moreover, a congestion-free EC node results in a significantly low packet drop percentages in the EC node as shown in the bar plots.

In the *radio network*, there is no advantage compared to the default scenario as EC setup is affecting only the core network latency. Although a significant reduction in the packet latency is achieved in the core network compared to the default scenario, it is not visible in the end-to-end packet latency curve. This is because the reduction in core network latency is insignificant given the different order of magnitude of the radio network latency and hence the end-to-end packet latency.

As the radio network latency is the dominant component in the end-to-end latency for both the cases, a similar trend is observed for both the curves for *end-to-end* latency as the radio network latency.

In the bar charts, it can be observed that the difference in the levels for the latency percentiles and the average latency corresponds to the latency reduction achieved in the core network.

7.1.7 Combined impact of Edge Computing and differentiated scheduling on latency in DL

Figure 7-9 shows the combined impact of Edge Computing and differentiated scheduling on latency and packet drop percentage in DL.

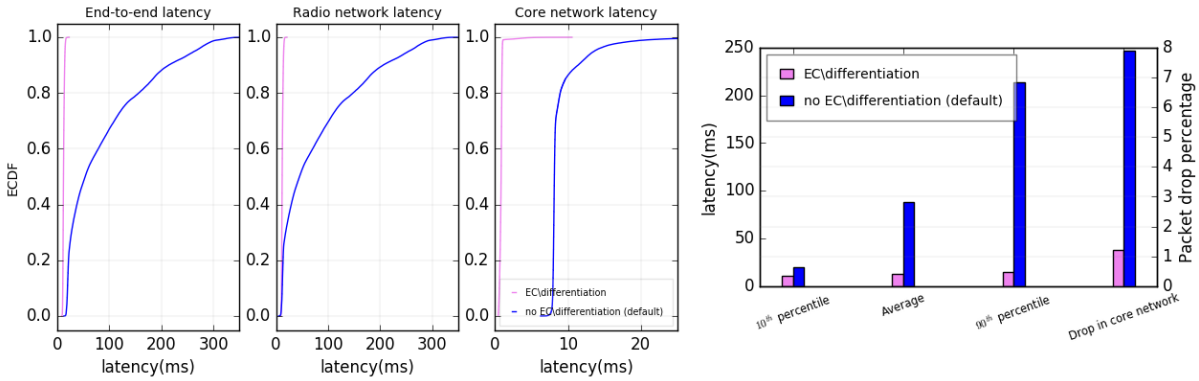

Figure 7-9: Combined impact of Edge Computing and differentiated scheduling on latency and packet drop percentage in DL

In this measurement, the combined impact of EC and the highest level of differentiation is evaluated to investigate the maximum achievable latency reduction in the DL with all other scenario aspects set to their default setting. This result will enable to identify if the considered latency reduction techniques are enough to reach the latency targets of 5G.

Both the plots indicate that there is a significant reduction in latency and packet drop percentage achieved in the DL compared to the default scenario. This is the best latency reduction achievable using the two techniques as both have been evaluated at their most favourable scenario i.e. for differentiation, only the tagged user is prioritised in the cell and for EC, the computing node is considered as congestion-free.

However, the result is still limited by the minimum processing delay for the packets in the radio network for DL. These delays cannot be reduced by any kind of latency reduction techniques unless the radio access technology is completely redesigned to meet the latency targets for 5G.

7.2 Results of measurements in the Uplink

In this section, the results and analysis of the packet latency measurements in the UL are presented. The measurement strategy considered for the measurements are as explained in the Section 6.1, where, to study the effect of a particular scenario aspect, a default scenario is considered and a unilateral deviation of the scenario aspect is considered. The results order are as follows:

- Network load variation
- User distance variation
- Packet size variation
- Packet rate variation
- Differentiation variation
- Edge Computing

7.2.1 Impact of load on latency in UL

Figure 7-10 shows the impact of load on latency and packet drop percentage in UL.

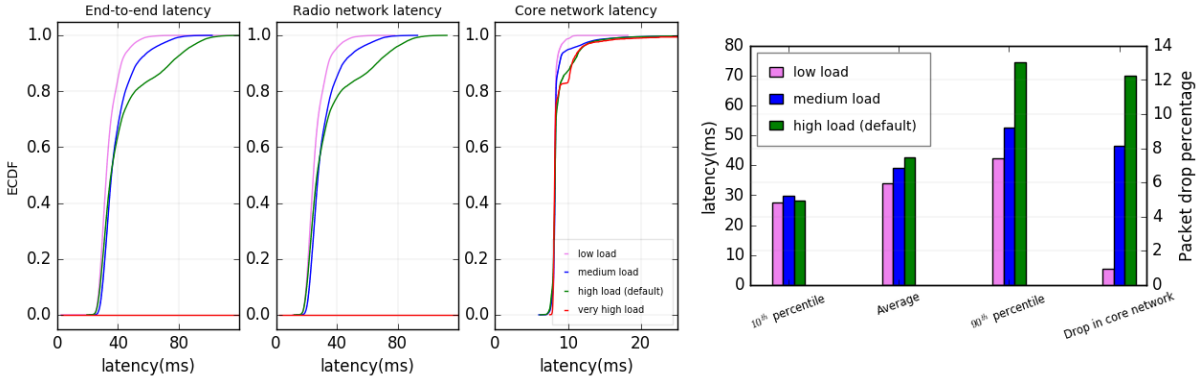

Figure 7-10: Impact of load on latency and packet drop percentage in UL

In the *core network*, results indicate that packet latency is insensitive to the load in the core network for percentile up to the 70th percentile similar to that in the DL.

As explained in the Section 7.1.1, the instantaneous resource availability in the core network for the tagged user decreases with an increase in load level. Therefore, some packets entering the core network during the period when the resources are fully utilised, are buffered and experience higher latencies. This is reflected in the higher latency percentiles. As the load level varies, the buffer occupancy in the core network also varies. Thus, some packets that arrive into the core network during a period when the buffer is fully occupied, are dropped. With an increase in the load level, the chances for the buffer to be fully occupied at an instant, also increases. Therefore, the number of packets dropped in the core network also increases with an increase in the load level. The percentage of packets dropped in the core network at different load levels are shown in the bar graph.

In the *radio network*, the results have the expected trend where the packet latency increases in the load. One important aspect to be noted is that the maximum packet latency observed for the default scenario (high load) is approximately 120 ms. Compared to the DL default scenario where the maximum packet latency is approximately 350 ms, the latency for the default scenario for UL is much lower.

