

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

Afonso Lopes Vieira de Carvalho

Thesis to obtain the Master of Science Degree in
Electrical and Computer Engineering

Supervisors: Prof. Luís Manuel de Jesus Sousa Correia

Prof. António Manuel Raminhos Cordeiro Grilo

Examination Committee

Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino

Supervisors: Prof. Luís Manuel de Jesus Sousa Correia

Members of Committee: Prof. António José Castelo Branco Rodrigues

Eng. Ricardo Dinis

November 2021

I declare that this document is an original work of my own authorship and that it fulfils
all the requirements of the Code of Conduct and Good Practices of the
Universidade de Lisboa.

To my beloved family and friends

Acknowledgements

First of all, I would like to take this opportunity to express my strongest thanks and gratitude to my thesis professor and supervisor Luís Manuel Correia, for always being available and helpful to me when I needed the most. The Monday meetings were, without any doubt, the funniest part of the day during the covid times, but also the period in which we learned the most and improved our work. Without the help and guidance of this amazing professor, the development of this document, as complex and detailed as it is, would not be possible.

I would also like to show my appreciation to professor António Grilo, for his unconditional help and guidance, which was crucial to improve the quality of my work.

To all GROW members, especially to my dear friends Beatriz Ferreira, Sofia Patrício, Marco Filho and Maria Ferreira, for their support and the good energy that they brought to all meetings.

To Eng. Ricardo Dinis from NOS, for always being available to answer my calls and my questions, and also for providing me with important suggestions and a better insight of the reality of telecommunications.

To all my family, my mother, Paula Vieira, my father, Valério Carvalho, my brother, António Carvalho, my sister, Carolina Carvalho, both my grandmothers, Maria Lopes and Ascensão Carvalho, both my grandfathers, Albertino Vieira and Valério Carvalho, and to my stepfather, João Pacheco, for always being there for me when I needed, for their care and attention, and for being such an amazing family.

To all my close friends from Instituto Superior Técnico, for being my company in what has been an amazing journey throughout these intense five years.

I would like to finish my thesis acknowledgements, with my last but not least expression of appreciation, to my friends João Pedro Fonseca, Francisco Ramirez, Gonçalo Calado and Marta Miranda for their trustful help and for always being by my side.

Abstract

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

Resumo

O objetivo principal desta tese é identificar e estudar as diferentes estratégias que reduzem efetivamente a latência entre as duas extremidades das redes de telecomunicações 4G e 5G. Esta redução na latência permitirá às operadoras fornecer serviços *URLLC*, como por exemplo cirurgia remota, sistemas inteligentes de transporte e serviços de automação nas fábricas. Para verificar a possibilidade de implementação dos referidos serviços usando as redes 4G e 5G, o modelo desenvolvido tem em conta as variáveis seguintes: opção de implementação do nó MEC, *splitting options*, técnicas de rádio e as diferentes arquiteturas da rede. A tecnologia MEC aparece como sendo a solução que permitirá a redução da latência para valores de 1 ms entre as duas extremidades da rede. Os resultados obtidos mostram que o sistema 4G não apresenta uma capacidade suficiente para garantir a existência dos serviços futuros. Mesmo com a implementação do nó MEC que permite minimizar a latência, a rede do sistema LTE não consegue fornecer aos utilizadores os serviços *URLLC* estudados. As simulações mostram que utilizando as estratégias de redução de latência e as técnicas rádio adequadas, o sistema 5G tem a capacidade necessária e latências suficientemente baixas para garantir o fornecimento dos serviços do futuro.

Palavras-chave

5G, 4G, Arquiteturas em Nuvem, Latência, MEC.

Table of Contents

Acknowledgements	vii
Abstract.....	ix
Resumo	x
Table of Contents.....	xi
List of Figures	xiv
List of Tables.....	xix
List of Abbreviations.....	xxi
List of Symbols.....	xxv
List of Software	xxix
1 Introduction	1
1.1 Overview.....	2
1.2 Motivation and Contents.....	4
2 Fundamental Aspects.....	7
2.1 5G aspects	8
2.1.1 Network Architecture.....	8
2.1.2 Radio Interface.....	10
2.2 Network Slicing and Virtualisation	14
2.3 Cloud and Edge Networks.....	16
2.3.1 Cloud Network	16
2.3.2 Edge Network	18
2.4 Services and Applications	20
2.5 Latency in the Network.....	22
2.6 State of the Art	25
3 Models and Simulator Description	27
3.1 Model Overview.....	28
3.2 Network Description	31
3.2.1 Network Architecture.....	31

3.2.2	Latency Contributions	35
3.2.3	Network Latency	40
3.2.4	Link Throughput	44
3.3	Model Implementation	49
3.4	Model Assessment	51
4	Results Analysis	55
4.1	Scenarios.....	56
4.2	Radio Characteristics Analysis	60
4.3	Total Node Latency and MEC Node Deployment Analysis.....	64
4.4	E2E Distance Analysis	71
5	Conclusions.....	77
Annex A.	User's Manual.....	83
A.1	Run the Simulator	84
A.2	Simulator Configuration	84
Annex B.	4G and 5G services	87
Annex C.	Latency Contributions Description	89
Annex D.	Data Centre and MEC node processing latency	92
Annex E.	Latency Adaptation Parameters.....	94
Annex F.	4G and 5G Link Throughputs.....	96
Annex G.	Target Code Rate R for the 4G system	99
Annex H.	Target Code Rate R for the 5G system.....	101
Annex I.	5G TDD slot formats	103
Annex J.	Relevant Data for the Scenarios.....	105
Annex K.	Reference Scenario Configuration	107
Annex L.	Radio Techniques Configuration	112
Annex M.	Urban ITS and A1 Highway Scenarios with 30% and 90% of the Maximum	114
Annex N.	MEC Coverage.....	118

References.....	121
-----------------	-----

List of Figures

Figure 1.1 – Expected global mobile data traffic evolution (extracted from [Eric20]).	2
Figure 1.2 – 5G specification requirements (extracted from [Thal20]).	3
Figure 1.3 – Edge Computing (extracted from [YZWW19]).	4
Figure 2.1 – 5G Non-standalone and Standalone versions (extracted from [ARTM19]).	8
Figure 2.2 – LTE Network Architecture (extracted from [ITUT19]).	8
Figure 2.3 – 5G Network Architecture (extracted from [3GPP19a]).	9
Figure 2.4 – 5G Frame Structure (adapted from [GMFF20]).	13
Figure 2.5 – OFDM cyclic prefix variation with subcarrier spacing (adapted from [AZRB17]).	13
Figure 2.6 – TTI scalability (extracted from [Qual16]).	14
Figure 2.7 – SDN Layers (extracted from [VGMS12]).	15
Figure 2.8 – NFV Architecture (extracted from [ETSI13]).	16
Figure 2.9 – 5G Network Slicing (extracted from [MIW15]).	16
Figure 2.10 – Representation of a C-RAN (extracted from [KMNT18]).	17
Figure 2.11 – 5G and 4G Edge Networking (extracted from [NHKA19]).	19
Figure 2.12 – Edge Deployment Scenarios for Latency Optimization (extracted from [ASCC18]).	20
Figure 2.13 – 3 main directions of 5G (extracted from [ITUT17]).	21
Figure 2.14 – Contributions of the delays to the latency (extracted from [VODA18]).	23
Figure 2.15 – 4G vs 5G transport network architectures (adapted from [NGOF18]).	23
Figure 2.16 – Latency and network evolutions (extracted from [VODA18]).	24
Figure 3.1 – Model Overview.	28
Figure 3.2 – 3GPP C-RAN split architectures (adapted from [ITUT18]).	32
Figure 3.3 – MEC node installation options for 4G architecture.	32
Figure 3.4 – MEC node installation options for 5G architecture.	33
Figure 3.5 – 5G node aggregation in the network.	35
Figure 3.6 – 4G contributions in the scenario without the MEC node.	40
Figure 3.7 – 4G latency contributions in the scenario with the MEC node installed in between the BBU and the Core of the Network.	41
Figure 3.8 – 4G latency contributions in the scenario with the MEC node installed in between the RRH and the BBU.	41
Figure 3.9 – 5G scenario without MEC node.	42
Figure 3.10 – 5G latency contributions in the scenario with the MEC node in between the CU and	

the Core of the Network.	42
Figure 3.11 – 5G latency contributions in the scenario with the MEC node in between the DU and the CU.	42
Figure 3.12 – Latency contributions in the scenario with the MEC node in between the RU and the DU.	43
Figure 3.13 – The latency contributions in the scenario with the MEC node in between the RU and the DU with collocated nodes.	44
Figure 3.14 – Model Flowchart.	50
Figure 3.15 – RU queuing latency plot while varying the FH throughput.	50
Figure 4.1 – Sta. Maria Hospital NOS sites.	56
Figure 4.2 – Espírito Santo de Évora Hospital NOS sites.	57
Figure 4.3 – Avenida da Liberdade NOS sites.	58
Figure 4.4 – Sacavém - São João da Talha section NOS sites.	59
Figure 4.5 – Autoeuropa factory NOS sites location.	59
Figure 4.6 – RRH/RU throughputs for the Santa Maria Hospital scenario.	61
Figure 4.7 – RRH/RU throughputs for the Espírito Santo de Évora Hospital Factory scenario. .	61
Figure 4.8 – RRH/RU throughputs for the Urban ITS scenario with 60% of the maximum traffic.	62
Figure 4.9 – RRH/RU throughputs for the A1 Highway scenario with 60% of the maximum traffic.	63
Figure 4.10 – RRH/RU throughputs for the Autoeuropa Factory scenario.	64
Figure 4.11 – Total node latencies for the Santa Maria Hospital scenario.	64
Figure 4.12 – Individual latency contributions for the RU-DU MEC node deployment in Santa Maria scenario.	65
Figure 4.13 – Total node latencies for the Espírito Santo de Évora scenario.	66
Figure 4.14 – Individual latency contributions for the CU-Core MEC node deployment in Espírito Santo de Évora scenario.	66
Figure 4.15 – Total node latencies for the Urban ITS scenario with 60% of the maximum traffic.	67
Figure 4.16 – Individual latency contributions for the CU-Core MEC node deployment in the Urban ITS scenario for the 60% traffic usage.	67
Figure 4.17 – Total node latencies for the Urban ITS scenario with increasing percentages of the maximum traffic for the splitting option 7.2.	68
Figure 4.18 – Total node latencies for the A1 Highway scenario with 60% of the maximum traffic.	68
Figure 4.19 – Individual latency contributions for the DU-CU MEC node deployment in the	

Highway scenario for the 60% traffic usage.	69
Figure 4.20 – Total node latencies for the A1 Highway scenario with increasing percentages of the maximum traffic for the splitting option 7.2.	69
Figure 4.21 – Total node latencies for the Autoeuropa Factory scenario.	70
Figure 4.22 – Individual latency contributions for the RU-DU MEC node deployment in the Autoeuropa factory scenario.	70
Figure 4.23 – Santa Maria Hospital scenario maximum E2E distances.	71
Figure 4.24 – Espírito Santo de Évora Hospital scenario maximum E2E distance.	72
Figure 4.25 – Urban ITS scenario (with 60% of the maximum traffic) maximum E2E distances.	73
Figure 4.26 – A1 Highway scenario (with 60% of the maximum traffic) maximum E2E distances.	74
Figure 4.27 – Autoeuropa Factory scenario maximum E2E distances.	75
Figure D.1 – Data Centre and MEC node processign latency for a single functionality.	93
Figure M.1 – RRH/RU throughputs for the Urban ITS scenario with 30% of the maximum traffic.	115
Figure M.2 – RRH/RU throughputs for the Urban ITS scenario with 90% of the maximum traffic.	115
Figure M.3 – RRH/RU throughputs for the A1 Highway scenario with 30% of the maximum traffic.	115
Figure M.4 – RRH/RU throughputs for the A1 Highway scenario with 90% of the maximum traffic.	116
Figure M.5 – Latency results for the Urban ITS scenario with 30% of the maximum traffic.	116
Figure M.6 – Latency results for the Urban ITS scenario with 90% of the maximum traffic.	116
Figure M.7 – Latency results for the A1 Highway scenario with 30% of the maximum traffic.	117
Figure M.8 – Latency results for the A1 Highway scenario with 90% of the maximum traffic.	117
Figure N.1 – Santa Maria scenario MEC node coverage.....	119
Figure N.2 – A1 Highway MEC node coverage.....	120

List of Tables

Table 2.1 – NR Frequency Bands List (adapted from [5GOB20]).	12
Table 2.2 – Supported Transmission Numerologies and Frame Structure (adapted from [3GPP19b]).	12
Table 2.3 – Expected 5G services characteristics (adapted from [3GPP19c]).	21
Table 2.4 – Latency Critical IoT application list (adapted from [PSMM17]).	22
Table 3.1 – Network specifications.	29
Table 3.2 – User Specification.	30
Table 3.3 – Service specifications.	30
Table 3.4 – 4G and 5G UE processing delay ratios (extracted from [DMAG18]).	37
Table 3.5 – RU, DU and CU processing latency ratios (adapted from [SeDo19]).	38
Table 3.6 – Overhead for control channels in 5G (adapted from [ETSI18c]).	45
Table 3.7 – Number of Resource Blocks depending on the Bandwidth (extracted from [Corr20]).	46
Table 3.8 – Maximum NRBs for each transmission bandwidth and subcarrier spacing (adapted from [ETSI18b] and [3GPP18b]).	46
Table 3.9– Typical NG Backhaul and Transport Network throughputs (extracted from [ITUT18]).	49
Table 3.10 – Model Assessment Tests.	52
Table 4.1 – Scenario Comparison.	56
Table 4.2 – Santa Maria Hospital maximum E2E distances for each MEC node Deployment and splitting option.	72
Table 4.3 – Espírito Santo Évora Hospital scenario maximum E2E distances for each MEC node Deployment and splitting option.	73
Table 4.4 – Urban ITS scenario (with 60% of the maximum traffic) maximum E2E distances for each MEC node Deployment and splitting option.	74
Table 4.5 – A1 Highway scenario (with 60% of the maximum traffic) maximum E2E distances for each MEC node Deployment and splitting option.	75
Table 4.6 – AutoEuropa Factory scenario maximum E2E distances for each MEC node Deployment and splitting option.	76
Table A.1 – Network Specifications input parameters definition.	84
Table A.2 – User specification input parameters definition.	85
Table A.3 – Service specification input parameters definition.	86
Table B.1 – Service requirements.	88

Table C.1 – Latency contributions description.	90
Table E.1 – Latency adaptation parameter.	95
Table F.1 – 4G and 5G link throughputs depending on the splitting options.	97
Table G.1 – R parameter for 4G radio throughput (extracted from [3GPP17]).	100
Table H.1 – R parameter for 5G radio throughput (extracted from [ETSI18d]).	102
Table I.1 – 5G TDD slot formats.	104
Table J.1 – Link type and UE distances for the simulated scenarios.	106
Table K.1 – Average number of users per node (both receiver and transmitter sides) on the Santa Maria Hospital scenario.	108
Table K.2 – Santa Maria Hospital scenario receiver and transmitter RU/RRH service mix.	108
Table K.3 – Average number of users per node on the Espírito Santo de Évora Hospital scenario.	108
Table K.4 – Espírito Santo de Évora Hospital scenario transmitter RU/RRH service mix.	109
Table K.5 – Average number of users per node on the Urban ITS scenario for 30 %, 60 % and 90% of network usage.	109
Table K.6 – Urban ITS scenario RU (both transmitter and receiver sides) service mix.	110
Table K.7 – Average number of users per node on the A1 Highway for 30 %, 60 % and 90 % of the network usage.	110
Table K.8 – A1 Highway scenario RU (transmitter and receiver) service mix.	111
Table K.9 – Average number of users per node on the AutoEuropa factory scenario.	111
Table K.10 – AutoEuropa factory scenario RU (on both the transmitter and receiver side) service mix.	111
Table L.1 – 4G Radio Characteristics for each scenario.	113
Table L.2 – 5G Radio Characteristics for each scenario.	113
Table N.1 – Santa Maria scenario MEC node coverage hospitals list.	119

List of Abbreviations

3GPP	3 rd Generation Partnership Project
4G	4 th Generation of Mobile Communication Systems
5G	5 th Generation of Mobile Communication Systems
AF	Application Function
AMF	Access and Mobility Management Function
AUSF	Authentication Server Function
AR	Augmented Reality
BBU	Baseband Unit
BH	Backhaul
CQI	Channel Quality Indicator
C-RAN	Cloud Radio Access Network
CU	Centralised Unit
DL	Downlink
DU	Distributed Unit
eMBB	Enhanced Mobile Broadband
eNB	Evolved Node B
EC	Edge Computing
EPC	Evolved Packet Core
EU	European Union
E-UTRAN	Evolved-UMTS Terrestrial Radio Access Network
FDD	Frequency Division Duplexing
FH	Fronthaul Link
HSS	Home Subscriber Service
IoT	Internet of Things
ITS	Intelligent Transport Systems
LTE	Long Term Evolution
mmWave	Millimetric Wavelength
mMTC	Massive Machine Type Communications
MANO	Management and Orchestration
MEC	Multi-Access Edge Computing
MIMO	Multiple Input Multiple Output

