Analysis of Strategies for Minimising End-to-End Latency in 4G and 5G Networks

Afonso Carvalho
Luís M. Correia & António Grilo
Instituto Superior Técnico / INESC-ID
Lisbon, Portugal
afonsolxp@gmail.com
luis.m.correia@tecnico.ulisbo.pt

Ricardo Dinis
NOS SGPS
Lisbon, Portugal
ricardo.dinis@nos.pt

Abstract— The main purpose of this thesis is to identify and study a variety of strategies that effectively reduce the end-to-end latency in both 4G and 5G networks. This latency reduction will allow the operators to provide URLLC services to the users, such as: remote surgeries, the Intelligent Transport Systems and factory automation services. To verify if those services can be implemented using the 4G and the 5G systems, the developed model considers several variables: the MEC node deployment option, the functionality splitting options, the radio techniques, and the network architectures. The MEC technology appears in the thesis as the solution that allows the end-to-end latency values to reach values below 1 ms, which are required for some of the URLLC services. The results obtained show that the 4G system does not have enough capacity to allow the existence of the upcoming services. Even with the MEC node deployment that minimizes the latency, the LTE network is not able to provide the URLLC services under study. The simulations show that using the adequate latency reduction strategies and radio techniques, the 5G system has enough capacity and sufficiently low latencies to provide the upcoming services.

Keywords- 5G, 4G, Cloud Architecture, Latency, MEC.

I. INTRODUCTION

Approximately four decades ago, the 1st generation of mobile communications emerged. Since then, there was a prominent evolution in this area that allows the users to perform more complex activities. In the beginning, users were able to perform voice calls, but in the modern era of the mobile communication, they can perform videoconferences, watch high-definition videos, access the internet almost everywhere, and stay constantly connected with other people using social media, amongst other activities. All these capabilities demand a high data traffic, and it is expected that the quantity of this traffic will continue to increase.

The existence of these capabilities (videoconferences and high-definition videos, amongst many others) became possible along with the 4th generation (4G) of mobile communication systems. The popularization of this technology led to the increase of the data consumption every year, mainly because of the video and music streaming popularity. Consequently, the existing spectrum bands are becoming congested, leading to breakdowns in service, particularly when lots of people in the same area try to access online mobile services at the same time. The 5th generation of mobile communication systems (5G) appears as the solution to these limitations with 3 main objectives: increase the data rates, reduce the latency, and provide higher capacity. The NR lowest end-to-end (E2E) planned latency is 1 ms (extracted from [1]), which is more than 10 times lower than the previous implemented system.

Latency is measured as the delay in the packet transmission, propagation, queuing and processing, and it is one of the important parameters to have into account in mobile communications. Typically, the network latency is measured from the device, up to the radio, down into the radius of the baseband processor, back out into the core, and then to the data centre itself. In recent years, 5G has been tied to the Edge Computing (EC) architecture, since this technology might be the solution for the network revolution and the unprecedented technical requirements specified in [1]. Basically, EC brings the service infrastructure to the edge of the network, where the "edge" can be defined as an arbitrary location along the path in between the service user and the service host (traditionally located in a remote data centre). The main goal behind this approach is to reduce the physical and the logical distance that separates both ends of the service path, thereby reducing the E2E latency.

The paper is organized as follows. Section II presents the state of the art. Section III presents the model development, starting with the model parameters, following with the specification of the network scenarios, the list of the service requirements, the network latency, link throughputs and finally the model implementation. Section IV contains the results analysis, with the description of the scenarios, followed by the impact of different studied variables on the E2E latency and distance. In Section V, the most important conclusions of this work are presented.

II. STATE OF THE ART

In this subsection it is presented the information given by the most current research about the 5G and 4G systems regarding the topic of the thesis. The main topics of the presented research are the latency related subjects. This research is mainly based on academic works, corporations involved in mobile communications (such as Qualcomm), and other entities like 3GPP.

The virtualization process is one of the evolutions that take responsibility for the network latency reduction since it takes an important role in the SDNs and VNF. Both these virtualization-based technologies are revolutionary because they influence the Cloud and Edge Network deployments, which are two of the studied strategies for minimizing E2E latency in both the 4G and 5G networks in this thesis.

The existence of many contributions to the delay is inconvenient and makes the systems unsuitable for real-time applications. To cope with the delay problem, a new emerging concept, known as Multi-Access Edge Computing (MEC), has been introduced. The Edge Network approach is a computer paradigm that provides computer and storage resources by bringing the network closer to the end users, according to [2]. This process will reduce the latency by shortening the path in between the user and the network resources that the signal needs to reach, by bringing them to the edge of the network. It is possible to implement the Edge nodes in different parts of the network, and each one of the possible deployments has different impacts in terms of the latency according to [3]. In the 4G architecture, the MEC node can be physically installed in between the RRH (Remote Radio Header) and the BBU (Baseband Unit), in between the BBU and the Core or in between the Core and the data centre. From the latency view point the first referred deployment is the one that can minimize the latency, although the other 2 possible deployments can also have impacts on the latency reduction.

The 5G network architecture present in [4] is the one proposed by 3GPP, in which instead of having the separation of the BS functions by the RRH and the BBU, the BBU functionalities will also be split by 3 nodes, to provide higher flexibility and scalability. Therefore, the nodes of the 5G-RAN architecture and their respective functionalities are defined as the following: the RU that keeps the functionalities of the RRH, and the BBU functionalities are split by the RU (Radio Unit), DU (Distributed Unit) and the CU (Centralized Unit). From the latency point-of-view, the evolution of the proposed architecture is beneficial because the MEC node deployment becomes more flexible.

There are 4 main contributions to the E2E latency given by [5] which are: the transmission delay, the propagation delay, the processing delay, and the queuing delay.

The radio interfaces of the mobile communication systems had a fast evolution over the last years, and it is expected that they continue to improve to achieve ultra-low latencies. The evolution of these radio interfaces is shown by [5].

III. Model Development

A. *Model Overview*

One of the objectives of this thesis is to create a model that can adapt itself to the characteristics of the network, the profile of the users, and the services provided to the users, in order to calculate the E2E latency and provide an analysis of several solutions to reduce this parameter in the 4G and 5G networks. The scheme of model overview, including the input parameters, the intermediate calculations that occur during the program execution, and output parameters are represented in Figure 1.

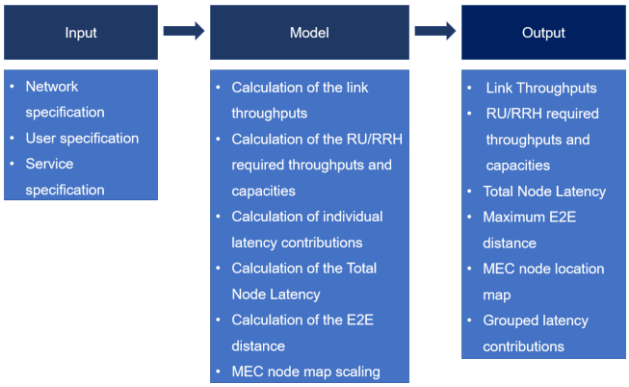

Figure 1. Model Overview.