In the comparison of channel quality degradation for the default load level with respect to the channel quality under a low load (no interference), for both UL and DL, it is observed that the degradation for the DL is much larger than the UL. This is because the default load level in the UL and DL are not matching, where in the DL, the default load level considers the interference effect, which is 50% of the maximum possible. However, in the UL, the default load level considers the effect of interference at a level of 3 dB noise floor rise, where theoretically the maximum possible noise floor rise is infinity.

Thus, the significantly high latency in the DL for default load level is due to the fact that the channel degradation is much larger than in the UL, resulting in higher transmission latency and consequently higher scheduling latency for the subsequent packets. Besides this increased channel degradation, the intra-cell *time-domain* channel sharing existing only in the DL in contrast to the UL, also contributes to the transmission latency. Thus, the higher

transmission and scheduling latency causes the overall latency much higher for the DL than in the UL for default load level.

However, if a comparison is performed between the UL and DL latency for the low load levels, where the load levels are matching due to no interference, the results are the other way around. Due to a higher transmit power in the DL and due to no inter-cell interference in the low load level, it will lead to higher SINR and consequently higher bit rates. Besides the higher transmit power in the DL, a user with good channel would be assigned with higher modulation order available only in the DL, resulting in higher bit rates. The combined effects of a much higher bit rate and due to the absence of intra-cell *time-domain* channel sharing, in a low load level, the latency in the DL is significantly lower than that in the UL.

In the UL, as the load level is increased, the increased interference results in lower SINRs and consequently lower channel bit rates. This will cause packets to experience higher transmission latency and consequently higher scheduling latency to the subsequent packets. Moreover, periods of unfavourable channel conditions (multipath fading dips) results in a relatively lower SINR. Thus, packets arriving to the UE buffer during such a period, will experience higher transmission latency. This higher transmission latency for packets will consequently cause higher scheduling latency for the subsequent packets. Moreover, due to a frequency-domain based resource allocation in the UL, a higher load level lead to lesser frequency-domain resource assigned per user. Therefore, as the load level is increased, packets experience higher latencies.

For the extreme case of a very high load level, similar to the case in the DL, the latency curve results in a horizontal line.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

In the bar plots, the periods of relatively favourable channel conditions (multi path fading spikes) that occur equally often across all load levels and likely to determine the lower packet latency percentiles, are favourable enough to causes a few packets arriving during those periods a similar low latency even at a high load level, as compared to a low load level resulting in a similar 10^{th} latency percentile across the different load levels.

At high load level, the periods of relatively unfavourable channel conditions (multi path fading dips) that are likely to determine the higher latency percentiles, causes a major fraction of the total packets, a higher transmission latency and consequently higher scheduling latencies to the subsequent packets. This results in a high 90^{th} latency percentile. However, a few packets arriving during the periods of more favourable channel conditions, experiences lower latencies as explained above, determining the 10^{th} latency percentile. Therefore, a significant difference in the levels between 10^{th} and 90^{th} percentile is observed at higher load levels. This will consequently lead to a significant difference in the average latency.

7.2.2 Impact of user distance on latency in UL

Figure 7-11 shows the impact of user distance on latency and packet drop percentage in UL.

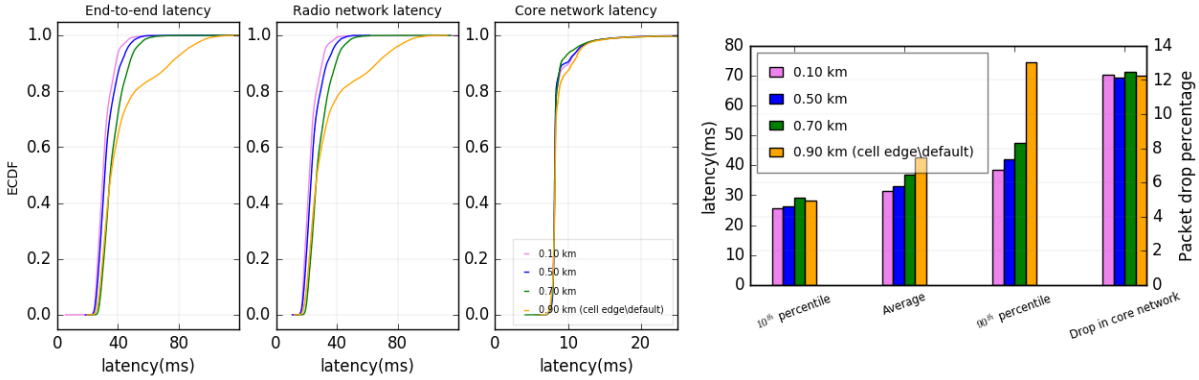

Figure 7-11: Impact of user distance on latency and packet drop percentage in UL

In the *core network*, similar to the case in the DL, the results suggest that packet latency is hardly affected by user distance, in the core network. However, a small variation in the packet latency is noticed for higher latency percentiles in the core network and this is due to the sporadic variation in the resource availability in the core network. As the load level considered in this experiment is kept the same (high load) for all the cases, the packet drop percentage remains almost the same in all cases, as shown in the bar plots.

As the traffic characteristics and the load level are remaining the same for various distances considered, the variation in latency in the radio network is only due to the channel quality variation. As the distance of the user increase from the eNB, the SINR decreases due to an increased path loss, consequently causing lower bit rates. Moreover, during unfavourable channel conditions (multipath fading dips), the bit rates are further reduced due to a further reduction in the SINR. This will cause packets to experience higher transmission latency and consequently higher scheduling latency for the subsequent packets. Therefore, latency increases as the distance is increased.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

In the bar plots, the periods of relatively favourable channel conditions (multi path fading spikes) occurring equally often across all the distances considered for this measurement are favourable enough such that it causes a few packets of even 1200-byte arriving during those periods, a similar low latency as compared to packets of smaller size thus resulting in a similar 10^{th} latency percentile across different packet size.

At far distance, the periods of relatively unfavourable channel conditions (multi path fading dips) that are likely to determine the higher latency percentiles, causes higher transmission latencies and consequently higher scheduling latencies to the subsequent packets. This result in a high 90^{th} latency percentile. However, a few packets arriving during the periods of more favourable channel conditions, experiences lower latencies as explained above, determining the 10^{th} latency percentile. Therefore, a significant difference in the levels between 10^{th} and

90th percentile is observed at far distances. This will consequently lead to a significant difference in the average latency.