MME	Mobility Management Entity
MCS	Modulation and Coding Scheme
MH	Middlehaul Link
MSC	Mobile Switching Centre
NEF	Network Exposure Function
NFV	Network Function Virtualisation
NFVI	Network Function Virtualisation Infrastructure
NR	New Radio
NRF	Network Repository Function
NSA	Non-Standalone
NSSF	Network Slice Selection Function
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OSS	Operation Support System
PCF	Policy Control Function
PCFR	Policy and Charging Rules Function
PDCF	Packet Data Convergence Protocol
P-GW	Packet Data Network Gateway
PST	Packet Sending Time
QoS	Quality of Service
RAN	Radio Access Network
RB	Resource Block
RLC	Radio Link Control
RRH	Remote Radio Header
RRU	Remote Radio Unit
RU	Radio Unit
SA	Standalone
SC-FDMA	Single-Carrier Frequency Division Multiple Access
SCS	Subcarrier Spacing
SDN	Software Defined Network
S-GW	Serving Gateway
SMF	Session Management Function
SMS	Short Message Service
TDD	Time Division Duplexing

TL	Transport Link
TTI	Transmission Time Interval
UE	User Equipment
UDM	Unified Data Management
UDR	Unified Data Repository
UL	Uplink
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communication
V2N	Vehicle to Network
VNF	Virtualisation Network Function
VR	Virtual Reality

List of Symbols

δ_{AL_Rx}	Air link propagation latency on the receiver side
δ_{AL_Tx}	Air link propagation latency on the transmitter side
δ_{App}	Maximum latency of the chosen application
δ_{BBU_Proc}	BBU processing latency
δ_{BBU_Queu}	BBU queuing latency
δ_{BBU_Rx}	Latency of the BBU node on the receiver side
δ_{BBU_Trans}	BBU transmission latency
δ_{BBU_Tx}	Latency of the BBU node on the transmitter side
δ_{BH_Rx}	Propagation latency of the BH link on the receiver side
δ_{BH_Tx}	Propagation latency of the BH link on the transmitter side
δ_{Core_Proc}	Core processing latency
δ_{Core_Rx}	Latency of the Core of the network on the receiver side
δ_{Core_Trans}	Core transmission latency
δ_{Core_Tx}	Latency of the Core of the network on the transmitter side
δ_{CU_Proc}	CU processing latency
δ_{CU_Queu}	CU queuing latency
δ_{CU_Rx}	Latency of the CU node on the receiver side
δ_{CU_Trans}	CU transmission latency
δ_{CU_Tx}	Latency of the CU node on the transmitter side
δ_{DU_Proc}	DU processing latency
δ_{DU_Queu}	DU queuing latency
δ_{DU_Rx}	Latency of the DU node on the receiver side
δ_{DU_Trans}	DU transmission latency
δ_{DU_Tx}	Latency of the DU node on the transmitter side
δ_{E2E}	End-to-end latency
δ_{EDC}	Latency of the EDC node

$\delta_{\text{EDC_Proc}}$	EDC processing latency
$\delta_{\text{EDC_Trans}}$	EDC transmission latency
$\delta_{\text{FH_Rx}}$	Propagation latency of the FH link on the receiver side
$\delta_{\text{FH_Tx}}$	Propagation latency of the FH link on the transmitter side
δ_{MEC}	Latency of the MEC node
$\delta_{\text{MEC_Proc}}$	MEC processing latency
$\delta_{\text{MEC_Trans}}$	MEC transmission latency
$\delta_{\text{MH_Rx}}$	Propagation latency of the MH link on the receiver side
$\delta_{\text{MH_Tx}}$	Propagation latency of the MH link on the transmitter side
δ_{Prop}	Propagation latency
δ_{Queu}	Queuing latency
$\delta_{\text{RRH_Proc}}$	RRH processing latency
$\delta_{\text{RRH_Queu}}$	RRH queuing latency
$\delta_{\text{RRH_Trans}}$	RRH transmission latency
$\delta_{\text{RRH_Tx}}$	Latency of the RRH node on the transmitter side
$\delta_{\text{RRH_Rx}}$	Latency of the RRH node on the receiver side
$\delta_{\text{RU_Proc}}$	RU processing latency
$\delta_{\text{RU_Queu}}$	RU queuing latency
$\delta_{\text{RU_Trans}}$	RU transmission latency
$\delta_{\text{RU_Tx}}$	Latency of the RU node on the transmitter side
$\delta_{\text{RU_Rx}}$	Latency of the RU node on the receiver side
$\delta_{\text{TL_Rx}}$	Propagation latency of the TL on the receiver side
$\delta_{\text{TL_Tx}}$	Propagation latency of the TL on the transmitter side
δ_{Trans}	Transmission latency
$\delta_{\text{UE_Proc}}$	UE processing latency
$\delta_{\text{UE_Rx}}$	User equipment latency on the receiver side
$\delta_{\text{UE_Tx}}$	User equipment latency on the transmitter side
$\delta_{\text{UE_Trans}}$	UE transmission latency
μ_s	Subcarrier utilisation

ρ_{CU}	Number of functionalities performed by the CU node
ρ_{DU}	Number of functionalities performed by the DU node
ρ_{func}	Number of functionalities performed by the MEC node
ρ_{lat}	Latency adaptation parameter
ρ_{RU}	Number of functionalities performed by the RU node
ρ_{UE}	Latency parameter of the UE
A_f	Average Factor
B	System bandwidth
B_c	Bandwidth for control signals
d	Link Length
d_{E2E}	End-to-end distance
D	Data Packet Size
$D_{serv,p}$	Data Packet Size for a service with priority p
D_{ur}	Downlink Usage Ratio
f	5G scaling factor
F_{DL}	Fraction of the slot reserved for the DL
F_{UL}	Fraction of the slot reserved for the UL
M	Modulation order
M_{info}	Information in the MAC
M_c	Modulation order for control signals
M_{Pserv}	Number of users connected to the node using services with a higher or equal priority than the studied user
N_A	Number of antenna ports
N_L	Number of layers in the system
$N_{L,c}$	The number of layers for control signalling
N_Q	Bitwidth in bits
N_{RB}	Number of Resource Blocks
N_s	Number of symbols per resource block
N_{SC}	Number of subcarriers used in the system
N_{SY}	Number of symbols

N_u	Number of users connected to the RU/RRH
O	Overhead
Q_m	Average Modulation order
R	Maximum throughput offered by the RU/RRH
R_8	FH throughput for Splitting option 8
$R_{7.1_DL}$	FH throughput for Splitting option 7.1 in the DL
$R_{7.1_UL}$	FH throughput for Splitting option 7.1 in the UL
$R_{7.2}$	FH throughput for Splitting option 7.2
$R_{7.3}$	FH throughput for Splitting option 7.3
R_6	FH throughput for Splitting option 6
R_c	Signalling rate
R_{code_max}	Code rate value
$R_{FDD/DL}$	Maximum theoretical FDD throughput in the DL
$R_{FDD/UL}$	Maximum theoretical FDD throughput in the UL
R_{max}	Maximum throughput offered by the link
R_p	Peak rate
R_s	Data rate of the services provided by the RU/RRH
R_u	RU/RRH used throughput
$R_{TDD/DL}$	Maximum theoretical TDD throughput in the DL
$R_{TDD/UL}$	Maximum theoretical TDD throughput in the UL
S_r	Sampling rate in samples per second
T_s^μ	Average subframe OFDM symbol duration for the 5G system
U_{ur}	Uplink Usage Ratio
v_{Layers}	Number of MIMO layers
v	Signal velocity in the link

List of Software

MATLAB R2019a

Numerical computing software

Microsoft Excel

Spreadsheet application

Microsoft PowerPoint

Presentation and slide program

Microsoft Word

Text Editor Software

Chapter 1

Introduction

This chapter gives a brief introduction to the thesis. In the beginning, an overview of the 4th and 5th generations of mobile communications is given, as well as the importance behind this evolution, followed by the scope and the motivation of the work. At the end of the chapter, the work structure is provided.

1.1 Overview

Approximately four decades ago, the 1st generation of mobile communications emerged. Since then, there was a prominent evolution in this area, allowing users to perform more complex activities.

In the beginning, users were able to perform voice calls, but in the modern era of 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, being expected that the quantity of this traffic will continue to increase. In Figure 1.1, one presents the expected evolution of the global mobile data traffic.

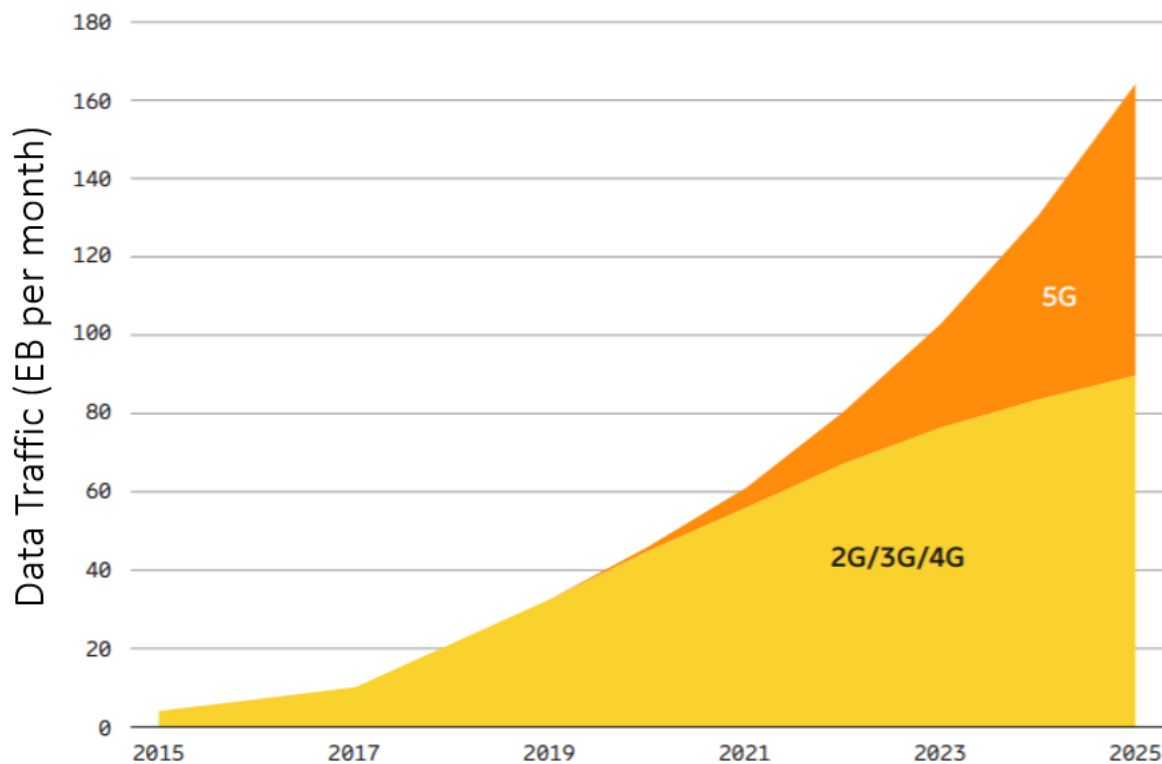


Figure 1.1 – Expected global mobile data traffic evolution (extracted from [Eric20]).

The existence of these capabilities (videoconferences and high-definition videos, among many others) became possible with the 4th generation (4G) of mobile communication systems. The popularisation of this technology led to the increase of the data consumption every year, mainly because of 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. 5G networks will be different from traditional 4G ones, allowing a greater optimisation of network traffic and a smoother handling of usage spikes. These objectives can be achieved by increasing the system data rate and bandwidth, otherwise congestions will continue to occur. Another limitation of 4G is its latency, because there are many services that demand very low delays, otherwise they cannot be performed, such as remote surgery or self-driving cars.

The 5th generation of mobile communication systems (5G) will be implemented with 3 main objectives: increase data rates, reduce latency and increase capacity. The increase in the data rate of mobile

communication systems to about 10 Gbps [Thal20] will allow users to engage in higher demanding activities than the ones that are presently available with the 4G system implementation, since 5G peak data rate is 10 times higher than the previous system.

5G capacity strongly depends on the number of users and on their type of service usage, and the number of users that can be connected to the system simultaneously is expected to be 100 times higher (per unit of area) than the previous system [Thal20]. 5G's lowest E2E planned latency is 1 ms [Thal20], which is more than 10 times lower than the previous implemented system. 5G specification requirements are represented in Figure 1.2.

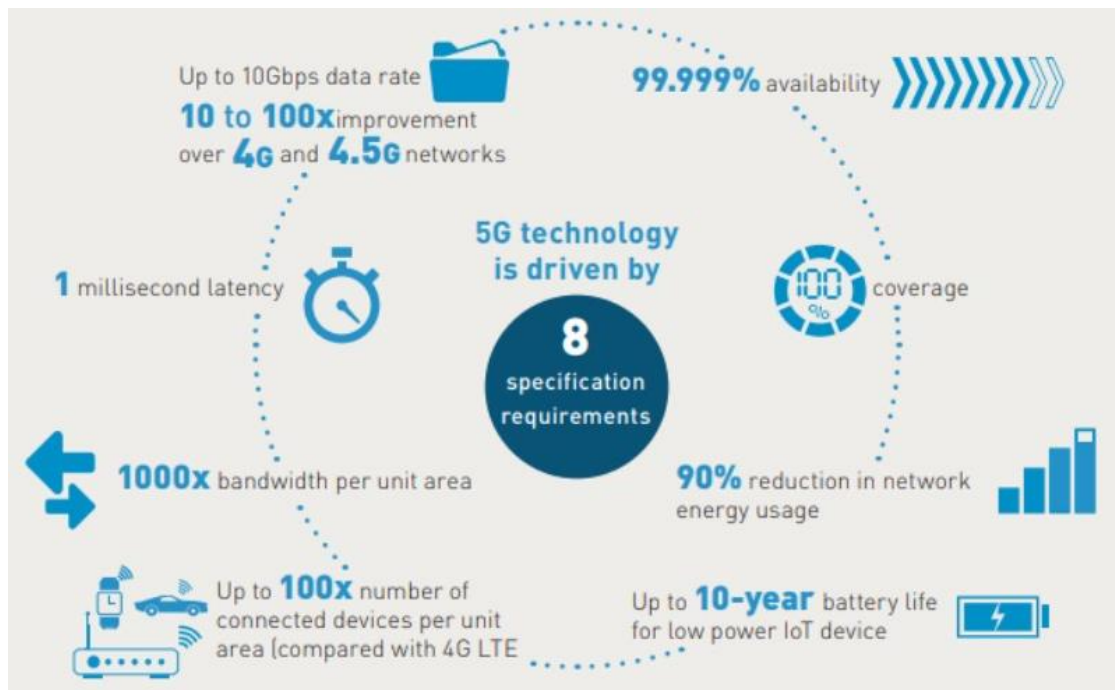


Figure 1.2 – 5G specification requirements (extracted from [Thal20]).

Latency is measured as the delay in packet transmission, propagation, queuing and processing, and it is one of the important parameters to take into account in mobile communications. Nowadays, it is almost a requirement to have real-time communications between the client and the server, and this makes the problem of reducing this parameter a substantial subject. Low delays achieved by the development of 5G mobile networks will open the way to radically new experiences and opportunities, including mobile gaming, virtual reality experiences, factory robots, self-driving cars and other applications. This delay can be caused by many factors such as network protocols, hardware, location of servers, etc. Typically, network latency is measured from the device, up to the radio, down to the baseband processor, then into the core, and further 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 the Figure 1.2. 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 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 need for

moving the cloud, or more precisely, extending it to the edge, is indisputable in 5G very low latency and real time ecosystems. Furthermore, delay-sensitive critical communications and traffic safety impose additional complexities, as they require top-level reliability and availability while ensuring very low latencies. Edge Computing emerges as the core solution to these problems, as it minimises the delay accumulated along the network by placing the service infrastructure at the network edge. A representation of the location of Edge Computing nodes is represented in Figure 1.3.