The calculations that are performed during the program execution are based on the input parameters. The link throughputs are calculated based on the input splitting options present in the network specification parameters. The RU/RRH required throughputs are calculated based on the number of active users connected to the radio node and the data rate required to provide the user services with quality. The individual latency contributions are calculated based on the packet size, the data rates, and the simulated traffic in the network nodes (present in the network, service, and user specification respectively), and the sum of all these latency contributions (except the propagation latency) is the Total Node Latency. It is possible to calculate the maximum E2E distance with the maximum allowed E2E latency for the service, since it depends on the margin in between the maximum E2E latency and the Total Node latency, because the latency that is not accumulated in the nodes corresponds to a certain link length (and consequently an associated maximum propagation latency). The maximum E2E distances are represented in a map, as one of the program outputs.

B. *Network Architecture*

As a general approach, the 5G network contains a Fronthaul, a Middlehaul and a Backhaul, but not every network will have the latency contribution of these 3 links. The CU, the RU and the DU can be installed in a collocated way in different combinations, and these approaches have different impacts in terms of the latency. The 4G C-RAN network has a standardized architecture, and the nodes are not collocated, since the network is already installed. 3GPP has identified 4 possible network deployment scenarios for the 5G C-RAN:

- Independent RU, DU and CU deployment, which is the scenario where the Fronthaul, the Middlehaul and the Backhaul links are present in the network.
- Collocated DU and CU and independent RU deployment, in which the CU and DU are located together, and consequently the Middlehaul link is not present in the architecture.
- Collocated RU and DU and independent CU deployment, in which the RU and the DU are located together. In this architecture the distance in between the RU and the DU is in the order of the hundreds of

meters, which reduces the propagation latency and the cost since they can be connected through optical fibre and no transport equipment is needed.

- Collocated RU, DU and CU in which the 3 nodes are located together. This structure may be used for small cell and hot-spot scenarios, and in this case the network only has a Backhaul link.

The MEC node deployment options for both 4G and 5G systems are present in Figure 2.

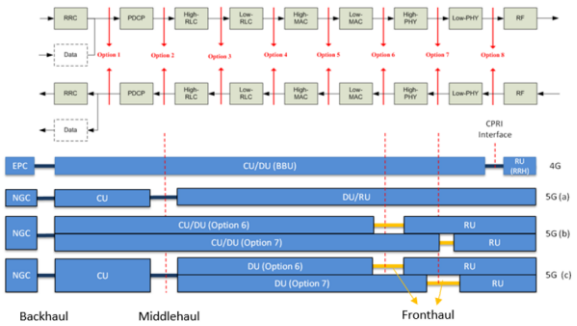

Figure 2. Network Architectures (extracted from [6]).

The 4G MEC node deployment options are present in Figure 3.

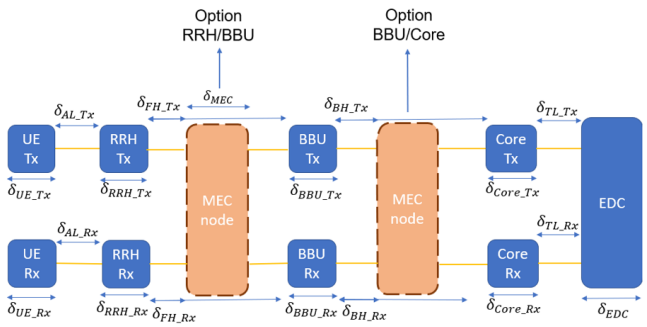

Figure 3. 4G MEC deployment options.

For the 4G system scenario there are 2 possible MEC node deployment options studied:

- The installation of the MEC node in between the RRH and the BBU (Option RRH/BBU).
- The installation of the MEC node in between the BBU and the Core of the Network (Option BBU/Core).

The 5G MEC node deployment options are present in Figure 4.

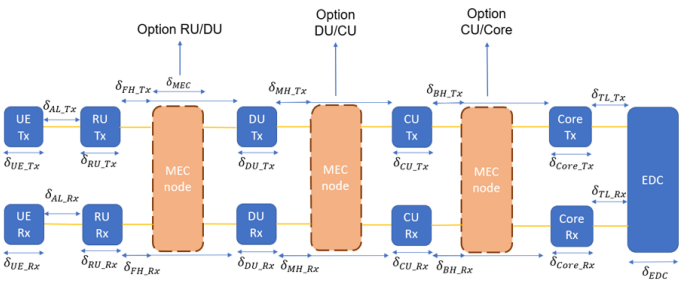

Figure 4. 5G MEC deployment options.

For the 5G system scenario there are 3 possible MEC node deployment options studied:

- The installation of the MEC node in between the RU and the DU (Option RU/DU).
- The installation of the MEC node in between the DU and the CU (Option DU/CU).
- The installation of the MEC node in between the CU and the Network Core (Option CU/Core).

C. Services specification

The traffic demand on each node is present in the user specification input parameters, which are necessary for the computation of the queuing delay in each node and the RU/RRH required throughputs.

Table 1. List of the service requirements (adapted from [7],[8] and [9]).

Service	Service Class	Latency [ms]	Data Rate [Mbps]	Packet Size [B]	Priority
Int. Cont. Manip.	Conv. URLLC	1	0.512	20	1
Vid. Str.		100	10	188	
Hap. Feed.		3	0.400	20	
Ext. Cont. Manip.		3	0.512	20	
Traffic Info.	Conv. URLLC eMBB	100	2	300	2
Rem. Driving		5	25	1600	
Net. Bas. Sen. Shar.		3	20	1000	
Mach. Tools	Conv. URLLC	0.25	1	10	2
Print. Mach.		1	1	30	
Pack. Mach.		2.5	1	15	
Aug. Real.	Stream. URLLC	15	600	650	4
Voice	Conv.	100	0.032	218	3
Video Conf.	Conv.	150	2	800	5
Web Brow.	Interac.	300	0.5	512	9
Email	Back.	300	0.512	128	12
Soc. Net.	Interac.	300	2	1000	8
File Transf.	Interac.	300	1	4096	10