7.2.3 Impact of packet size on latency in UL

Figure 7-12 shows the results for the impact of packet size on latency and packet drop percentage in UL.

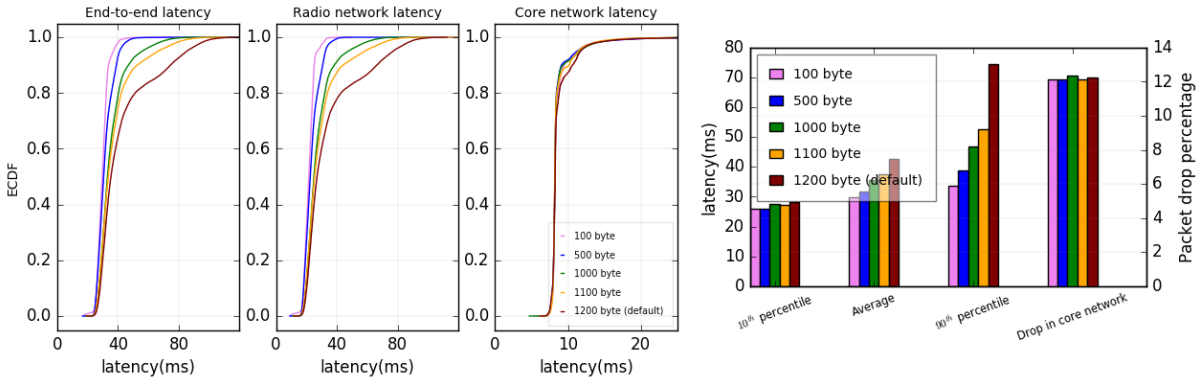

Figure 7-12: Impact of packet size on latency and packet drop percentage in UL

In the *core network*, similar to the DL, the packet latency distribution for packets of different sizes are overlapping, suggesting that core network latency is hardly impacted by packet size. As the load level in the network is set to the default load level of high load for all the cases, the small variation in the packet latency observed for higher latency percentiles in the results is due to the sporadic variation in the resource availability in the core network. Due to this, the packet drop percentage remains almost the same in all cases, as shown in the bar plots.

The results indicate that the packet size affects the *radio network* latency. For a large sized packet, as it requires more resources to be cleared completely from the UE buffer, it will experience higher transmission latency compared to a small packet. Moreover, during the period of multipath fading dips, the larger packets will experience even higher transmission latency due to a lower channel bit rate during that period. This can consequently cause higher scheduling latency to the subsequent packets. Therefore, as packet size is increased, the latencies experienced by the packets also increase.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

In the bar plots, the periods of relatively favourable channel conditions (multi path fading spikes) are favourable enough to cause a similar low latency to a 1200-byte packet, as compared to a small packet.

For large packet size, the periods of relatively unfavourable channel conditions (multi path fading dips) causes a packet of 1200-byte, higher transmission latencies and consequently higher scheduling latencies to the subsequent packets. This results in a high 90th latency percentile. However, a few packets arriving during the periods of more favourable channel conditions, experiences lower latencies as explained above, determining the 10th latency percentile. Therefore, a significant difference in the levels between 10th and 90th percentile is observed at far distances. This will consequently lead to a significant difference in the average latency.

7.2.4 Impact of packet rate on latency in UL

Figure 7-13 shows the results for the impact of packet rate on latency and packet drop percentage in UL.

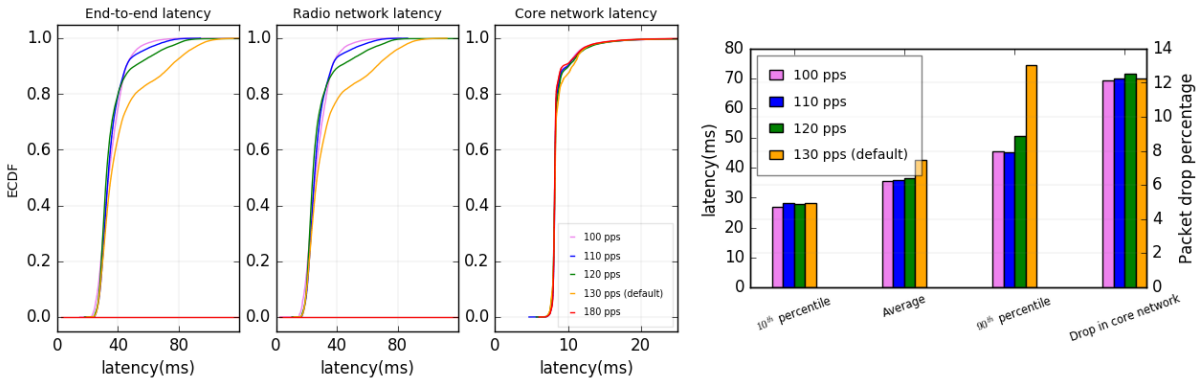

Figure 7-13: Impact of packet rate on latency and packet drop percentage in UL

In the *core network*, similar to the DL, the packet latency distribution for packets of different rate are overlapping suggesting that core network latency is hardly impacted by the packet rate variation of the tagged user. As the load level in the network is considered as the default load level of high load for all the cases, the small variation in the packet latency observed for higher latency percentiles in the results is due to the sporadic variation in the resource availability in the core network. Due to this, the packet drop percentage remains almost the same in all cases, as shown in the bar plots.

In the *radio network*, the packet latency increases with increase in the packet rate. As the channel quality is remaining the same as in the default scenario, this variation in packet latency is due to the fact that when packet rate is increased, the packets arriving into the UE buffer experience increased scheduling latency due to an increased number of packets already in the buffer.

As the average bit rate of the channel bit rate trace for the default load level and default position of the user is 2.2 Mbps, a 1200-byte packet will experience a transmission latency of 5 ms. Moreover, during a period of unfavourable channel conditions (multipath fading dips), due to a lower SINR and consequently a lower bit rate, the transmission latency increases even further. Therefore, an increased packet rate means that for a packet arriving to the UE buffer, there is an increased number of packets already in the UE buffer. Therefore, the arriving packet will experience a higher scheduling latency.

For the extreme case of a packet rate of 180 pps, due to the similar reasons as in the DL, where the latency determined from the python scripts indicate infinite packet latencies, it is represented as a horizontal line.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

In the bar plots, the periods of relatively favourable channel conditions (multi path fading spikes) are favourable enough to cause a similar low latency to a few packets arriving during those periods even at a higher packet rate, as compared to a lower packet rate, resulting in a similar 10^{th} latency percentile across different packet rate.