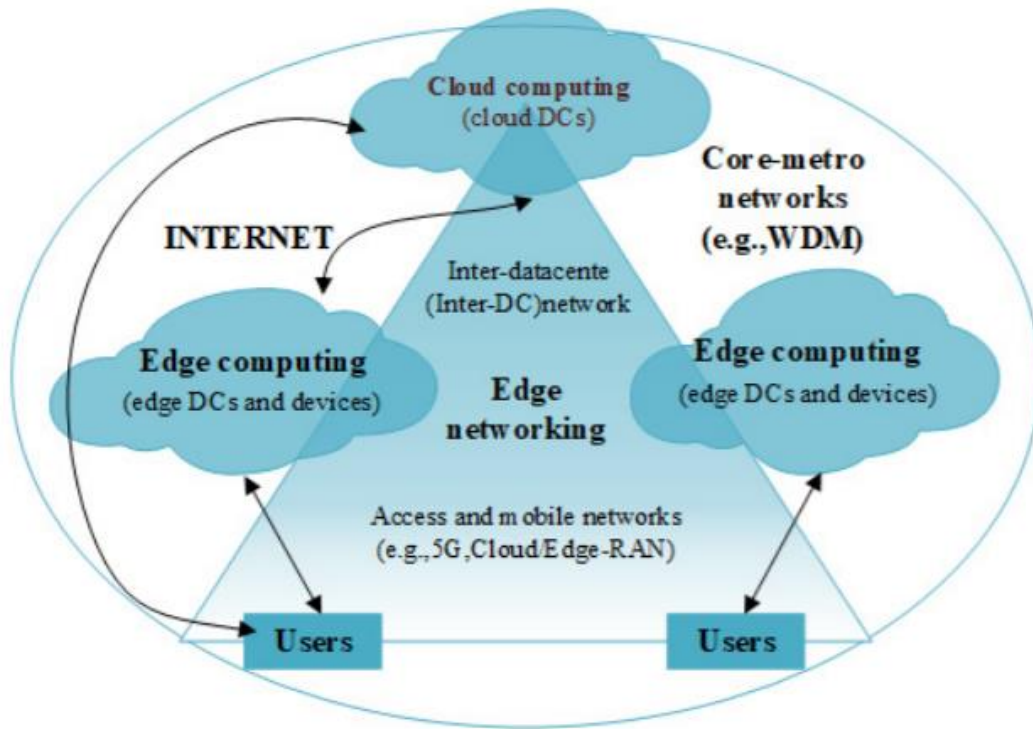


Figure 1.3 – Edge Computing (extracted from [YZWW19]).

1.2 Motivation and Contents

Mobile communication systems are evolving fast, and one of the objectives of the emerging 5G system is to allow the existence of mobile network supported URLLC (Ultra Reliable Low Latency Communication) services. The latency reduction will not only allow the emergence of new services but also improve the QoS of the already existent ones. Having a real-time connection between users and servers is almost a requirement to allow delay-sensitive activities, like factory robots, self-driving cars and other applications for which a quick response is not optional at all but rather a strong prerequisite. Some of the already available services that will be improved with the latency reduction are online gaming, video chatting and web browsing.

This thesis is motivated by the need to study methods to reduce latency in 4G and 5G networks. Having this purpose into account, a mathematical model to account for all latency contributions along the 4G and 5G systems was developed and implemented so that the different strategies to minimise E2E latency can be tested and evaluated. It is also important to verify if the radio techniques present in both

systems are enough to guarantee the required usage throughputs, so that latency can be kept at low levels. The model also analyses the latency performance of the network depending on the node positioning and the deployment options of the Edge technology.

The presented work consists of five chapters. The first one, which contains an introduction to the work, and the motivation behind the thesis along with its contents.

The second chapter contains the fundamental concepts that are crucial for the development of the model, such as the 5G and 4G architectures, the 5G radio interface, the Network Slicing and Virtualisation concept, the Cloud and Edge Networks, the eMBB, URLLC and mMTC services and the comparison between the 4G and 5G transport networks, as well as the network latency characterisation. In the end of the chapter, one provides the State of the Art with the references and explanations of the important literature and documents for the model and thesis development.

The third chapter contains the model definition and development. The input and output parameters are defined, as well as the intermediate parameters and calculations that lead to the outputs. It is also in this part of the document that the network architectures are detailed, along with the latency contributions and throughput calculation expressions. At the end of the chapter, one shows the assessment tests used to validate the model.

The fourth chapter presents the results and contains their analysis. In the beginning, it gives a description of the simulated scenarios, and afterwards, the outputs of the program are analysed for each simulation.

The last chapter summarises the thesis conclusions and presents some suggestions and ideas for the future work.

Chapter 2

Fundamental Aspects

This chapter provides an overview of the 4G and 5G systems. The topics approached are the network architecture, the radio interface, network slicing and virtualisation, cloud and edge networks, services and applications, the latency in the network, ending with the state of the art.

2.1 5G aspects

2.1.1 Network Architecture

5G (NR) systems are being implemented in 2 versions: the Non-standalone and the Standalone ones, both being represented in Figure 2.1.

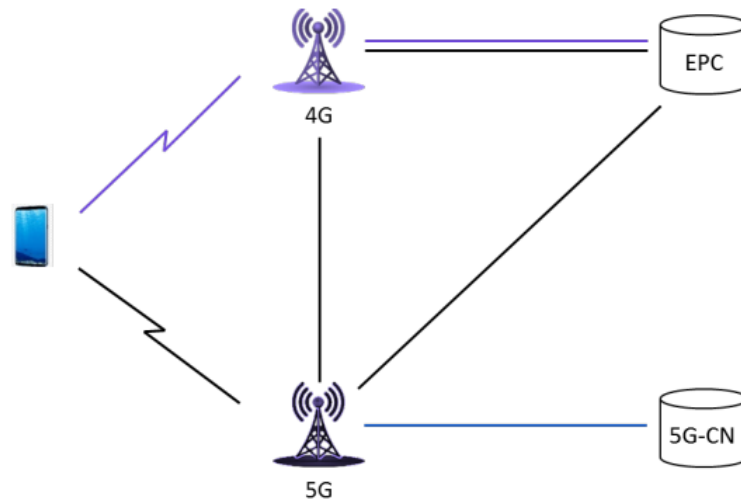


Figure 2.1 – 5G Non-standalone and Standalone versions (extracted from [ARTM19]).

5G implementations are divided in two main deployment versions [ARTM19]:

- The NSA (Non-standalone) – the 5G architecture is built upon an existing 4G infrastructure. In this version, the radio access network is not shared by 4G and 5G systems, but the architecture uses the same Core Network from 4G (LTE). In this deployment, eNB nodes (4G base stations) are connected to gNB nodes (5G base stations).
- The SA (Standalone) – the 5G architecture is not built upon an existing 4G infrastructure. The radio access network and the core network use only 5G systems.

Figure 2.2 presents the Network Architecture of 4G.

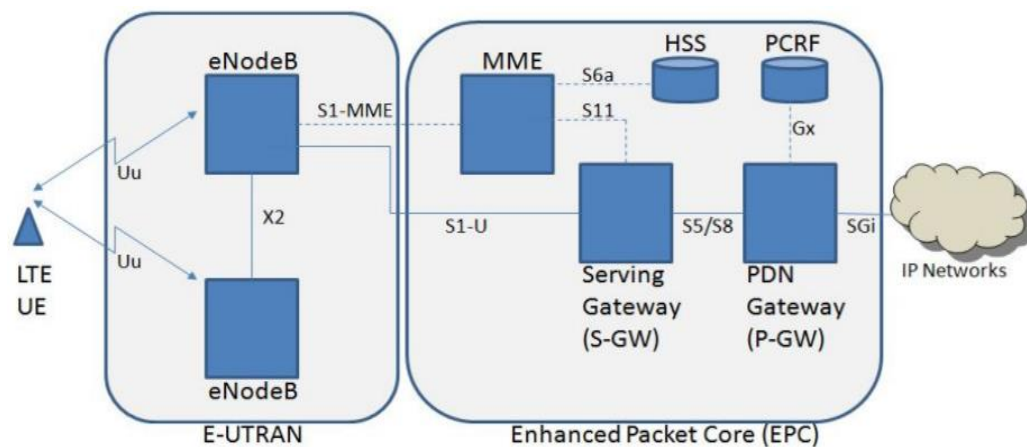


Figure 2.2 – LTE Network Architecture (extracted from [ITUT19]).

The LTE Network Architecture is divided in 2 main components [ITUT19]: the radio network Evolved-UTRAN (E-UTRAN) and the core network EPC (Evolved Packet Core). The E-UTRAN consists of a set

of base stations, called eNodeB (eNB), which are directly connected to the UE (user equipment) via radio interfaces. The eNBs are interconnected by the X2 interface, but also connected to the EPC, because these nodes are responsible for sending the data traffic to the Core Network. Therefore, base stations are connected to the MME (Mobility Management Entity) through the S1-MME interface, and to the S-GWs (Serving Gateways) through S1-U interface.

The EPC is composed of the following nodes with their functionalities and interconnections identified:

- Home Subscriber Server (HSS), which consists of the subscribers' database that contains information about subscription and location of the home operator's mobile users. It also contains the necessary parameters to authenticate and communicate with users.
- Mobility Management Entity (MME), which consists of the main control node in LTE responsible for handling several control functionalities, especially security ones. The MME interacts with the HSS to authenticate users and then handles the security concepts of each one. The MME uses interfaces S11 and S6a to establish a connection with the SGW and HSS nodes, respectively.
- Serving Gateway (S-GW) has the function of managing the user plan mobility by acting as a router of user data between the eNB and the P-GW. The S-GW is connected to the eNB via the S1-U interface and to the PG-W via the S5/S8 one.
- Packet Data Network Gateway (P-GW) is the contact point of the EPC subsystem with the outside packet networks in order to exchange the users' traffic. It is connected to the PCRF via the Gx interface and to IP Networks via the SGi interface.
- Policy and Charging Rules Function (PCRF) determines the policy and charging rules for each service data flow and decides how to handle services in terms of Quality of Service (QoS).

Figure 2.3 shows 5G standalone Network Architecture.

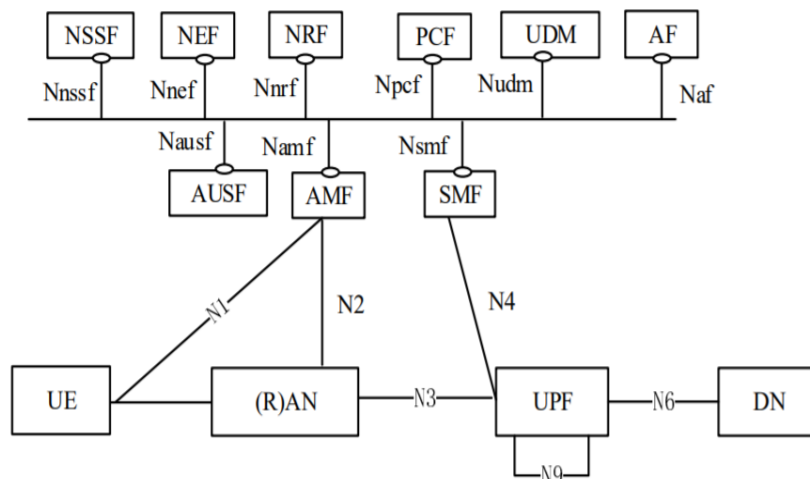


Figure 2.3 – 5G Network Architecture (extracted from [3GPP19a]).

The Standalone version of the 5G system has a different Network Architecture. The UE is connected to the core network through the NG-RAN (New Generation Radio Access Network) via gNBs (radio network node for 5G). The 5G architecture consists of the following network functions [3GPP19a]:

- Access and Mobility Management Function (AMF) manages registration, connection and reachability. This module is responsible for the access, authentication and authorisation

processes, and it is in charge of managing handovers between gNBs, within the NG-RAN. It also marks the termination of the RAN interface (N2).

- User Plane Function (UPF) is responsible for packet routing and forwarding, packet inspection, QoS handling, external PDU session for interconnecting DNs (Data Networks, e.g., operator services, Internet access or third-party services), and it is the anchor point for Intra/Inter Radio Access Technology (RAT) in the 5G architecture.
- Authentication Server Function (AUSF) supports authentication for 3GPP access and untrusted non-3GPP access.
- Application Function (AF) interacts with the 3GPP Core Network via the NEF in order to access network.
- Network Exposure Function (NEF) stores information as structured data using a standardised interface (Nudr) to the Unified Data Repository (UDR) and handles masking of network and user sensitive information to external AFs according to the network policy.
- Network Slice Selection Function (NSSF) selects the set of Network Slice instances serving the UE, and determines the AMF set to be used to serve the UE, or based on configuration, a list of candidate AMFs, possibly by querying the NRF.
- Network Repository Function (NRF) is one of the key components of the 5G Service Based Architecture. The NRF maintains an updated repository of all 5G elements in the network along with the services provided by each of the elements in the 5G core. It is responsible for maintaining the profiles of the network function instances and their supported services, and allows network function instances to track the status of other network function instances.
- Policy Control Function (PCF) supports unified policy framework to govern network behaviour, provides policy rules to Control Plane functions to enforce them, and accesses subscription information relevant for policy decisions in a Unified Data Repository (UDR).
- Session Management Function (SMF) is responsible for allocating the IP addresses for the UE and controlling the packet data unit session.
- Unified Data Management (UDM) is responsible for handling the user identification, access authorisation based on subscription data (e.g., roaming restrictions), UEs Serving Registration Management (e.g., storing serving AMF for UE), and also takes care of the subscription and SMS management.
- Unified Data Repository (UDR) is responsible for storing data such as: subscription data, policy data, structured data for exposure and application data.

2.1.2 Radio Interface

The 5G radio interface uses both duplexing modes: FDD (Frequency Division Duplexing) and TDD (Time Division Duplexing). FDD has a lower latency in comparison with TDD, since in the former both UL and DL are transmitting and receiving simultaneously, while in latter simultaneous transmissions are not possible.

5G spectrum is divided into 3 main categories: low-bands (below 2 GHz), mid-bands (from 2 GHz up to 8 GHz) and high-bands (above 24 GHz). Low FDD bands (600 MHz, 700 MHz, 800 MHz, etc.) can be used in combination with mid-bands to provide wider and deeper indoor 5G coverage, and to improve latency performance. Mid-bands can meet eMBB (enhanced mobile broadband) service requirements for the initial stage of 5G, and therefore the release of these bands as the first step is essential for 5G development and business success. They are also crucial to support most 5G usage scenarios in wide-areas, and therefore the 1.5 GHz up to 2.1 GHz bands may be used in both FDD and TDD. The unpaired TDD bands such as 2.4 GHz and [3.3, 4.2] GHz deliver wide-area coverage and high capacity. Comparing with mid-bands, the high frequency bands suffer from poor radio propagation (these bands have a very high free space attenuation) and significant outdoor to indoor penetration loss (these bands are easily blocked by buildings), which make continuous citywide coverage not economically viable. Thus, the mm Wave usage scenarios are limited to fixed wireless access (FWA) and hotspot coverage, which do not require continuous national coverage [HUAW20]. 5G uses OFDMA (Orthogonal Frequency Division Multiple Access) as multiple access technique for DL (Downlink) and SC-FDMA (Single Carrier Frequency Division Multiple Access) for UL (Uplink) [Corr20].

In contrast with LTE physical channels NR ones are flexible and scalable, because they were developed to adapt to certain requirements, such as latency, data rate and reliability (which may vary depending on the service). Therefore, the subcarrier spacing (numerology configuration bandwidth) of 5G is given by [3GPP19b]:

$$\Delta f_{[\text{kHz}]} = 2^{\mu} \times \Delta f_{\text{ref} [\text{kHz}]} \quad (2.1)$$

where:

- μ is the subcarrier spacing configuration parameter (integer number), i.e., the numerology;
- Δf_{ref} is the reference frequency of 15 kHz (subcarrier spacing when μ is 0).

The lower- and mid-bands of the spectrum are already being used in previous systems, therefore, there is not a wide band spectrum left in this band. This spectrum limitation requires the use of narrower subcarrier spacing in this frequency ranges (600 MHz to 6 GHz), such as 15 kHz and 30 kHz.

When mobile communication systems use high frequency bands the Doppler Effect increases, which causes a higher inter-carrier interference and reduces the data rates when the terminal is moving. This motivates the use of wider subcarrier spacing in the higher frequency bands. In this case, 5G uses a 120 kHz and 240 kHz subcarrier spacing for the high frequency bands. Table 2.1 shows a list with some of the frequency bands for 5G in Europe.

By increasing the subcarrier spacing configuration parameter, the process of sending and receiving slots becomes faster [3GPP19b], because the frequency band available to send slots is wider, resulting in the reduction of latency. For example, if it is required to support an URLLC (Ultra Reliable Low-Latency Communication) application, the system can use a higher subcarrier spacing at a certain frequency band. On the other hand, if the service to be provided does not require very low latency, the system can use a lower subcarrier spacing. This allows 5G to support different frequency bands and heterogenous services in a cost-effective and efficient way. In Table 2.2, one presents a list of the numerologies that

are supported by 5G, and the relation of the subcarrier spacing configuration parameter with the number of slots per frame and subframe.

Table 2.1 – NR Frequency Bands List (adapted from [5GOB20]).