D. Network Latency

The latency contributions present in Figures 3 and 4 are calculated using the following equations:

$$\delta_{UE\_Tx} = \delta_{UE\_Proc} + \delta_{UE\_Trans} \quad (3.1)$$

$$\delta_{UE\_Rx} = \delta_{UE\_Proc} \quad (3.2)$$

$$\delta_{RU\_Tx} = \delta_{RU\_Rx} = \delta_{RU\_Proc} + \delta_{RU\_Queu} + \delta_{RU\_Trans} \quad (3.3)$$

$$\delta_{RRH\_Tx} = \delta_{RRH\_Rx} = \delta_{RRH\_Proc} + \delta_{RRH\_Queu} + \delta_{RRH\_Trans} \quad (3.4)$$

$$\delta_{DU\_Tx} = \delta_{DU\_Rx} = \delta_{DU\_Proc} + \delta_{DU\_Queu} + \delta_{DU\_Trans} \quad (3.5)$$

$$\delta_{CU\_Tx} = \delta_{CU\_Rx} = \delta_{CU\_Proc} + \delta_{CU\_Queu} + \delta_{CU\_Trans} \quad (3.6)$$

$$\delta_{BBU\_Tx} = \delta_{BBU\_Rx} = \delta_{BBU\_Proc} + \delta_{BBU\_Queu} + \delta_{BBU\_Trans} \quad (3.7)$$

$$\delta_{Core\_Tx} = \delta_{Core\_Rx} = \delta_{Core\_Proc} + \delta_{Core\_Trans} \quad (3.8)$$

$$\delta_{EDC} = \delta_{EDC\_Proc} + \delta_{EDC\_Trans} \quad (3.9)$$

$$\delta_{MEC} = \delta_{MEC\_Proc} + \delta_{MEC\_Trans} \quad (3.10)$$

The formula used in the thesis to calculate the transmission latency is:

$$\delta_{Trans\,[\mathrm{ms}]} = \frac{8\,D_{[\mathrm{Bytes}]}}{R_{[\mathrm{Gbits/s}]}} 10^{-6} \quad (3.11)$$

where:

- $D_{[\mathrm{Bytes}]}$ – Packet size in bytes.
- $R_{[\mathrm{Gbits/s}]}$ – Data rate/throughput provided by the link.

Table 2. 4G and 5G UE processing delay ratios (extracted from [10]).

	4G LTE	5G NR			
Subcarrier Spacing (kHz)	15	15	30	60	120
ρ_{UE}	$\frac{21}{14}$	$\frac{2}{14}$		$\frac{3}{14}$	$\frac{4}{14}$

The processing delay in the UE is given by (adapted from [10]):

$$\delta_{UE\_Proc} = \delta_{UE\_Trans}\,\rho_{UE} \quad (3.12)$$

The processing delay in the RRH is given by (adapted from [DMAG18]):

$$\delta_{RRH\_Proc} = 1.5 \cdot \delta_{UE\_Trans}\,\rho_{RU} \quad (3.13)$$

where:

- ρ_{RU} – the ratio of functionalities attributed to the radio node in the splitting option 8.

The URLLC adaptation parameters for each service are present in Table 3.

Table 3. URLLC adaptation parameter (ρ_{lat}).

Service	**Service Number**	$\boldsymbol{\rho_{lat}}$
Internal Control Manipulations	1	$\frac{1}{100}$
Video Streaming	2	1
Haptic Feedback	3	$\frac{3}{100}$
External Control Manipulations	4	$\frac{3}{100}$
Traffic Information	5	1
Remote Driving	6	$\frac{1}{20}$
Network Based Sensor Sharing	7	$\frac{3}{100}$
Machine Tools	8	$\frac{1}{10}$
Printing Machines	8	$\frac{4}{10}$
Packaging Machines	10	1
Augmented Reality	11	1
Voice	12	1
Video Conference	13	1
Web Browsing	14	1
Email	15	1
Social Networking	16	1
File Transfer	17	1

The node processing latency ratios are present in Table 4.

Table 4. RU, DU and CU processing latency ratios (adapted from [11]).

	Fronthaul Splitting Option				
	8	7.3	7.2	7.1	6
ρ_{RU}	1	$\frac{25}{11}$	$\frac{19}{11}$	$\frac{15}{11}$	3
ρ_{DU}	6	$\frac{52}{11}$	$\frac{58}{11}$	$\frac{62}{11}$	4
ρ_{CU}	2	2	2	2	2

The processing delay in the BBU is given by:

$$\delta_{BBU\_Proc} = \delta_{RRH\_Proc}\,(\rho_{CU} + \rho_{DU})\,\rho_{lat} \quad (3.14)$$

where:

- ρ_{DU} – the ratio of functionalities attributed to the DU.
- ρ_{CU} – the ratio of functionalities attributed to the CU.
- ρ_{lat} – the parameter used to adapt the processing resources to the latency.

The processing delay in the RU is given by (adapted from [10]):

$$\delta_{RU\_Proc} = \delta_{UE\_Trans}\,\rho_{RU}\,\rho_{lat} \quad (3.15)$$

The processing delay in the DU is given by:

$$\delta_{DU\_Proc} = \delta_{UE\_Trans}\,\rho_{DU}\,\rho_{lat} \quad (3.16)$$

The processing delay in the CU is given by:

$$\delta_{CU\_Proc} = \delta_{UE\_Trans}\,\rho_{CU}\,\rho_{lat} \quad (3.17)$$

The processing delay in the Core is given by (adapted from [12]):

$$\delta_{Core\_Proc\,[\mathrm{ms}]} = \frac{4}{2385} D_{[\mathrm{Bytes}]} + \frac{469}{477} \quad (3.18)$$

The processing delay in the MEC and the data centre is given by (adapted from [12]):

$$\delta_{MEC\_Proc\,[\mathrm{ms}]} = 4 \cdot 10^{-5} \cdot D_{[\mathrm{Bytes}]} \cdot \rho_{func} \quad (3.19)$$

$$\delta_{EDC\_Proc\,[\mathrm{ms}]} = 1.33 \cdot 10^{-5} \cdot D_{[\mathrm{Bytes}]} \quad (3.20)$$

where:

- ρ_{func} - number of functionalities that the MEC node needs to execute.

The queuing latency is calculated using the formula:

$$\delta_{Queu\,[\mathrm{ms}]} = 10^3 \sum_{p=1}^{M_{Pserv}} \frac{8\,D_{serv,p\,[\mathrm{Bytes}]}}{R_{max\,[\mathrm{bps}]}} \quad (3.21)$$

where:

- $D_{serv,P}$ – Packet size in bytes for a specific service with priority p.
- R_{max} - Maximum throughput offered by the following link.
- M_{Pserv} – Number of users connected to the node, using services with a higher or equal priority than the studied user.