For high packet rates, the periods of relatively unfavourable channel conditions (multi path fading dips) are such that it causes a higher scheduling latency to the major fraction of the packets arriving during those period, due to an increased number of packets already in the buffer. This results in a high 90th latency percentile. However, a few packets arriving during the periods of more favourable channel conditions, experiences lower latencies as explained above, determining the 10th latency percentile. Therefore, a significant difference in the levels between 10th and 90th percentile is observed at far distances. This will consequently lead to a significant difference in the average latency.

7.2.5 Impact of differentiated scheduling on latency in UL

Figure 7-14 shows the results for the impact of differentiated scheduling on latency and packet drop percentage in UL.

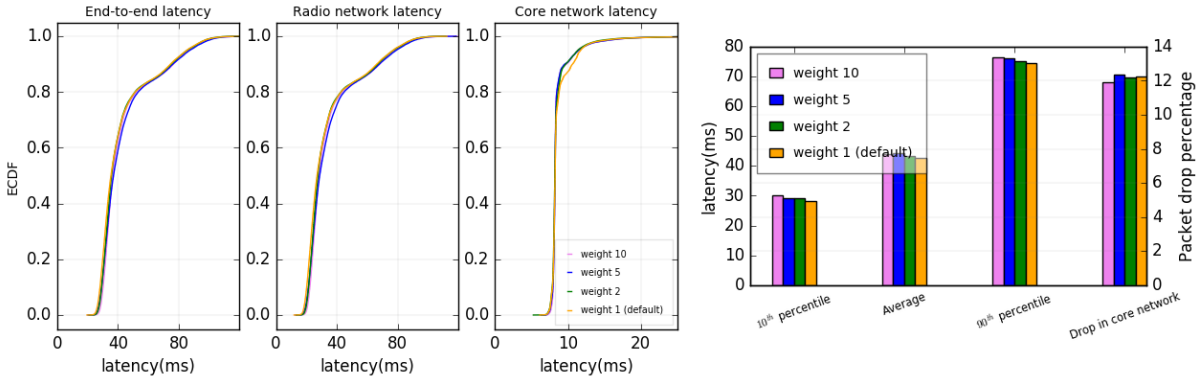

Figure 7-14: Impact of differentiated scheduling on latency and packet drop percentage in UL

As explained in the Section 6.2.1.2, the tagged user is prioritised over the other users for the frequency-domain resource assignment. Only the tagged user is prioritised while the rest of the users are assigned resources with the default priority of weight 1 (meaning no prioritisation).

In the *core network*, packet latency distribution for packets with different levels of differentiation are overlapping each other suggesting that the differentiation in the radio network is not impacting the latency in the core network. As the load level in the network is considered as the default load level of high load for all the cases, the small variation in the packet latency observed in the results is due to the sporadic variation in the resource availability in the core network. Due to this, the packet drop percentage remains almost the same in all cases, as shown in the bar plots

Assigning the tagged user with a weight of 5 might lead one to expect a reduction of the packet latency of the user compared to the default case with a weight of 1. However, such a reduction is not observed. As explained in the Section 6.2.1.2, due to the power limitation in the UE, advantage of assigning more PRBs according to the weight considered may be negligible. What this means is that even if the user is prioritised over other users to be assigned with more PRBs, in reality the extra PRBs may not be used (and hence remain available for other users). Even with a weight assignment of 2, it was observed that in a vast majority of TTIs, the user cannot be sensibly assigned more PRBs than the default case. Thus, the average channel quality for a weight of 2 is remaining the same as for the default case. Therefore, no improvement in packet latency is possible with a weight of 2 and similarly for other higher weights. Therefore, the curves are overlapping. Hence, we see no gain from differentiation in the UL.

As the radio network latency is the dominant component in the end-to-end latency for all the cases, a similar trend is observed for all the curves for *end-to-end* latency as the radio network latency.

As the latency experienced under different prioritisation weights are remaining the same, a similar trend is observed in the bar graphs, where the different latency percentiles and the average latency are similar to the default setting.

Similar to the case of DL, it is important to assess how differentiated scheduling is affecting the latency of the other users in the network. In order to also quantify this downside of differentiated scheduling, the packet latency distribution for the other user in the network is as shown in Figure 7-15.

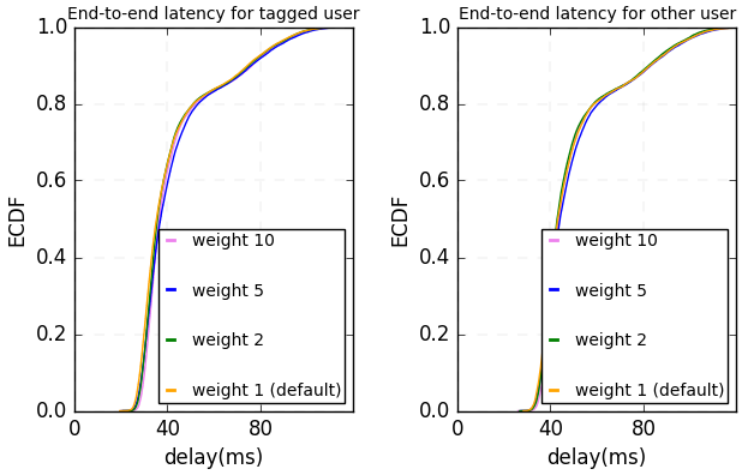

Figure 7-15 : Impact of differentiated scheduling of the tagged user for the packet latency for the other users

The ECDF plot on the left is the end-to-end latency distribution for the tagged user for different weights of differentiation. This is the same plot as shown in Figure 7-14. The plot on the right shows the end-to-end packet latency distribution for the other users in the system. All users in the cell are assumed to be in the default scenario conditions with the exception that different prioritisation levels are considered for the tagged user.

In the default prioritisation weight of 1, where all the users are assigned the resources with an equal priority, the latency distribution for the other users is identical to the tagged user. Moreover, as the resource utilisation of the tagged user is remaining the same as the default setting due to the reasons explained above, the resource availability for the other users in the radio network are also remaining the same as the default setting resulting in an identical latency distribution across the different prioritisation weights.

7.2.6 Impact of Edge Computing on latency in UL

Figure 7-16 shows the impact of Edge Computing on latency and packet drop percentage in UL.