Band	Band [MHz]	Uplink [MHz]	Downlink [MHz]	Duplex Mode	Region
N1	2 100	1 920 – 1 980	2 110 – 2 170	FDD	Global
N3	1 800	1 710 – 1 785	1 805 – 1 880	FDD	Global
N7	2 600	2 500 – 2 570	2 620 – 2 690	FDD	EU and others
N8	900	880 – 915	925 – 960	FDD	Global
N20	800	832 - 862	791 - 821	FDD	EU and others
N28	700	703 - 748	758 - 803	FDD	EU
N34	2 000	2 010 – 2 025		TDD	EU and others
N38	2 600	2 570 – 2 620		TDD	EU and others
N41	2 500	2 496 – 2 690		TDD	Global
N78	3 500	3 300 – 3 800		TDD	Global
N257	28 000	26 500 – 29 500		TDD	Global
N258	26 000	24 250 – 27 500		TDD	Global
N259	41 000	37 000 – 40 000		TDD	Global
N260	39 000	27 500 – 28 350		TDD	Global

Table 2.2 – Supported Transmission Numerologies and Frame Structure (adapted from [3GPP19b]).

μ	$2^\mu \times 15$ [kHz]	Cyclic prefix	Number of slots per frame (10 ms)	Number of slots per subframe (1 ms)
0	15	Normal	10	1
1	30	Normal	20	2
2	60	Normal, Extended	40	4
3	120	Normal	80	8
4	240	Normal	160	16

In Figure 2.4, one illustrates the process described in Table 2.2. As the subcarrier spacing configuration parameter increases by 1 unit (going from the bottom to the top of the figure) the time taken to send a slot gets reduced by half.

The cyclic prefix protects from inter-symbol interference by acting as a guard interval between symbols. After sending a symbol, the end part of that same symbol is retransmitted (cyclic prefix) acting as a buffer region between both symbols. This provides robustness to the signal, since the data that is retransmitted can be used if required, and also makes the system more reliable because of the inter-

symbol interference reduction. Latency reductions require reductions in the length of the symbols, and the addition of the cyclic prefix adds latency to the transmission. The cyclic prefix represents approximately 7% of the symbol duration, and since the symbol duration changes according to the subcarrier spacing, the cycle prefix also varies with the subcarrier spacing [AZRB17]. The relation between subcarrier spacing and the cycle prefix is evidenced in Figure 2.5.

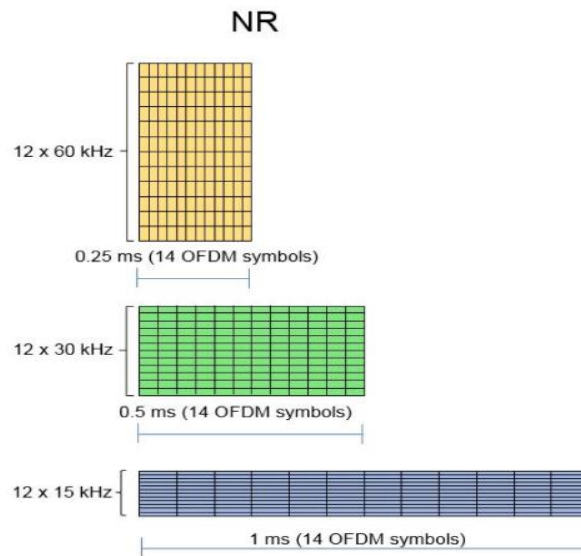


Figure 2.4 – 5G Frame Structure (adapted from [GMFF20]).

Subcarrier spacing	15kHz	30kHz (2 x 15kHz)	60kHz (4 x 15kHz)	15×2^n kHz, ($n = 3, 4, \dots$)
OFDM symbol duration	66.67 μ s	33.33 μ s	16.67 μ s	$66.67/2^n$ μ s
Cyclic prefix duration	4.69 μ s	2.34 μ s	1.17 μ s	$4.69/2^n$ μ s
OFDM symbol including CP	71.35 μ s	35.68 μ s	17.84 μ s	$71.35/2^n$ μ s
Number of OFDM symbols per slot	7 or 14	7 or 14	7 or 14	14

Figure 2.5 – OFDM cyclic prefix variation with subcarrier spacing (adapted from [AZRB17]).

Another important parameter related to the time structure of both 4G and 5G networks is the TTI (Transmission Time Interval), which corresponds to the duration of a transmission in the radio link. 5G needs to enable a scalability of latencies to lower values than the previous system, which is achieved by using a flexible frame structure, where the TTI can scale up and down depending on the expected service and its respective required latency [DMAG18]. With this flexibility the air interface can present lower latencies by using a shorter TTI, since this will allow the transmission to begin as soon as the previous transmission is completed, instead of having to wait until the next subframe. Figure 2.6 is a representation of the TTI scalability that is present in 5G.

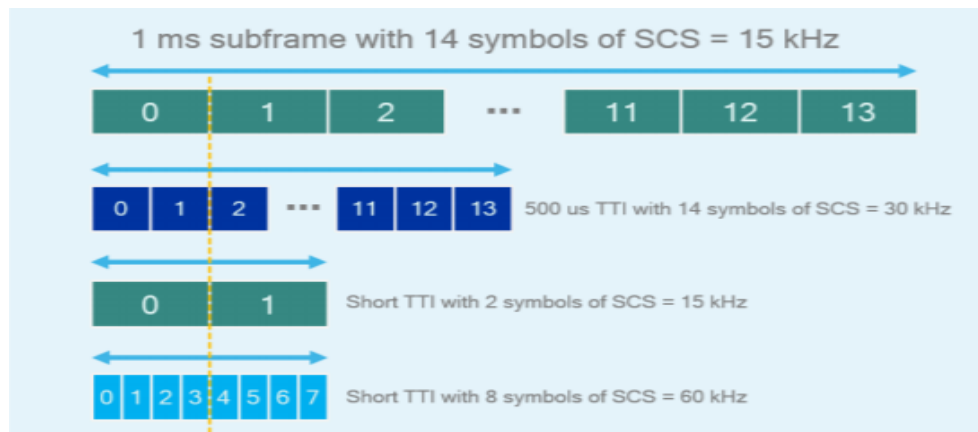


Figure 2.6 – TTI scalability (extracted from [Qual16]).

2.2 Network Slicing and Virtualisation

A SDN (Software Defined Network) is a network architecture approach that enables the network to be controlled by software applications, and it represents a key aspect in virtualisation and network slicing. As a consequence of the SDN technology, the control of the network can reside beyond the devices that provide physical connectivity (by decoupling the hardware from the software) and operators can control the network to provide different services, or even individual customer services. The SDN technology has been used to upgrade the quality of the data centres performance, because the software applications have been evolving throughout the years, allowing these applications to define certain network functionalities.

SDN architectures consist of 3 main layers [VGMS12]:

- The application layer, which is responsible for containing the programs and software required to control the network. Many types of applications can be used in this layer, such as network automation, network configuration and management, and network monitoring.
- The control/orchestration layer, which creates the connection between the application and the infrastructure layers. It is responsible for processing the instructions and requirements created in the application layer, and to proceed them to the networking components.
- The infrastructure layer, which consists of the network devices/elements that control the forwarding and data processing capabilities of the network. It represents the physical layer that handles packets based on the control layer instructions.

Figure 2.7 illustrates the layer representation of a SDN and the interconnections between layers.

The NFV (Network Function Virtualisation) is a network architecture concept that is used to bring the network to a virtual level. The difference between SDN and NFV is that the former revolves around automation and software to decouple the hardware from the software while the latter revolves around the virtualisation of network components. Figure 2.8 represents the structure of an NFV architecture.

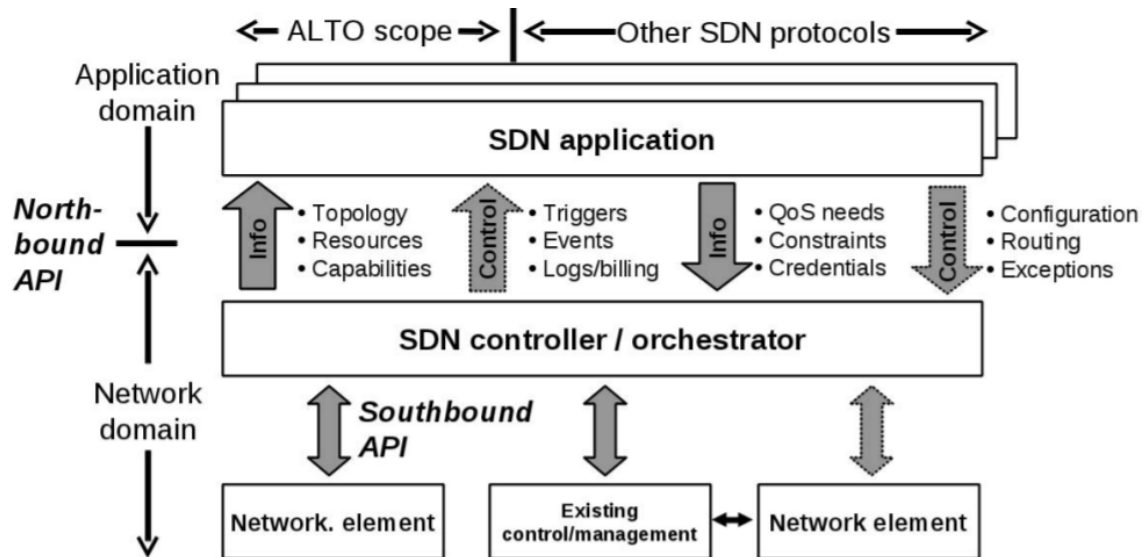


Figure 2.7 – SDN Layers (extracted from [VGMS12]).

The NFV architecture is divided in 4 layers [HMRS16]:

- The VNF (Virtualisation Network Function) layer, which is composed of the Virtualisation Network Function and by the EMS (Element Management System). It is the basic block of a NFV architecture. VNFs are virtualised networking services running on open computer platforms and can replace physical hardware.
- The NFVI (Network Function Virtualisation Infrastructure) layer, which represents the totality of the hardware and software resources in which the VNF are deployed, managed, and executed.
- The OSS (Operation Support Subsystem) and BSS (Business Support System) layer, in which the OSS deals with network management, fault management, configuration management and service management and the BSS takes care of the customer management, product management and order management, amongst other functionalities.
- The MANO (Management and Orchestration) layer, which includes 3 components: the virtualised infrastructure manager, the orchestrator and the VNF manager. The virtualised infrastructure manager performs the inventory of the software, computing and storage, among others. The VNF manager is responsible for the VNF lifecycle management which includes installation, updates, up and down scaling and termination. The orchestrator orchestrates and manages infrastructure and software resources.

5G network slicing is based on the partition of the physical network into multiple independent virtual slices that provide different quantity of resources to different traffic types, thereby adjusting the network to meet the heterogenous services that NR is supposed to provide. The way to achieve a sliced network is to transform the network into a set of logical networks on top of a shared infrastructure with the processes that were described before. 5G has a large application range, and the specifications for each service differ from each other. To handle this requirement, 5G needs to provide service differentiation, flexibility, and better resource efficiency than 4G. A possible solution for this problem is network slicing [MIW15]. Figure 2.9 illustrates the division of 3 slices in 5G.

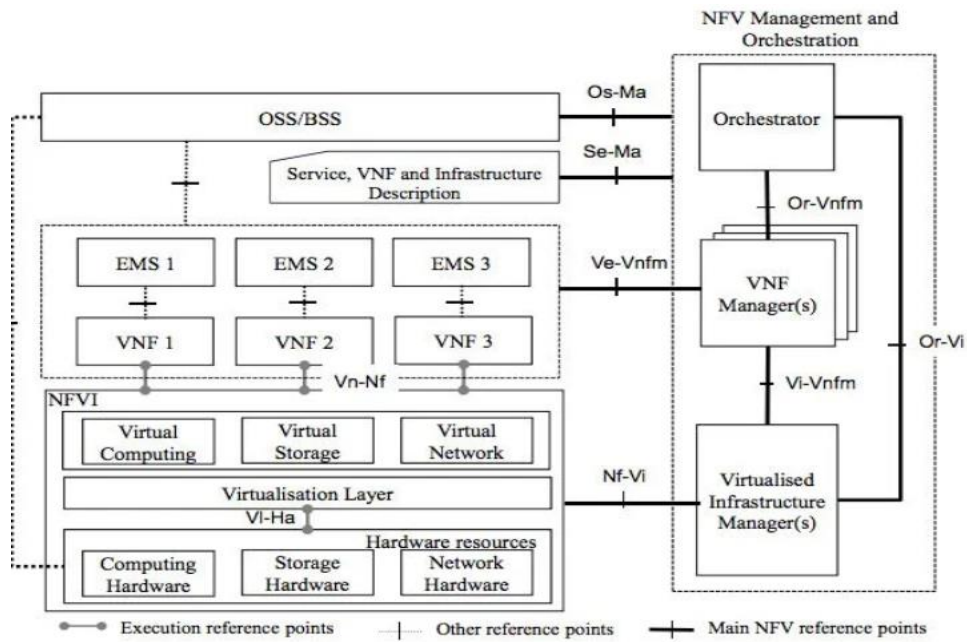


Figure 2.8 – NFV Architecture (extracted from [ETSI13]).

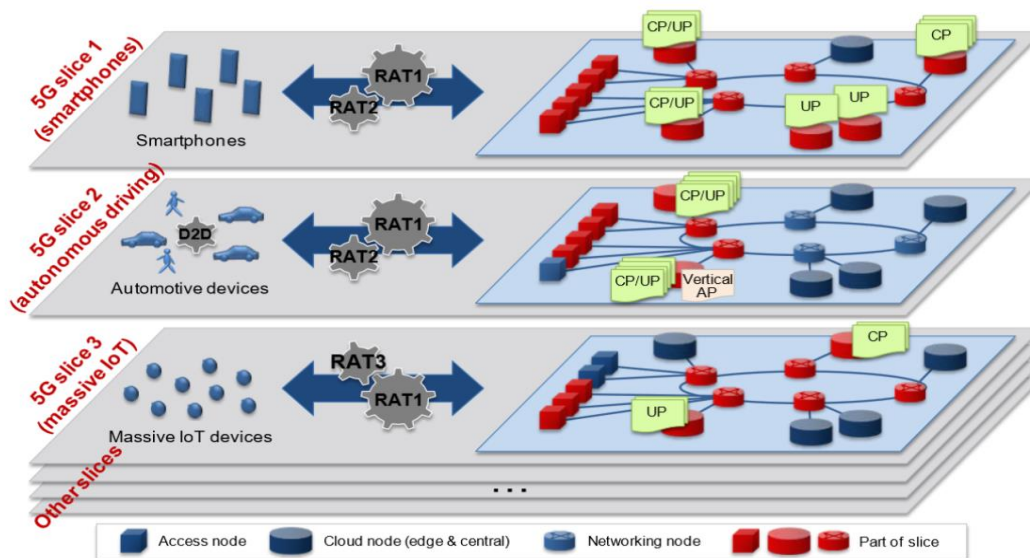


Figure 2.9 – 5G Network Slicing (adapted from [MIIW15]).

2.3 Cloud and Edge Networks

2.3.1 Cloud Network

The Cloud Radio Access Network (C-RAN) is a centralised cloud computing-based architecture for radio access networks that enables large-scale deployment, collaborative radio technology support, and real-time virtualisation capabilities.

It is also important to refer that, apart from centralising, C-RANs also take a part in the process of

virtualising the network. In the C-RAN architecture, the traditional BSs are functionally separated into two parts: the BBU (Base-band Unit) and the RRH (Remote Radio Head) [KMNT18]. The RRH is responsible for the transmission and the reception of radio signals, and their filtering and amplification. The RRHs are grouped into clusters and assigned to one BBU through a Fronthaul link (typically an optical fibre, since it meets both the demanding bandwidth and latency requirements). The BBUs are responsible for the generation and processing of digitised baseband signals, and they are clustered in a centralised data and processing centre (the BBU pool), which provides high computational and storage capabilities.

C-RAN has the capability to reduce latency in performing various operations. For instance, the time needed to perform handovers can be reduced (as it can be done inside a cloud instead of between BSs) as well as the failure rate in handover and the fronthaul link that connects the RRHs with the BBUs can be adapted in order to achieve lower latencies, while still guaranteeing the required data rates.

The BBUs and RRHs can perform 3 main layer functions: the physical layer (radio) functions, the media access control layer functions and the network layer (routing) functions. The distribution of these functions must balance the advantages and disadvantages of assigning certain responsibilities to the BBU or the RRH. Therefore, 2 types of C-RANs emerged: the Fully and the Partially Centralised ones.

The C-RANs can be divided into 2 types [HMRS16]:

- The Fully Centralised type assigns all functionalities of the physical layer, the media access control layer and the network layer to the BBU. Although this structure can benefit from easy resource sharing between BBUs and easy maintenance, this solution requires a high bandwidth connection between the BBU and the RRH. This C- type creates strict requirements in terms of bandwidth and latency, resulting in an optical fronthaul that must be carefully planned and dimensioned.
- The Partially Centralised type has the physical layer functions performed by the RRHs and the media access layer and the network layer functions by the BBUs. The great disadvantages of using this type of C-RAN are that it adds complexity to the RRH, resource sharing becomes considerably reduced and advanced features cannot be effectively supported. On the other hand, it reduces the burden in terms of bandwidth on the optical transport links.