For the 4G scenario with the BBU/Core MEC node deployment option, the Total Node latency is described by the equation:

$$\delta_{Tot\_Node\,[\mathrm{ms}]} = \delta_{UE\_Tx} + \delta_{AL\_Tx} + \delta_{RRH\_Tx} + \delta_{BBU\_Tx} + \delta_{MEC} + \delta_{BBU\_Rx} + \delta_{RRH\_Rx} + \delta_{AL\_Rx} + \delta_{UE\_Rx} \quad (3.22)$$

For the 4G scenario with the RRH/BBU MEC node deployment option, the Total Node latency is described by the equation:

$$\delta_{Tot\_Node\,[\mathrm{ms}]} = \delta_{UE\_Tx} + \delta_{AL\_Tx} + \delta_{RRH\_Tx} + \delta_{MEC} + \delta_{RRH\_Rx} + \delta_{AL\_Rx} + \delta_{UE\_Rx} \quad (3.23)$$

For the 5G scenario with the CU/Core MEC node deployment option, the Total Node latency is described by the equation:

$$\delta_{Tot\_Node\,[\mathrm{ms}]} = \delta_{UE\_Tx} + \delta_{AL\_Tx} + \delta_{RU\_Tx} + \delta_{DU\_Tx} + \delta_{CU\_Tx} + \delta_{MEC} + \delta_{CU\_Rx} + \delta_{DU\_Rx} + \delta_{RU\_Rx} + \delta_{AL\_Rx} + \delta_{UE\_Rx} \quad (3.24)$$

For the 5G scenario with the DU/CU MEC node deployment option, the Total Node latency is described by the equation:

$$\delta_{Tot\_Node\,[\mathrm{ms}]} = \delta_{UE\_Tx} + \delta_{AL\_Tx} + \delta_{RU\_Tx} + \delta_{DU\_Tx} + \delta_{MEC} + \delta_{DU\_Rx} + \delta_{RU\_Rx} + \delta_{AL\_Rx} + \delta_{UE\_Rx} \quad (3.25)$$

For the 5G scenario with the RU/DU MEC node deployment option, the Total Node latency is described by the equation:

$$\delta_{Tot\_Node\,[\mathrm{ms}]} = \delta_{UE\_Tx} + \delta_{AL\_Tx} + \delta_{RU\_Tx} + \delta_{MEC} + \delta_{RU\_Rx} + \delta_{AL\_Rx} + \delta_{UE\_Rx} \quad (3.26)$$

The total propagation latency for the 4G scenario is given by the equation (not all the links have to be considered, depending on the MEC node deployment):

$$\delta_{Prop\_4G\,[\mathrm{ms}]} = \delta_{FH\_Tx} + \delta_{BH\_Tx} + \delta_{TL\_Tx} + \delta_{TL\_Rx} + \delta_{BH\_Rx} + \delta_{FH\_Rx} \quad (3.27)$$

The total propagation latency for the 5G scenario is given by the equation (not all the links have to be considered, depending on the MEC node deployment and the chosen architecture):

$$\delta_{Prop\_5G\,[\mathrm{ms}]} = \delta_{FH\_Tx} + \delta_{MH\_Tx} + \delta_{BH\_Tx} + \delta_{TL\_Tx} + \delta_{TL\_Rx} + \delta_{BH\_Rx} + \delta_{MH\_Rx} + \delta_{FH\_Rx} \quad (3.28)$$

The E2E latency is calculated using the formula:

$$\delta_{E2E\,[\mathrm{ms}]} = \delta_{Tot\_Node\,[\mathrm{ms}]} + \delta_{Prop\,[\mathrm{ms}]} \quad (3.29)$$

The maximum E2E distance is a parameter that will be calculated by calculating the latency margin without accounting the propagation latency, and it can be calculated by the formula (the 1.67 factor exists because the optical fibers are not installed in a straight line):

$$d_{E2E\,[\mathrm{km}]} = \left(\delta_{App\,[\mathrm{ms}]} - \delta_{Tot\_Node\,[\mathrm{ms}]}\right) \frac{v_{\,[\mathrm{km/s}]}}{2\times 1.67} 10^{-3} \quad (3.30)$$

where:

- $\delta_{App\,[\mathrm{ms}]}$ – Maximum latency depending on what application is chosen.
- $v_{\,[\mathrm{km/s}]}$ – Propagation speed in the link.

E. Link Throughputs

The maximum throughput of the 4G radio link in the FDD mode can be calculated using the formula:

$$R_{FDD\,[\mathrm{Mbits/s}]} = \frac{2\cdot 10^{3}_{[s^{-1}]}\, v_{Layers}\; N_{RB\,[\mathrm{RB}]}\, N_{s\,[\mathrm{symbols/RB}]}\, Q_{m[\mathrm{bits/symbol}]}\, R_{code\_max}\,(1-O)}{10^{6}} \quad (3.31)$$

The maximum throughput of the 5G radio link in the FDD mode can be calculated using the formula:

$$R_{FDD\,[\mathrm{Mbits/s}]} = \frac{v_{Layers}\; Q_{m\,[\mathrm{bits/symbol}]}\, f\, R_{code\_max}\, \frac{12\cdot N_{PRB}^{BW,\mu}[\mathrm{symbols}]}{T_s^{\mu}\,[\mathrm{s}]}(1-O)}{10^{6}} \quad (3.32)$$

where:

- v_{Layers} – Number MIMO of layers.
- $N_{RB}/N_{PRB}^{BW,\mu}$ – Number of Resource Blocks.
- $N_{symbols/RB}$ – Number of symbols per Resource Block in the 4G scenario.
- Q_m – Average modulation order.
- R_{code_max} – Constant defined by 3GPP that depends on the modulation order.
- $T_s^{\mu} = \frac{10^{-3}}{14\times 2^{\mu}}$ – The average OFDM symbol duration in a subframe for the numerology parameter μ, for the 5G system.
- f – 5G Scaling factor.
- O –The overhead for control channels.

The Overhead for the 5G system is represented in Table 5, while the 4G one is considered 0.25.

Table 5. Overhead for control channels in 5G (adapted from [13]).

5G Band	UL	DL	Average Overhead
FR1	0.08	0.14	0.11

The number of available RBs (Resource Blocks) for the 4G system is represented in Table 6.

Table 6. Number of Resource Blocks depending on the Bandwidth (extracted from [14]).

Bandwidth [MHz]	1.4	3	5	10	15	20
Number of RBs, N_{RB}	6	15	25	50	75	100

The number of maximum available RBs (Resource Blocks) for the 5G system is present in [15] and [16].

The average throughput of the radio in the 5G UL in the TDD mode can be calculated using the formula:

$R_{TDD/UL\,[\text{bits/s}]} = R_{FDD\,[\text{bits/s}]}\, F_{UL}\, A_f$ (3.33)

where:

- F_{UL} – The fraction of the slot that is reserved for the UL.
- A_f – The average factor that accounts for the throughput losses in the 5G system (because the calculated theoretical throughput is higher than the real throughputs).