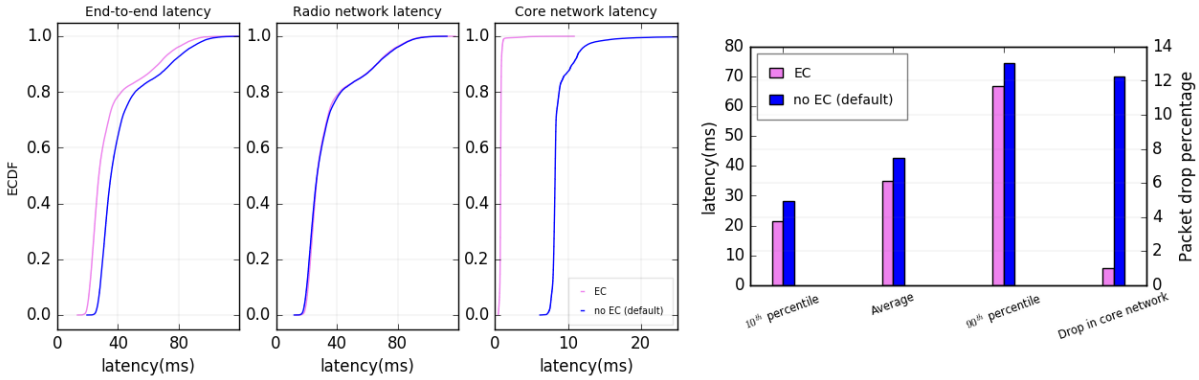

Figure 7-16: Impact of Edge Computing on latency and packet drop percentage in UL

In EC, the data service is provided within the radio access network without having to be passed through the whole core network as explained in the Section 5.6.

Similar to the results for DL EC, there is a significant reduction in latency and packet drop percentage the core network.

In the *radio network*, there is no advantage compared to the default scenario as EC setup is affecting only the core network latency. However, due to a lower order of magnitude in the packet latency compared to that in the DL, the reduction in the core network latency is visible in the end-to-end latency curve also.

As the radio network latency is the dominant component in the end-to-end latency for both the cases, a similar trend is observed for both the curves for *end-to-end* latency as the radio network latency.

In the bar charts, it can be observed that the difference in the levels for the latency percentiles and the average latency corresponds to the latency reduction achieved in the core network.

7.2.7 Combined impact of Edge Computing and differentiated scheduling on latency in UL

Figure 7-17 shows the combined impact of Edge Computing and differentiated scheduling on latency and packet drop percentage in UL.

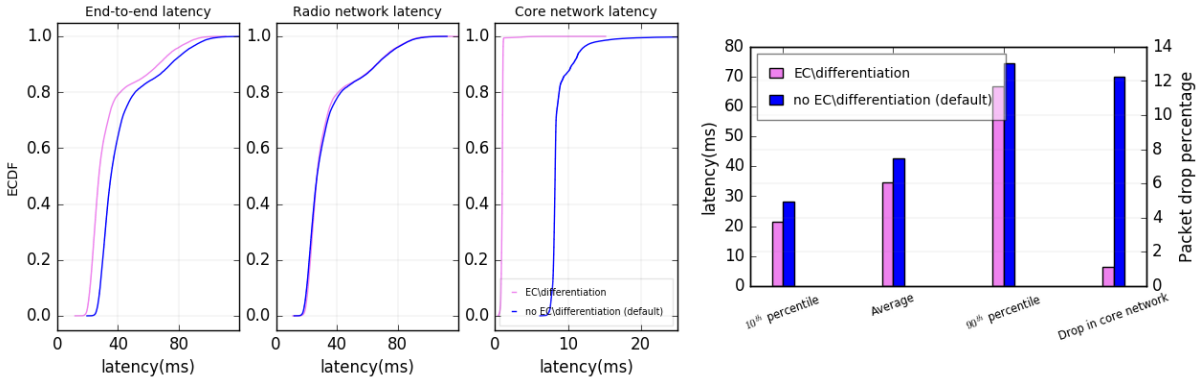

Figure 7-17: Combined impact of Edge Computing and differentiated scheduling on latency and packet drop percentage in UL

In this measurement, the combined impact of EC and the highest level of differentiation is evaluated to investigate the maximum achievable latency reduction in the UL with all other scenario aspects set to their default setting. This is the best latency reduction achievable using the two techniques as both have been evaluated at their most favourable scenario. This result will enable to identify if the considered latency reduction techniques are enough to reach the latency targets of 5G.

Due to the reasons explained for the case of UL differentiation, it is not possible to reduce packet latency in the radio network by various levels of differentiation. However, a reduction in latency in the core network is possible by EC. As the radio network latency is the dominant components in the end-to-end latency, a significant reduction in packet latency is not possible in the UL, using differentiated scheduling and EC combined as compared to the DL.

7.3 Processing delay in the radio network

In this section, results of the processing delays involved in the radio network for UL and DL, due to the different signalling processes depending on the direction of packet transmission, are presented in Figure 7-18. The results presented in Sections 7.1 and 7.2 are already including the processing delay in the determined packet latencies, besides the effects due to the various scenario aspects. Therefore, to identify the processing delay caused in an LTE radio network due to the different signalling processes in the UL and DL, independently without any influence from any of the scenario aspects, measurements are performed on the measurement setup without any realistic effects from the various scenario aspects.

The results show that the average processing delay in the DL is approximately 8 ms and that for UL is approximately 21 ms. As the delays due to the processing in the various layers of the protocol stack is identical for DL and UL, a significant difference in the average values suggest fundamental difference in the signalling procedure in the DL and UL.

In the DL, the average processing delay is very close to the expected value of 7.5 ms as mentioned in the Section 3.3.1.2. However, in the UL, the average processing delay is much higher than the expected value of 16 ms as mentioned in the Section 3.3.1.1. The results indicate that there are packets in the UL, where the processing delay experienced are much higher than the expected 16 ms. This results in a higher average value of 21 ms than the expected value.

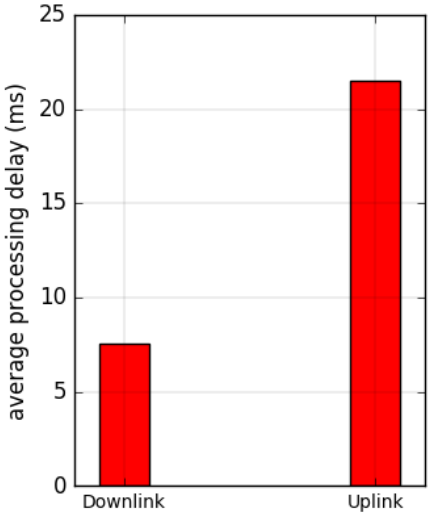

Figure 7-18 : Average processing delay in UL and DL

Chapter 8. Conclusion and future work

In this thesis, a consolidated experimental study on all the potential factors that affect the packet latency in an LTE network has been presented. The results helped to identify the dominant components affecting packet latency. The driver for this thesis work is the ultra-low latency requirements of future services. Such stringent requirements triggered the need for understanding the contributing factors of packet latency in a 4G network. Such a knowledge is crucial because without knowing the causes of latency in 4G, the components or aspects that require a new design in 5G network technology cannot not be identified.