Figure 2.10 represents a Cloud Radio Access Network.

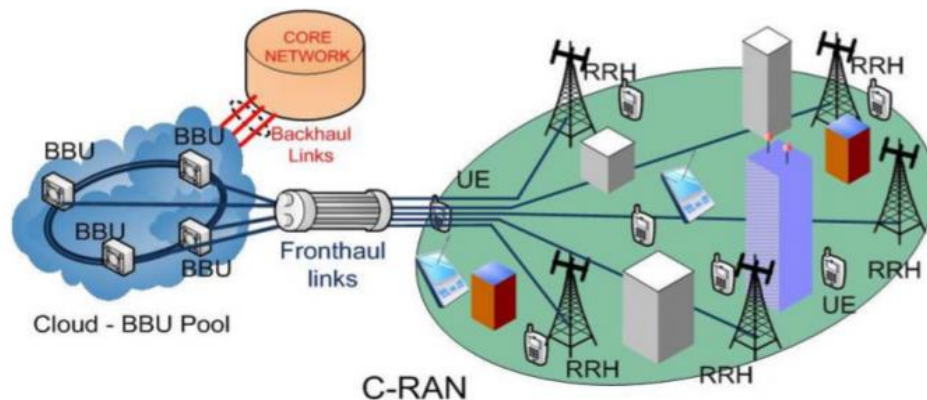


Figure 2.10 – Representation of a C-RAN (extracted from [KMNT18]).

C-RANs' implementation has many advantages such as [SeDo19]:

- Flexibility and scaling, since BBUs are dynamically scaled according to network requirements; if the network traffic increases, a virtual BBU can use more computing resources, and the network is also able to solve the opposite situation by using less resources. Since there are 2 types of C-RANs, there are also different solutions that provide flexibility according to network requirements.
- Lower Latencies, because the general amount of signalling information sent to the core network can be reduced in C-RANs (due to the centralisation in the BBU pools), and therefore the latency gets reduced. The solutions found to provide this latency reduction put a lot of pressure in the fronthaul of the network, but a well dimensioned and planned solution can reduce the impacts of this problem.
- Energy efficiency and power cost reduction, since the process of centralising the network might reduce the number of base station sites, also reducing the air conditioning and other on-site power-consuming equipment.
- Convenience of operation, installation, and maintenance, since there is a centralisation of functionalities in the BBU all these activities are performed in the same place.
- Increased Security, because since the operations are centralised, by protecting efficiently the BBU pools it protects the network from cyber-attacks.

2.3.2 Edge Network

Edge Computing (EC) is a computer paradigm that provides computer and storage resources by bringing the network closer to end users [NGKA19]. EC is a technique that proposes several geographically distributed nodes that are located closer to the end user than data centres, reducing the number of processing nodes and also the physical distance between the user and the network, and the required data that is brought closer to the edge of the network. Therefore, 5G uses this type of network in order to be able to fulfil the heavy requirements of services to be provided, namely those with ultra-low latency requirements, making 5G Edge Networking a viable solution for the problem in hands. Cloud and Edge Networks can be used together for a better performance of the network, and that is how operators anticipate taking advantage of both architectures. Logically, MEC hosts are deployed in the edge or central data network, and it is the User Plane Function (UPF) that takes care of steering the user plane traffic towards the targeted MEC applications in the data network. Figure 2.11 is a representation of Edge implementation in 4G and 5G.

Figure 2.12 represents the possible Edge Nodes deployment in the 4G C-RAN. There are 4 possible Edge node deployments with different latency impacts [ASCC18]:

- With the deployment of the Edge node between the Core and the Data Centre, the latency can be reduced due to the reduction in the traffic centralisation, meaning that the Edge technology adds processing capacity to the network. If the Edge node is physically closer to the Core of the Network than the Data Centre this deployment causes a reduction of the propagation delay, culminating in a lower E2E latency.

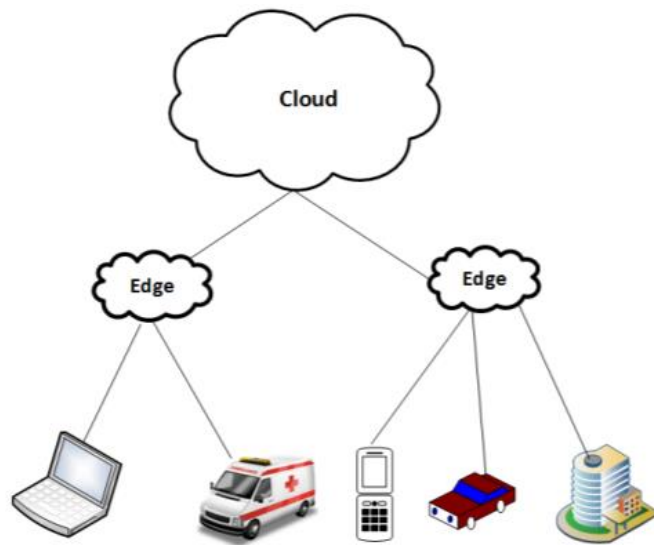


Figure 2.11 – 5G and 4G Edge Networking (extracted from [NGKA19]).

- The scope A, with the deployment of Edge nodes between the Core and the BBU (the Core components are virtualised and deployed in a distributed manner within the Edge node set), both the physical and logical distance between end users and the needed resources are reduced, since the Edge technology can aggregate the Data Centre and the Core functionalities in a single node. This implementation represents a solution with low reductions in terms of the latency, since the only delay contributions that are effectively eliminated are the ones associated with the Core Network, the Data Centre itself and the propagation between both nodes. Within this scope, from a functional viewpoint, the communication between the end user and the Edge Node could occur without the interference of any Core entities, thereby excluding most of the EPC/5G-Core associated delay. The Fronthaul and the Backhaul can be optimised to the circumstances to reduce the latency and still comply with the required data rate.
- The scope B, with the deployment of the Edge nodes comprising the Data Centre, the Core Network components and the BBU. In this deployment case, the latency associated with the Backhaul and the latency associated with the core network and the data centre interconnection are minimised. Only the Fronthaul latency can be optimised through an efficient Edge Node location. This scope is the most cost-effective deployment scheme regarding latency optimisation, since it has a lower cost than the scope C and a higher latency reduction than the scope A solution.
- The scope C, with the deployment of the Edge nodes comprising the data centre, the Core Network components, the BBU and the RRH. Since this strategy uses a pure co-location, it is not possible to adapt any part of the network to optimise it, like in scopes A and B. This deployment offers the best latency reduction possible from the studied deployments, but it has high costs and does not comply with the scalability required. Therefore, this deployment is not considered as a feasible option to implement.

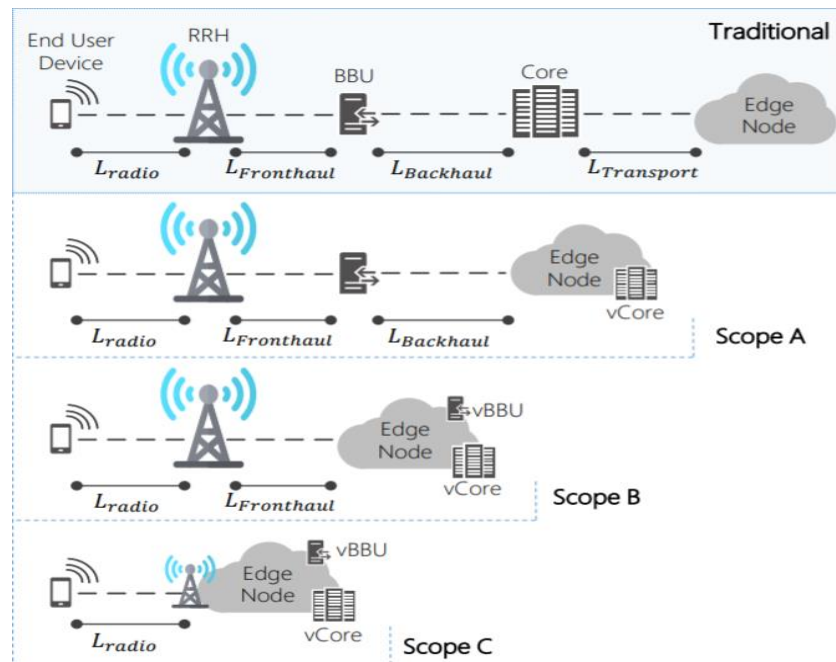


Figure 2.12 – Edge Deployment Scenarios for Latency Optimisation (extracted from [ASCC18]).

2.4 Services and Applications

5G is being developed in order to guarantee 3 main types of services, with different requirements: eMBB, URLLC and mMTC [ITUT17]:

- Enhanced mobile broadband (eMBB), through massive MIMO and mm waves, 5G will achieve high bit rates, offering users new services, such as Augmented and Virtual Realities.
- Ultra-reliable low-latency communications (URLLCs), which is based on services that require very low latencies, such as factory automation, autonomous driving, industrial internet and robotic surgeries. These services require latencies below 10 ms [Thal20].
- Massive machine-type communications (mMTCs), which provides connection to a large number of devices that intermittently transmit small amounts of traffic. 5G will have to support a high density of devices (200 000 devices per km²) using services with low data rates (1-100 kbps).

The services provided can be classified according to 4 classes:

- Conversational Services, such as voice calls and real-time multimedia messaging, which demand low latencies due to the bidirectional flow of data.
- Streaming Services, which is an alternative to the download of data, consisting of performing activities that require data transmission in real-time, like music listening or video watching.
- Interactive Services, which are services in which the user can interact directly with the application, such as online gaming and web browsing.
- Background Services, which are characterised as the process of sending and receiving data in the background, not requiring an interface for the user. Some examples of this class of services are SMS, e-mail and databases download.

Figure 2.13 represents the 3 main directions of 5G.

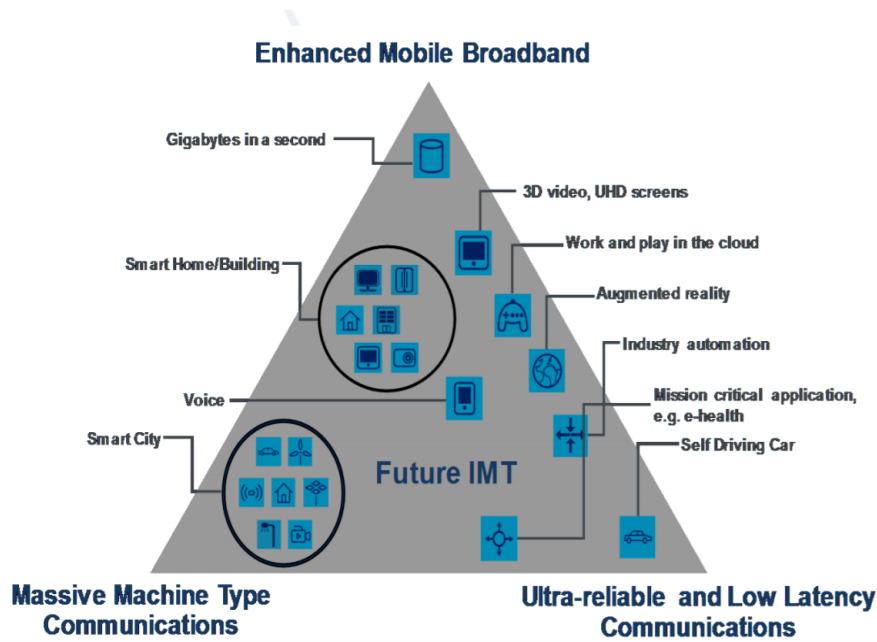


Figure 2.13 – 3 main directions of 5G (extracted from [ITUT17]).

5G presents an enlargement of the provided services comparing with 4G. Higher data rates, lower latencies among other characteristic improvements are essential to achieve the expected goals. Table 2.3 represents some of the services that 5G provides to users. Applications in which latency is critical are included in the range of the services provided by the evolution that 5G present in comparison with the previous mobile system generations.

Table 2.3 – Expected 5G services characteristics (adapted from [3GPP19c]).

Service Type	Use Case	Service Class	Latency [ms]	Data Rate [Mbps]	Source
URLLC	Vehicle Automation	Streaming Service	<5	5.1-3500 (Depends on the sensor data)	[3GPP19c]
URLLC	Telepresence	Conversational Service	<100	25	[Qual18] and [3GPP19c]
URLLC	Real-Time Remote Surgery	Streaming Service	<10	Depends on the quality of multimedia device used	[3GPP19c]
eMBB	Virtual reality	Interactive Services	10	50 (depending on the quality)	[Qual18] and [3GPP19c]
eMBB	Augmented Reality for Gaming	Interactive Services	20	50 (Depending on the quality)	[Qual18] and [3GPP19c]
mMTCC	Smart Grid	Streaming Service	5	Depends on the grid characteristics	[3GPP19c]

Table 2.4 lists some of these latency critical IoT applications.

Table 2.4 – Latency Critical IoT application list (adapted from [PSMM17]).

Use Case	Latency [ms]	Update Time [ms]	Device Density	Communication Range [m]	Mobility [km/h]
Factory Automation	0.25 to 10	0.5 to 50	0.33 to 3 devices/m ²	50 to 100	< 30
Process Automation	50 to 100	100 to 5 000	10 000 devices/plant	100 to 500	< 5
Smart Grids	3 to 20	10-100	2 000 devices/km ²	a few m to km	0
Intelligent Transport System	3 to 100	100	3 000 devices/km ²	2 000	< 50
Professional Audio	2	0.01 to 0.5	up to 1/m ²	100	< 5

2.5 Latency in the Network

The E2E latency can be calculated by summing the contributions along the network of [VODA18]:

- The transmission delay, which is created by the process of transmitting bits to links. It is possible to reduce the contribution caused by this delay by increasing the throughput in the system.
- The propagation delay, which is the delay added by the time that a packet takes to travel between two points; in this case, since the problem under study is the end-to-end latency, these points are the origin and the destination. This contribution to the latency can be reduced by shortening the physical distance between the user and the destination.
- The processing delay, which is the time that the system takes to process the header, determine errors and route as well as other processing requirements. This delay can be reduced by using more adequate multiplexing techniques and algorithms to track errors and route properly.
- The queuing delay, which is characterised by the time to wait until the packets that are in the queue are transmitted. This contribution to the latency can be reduced by implementing network slicing and by guaranteeing a good link throughput.

The latency contributions are represented in Figure 2.14.

In order to specify where the latency will be added along the entire network it is important to have a general picture of the end-to-end architecture, because the end-to-end latency is obtained by summing all its contributions along the network.





Contributors		Typical values	Cure
Transmission (Serialization)  Push bits into link ...0110101100010		Bandwidth 1518 byte 9600 byte 	Bandwidth
	1Gbps	12.1 μ s	
	10Gbps	1.21 μ s	
Propagation  Travel from origin to destination		Medium μ s/km km/ms 	Short distance
	Fibre	5	
	MW	3.5	
Processing  Process header, determine errors and route ...0110101100010		10-30 μ s (DWDM, MW) 0...500 μ s (IP-NE)	Better HW Low # of NEs
Queuing  Waiting until the packet gets transmitted		0...10 ms	No congestion !!! → QoS, network slicing, bandwidth

Figure 2.14 – Contributions of the delays to the latency (extracted from [VODA18]).

The 5G C-RAN architecture defines the separation of the baseband functions into 3 main entities (instead of just 2, as in the 4G RAN [NGOF18]): the Centralised (CU), the Distributed (DU) and the Radio Units (RU). This split allows a more flexible deployment of the baseband functions, and an intelligent placement of the CUs and DUs is expected to bring advantages in terms of cost, power consumption and service blocking. The BBU functions described in previous sections split between the CU, the DU and the RU, while the RU keeps the RRH functionalities. According to the 5G transport network architecture, there are 2 main possible cases for the latency contributions: the case in which the needed resources are provided by the MEC (Multi-Access Edge Computing) node, and the case without MEC node in which the contributions for the latency are different than the previous case. This node can be installed between every pair of network nodes, even if they are collocated (e.g., a CU+DU architecture).

Figure 2.15 compares the 4G transport network architecture with the 5G one.