The average throughput of the radio in the 5G DL in the TDD mode can be calculated using the formula:

$R_{\,TDD/DL\,[\text{bits/s}]} = R_{FDD\,[\text{bits/s}]}\, F_{DL}\, A_f$ (3.34)

where:

- F_{DL} – The fraction of the slot that is reserved for the DL.

The average throughput of the radio in the DL in the 4G system can be calculated using the formula:

$R_{FDD/DL\,[\text{bits/s}]} = R_{FDD\,[\text{bits/s}]}\, D_{ur}$ (3.35)

where:

- $R_{FDD/DL\,[\text{bits/s}]}$ – Average FDD capacity in the DL.
- D_{ur} – DL usage ratio.

The average throughput of the radio in the UL in the 4G system can be calculated using the formula:

$R_{FDD/UL\,[\text{bits/s}]} = R_{FDD\,[\text{bits/s}]}\, U_{ur}$ (3.36)

where:

- $R_{FDD/UL\,[\text{bits/s}]}$ – Average FDD capacity in the UL.
- U_{ur} – UL usage ratio.

The RU/RRH required throughput is calculated by summing the data rates of the services provided by the RU/RRH. The following formula is used to calculate the used throughputs of the RU/RRH in both the receiver and transmitter side:

$R_{u\,[\text{Mbits/s}]} = \sum_1^{N_u} R_{s\,[\text{Mbits/s}]}$ (3.37)

where:

- N_u – Number of users connected to the RU/RRH.
- R_s – Data rates of the services offered by the RU/RRH.

F. Model Implementation

The model for the estimation of the E2E latency used in this thesis was developed using the MATLAB program. The program starts by calculating the link throughputs and the radio node capacities, followed by the used throughputs calculations. Some of the latency contributions are common to all the architectures and MEC node deployments, and therefore these initial latency calculations are performed. After the initial calculations, the program calculates the remaining latency contributions. After calculating each one of the individual latency contributions the program will compute the sum of all the contributions obtaining as the result the total node latency. This mathematical value does not account the propagation latency since the lengths of the links still have to be dimensioned. This part of the program is particularly important because it provides an estimation of how the network should be implemented and the distance in between the nodes (E2E distance calculation) for each service group. After obtaining the mathematical results described above, the program is also able to print a map with the maximum distance limits of where the MEC nodes should be implemented to provide a certain service. Figure 5 represents the Model flowchart.

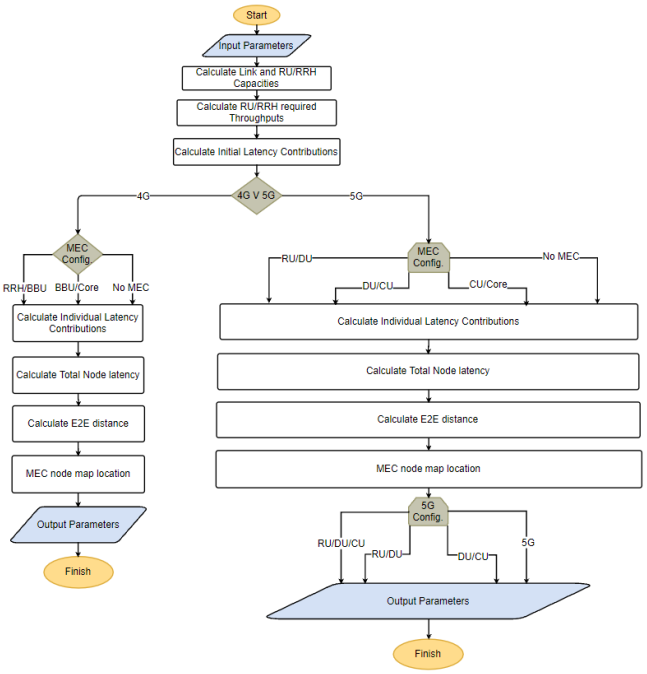

Figure 5. Model Flowchart.

IV. Results Analysis

A. Scenarios

The considered scenarios are the implementation of Internal Remote Surgery inside the Santa Maria Hospital in Lisbon, the implementation of the Network Based Sensor Sharing service in the A1 Highway and the implementation of the Factory Automation services in the Autoeuropa factory.

B. Radio Techniques Analysis

As it is possible to analyze from Figure 6, the 4G indoor RRH is not able to provide the required UL and DL capacity for the Internal Remote Surgery service. This result is obtained because the typical UL capacity of the 4G system is low, even for indoor Base Stations, with values around 15 Mbps on average (depending on the average CQI), which is insufficient in comparison with the required 197.79 Mbps. The 4G DL capacity in the simulation is also below the required 295.18 Mbps in the scenario, with a total capacity of 60.93 Mbps. According to the obtained results, the 5G system provides enough capacity in both the UL and the DL to allow the existence of the Internal Remote Surgery in the hospital, even without using the Radio Techniques that maximize the throughputs, such as the maximum Bandwidth or the maximum number of MIMO layers (which can go up to 400 MHz and 8 layers respectively). The total DL and UL capacities are 723.39 Mbps and 545.72 Mbps respectively, higher than the DL and UL required throughputs.

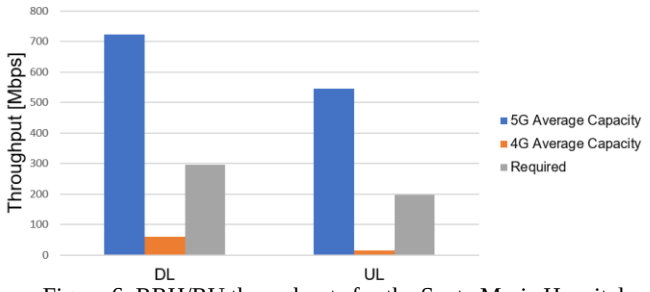

Figure 6. RRH/RU throughputs for the Santa Maria Hospital scenario.

From the analysis of the graph present in Figure 4.7, it is possible to conclude that the 4G system does not provide enough capacity in both the UL and the DL to guarantee the connectivity of all the users. For the 4G system, the DL and UL average capacities are 59.16 Mbps and 15.48 Mbps respectively, which are below the 463.56 Mbps and 459.82 Mbps DL and UL required throughputs. In the simulation, the 5G system provides enough capacity in the DL with 508.15 Mbps, but not enough UL capacity (383.34 Mbps), because the data rate requirements of each considered service are at the high range of the possible interval, and therefore, the required UL throughputs that would exist in a real scenario would be lower and satisfied by the 5G system.