The results of the latency reduction techniques helped to verify whether such techniques alone could achieve the latency targets of a 5G network, as an improvement to a 4G network. As the measurements of the latency reduction techniques were evaluated for the best case i.e. only the tagged user prioritised and a congestion-free edge computing setup, it enables to identify that as the best case is not sufficient to achieve the latency targets for 5G, a redesign of the technology is certainly required.

The outline of this final chapter is as follows. In Section 8.1, the main conclusions from the results of the experiments are derived. In Section 8.2, recommendations for possible extension of this thesis work are provided.

8.1 Conclusion

The results of the measurements done for UL and DL indicate that latency contribution from the radio network is the dominant component in the end-to-end packet latency.

In the *radio network*, the latency is caused by the processing delay due to the different signalling procedure, transmission latency due to varying radio channel bit rates caused by the SINR variations and the scheduling latency due to buffered packets. The processing delay in the UL is much higher than the DL (by about 9 ms) due to more operations to be performed to transmit a packet in the UL as discussed in the Section 3.3.1. As the results in Sections 7.1 and 7.2 indicate, the transmission and scheduling latency is affected by the six scenario aspects.

In the *core network*, the latency is caused by the transport delay, processing latency and the buffering latency. Processing delay in the core network is caused due to the packet processing from the SGW-U and PGW-U functionalities, running inside the VM sw. The buffering latency is caused due to the buffering of some of the packets of the tagged user due to the instantaneous resource unavailability in the core network.

As explained in the Section 7.2.1, a comparison between packet latency in the *radio network* for UL and DL can only be performed for a low load level. In the radio network, due to a higher transmit power along with the possibility for having higher modulation in the DL, it leads to a much higher channel bit rate compared to the UL. This significantly higher bit rates will cause a lower transmission and scheduling latencies to the packets. Combined with the low transmission latency, scheduling latency and significantly lower processing delay, the DL latency is lower than UL. In the core network, the latency experienced is similar for UL and DL.

With an increase in the load in the radio network, the packet latency increases due to a lower PRB assignment per user, while the different loading levels furthermore cause different levels

of interference in the radio network, affecting the radio channel quality. The increased interference due to increased load level causes an increase in the packet latency due to a reduction in SINR and, consequently, bit rates. In the core network, with an increase in the load level, some packets are getting buffered in the core network, causing an increase in packet latency. Also, as the load level is increased in the core network, the packet drop percentage in the core network also increases. As such high packet drop percentage are due to the performance limitations of the test bed used for this thesis, this is not representative of a practical core network platform.

The *position* of the user affects the packet latency in the radio network. As the distance of the user increases, the channel quality degrades due to an increased path loss and a higher inter-cell interference (DL only), consequently causing higher packet latencies. However, depending on the traffic characteristics, as long as the user is close enough so as to have a particular level of channel quality which is high enough, the packet latency remains low. Beyond that distance, the packets experience higher latency. For the UL, the maximum distance beyond which the packet latency increases, as identified from the experiment result is 0.7 km and for the DL it is 0.5 km.

In the radio network, for both the UL and DL, packet latency increases in the *packet size*. However, below a certain packet size, the considered channel bit rate is favourable enough to cause lower packet latency. For larger sizes, the packets experience higher latency. The results indicate that up to a packet size of 500 bytes, the packets experience a lower latency compared to larger packets. In the core network, the packet latency is hardly impacted by the packet size in both the UL and DL. As packet size of only the tagged user is varied, this variation is insignificant to cause any noticeable variation in the buffer utilisation in the core network.

In the radio network, the packet latency increases in the *packet arrival* rate for both the UL and DL. As packet arrival rate is increased, more packets arrive into the UE and eNB buffer, respectively, while there might be already some packets queued in the buffer. This causes higher scheduling latency for the arriving packets, increasing the packet latency. In the core network, the packet latency is hardly impacted by the packet rate in both the UL and DL. As packet rate of only the tagged user is varied, this variation is insignificant to cause any noticeable variation in the buffer utilisation in the core network.

Results indicate that it is possible to achieve a significant latency reduction in radio network for DL using *differentiated scheduling*. However, such a significant reduction is not really effective in the UL due to the fact that the UE power limitation makes that a more generous PRB assignment would lead to reduced SINRs and hence per-PRB bit rate; in many cases this nullifies the potential gain from the higher number of assigned PRBs. The packet latency in the core network is not affected by differentiated scheduling in the radio network.

Results indicate that it is possible to achieve a significant latency reduction in the core network for both UL and DL using *EC*. The EC is not affecting the packet latency in the radio network.

The combined results of the two latency reduction techniques, i.e. differentiated scheduling and EC, indicate that a significant reduction in the end-to-end latency is possible only in the DL. Although a significant reduction is possible in the DL, the reduction is limited by the processing delay in the DL in the radio network. Even if a significant reduction in the end-to-end latency would be achieved as in theory for the UL, the reduction would still be limited by the processing delay in the UL. As the latency targets for 5G are much lower than these bottom limits, a new radio access technology is required for future mobile network.

Although 3GPP has defined the user plane latency target for an LTE network as 5 ms in the radio network, the packet latency observed in an actual network is higher than the specified targets as indicated in the results. As a significant influence is observed on the latency from the various realistic scenarios that can occur in a network, the very low latency target of 5 ms is hardly observed in an actual network. Moreover, the lowest of the processing delay i.e. the processing delay in the DL, experienced by all the packets in the radio network irrespective of the network condition, is by itself higher than the target value by 3 ms.

8.2 Future work

As indicated by the latency impact seen for the scenarios incorporating latency reduction techniques and also considering the processing delays in the UL and DL, the minimum packet latency achievable in a 4G network is still much higher than the latency targets for a 5G network. Thus, to achieve the targets of a 5G network, a completely new radio access technology is required with a focus on latency reduction.