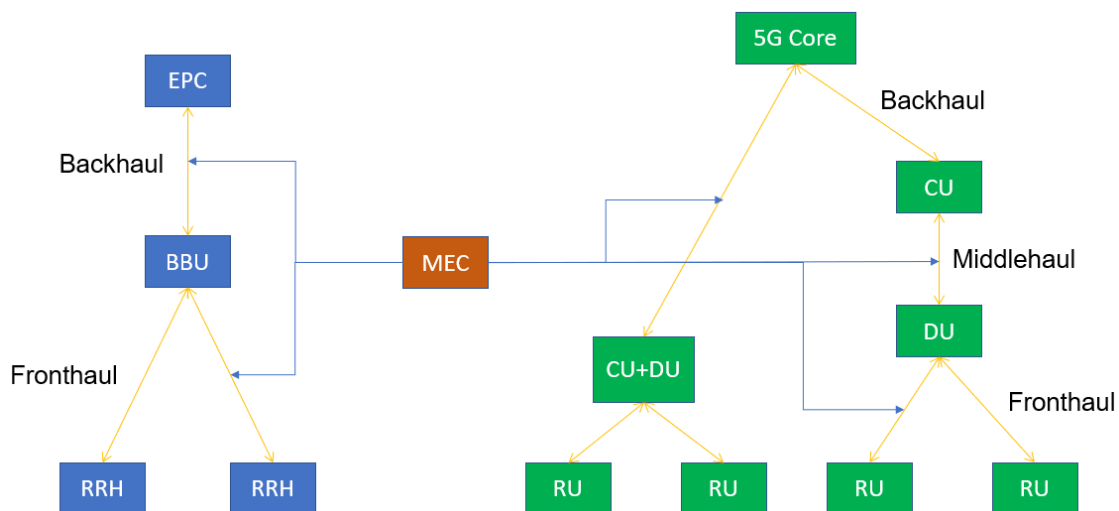


Figure 2.15 – 4G vs 5G transport network architectures (adapted from [NGOF18]).

In the absence of the MEC node, the accumulated latency along the network is divided into the following contributions [SeDo19]:

- Processing in the UE.

- Transmission latency to the links (Air Link, Fronthaul, Middlehaul, Backhaul and Transport Link).
- Propagation latency in the links (Air Link, Fronthaul, Middlehaul, Backhaul and Transport Link).
- Queuing latency in the nodes (Radio Unit, Distributed Unit, Centralised Unit and Network Core).
- Processing latency in the nodes (Radio Unit, Distributed Unit, Centralised Unit, Network Core, and External Data Centre).

For the case with the MEC technology, it is important to understand where the node is installed, because according to its position in the network the number of contributions to the E2E latency may vary. The best case possible from the latency reduction viewpoint is the deployment of the MEC node installed logically as the DU, since the deployment of the node as the RU is not possible due to its high cost and scalability requirements. For the case with MEC, the differences in the latency are:

- The latency accumulated between the RU and the MEC node that may vary according to its position along the network. This technology can prevent the process of sending packets forward in the network, reducing the queuing, processing and transmission delays of the nodes that are installed after it.
- The MEC node processing delay, which is used to account for the time that the node processor takes to process the received data.

The processing delay in a certain node has into account three main contributions [SeDo19]:

- The number of functionalities that are addressed in the node.
- The complexity of the functionalities that are addressed in the node.
- The processing capacity of the node.

The queuing delay of the nodes depends on [SeDo19]:

- The throughput in the output of the node.
- The traffic aggregated in the node.
- Priority level of the provided service.

The radio interface has been evolving to allow the systems to achieve ultra-low latencies, which plays an important role in the latency reduction, namely in the reduction of the radio latency contribution. This process is represented in Figure 2.16.

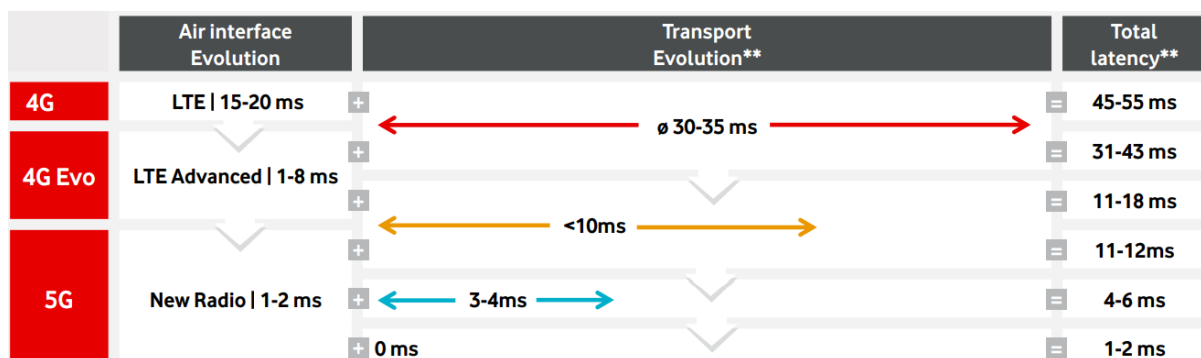


Figure 2.16 – Latency and network evolutions (extracted from [VODA18]).

2.6 State of the Art

In this section, one presents the information given by the most current research about 5G and 4G regarding the topic of the thesis, related to latency. This research is mainly based on academic works, corporations involved in mobile communications (such as Qualcomm) and also other entities like 3GPP.

5G spectrum bands are divided into 3 different groups [HUAW20]: low-frequency bands (below 2 GHz), mid-frequency bands (from 2 GHz up to 8 GHz) and high-frequency bands (above 24 GHz). This is an improvement in relation to LTE, because by using higher frequencies 5G is able to achieve higher data rates and lower latencies. By having frequency bands in less congested parts of the spectrum it is possible to use higher bandwidths leading to lower transmission times, following 5G numerology and time structure [AZRB17] and [GMFF20]. 3GPP describes numerology as a set of parameters that are used for setting the configuration of the waveform, building a Sub-Carrier Spacing in the frequency domain which dependent on the Cyclic Prefix and the symbol duration of the frame structure. 5G NR numerology is characterised by its flexible frame structure, and this flexibility is implemented not only in the frequency domain with OFDM numerology but also with Transmission Time Interval (TTI) in the time domain [GMFF20]. By using higher values of the numerology parameter, it is possible to reduce the TTI, and consequently the latency associated with the Radio Interface gets reduced, process that is crucial to achieve the low values for the latency that are under study. The virtualisation process is one of the evolutions that take responsibility for the network latency reduction since it takes an important role in SDN and VNF. Both these virtualisation-based technologies are revolutionary because they influence Cloud and Edge Network deployments, which are two of the studied strategies for minimising E2E latency in both 4G and 5G networks in this thesis. The Cloud Networks centralise the network resources into BBU pools, which will be easier to access than the Core network [KMNT18]. Since BBU pools are not only physically but also logically closer to users than the Core the implementation of this architecture becomes of high relevance in terms of the latency reduction.

The existence of many contributions to delay is inconvenient and makes systems unsuitable for real-time applications. To cope with this 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 [NHKA19]. This process reduces latency by shortening the path between the user and network resources that the signal needs to reach, by bringing them to the edge of the network. Edge Computing resources can be exploited also by operators and third parties for specific purposes. In the 4G architecture, it is possible to implement MEC nodes in different parts of the network, and each one of the possible deployments has different impacts in terms of the latency [ASCC18]. From a cost and scalability viewpoint the replacement of radio functionalities by a MEC node is not advisable, and therefore this possible deployment is not considered in this document, while the other possible deployments are considered and studied. The Edge technology can be physically installed between the RRH and the BBU, between the BBU and the Core or between the Core and the data centre. From the latency view point the first referred deployment is the one that can minimise the latency, although the other two can also have impact on the latency reduction.

The offload of data to Edge nodes can be done in two regimes [PaMa17]: full and partial offloadings. The former is implemented to minimise the execution delay and the energy consumption while fulfilling the delay constraints, while the latter is focused on reaching the best possible trade-off between energy consumption and execution delay. From the viewpoint of latency reduction the full offload is better to ensure the reliability of the very low latencies.

In [YCNP18], the features of support for MEC in a 5G network are presented, and the working mechanisms are explained for application allocation, user plain selection, traffic local routing and steering. The UPF is the main component of the network responsible for the traffic routing and therefore it is the node where the decision about the direction of packets is taken. An Application Function may influence UPF selection and traffic routing to reduce the latency, and the 5G Core Network will also select the UPF to route the user traffic to the Local Data Network where the Edge nodes are installed.

Many of the services and the applications that users expect 5G to be able to implement fall into the URLLC category and require very low latencies, such as real-time remote surgery, self-driven vehicles and factory automation. These are just some examples of the possible applications that this latency reduction will provide to users [3GPP19c].

There are 4 main delay contributions to latency [VODA18]: transmission, propagation, processing and queuing. In order to achieve low values for this parameter, it is important to understand how these delays are distributed in the network, and consequently it is mandatory to understand how end-to-end network architectures are designed,

The architecture of the network present in [NGOF18] 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 are also 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 follows: the RU that keeps the functionalities of the RRH, and the BBU functionalities are split by the RU, DU and the CU. From the latency viewpoint, the evolution of the proposed architecture is beneficial because the MEC node deployment becomes more flexible, since the node can be installed in more positions than in the LTE network.

There are 2 possible scenarios to account for latency regarding MEC nodes [SeDo19], in the sense that they can be used or not. The first scenario allows MEC technology to perform the functionalities of certain nodes of the network reducing both the physical and the logical distance between both ends of the network, which may result in a substantial latency reduction. The second scenario has the contributions of more nodes and distances to the latency, which does not allow a flexible latency reduction, and does not present itself as a viable solution for the study in hands.

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. For the 5G system to achieve E2E latencies in the order of 5 ms and lower, the radio interface cannot allow its contributions to these delays to be higher than 1 to 2 ms. The evolution of these radio interfaces is presented in [VODA18].

Chapter 3

Model and Simulator Description

This chapter provides a detailed description of the model used in the thesis, namely its mathematical expression and its implementation. In the end of the chapter, one provides the assessment of the developed simulator.

3.1 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 also the services provided to users, in order to calculate the E2E latency and provide an analysis of several solutions to reduce this parameter in 4G and 5G networks. The scheme of the model overview, including input parameters, intermediate calculations that occur during the program execution, and output parameters are represented in Figure 3.1.

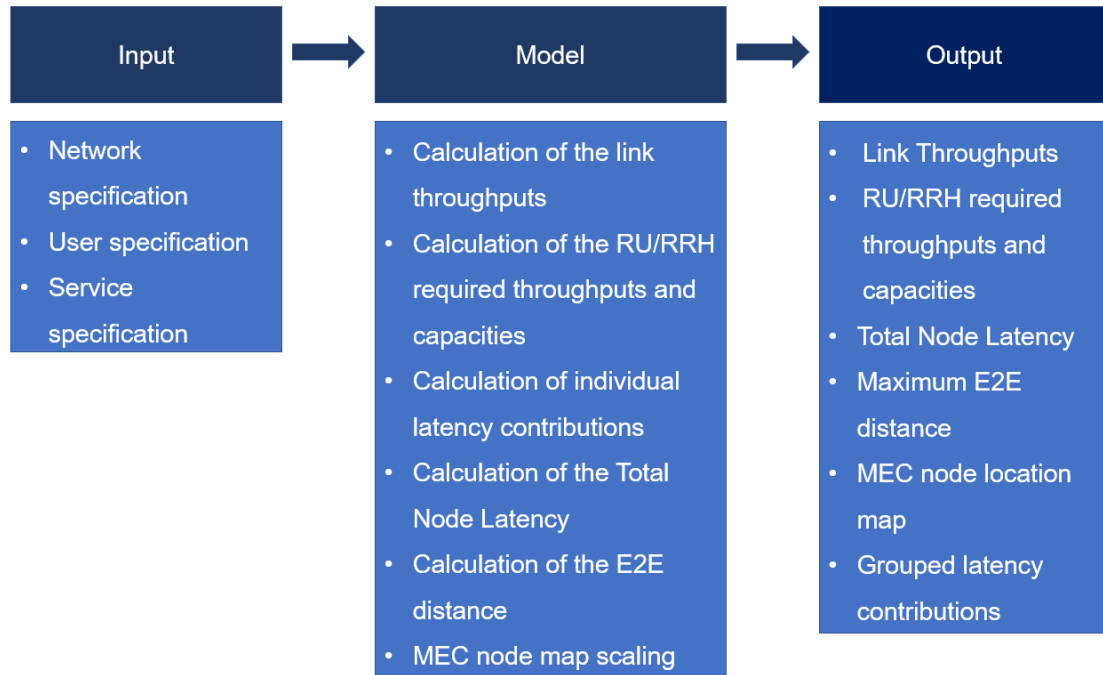


Figure 3.1 – Model Overview.

The input specification parameters of the model are divided into 3 main groups: network, user and service. These parameters are saved into program variables and used in mathematical expressions and calculations to produce the program outputs.

Network specification parameters consist of all definitions related to the network required to calculate the Total Node Latency in a certain scenario, such as: network type (4G or 5G), network architecture, MEC node deployment option and link type interconnecting nodes. There are also other parameters, such as packets size and the priority assigned to a certain service, that are also determined by the network.

User specification parameters consist of the user profile for the specified scenario, with variables such as the number of users in the RU/RRHs, DUs and CU/BBUs, and the percentages of users that are using a certain service in a certain node, among others. It is also in this input parameter category that the program user can define which service is tested in the simulation.

Service specification parameters consist of the characteristics of the services that can be provided by the network to users, such as the maximum E2E latency and the data rate.

Table 3.1 represents the definition of the network specification parameters.

Table 3.1 – Network specifications.

Network Architecture		Network architecture used in the simulation.
MEC specification		MEC node position in the network.
FH links		FH splitting options, used to define the RAN node functionalities.
MH links		MH splitting option, used to define the MH throughput.
BH links		BH link throughput.
Transport links		Transport link throughput.
Initial Link Type		The definition of the initial link type for each service (UL or DL).
Packet Size		The size of the packets for the services under study.
Latency Adaptation Parameter (ρ_{lat})		Assignment of processing capabilities of the nodes to a certain URLLC service.
Priority		The priorities of the services under study.
RU/RRH Specification	MIMO	The number of MIMO layers.
	Numerology parameter	The numerology parameter is used to define the RU/RRH throughput.
	Bandwidth	The radio node Bandwidth defines the Number of Resource Blocks.
	CQI	The Channel Quality Indicator determines the Rmax and the Modulation Order.
	Scaling Factor	5G system parameter that scales the RU throughput.
	DL Frame Structure	Percentage of the frame that is used for DL (and consequently UL) in the 5G TDD band.
	Average Factor/ DL UL ratios	Factor that accounts the throughput losses of the theoretical model in 5G, or DL and UL usage ratios in 4G (since the considered 4G bands use FDD).

User specifications represent the input parameters related to the usage profile in the scenario. Some of the defined parameters in this category are the distances between UEs and BSs, the number of users connected to each node and the profile of the users (service mix) present in each node. Table 3.2 represents the definition of the user specification parameters.

Service specification parameters represent a list of the services that can be provided by the operators and their specific requirements in terms of maximum E2E latency and data rate. The services under

study are mainly real-time communication ones, and thereby, the service list used for the thesis has latency critical services. This list is presented in Annex B. Table 3.3 represents the definition of the service specification parameters.

Table 3.2 – User Specification.

Distance	Distance between the UE and the connected RU/RRH, used to calculate the propagation latency in the air link.
Number of users	Total number of users connected to each node.
RU/RRH Service Mix	Percentage of total users in the cell that are using a certain service, including the reference user.
DU Service Mix	Percentage of total users connected to the DU (through the aggregation of the RUs) that are using a certain service, including not just the RU to which the reference user is connected.
CU/BBU Service Mix	Percentage of total users connected to the CU/BBU (through the aggregation of the DUs or RRHs) that are using a certain service, including not just the DU/RRH to which the reference user is connected.

Table 3.3 – Service specifications.

Maximum Latency	Maximum E2E latency for the services.
Data Rate	Data rate requirements for the services.

For each architecture and positioning of the MEC node, the output parameters of the model are:

- **Total Node Latency** – It is the parameter that is computed to calculate the sum of all latency contributions, except the propagation one, and it cannot exceed the maximum E2E latency.
- **Maximum E2E Distance** – The 5G network has not yet been installed and the position of the nodes is not yet defined, as well as the lengths of the links that interconnect network nodes. In the 4G case, the network is already installed, but it is also important to understand where the MEC nodes and the other nodes need to be installed in order to fulfil certain latency requirements. One of the output parameters of the model is the Maximum E2E Distance, which is useful to create estimations of the maximum link lengths to achieve a certain required latency.
- **Link throughputs** – This parameter allows to understand if it is possible to dimension network parameters in order to achieve lower latencies and to dimension the network in a more effective way. This parameter also dictates the processing capacity of the nodes, meaning that if a link provides a higher throughput, the following node usually has a higher processing capability.
- **RU/RRH required throughput** – After calculating the RU/RRH capacity, the program calculates the required throughput after providing the required data rates to the users that are connected to the RU/RRH of the simulated scenario.

- MEC node location map – Each service in Annex B has latency requirements that need to be achieved to guarantee QoS. One of the output parameters is a map that shows the maximum distance where the MEC nodes can be installed to achieve certain latency requirements.
- Grouped Latency Contributions – The E2E latency is the sum of 4 latency contributions: the transmission, propagation, queuing and processing ones, and one of the program outputs is the delay in each contribution.