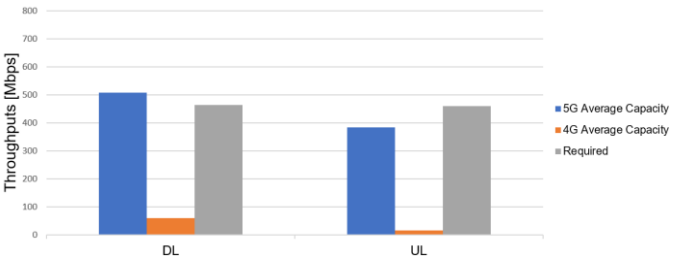

Figure 7. RRH/RU throughputs for the A1 Highway scenario.

As it is possible to analyze from Figure 8, the 4G indoor RRH is not able to provide the required UL and DL capacity for the Factory Automation service. The typical UL capacity of the 4G system has values around 15 Mbps on average (depending on the average CQI), which is not enough to satisfy the required 197.28 Mbps. The 4G DL Capacity in the simulation is also below the required 256.92 Mbps in the scenario, with a total capacity of 60.93 Mbps. According to the obtained results, the 5G system provides enough capacity in both the UL and the DL to allow the existence of the factory automation services in the Autoeuropa factory, even without using the Radio Techniques that maximize the throughputs, such as the maximum Bandwidth or the maximum number of MIMO layers. The total DL and UL capacities are 723.39 Mbps and 545.72 Mbps respectively.

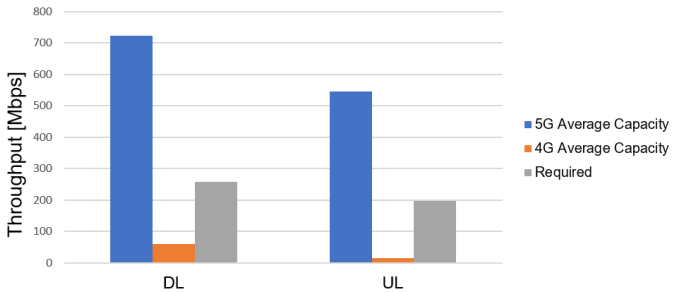

Figure 8. RRH/RU throughputs for the Autoeuropa scenario.

C. *Total Node Latency and MEC Node Deployment Analysis*

As it is possible to analyze from the graph present in Figure 9, some of the E2E latency values of the CU-Core MEC node deployment option are above the 0.9 ms limit, which are considered too high for the required availability and quality of service. The DU-CU MEC node deployment option is also close to the line that marks the margin of the E2E latency, and adding the information that the simulated traffic for the scenario is not considering an intense usage of the network, the chosen MEC node deployment should be in between the RU and DU. The 4G system does not provide sufficiently low latency values for the service implementation, since the minimum E2E latencies of the system are around 8 ms. The points of the graph are close to each other independently of the MEC node deployment option, which happens because the queuing latency added to the E2E latency is only due to the remote surgery packets, since this is a high priority service.

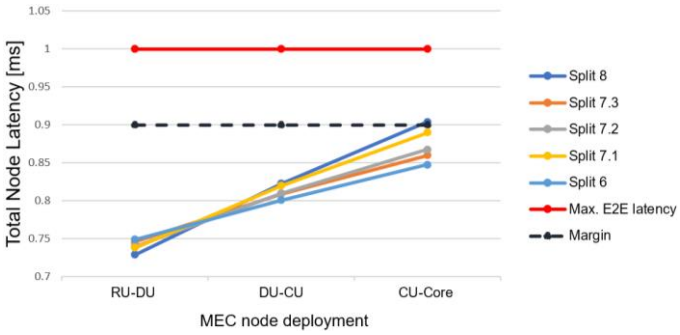

Figure 9. Total node latencies for the Santa Maria Hospital scenario.

Figure 10 illustrates the total node latency results obtained for the A1 Highway with 60% of the maximum traffic scenario, in which the maximum allowed latency value is 3 ms, since the studied service in this scenario is the Network Based Sensor Sharing, which allows the network to participate actively in the achievement of the High Level of Automation for the ITSs. For safety and insurance purposes, it is considered a margin of 10% regarding the maximum E2E latency.

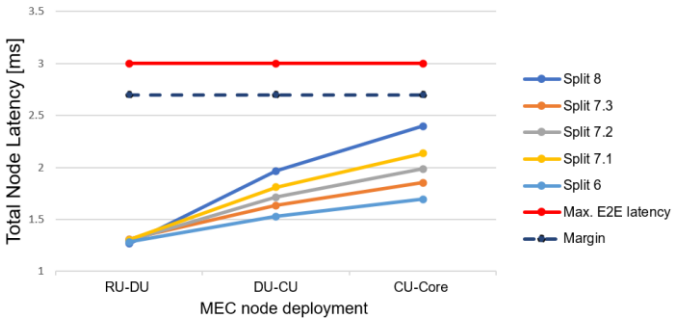

Figure 10. Total node latencies for the A1 Highway scenario with 60% of the maximum traffic.

After analyzing the graph present in Figure 11 it is possible to conclude that the MEC node deployment in between the DU and the CU (in the 5G system) is the most adequate installation to keep the E2E latency below the margin. The installation of the MEC node in between the CU and the Core keeps the latency too close to the margin for the 60% percentile of the maximum traffic. Figure 11 presents the E2E latency values for 30%, 60% and 90% of the maximum traffic load for the splitting option 7.2.

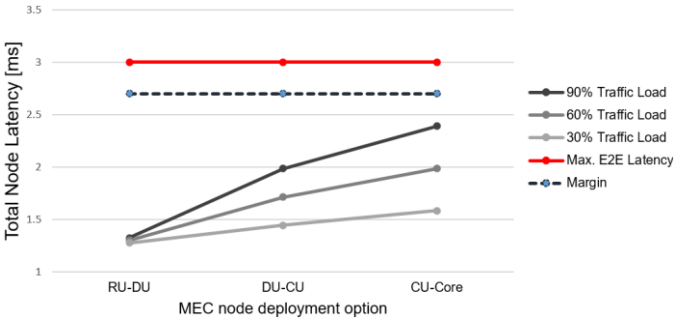

Figure 11. Total node latencies for the A1 Highway scenario with increasing percentages of the maximum traffic for the splitting option 7.2.

After analyzing the graph in Figure 12, it is possible to understand that the DU-CU and CU-Core MEC node deployments are not suited to keep the latency at sufficiently low levels that allow the availability of the Machine Tools service. Therefore, it is required the positioning of the MEC node in between the RU and the DU nodes to reduce the probability that the E2E latency is kept above the margin, and in some intense traffic scenarios the latency may even exceed the margin. The 4G system is not able to provide sufficiently low levels of the E2E latency to guarantee the quality of the service, since the minimum 4G air latency is 8 ms.