One approach to achieve the low latency target of a 5G network is to have a shorter frame duration. The different approaches to reduce the latency by shortening the frame duration are:

- A shorter frame duration can be achieved with a shorter TTI [51]. TTI length can be shortened by reducing the number of OFDM symbols per TTI. As a shorter TTI is directly proportional to the air interface latency and can be implemented with backward compatibility and usability in the existing LTE band, such a technique can be adopted as an improvement for a 4G network to achieve the latency targets. However, the downside of this approach is the increased control signal overhead [52]. The control signal overhead rise is more significant for a system with lower bandwidth.
- The second approach to shorten the frame duration can be achieved by keeping the same number of OFDM symbols per TTI [53]: by increasing the subcarrier spacing the OFDM symbol duration is shortened. With this shortened OFDM symbol duration, a shorter subframe with the same number of OFDM symbols as in an LTE frame structure can be designed. However, the downside of this approach is a degraded spectral efficiency [54]. A reduced symbol time also need a proportional increase in the cyclic prefix rate resulting in reduced spectral efficiency.

Another approach to reduce the packet latency is by eliminating the need for a scheduling request for every UL packet [19]. Applying Semi-Persistent Scheduling (SPS), the eNB configures the UE to have an UL grant at every time interval, even as frequently as in every TTI. This can enable an UL transmission in consecutive subframe, without the need to send a scheduling request and wait for the corresponding UL grant. The downside of this approach is that the resources are wasted even if there is no data to be transmitted due to the already provided UL grant.

Another approach to reduce the latency is by contention-based PUSCH transmission [19]. In this approach multiple UEs share the same PUSCH resource. Therefore, the UE can transmit in the UL, as soon as some data becomes available in the buffer without having to wait for an UL grant. However, the drawback of this approach is that collisions may happen if two or more UEs sharing the same PUSCH resource perform the PUSCH transmission at the same time resulting in increased latency due to retransmissions.

References

[1] B. Murara, "IMT-2020 Network high level requirements, how African countries can cope," ITU, 2017.

[2] T. K. Sawanobori, "CTIA Everything Wireless," [Online]. Available: https://www.ctia.org/docs/default-source/default-document-library/5g_white-paper_web2.pdf.

[3] "huawei.com," Huawei, [Online].
Available: http://www.huawei.com/minisite/5g/en/touch-internet-5G.html.

[4] P. Lawlor, "VR and AR pushing connectivity limits," Qualcomm, 2017.

[5] "TR 25.913 Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," 3GPP, 2010.

[6] H. Ramazanali, A. Vinel, E. Yavuz and M. Jonsson, "Modeling of LTE DRX in RRC Idle State,"
in *2017 IEEE 22nd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Lund, 2017.

[7] "TS 22.261 Service requirements for the 5G system," 3GPP, 2017.

[8] "M.2134 Guidelines for evaluation of radio interface technologies for IMT-Advanced," ITU-R, 2009.

[9] R. E. Hattachi, NGMN, [Online].
Available: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN_5G_White_Paper_V1_0.pdf.

[10] "ericsson.com," Ericsson AB, [Online]. Available: https://www.ericsson.com/assets/local/publications/white-papers/wp-5g-systems.pdf.

[11] "TR 36.913 Requirements for Further Advancements for E-UTRA," 3GPP, 2010.

[12] "onestore.nokia.com," Nokia, [Online]. Available: https://onestore.nokia.com/asset/200176/Nokia_LTE-Advanced_Pro_White_Paper_EN.pdf.

[13] "M.2083: IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," ITU-R, 2015.

[14] "TR 38.913 Study on Scenarios and Requirements for Next Generation Access Technologies," 3GPP, 2017.

[15] M. Nohrborg, “3gpp.org,” 3GPP, [Online]. Available: http://www.3gpp.org/technologies/keywords-acronyms/98-lte.

[16] F. Firmin, "3gpp.org," [Online]. Available: http://www.3gpp.org/technologies/keywords-acronyms/100-the-evolved-packet-core.

[17] "TS 23.002 Release 12 LTE Network architecture," 3GPP, 2014.

[18] "LTE in a Nutshell: Protocol Architecture," Telesystem Innovations, 2010.

[19] "TR 36.881 Study on latency reduction techniques for LTE," 3GPP, 2016.

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

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

[22] 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.

[23] Y. 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.

[24] 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.

[25] 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.

[26] S. Wang, D. Xu and S. Yan, "Analysis and Application of Wireshark in TCP/IP Protocol Teaching," in *2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT)*, Shenzhen, 2010.

[27] E. Brosh, S. A. Baset, V. Misra, D. Rubenstein and H. Schulzrinne, "The Delay-Friendliness of TCP for Real-Time Traffic," in *IEEE/ACM Transactions on Networking*, 2010.

[28] P. Srivats, "ostinato.org," [Online]. Available: https://ostinato.org/.

[29] "spirent.com," [Online]. Available: https://www.spirent.com/-/media/Datasheets/Broadband/PAB/SpirentTestCenter/Spirent_C50_WLAN_Wave-2_Appliance_datasheet.pdf.

[30] P. Srivats, "ostinato.org," [Online]. Available: https://userguide.ostinato.org/Architecture.html.

[31] J. Garcia, P. Hurtig, "KauNetEm: Deterministic Network Emulation in Linux," in *Proceedings of NetDev 1.1: The Technical Conference on Linux Networking*, Seville, 2016.

[32] A. Jurgelionis, J. P. Laulajainen, M. Hirvonen and A. I. Wang, "An Empirical Study of NetEm Network Emulation Functionalities," in *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, Maui, 2011.

[33] "apwpt.org," Motorola, [Online]. Available: https://www.apwpt.org/downloads/realistic_lte_experience_wp_motorola_aug2009.pdf.

[34] J. M. Molina-Garcia-Pardo, M. Lienard, A. Nasr and P. Degauque, "Wideband Analysis of Large Scale and Small Scale Fading in Tunnels," in *2008 8th International Conference on ITS Telecommunications*, Phuket, 2008.

[35] R. Lacoste, "Multipath fading effects on short range indoor RF links," ALCIOM, 2010.

[36] P. Chang, Y. Chang, Y. Han, C. Zhang, D. Yang, "Interference Analysis and Performance Evaluation for LTE TDD System," in *2010 2nd International Conference on Advanced Computer Control*, Shenyang, 2010.

[37] Y. Okumura, E. Ohmori, T. Kawano, K. Fukuda, Field Strength and Its Variability in VHF and UHF Land-Mobile Radio Service, vol. 16, 1968, pp. 825-873.