3.2 Network Description

3.2.1 Network Architecture

The latency in the network depends on its physical installation and the deployed architecture, and the 3GPP C-RAN architecture has several deployment options.

As a general approach, the 5G network contains a Fronthaul, a Middlehaul and a Backhaul, but the architecture may be deployed in a different way, and therefore 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. For example, if the RU and DU nodes are collocated, the Fronthaul link propagation latency is approximately 0 ms. On top of the architecture, it is important to understand how MEC nodes can be deployed in the network according to the type of architecture, because the implementation of this technology must be considered in the network structure. The 4G C-RAN network has a standardised architecture, and nodes are not collocated, since the network is already installed.

3GPP has identified 4 possible network deployment scenarios for 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. From the latency reduction viewpoint this implementation is not recommended because it adds many contributions to the propagation latency.
- 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. This type of architecture is typical of 4G, in which the BBU takes the functionalities of the DU and CU, and the RRH is only responsible for the radio functionalities, but it can also be implemented in 5G.
- Collocated RU and DU and independent CU deployment, in which the RU and DU are located together. In this type of architecture, the latency can reach a minimum value if the MEC node is installed logically between the RU and the DU. In this architecture the distance between the RU and the DU is in the order of hundreds of metres, which reduces propagation latency and 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.

Figure 3.2 illustrates the several 3GPP C-RAN architectures.

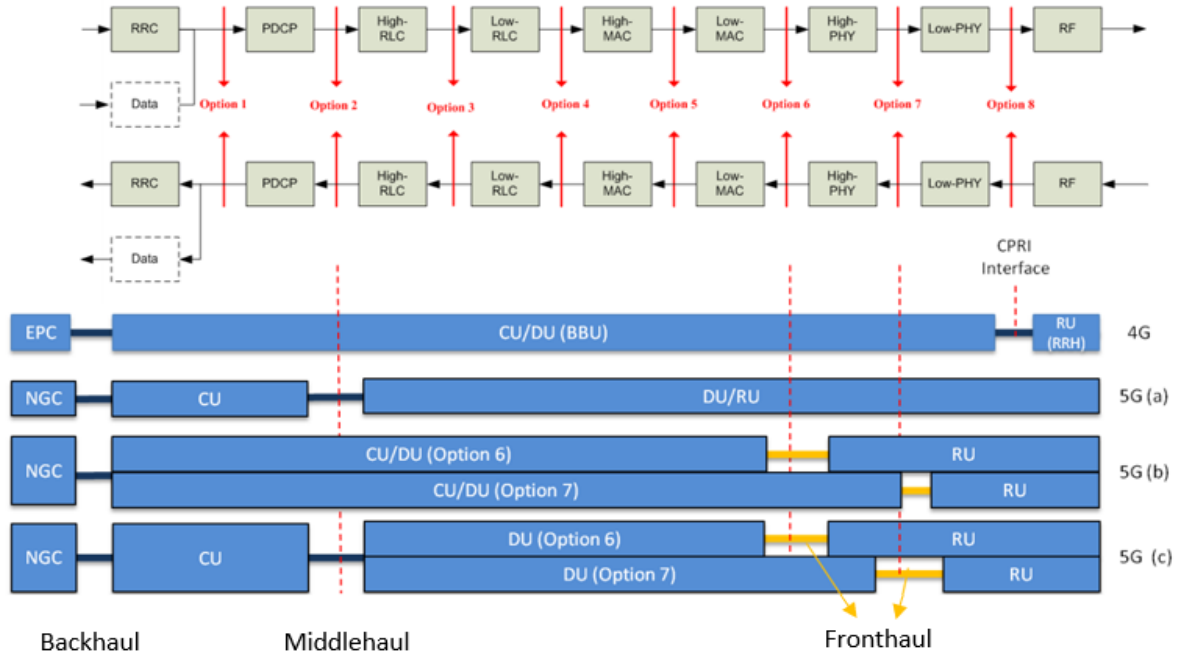


Figure 3.2 – 3GPP C-RAN split architectures (adapted from [ITUT18]).

MEC nodes are part of the network architecture, and they can be deployed in different places along the network links, between nodes, and their installation has a major impact in the E2E latency, since the purpose of this technology is to guarantee that the service is provided without the packets being sent forward in the network.

MEC node installation options studied in this thesis for the 4G architecture are represented in Figure 3.3. The definitions of all latency contributions present in the network architectures are present in Annex C.

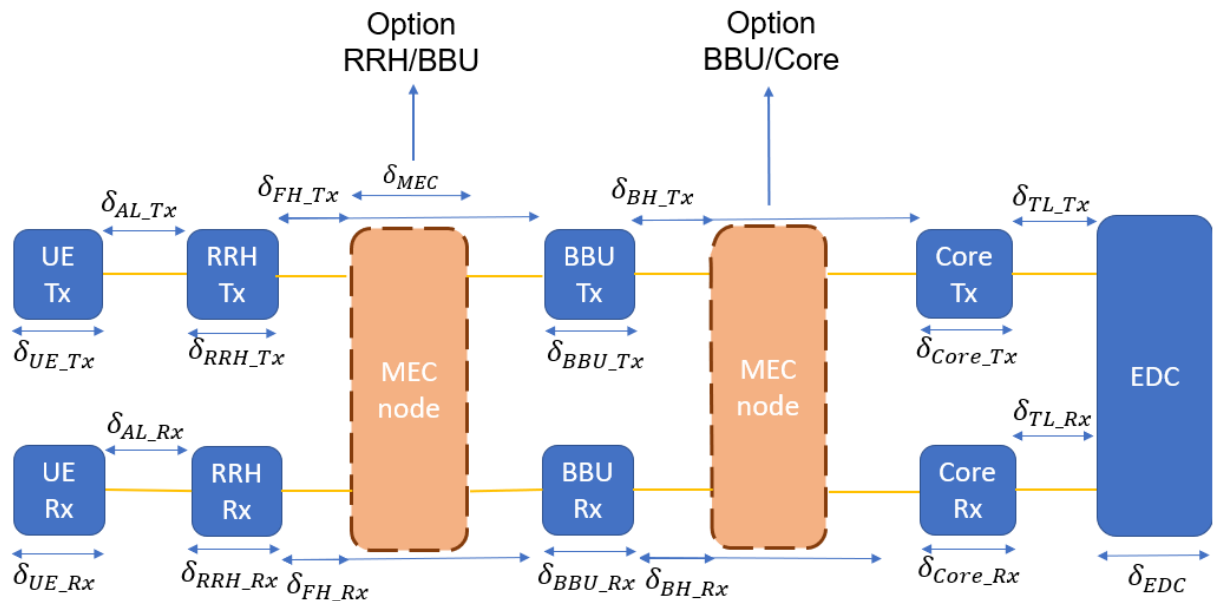


Figure 3.3 – MEC node installation options for 4G architecture.

In this latency model, one considers that there is symmetry between links, meaning that if the path of the packet is UE Tx – RU Tx - MEC node, to reach the UE Rx, the rest of the path is MEC node - RU Rx - UE Rx. It is useful for the rest of the nodes in the network to share MEC nodes, although it is not mandatory that this symmetrical installation is always present for every network branch. All MEC node installation options studied in this thesis for the 5G architecture are represented in Figure 3.4.

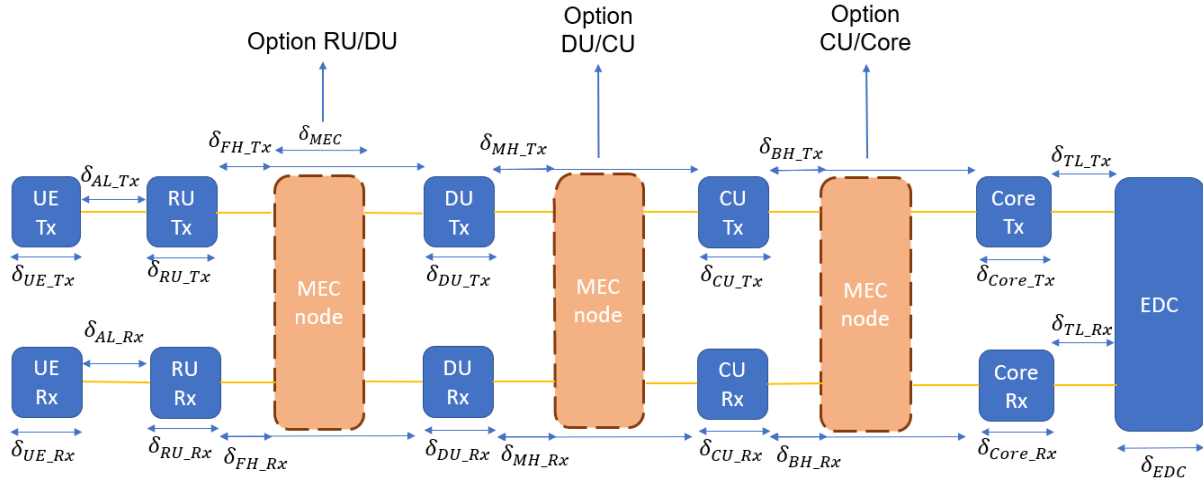


Figure 3.4 – MEC node installation options for 5G architecture.

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

- The installation of the MEC node between the RRH and the BBU (Option RRH/BBU) is the optimal solution to reduce latency, since the MEC node cannot replace RRH functionalities. By optimising the latency contributions associated with the radio link (guaranteeing data rate requirements and reducing propagation latency in the air link), increasing the Fronthaul throughput, and reducing the physical distance between the RRH and the MEC node, it is possible to achieve low latency values (in the order of 10 ms).
- The installation of the MEC node between the BBU and the Core of the Network (Option BBU/Core) presents itself as a viable solution, since the Backhaul has typically long lengths, which increases the accumulated latency. The optimisation of both the Fronthaul and Backhaul interfaces is important to maximise throughputs and reduce transmission latency, and the study of the traffic generated in the RRHs and the BBUs is mandatory to understand the queuing latency contribution. Depending on the criteria of the service to be implemented (e.g., maximum allowed latency, reliability and data rate) this solution may be considered suitable for some applications, but it is not the best possible implementation in terms of latency reduction.

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

- The installation of the MEC node between the RU and the DU (Option RU/DU) is the optimal solution to reduce latency, but it is not always possible to implement the node in this position, because of the distribution of the MEC nodes, since the MEC node is installed in the network and therefore, the node may be connected to a certain RU and to a DU or CU in another network branch due to the chosen network deployment. By optimising radio latency and Fronthaul throughput and reducing the physical distance between the RU and the MEC node it is possible to achieve very low E2E latency values (below 1 ms).

- The installation of the MEC node between the DU and the CU (Option DU/CU) presents itself as a viable solution, since the Middlehaul has typically long lengths and the MEC technology can be installed close to the DU, reducing the propagation latency associated with this implementation. The optimisation of the Middlehaul and Fronthaul study is important to maximise throughputs and reduce transmission latency, and the study of the generated traffic in the RUs and the DUs is mandatory to understand the queuing latency contribution. Depending on the criteria of the service to be implemented (e.g., maximum allowed latency, reliability and data rate) this solution may be considered suitable for some applications, but it is not the best possible implementation in terms of latency reduction.
- The installation of the MEC node between the CU and the Network Core (Option CU/Core) is another possible deployment under study. By saving the processing delay of the Network Core and shortening the contribution of the Backhaul length to propagation latency, it is possible to achieve a reduction of this parameter. But this option is not optimal, because there are still 3 nodes between the UE and the MEC node, and the CU is a point of aggregation of traffic since many DUs are connected to it. The improvements in the Backhaul capacity can reduce transmission latency as a viable way of optimising this implementation.

One of the other possible installations of the MEC node that is not shown in Figures 3.3 and 3.4 is the installation of the node between the Core and the External Network, which reduces latency because the transport network has typically long lengths, causing the accumulated propagation delay to decrease. On the other hand, the delay associated with the Core and the nodes closer to the user are still present and keep latency in values that are not suitable for the real-time services that need to be provided. Therefore, this option will not be the main one under study, since the other possibilities cause a higher latency reduction, which is the topic under study. Within the scope of the most general architecture specified by 3GPP, there are some scenarios where nodes are deployed in a collocated way. In these scenarios, the collocated nodes are installed within a few hundred metres of each other, which reduces the length of the link that connects the nodes. One of the possible achievements with these architectures is the possibility to install MEC nodes between the two collocated nodes, which is a very good solution to reduce latency, because propagation latency is reduced in that link of the network. It is possible to deploy MEC technology between a collocated RU and DU, but also between a collocated DU and CU, or even between the collocated RU, DU and CU. It is still important to have measures of the lengths of the links, but typically, between the collocated nodes it is possible to neglect the distances between them, which is the approach taken in this thesis. The 5G network has an architecture where the several sequential network node traffic is aggregated in the next network node, for example, the DU aggregates the traffic of the several connected RUs, and the CU aggregates the traffic of the several connected DUs. Figure 3.5 represents the 5G node aggregation in the network.

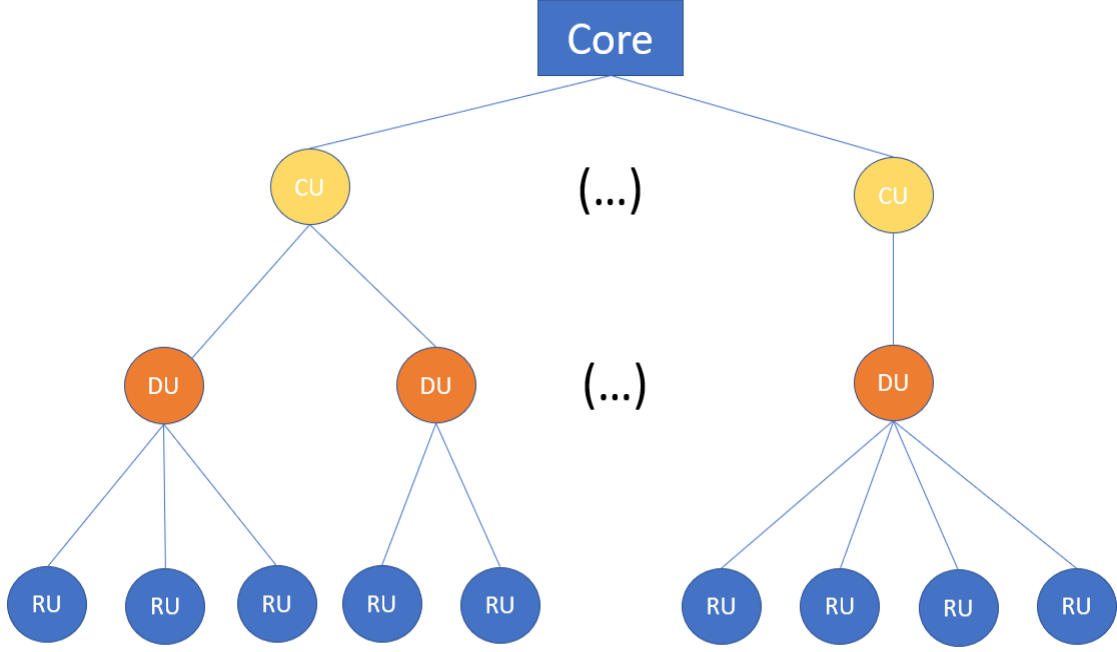


Figure 3.5 – 5G node aggregation in the network.