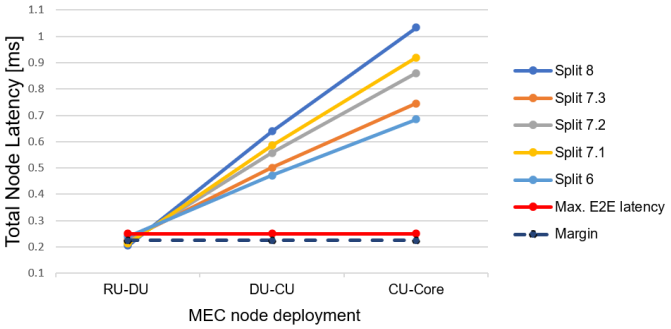

Figure 12. Total node latencies for the Autoeuropa factory scenario.

D. E2E Distance Analysis

Figure 13 represents the Maximum E2E distances in which the MEC node can be installed to guarantee that the E2E latency is below the maximum allowed value for the Santa Maria Hospital scenario. For this scenario only the Splitting Options 6, 7.2 and 8 are represented, because the circles from all the splitting options are close to each other, since the Total Node Latency has similar values for all the splitting options, which translates into similar maximum E2E distances.

Since the traffic that is considered is just a possible scenario, and the smallest radius of the circles 15.0 kms for the RU-DU MEC node deployment, the most viable implementation is to install the MEC node in the hospital facilities, to reduce the probability of exceeding the maximum E2E latency.

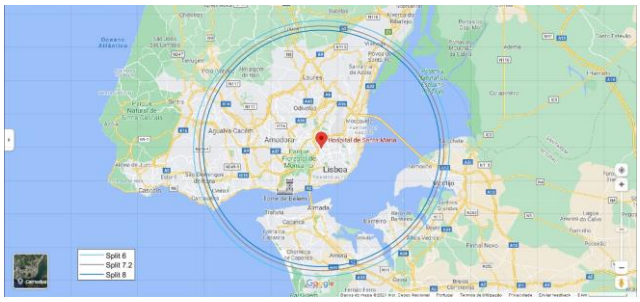

Figure 13. Santa Maria Hospital scenario maximum E2E distancies.

Figure 15 represents the Maximum E2E distance in which the MEC node can be installed to guarantee that the E2E latency is below the maximum allowed value for the A1 Highway scenario with 60% of the maximum traffic. The maximum E2E latency value considered for this scenario is 3 ms. The E2E distances presented in the map are in between 62.0 kms and 88.3 kms for Splitting Option 8 and Splitting Option 6 respectively. This difference exists because of the traffic aggregation, which increases the queuing latency and reduces the maximum distance in which the MEC node can be deployed to keep the latency below the maximum value allowed to perform the service.

Figure 14. A1 Highway scenario maximum E2E distancies.

Figure 15 represents the maximum E2E distance in which the MEC node can be installed to guarantee that the E2E latency is below the maximum allowed value for the Autoeuropa Factory scenario. The considered scenario has a low maximum allowed E2E latency value, in the order of 0.25 ms, delays that can only be fulfilled with the RU-DU MEC node deployment. According to the obtained results the MEC node should be deployed inside of the factory facilities in order to reduce the delay accumulation to fulfil the strict service requirements.

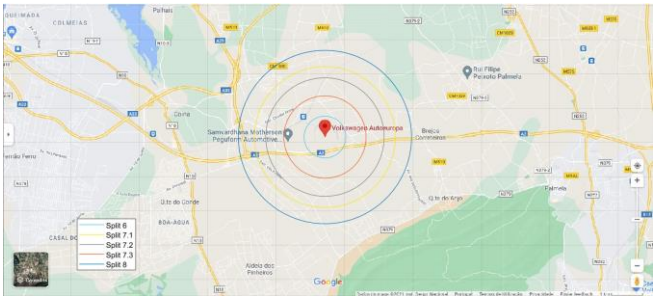

Figure 15. Autoeuropa factory scenario maximum E2E distancies.

Figure 16 represents the required MEC node coverage for the A1 Highway scenario. According to the calculations it is possible to achieve total coverage of the highway with 3 MEC nodes.

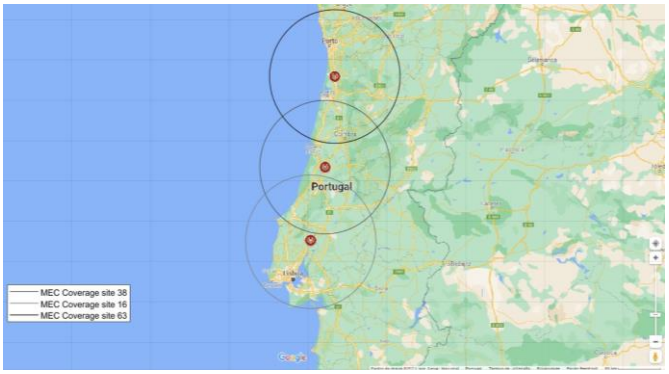

Figure 16. A1 Highway MEC node coverage.

V. CONCLUSIONS

For the indoor radio nodes, the average 5G capacity obtained in the simulations is 723.39 Mbps in the DL and 545.72 Mbps in the UL, as long as in the outdoor radio nodes, the average 5G capacity is 508.15 Mbps and 383.34 Mbps. The 4G system is already installed, and therefore there is a better knowledge about the traffic and the real system capacity. For the indoor capacity, the program calculates an average DL throughput of 60.93 Mbps and an average UL throughput of 15.72 Mbps. The outdoor capacity provided by LTE depends on the frequency bands with which the site works, for example, the A1 Highway site works with the 1800 and 800 MHz bands, according to the information provide by NOS. The calculated LTE system capacity in the A1 Highway scenario is 59.16 Mbps and 15.48 Mbps, for the DL and the UL respectively. After analyzing the results, it is clear that the 4G system is not able to guarantee enough capacity to provide the URLLC services that are expected to emerge in the following years. Nevertheless, LTE can be used to generate an offload of the NR system, which can be useful to the operators. For the Santa Maria Hospital scenario, the required throughputs in the DL and UL are 295.18 Mbps and 197.79 Mbps (the radio node that receives and transmits the packets is the same), respectively, which are below the radio node capacity. For the A1 Highway scenario with 60% of the maximum traffic the required throughputs in the DL and UL are 463.56 Mbps and 459.82 Mbps, respectively. Although the simulated NR capacity is not enough to guarantee the Network Based Sensor Sharing service in the scenario, the considered required data rates for the service was in the highest range of the possible values, meaning that on average the required throughputs will be lower, and therefore the 5G sites are expected to be able to provide these services. For Autoeuropa Factory scenario, the required throughputs in the DL and UL are 256.92 Mbps and 197.28 Mbps, respectively. As it is possible to understand by comparing the required throughputs with the 5G indoor and outdoor capacities, the NR system can provide capacity for the factory automation services in different scenarios. In the thesis it is presented an E2E latency study with a consequent E2E distance analysis. It is important to refer that this analysis is only performed for the best considered MEC node deployment option, after the total node latency analysis. Since the minimum E2E latencies are close to 8 ms, with the fixed 8 ms of transmission and processing delay in the radio, it is known that the LTE network is not prepared to achieve the low latency requirements of the studied services. Nevertheless, the obtained results mixed with the information available in papers leads to values in the 8 to 20 ms range for the RRH-BBU MEC node deployment and values in the 10 to 26 ms range for the BBU-Core MEC node deployment. For the Santa Maria scenario it is possible to conclude that the best MEC node deployment option is in between the RU and the DU, and even with this implementation it is possible for the E2E latency to exceed the maximum allowed latency for the service in some intense traffic scenarios. Therefore, although there is a margin in between the calculated Total Node latency and the maximum E2E latency that generates a maximum E2E distance of approximately 15 kms, it is advised that the MEC node is installed inside the hospital, to prevent the latency increase under certain circumstances. For the A1 Highway scenario one can conclude that the best MEC node deployment option is in between the DU and the CU, because even for intense network usage scenarios (90% of the