[38] "M.2135-1 Guidelines for evaluation of radio interface technologies for IMTAdvanced," ITU-R, 2009.

[39] U. Toseef, C. Goerg, T. Weerawardane and A. Timm-Giel, "Performance comparison of PDCP buffer management schemes in LTE system," in *2011 IFIP Wireless Days (WD)*, Niagara Falls, 2011.

[40] "open5gcore.org," Fraunhofer, [Online]. Available: https://www.open5gcore.org/.

[41] S. Mohan, R. Kapoor, and B. Mohanty, "Latency in HSPA Data Networks," Qualcomm, 2011.

[42] "ngmn.org," NGMN, [Online]. Available: https://www.ngmn.org/fileadmin/user_upload/NGMN_Optimised_Backhaul_Requirements.pdf.

[43] "Technical White Paper on Virtualization - Huawei Enterprise," HUAWEI TECHNOLOGIES CO., LTD., 2014.

[44] C. Mancaş, "Performance improvement through virtualisation," in *2015 14th RoEduNet International Conference - Networking in Education and Research (RoEduNet NER)*, Craiova, 2015.

[45] "libvirt.org," [Online]. Available: https://libvirt.org/formatnetwork.html#examplesRoute.

[46] "docs.vmware.com," [Online]. Available: https://docs.vmware.com/en/VMware-Workstation-Pro/12.0/workstation-pro-12-user-guide.pdf.

[47] "cisco.com," Cisco, [Online]. Available: https://www.cisco.com/c/en/us/support/docs/availability/high-availability/19643-ntpm.html.

[48] M. Patel, Y. Hu, P. Hédé, "etsi.org," [Online]. Available: https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing_-_introductory_technical_white_paper_v1%2018-09-14.pdf.

[49] N. Abbas, Y. Zhang, A. Taherkordi and T. Skeie, "Mobile Edge Computing: A Survey," *IEEE Internet of Things Journal,* vol. 99, pp. 1-1, 2017.

[50] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick and D. S. Nikolopoulos, "Challenges and Opportunities in Edge Computing," in *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, New York, 2016.

[51] E. Shin and G. Jo, "Uplink Frame structure of Short TTI system," in *2017 19th International Conference on Advanced Communication Technology (ICACT)*, Bongpyeong, 2017.

[52] X. Zhang, "Latency reduction with short processing time and short TTI length," in *2017 International Symposium on Intelligent Signal Processing and Communication Systems*, Xiamen, 2017.

[53] P. Guan et al, "Ultra-Low Latency for 5G - A Lab Trial," in *Proc. IEEE PIMRC*, 2016.

[54] Z. E. Ankarali, B. Peköz and H. Arslan, "Flexible Radio Access Beyond 5G: A Future Projection on Waveform, Numerology, and Frame Design Principles," *IEEE Access,* vol. 5, pp. 18295-18309, 2017.

Abbreviations

3GPP	:	3rd generation partnership project
4G	:	4th generation
5G	:	5th generation
5GPPP	:	5G infrastructure public private partnership
API	:	Application programming interface
AR	:	Augmented reality
AS	:	Application server
BLER	:	Block error rate
CMTC	:	Critical machine type communication
CPU	:	Central processing unit
DL	:	Downlink
DNS	:	Domain name server
DSP	:	Digital signal processing
DUT	:	Device under test
E2E	:	End to end
EC	:	Edge computing
ECDF	:	Empirical cumulative distribution function
EDGE	:	Enhanced data rates for GSM evolution
eMBB	:	Enhanced mobile broadband
eNB	:	Evolved NodeB
EPC	:	Evolved packet core
EPS	:	Evolved packet system
E-UTRA	:	Evolved universal terrestrial radio access
E-UTRAN	:	Evolved UMTS terrestrial radio access network
FDD	:	Frequency division duplex
FDMA	:	Frequency division multiple access
GPRS	:	General packet radio service
GSM	:	Global system for mobile communication
GUI	:	Graphical user interface
GW	:	Gateway
HARQ	:	Hybrid automatic-repeat-request
HSS	:	Home subscriber server
IEEE	:	Institute of electrical and electronics engineers
IMT	:	International mobile telecommunication
IoT	:	Internet of things
IP	:	Internet protocol
ITU	:	International telecommunication union
KPI	:	Key performance indicator
KVM	:	Kernel based virtual machine
kHz	:	Kilo hertz
LAN	:	Local area network
LTE	:	Long term evolution
MAC	:	Medium access control
MCS	:	Modulation and coding scheme

MHz	:	Mega hertz
MIMO	:	Multiple input multiple output
MME	:	Mobility management entity
MMTC	:	Massive machine type communication
MTP	:	Motion to photon latency
MTU	:	Maximum transfer unit
NAT	:	Network address translation
NFV	:	Network function virtualisation
NGMN	:	Next generation mobile networks
NTP	:	Network time protocol
OFDM	:	Orthogonal frequency division multiplexing
OFDMA	:	Orthogonal frequency division multiple access
OS	:	Operating system
OWD	:	One-way delay
PDCP	:	Packet data convergence protocol layer
PDN	:	Packet data network
PDU	:	Packet data unit
P-GW	:	PDN gateway
PRB	:	Physical resource block
PUCCH	:	Physical uplink control channel
PUSCH	:	Physical uplink shared channel
QAM	:	Quadrature amplitude modulation
QoS	:	Quality of service
QPSK	:	Quadrature phase shift keying
RAN	:	Radio access network
RAT	:	Radio access technology
RLC	:	Radio link control
RRC	:	Radio resource control
RTT	:	Round trip time
SCDMA	:	Synchronous code division multiple access
SC-FDMA	:	Single carrier frequency division multiple access
SDN	:	Software defined networking
S-GW	:	Serving gateway
SIM	:	Subscriber identity module
SINR	:	Signal to interference plus noise ratio
SM	:	Spatial multiplexing
SMS	:	Short message service
SPS	:	Semi persistent scheduling
SR	:	Service request
TCP	:	Transmission control protocol
TDD	:	Time division duplex
TDMA	:	Time division multiple access
TPC	:	Transmit power control
TTI	:	Transmission time interval
UDP	:	User datagram protocol
UE	:	User equipment

UL	:	Uplink
UMTS	:	Universal mobile telecommunications system
URLLC	:	Ultra-reliable low latency communications
USB	:	Universal serial bus
VM	:	Virtual machine
VNF	:	Virtual network functions
VR	:	Virtual reality