maximum traffic) the total node latency is considerably below the maximum allowed Network Based Sensor Sharing E2E latency of 5 ms, which does not occur for the deployment in between the CU and the Core. The E2E distance calculated for the best considered MEC node deployment has values in between 62.0 kms and 88.3 kms for the splitting options 8 and 6 respectively. For the Autoeuropa factory scenario one can conclude that the best and only possible MEC node deployment option is in between the RU and the DU, since the other options exceed the maximum allowed E2E latency. The E2E distance calculated for the best considered MEC node deployment has values in between 2.8 kms and 0.67 kms for the splitting options 8 and 6 respectively, which leads to the conclusion that the MEC node needs to be installed inside the factory facilities to guarantee the minimum delay.

Regarding future work, the performed simulations should be able to achieve better accuracy in terms of results when the processing capacity of the 5G network nodes is known, as well as the traffic profiles from the future URLLC services which is unpredictable based on the currently available information. The model used is a packet-based model, in which the E2E latency studied for 5G system is calculated based in a reference packet instead of the time frames. The processing delays were calculated based in the number of functionalities in the nodes and the packet size, which is an approximation that could be improved if the RU, DU, CU and MEC node processing capacities were known and studied. The model only allows the implementation of one architecture, but it should be able to simulate a hybrid architecture to obtain more realistic results and be adapted to all the possible network deployments.

REFERENCES

[1] Thales, Introducing 5G technology and Network, Dec.2020. Available : https://www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/inspired/5G.

[2] Najmul Hassan, Kok-Lim Alvin Yau and Celimuge Wu, "Edge Computing in 5G: A Review", *2015 IEEE Journals and Magazine*, Vol. 7, May.2015, pp 127276 –127289. Available: https://ieeexplore.ieee.org/document/8821283.

[3] Alejandro Santoyo-González, Cristina Cervelló-Pastor, "Edge Nodes Infrastructure Placement Parameters for 5G Networks", in *2018 IEEE Conference on Standards for Communications and Networking (CSCN)*, Paris, France, Oct.2018. Available: https://ieeexplore.ieee.org/document/8581749.

[4] Next Generation Optical Transport Network, *5G Oriented OTN technology White Paper*, Mar.2018. Available: http://www.ngof.net/download/5G.pdf.

[5] Vodafone, *Overview and Predictive Analysis for Latency Optimized Telecommunication Networks*, Hochschule Rhein Main Russelsheim University, Russelsheim, Germany, Oct.2018. Available: https://www.hs-rm.de/fileadmin/persons/khofmann/Gastvortraege/Vortragsfolien/20181026-Burk-Lemberg-vodafone-5G.pdf.

[6] ITU, *Transport network support of IMT-2020/5G*, Technical Report, Ciena, Canada Feb.2018 Available: https://www.itu.int/dms_pub/itu-t/opb/tut/T-TUT-HOME-2018-2-PDF-E.pdf.

[7] Imtiaz Parvez, Ali Rahmati, Ismail Guvenc, Arif Sarwat and Huaiyu Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions", *IEEE communications surveys & tutorials*, Vol. 20, No. 4, May.2018, pp. 3098 –3130. Available: https://ieeexplore.ieee.org/document/8367785.

[8] S. Domingues, *Analysis of the Performance of Multi-Access Edge Computing Network Slicing in 5G*, M.SC thesis, Instituto Superior Técnico, Lisbon, Portugal, Nov.2019. Available: https://grow.tecnico.ulisboa.pt/wp-content/uploads/2020/07/Thesis_SergioD_vPublic.pdf.

[9] Altice Labs, *5G Intelligent Communications for V2X ecosystems Whitepaper*, Jul.2020 Available: https://www.alticelabs.com/content/WP_5G_Intelligent_Communications.pdf.

[10] Daniel Maaz, Ana Galindo-Serrano, Salah Eddine Elayoubi, "URLLC User Plane Latency Performance in New Radio", in *IEEE 2018 25th International Conference on Telecommunications (ICT)*, Saint-Malo, France, Sep.2018. Available: https://ieeexplore.ieee.org/document/8464912.

[11] S. Domingues, *Analysis of the Performance of Multi-Access Edge Computing Network Slicing in 5G*, M.SC thesis, Instituto Superior Técnico, Lisbon, Portugal, Nov.2019. Available: https://grow.tecnico.ulisboa.pt/wp-content/uploads/2020/07/Thesis_SergioD_vPublic.pdf.

[12] Osama Al-Saadeh, Gustav Wikstrom, Joachim Sachs, "End-to-End Latency and reliability Performance of 5G in London", in *IEEE 2018 Global Communications Conference*, Abu Dhabi, United Arab Emirates, Feb.2019. Available: https://ieeexplore.ieee.org/document/8647379.

[13] ETSI, *5G NR; User Equipment (UE) radio access capabilities (Release 15)*, Internal Report TS 38.306, V15.3.0, Oct.2018. Available: https://www.etsi.org/deliver/etsi_ts/138300_138399/138306/15.03.00_60/ts_138306v150300p.pdf

[14] Luís Manuel Correia, Notes from Mobile Communication Systems course, Instituto Superior Técnico, Lisbon, Portugal, 2020.

[15] ETSI, *5G NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone*, Report TS 38.101-1, Release 15, V15.2.0, Jul.2018. Available: https://www.etsi.org/deliver/etsi_ts/138100_138199/13810101/15.02.00_60/ts_13810101v150200p.pdf.

[16] 3GPP, *5G NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone (Release 15)*, Report TS 38.101-2, V15.2.0, Jul.2018.