3.2.2 Latency Contributions

There are 4 main latency contributions: transmission, propagation, queuing and processing. These contributions have different mathematical expressions that allow to create the model that accounts for the latency along the network. Both processing and queuing latencies occur in the nodes, transmission latency occurs from the node to the link, and propagation latency occurs in the link [VODA18]. The following expressions represent the delays accumulated in the network nodes:

$$\delta_{UE_Tx} = \delta_{UE_Proc} + \delta_{UE_Trans} \quad (3.1)$$

$$\delta_{UE_Rx} = \delta_{UE_Proc} \quad (3.2)$$

$$\delta_{RU_Tx} = \delta_{RU_Rx} = \delta_{RU_Proc} + \delta_{RU_Queu} + \delta_{RU_Trans} \quad (3.3)$$

$$\delta_{RRH_Tx} = \delta_{RRH_Rx} = \delta_{RRH_Proc} + \delta_{RRH_Queu} + \delta_{RRH_Trans} \quad (3.4)$$

$$\delta_{DU_Tx} = \delta_{DU_Rx} = \delta_{DU_Proc} + \delta_{DU_Queu} + \delta_{DU_Trans} \quad (3.5)$$

$$\delta_{CU_Tx} = \delta_{CU_Rx} = \delta_{CU_Proc} + \delta_{CU_Queu} + \delta_{CU_Trans} \quad (3.6)$$

$$\delta_{BBU_Tx} = \delta_{BBU_Rx} = \delta_{BBU_Proc} + \delta_{BBU_Queu} + \delta_{BBU_Trans} \quad (3.7)$$

$$\delta_{Core_Tx} = \delta_{Core_Rx} = \delta_{Core_Proc} + \delta_{Core_Trans} \quad (3.8)$$

$$\delta_{EDC} = \delta_{EDC_Proc} + \delta_{EDC_Trans} \quad (3.9)$$

$$\delta_{MEC} = \delta_{MEC_Proc} + \delta_{MEC_Trans} \quad (3.10)$$

The transmission latency is generated by the process of transmitting the bits into a certain link, and it depends on the throughput / data rate and the amount of data to be transmitted. The general expression for this contribution is:

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

where:

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

Propagation latency is generated by the propagation of a signal through a link, which may occur in different types of links with different velocities, and consequently different latencies, since this delay depends mainly on the velocity of the signal in the link and the distance between the origin and the destination of the signal. It is important to study the possible solutions for the links, because the signal velocity in the optical fibre is different than in the air, but typically the Fronthaul, Middlehaul, Backhaul and Transport links in the networks are implemented with optical fibre, and the Air link is the exception with the typical velocity of 3×10^8 m/s. The general expression for this contribution is:

$$\delta_{Prop} [\text{ms}] = \frac{d_{[\text{m}]}}{v_{[\text{km/s}]}} \quad (3.12)$$

where:

- $d_{[\text{m}]}$ – Distance between the origin and the destination of the signal.
- $v_{[\text{km/s}]}$ – velocity of the signal in the link.

The processing latency is associated with the time taken to process the header of a packet, to determine where to route data and also to process the functionalities in the nodes, but it can also include other contributions, such as the time required to check errors. Typically, this contribution adds low values of delay to the E2E latency, when compared with the other latency contributions, but since the requirements are very strict due to the low latency and reliability it is important to take this contribution into account. The processing delay in the UE is related to the TTI, but in 5G the processing time is considered correlated to the time that one packet takes to be transmitted to the air interface, since the work is done at the packet level and not in a specific time frame. The increase in the data rate is related to a higher transmission rate and a higher processing capacity assigned to the packet processing. On the other hand, at the packet level, bigger sized packets have more bits to accommodate, making the processing of each packet a harder process. The 4G processing and transmission delays are known values present in the HARQ latency model in [DMAG18] (δ_{UE_Trans} is considered 1 ms for LTE). To calculate the other processing delays, one considers that the splitting option 8 is used in 4G and all the other options account as a possibility for 5G. In 4G deployment, the RRH keeps the radio functionalities, which accounts as one ninth of the entire processing delay, but the processing delay in the RU depends on the splitting option and therefore the functionalities distribution. The considered UE processing delay ratios can be calculated using the Table 3.4.

Table 3.4 – 4G and 5G UE processing delay ratios (extracted from [DMAG18]).

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

In [SeDo19], it is possible to understand that the processing latency is not influenced by the network architecture, which makes sense, since the nodes that are processing packets are the same and have the same functionalities if the same functional split is considered. Unlike the network architecture, the functional split has a direct impact on the processing delay of the nodes, because it assigns different functions to the different nodes, which have different processing capabilities, thus resulting in a different processing latency accumulation along the network. In the Option 8 split, the RU keeps only 1 of the 9 functionalities that are supposed to be distributed between network nodes, which are the RF functionalities, therefore, it is considered that the processing delay on the RU is one ninth of the total processing delay. In the other splitting options, one uses the same process to calculate the DU and CU processing delays. The splitting options 7.1, 7.2 and 7.3 divide partially the LOW-PHY and HIGH-PHY functionalities (which include 11 processes in total) between the RU and the DU, and the CU always keeps the RRC (Radio Resource Control) and PDCP (Packet Data Convergence Protocol) functionalities. By assuming that the functionalities that are distributed along the network nodes have all equal weights in terms of complexity and processing time, it becomes possible to obtain the equations that relate all processing delays.

The processing delay in the UE is given by [DMAG18]:

$$\delta_{UE_Proc} = \delta_{UE_Trans} \rho_{UE} \quad (3.13)$$

The processing delay in the RRH is given by [DMAG18]:

$$\delta_{RRH_Proc} = 1.5 \cdot \delta_{UE_Trans} \rho_{RU} \quad (3.14)$$

where:

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

The processing delay in the BBU is given by:

$$\delta_{BBU_Proc} = \delta_{RRH_Proc} (\rho_{CU} + \rho_{DU}) \rho_{lat} \quad (3.15)$$

where:

- ρ_{DU} – the ratio of functionalities assigned to the DU.
- ρ_{CU} – the ratio of functionalities assigned 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 [DMAG18]:

$$\delta_{RU_Proc} = \delta_{UE_Trans} \rho_{RU} \rho_{lat} \quad (3.16)$$

The processing delay in the DU is given by:

$$\delta_{DU_Proc} = \delta_{UE_Trans} \rho_{DU} \rho_{lat} \quad (3.17)$$

The processing delay in the CU is given by:

$$\delta_{CU_Proc} = \delta_{UE_Trans} \rho_{CU} \rho_{lat} \quad (3.18)$$

Latency adaptation parameters (for URLLC services) for each service are presented in Annex E. There is the possibility of implementing systems where instead of having a MEC node processing the information, there is an Edge computing layer where several MEC nodes divide the processing needs amongst each other. This way of processing information is good for high amounts of data, but not as good as the single MEC node approach for small data packets. Since the topic under study is the E2E latency in a specific network, for a specific service, one considered that only one MEC node is going to process the information. The residual data that cannot be processed in the MEC node is sent to the data centres, and the accumulated latency in this process depends on the communication between MEC nodes and the own computation delay of the Cloud computing layer (data centre or Cloud server). In [OAGW18], the typical latency performance values in Vodafone's London network are between 3 ms and 11 ms for the 4G RAN, Core and Core Gi-LAN. Assuming that the latency is split equally among the RAN, Core and Core Gi-LAN, the one-way Core latency is considered between 1 ms and 3.67 ms (for the biggest and smallest packets).

The processing delay in the Core is given by [OAGW18]:

$$\delta_{Core_Proc} [ms] = \frac{4}{2385} D_{[Bytes]} + \frac{469}{477} \quad (3.19)$$

The required parameters to calculate the processing delays of the RU, DU and CU nodes are presented in Table 3.5.

Table 3.5 – RU, DU and CU processing latency ratios (adapted from [SeDo19]).

	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 Cloud computing layer described in [GLJW18] has 5 Cloud servers, therefore it is considered that the computing delay in the data centre is 20% of the computing delay of the cloud computing layer, if all 5 processors are processing the same amounts of information. This causes an increase of the computation delay from 2 s to 10 s (considering 20 GB of data). Assuming the same proportionality, the processing capacity of a CPU is mainly dictated by the frequency of the processor, and in the paper it is considered that the processing frequency of the cloud data centre is 10 GHz when the usual value

tends to be around 2 GHz to 3 GHz (one considered the 3 GHz processing frequency for the data centre processor), which required an adaptation assuming the fact that the processing latency increases linearly with the decrease in the processor frequency. The MEC node processing frequency is usually around 1 GHz, giving a third of the processing capacity to the node, compared with the data centre. The processing latency of the MEC node (assuming the processing of a single functionality) and the data centres is represented in Annex D, and the processing delays in the MEC node and the data centre are:

$$\delta_{MEC_Proc} [\text{ms}] = 4 \cdot 10^{-5} \cdot D_{[\text{Bytes}]} \cdot \rho_{func} \quad (3.20)$$

$$\delta_{EDC_Proc} [\text{ms}] = 1.33 \cdot 10^{-5} \cdot D_{[\text{Bytes}]} \quad (3.21)$$

where:

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

The queuing delay of a certain packet depends on the earlier-arriving packets that are queued and waiting for transmission across the link, and also on the maximum throughput of the following link, since the existence of a higher throughput is related to a faster transmission rate, resulting in queue reduction. On the other hand, the number of services that are prioritised compared to the considered service increase queuing latency, as well as the size of the packets of the services that are higher in the priority list. In this model, one considers that the queuing delay is only accounted for in the RU, RRH, DU, CU, and BBU because the processing capacity Core of the Network will be adapted to network congestion. The queuing delay in a certain node, by accounting for the packets that arrive in the node and belong to a service that has an equal or higher priority than the considered reference user service, is

$$\delta_{Queue} [\text{ms}] = 10^3 \sum_{p=1}^{M_{Pserv}} \frac{8 D_{serv,p} [\text{Bytes}]}{R_{max} [\text{bps}]} \quad (3.22)$$

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.

The queuing expression is useful to represent the traffic aggregation generated by the simulated users, which allows the adaptation to the several scenarios. The simulated traffic for the scenarios is present in Annex K, regarding the number of users connected to a network node in a specific moment, which determines the queuing delay. The DU, the CU and the BBU traffic represent the aggregation of the previous network nodes, e.g., the DU traffic represents the traffic from the several aggregated RUs.

3.2.3 Network Latency

To sum up the latency contributions to calculate the E2E latency of 4G and 5G one divides the total latency into multiple segments to simplify the complexity and number of components of the model. This division is made by grouping the node latency contributions separated from the propagation ones. 5G

has not yet been installed, and therefore there are no specified distances between nodes, and to approach this problem in the model, propagation latencies are dimensioned to guarantee that the minimum latency requirements are achieved. In this process propagation latency contributions are considered equal to 0 ms in the expressions, and after calculating what has remained as a latency margin between the maximum E2E latency allowed for a service and the total node latency, the margin is equal to the maximum accumulated propagation latency in the network, and then the maximum lengths of the links to achieve certain latency requirements are determined.

MEC technology is going to be implemented along 5G and 4G networks, between nodes, as one of the strategies to minimise E2E latency. In this thesis, one considers the possibility, or not, of implementing a MEC node. Even if a certain branch of the network has a MEC node implemented, sometimes packets will still be sent forward, and therefore, both scenarios need to be studied and mathematically formulated in a realistic way. After being processed in the MEC node, data packets can be sent to another user in the network so that user information is updated in the UE receiver. A simple example of this process can be a remote surgery: while the manipulator is sending commands to the robot that is executing the surgery, packets need to be sent to the MEC first, processed in the node and only after that, packets can be sent to the surgery robot. The time taken by the process of sending data packets to the MEC and then to the receiver UE is the E2E latency.

For the 4G scenario without MEC nodes, the Total node latency is described by:

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RRH_Tx} + \delta_{BBU_Tx} + \delta_{Core_Tx} + \delta_{EDC} + \delta_{Core_Rx} + \delta_{BBU_Rx} + \delta_{RRH_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.23)$$

Latency contributions for the 4G scenario without the MEC node are represented in Figure 3.6.

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

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RRH_Tx} + \delta_{BBU_Tx} + \delta_{MEC} + \delta_{BBU_Rx} + \delta_{RRH_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.24)$$

Latency contributions for the 4G scenario with the BBU/Core MEC node option are represented in Figure 3.7.

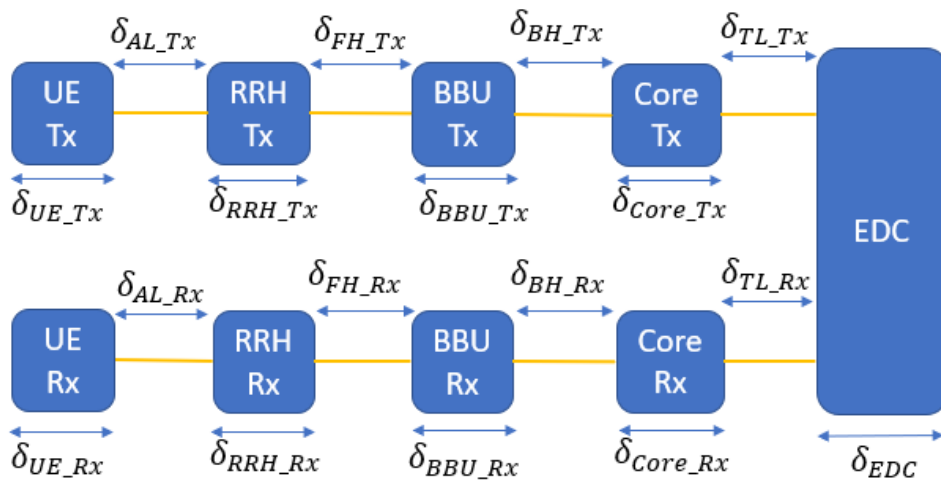


Figure 3.6 – 4G latency contributions without the MEC node.

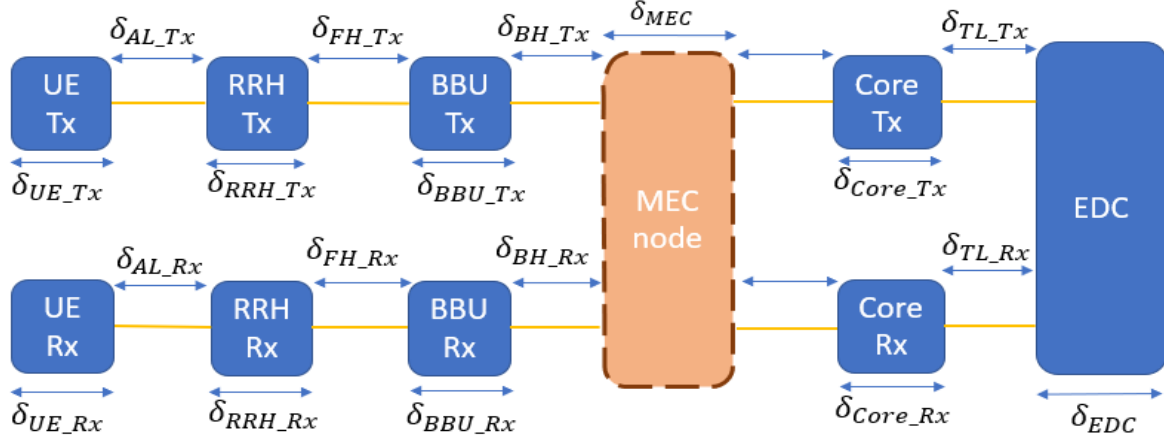


Figure 3.7 – 4G latency contributions with the MEC node installed between the BBU and the Core.

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

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RRH_Tx} + \delta_{MEC} + \delta_{RRH_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.25)$$

Latency contributions for the 4G scenario with the RRH/BBU MEC node option are represented in Figure 3.8.

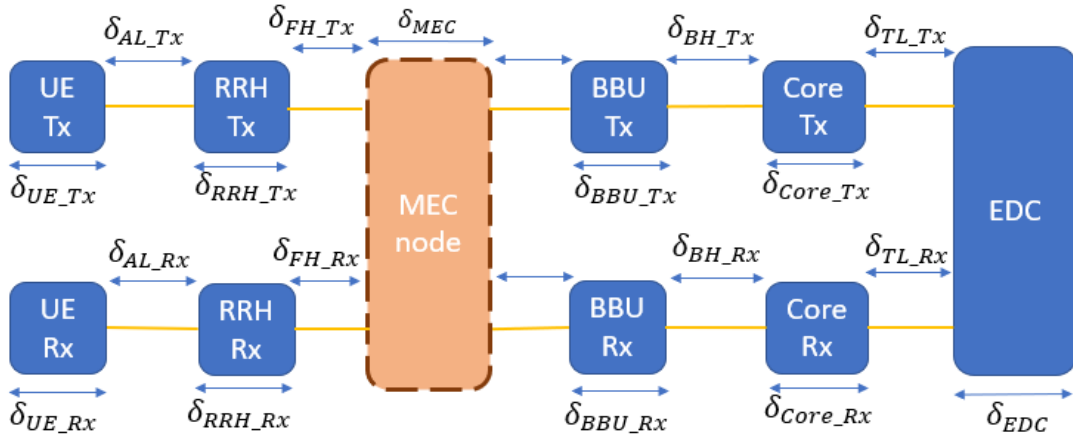


Figure 3.8 – 4G latency contributions with the MEC node installed between the RRH and the BBU.

For the 5G scenario without MEC nodes, the Total Node latency is described by:

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RU_Tx} + \delta_{DU_Tx} + \delta_{CU_Tx} + \delta_{Core_Tx} + \delta_{EDC} + \delta_{Core_Rx} + \delta_{CU_Rx} + \delta_{DU_Rx} + \delta_{RU_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.26)$$

Latency contributions in the 5G scenario without the MEC node are represented in Figure 3.9.

One of the possible deployment scenarios for the MEC technology in 5G is the one in which the MEC node is installed between the CU and the Core of the Network. For this specific scenario, the Total Node latency is described by:

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RU_Tx} + \delta_{DU_Tx} + \delta_{CU_Tx} + \delta_{MEC} + \delta_{CU_Rx} + \delta_{DU_Rx} + \delta_{RU_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.27)$$

Latency contributions in the 5G scenario with the MEC node between the CU the Core of the Network

are represented in Figure 3.10.

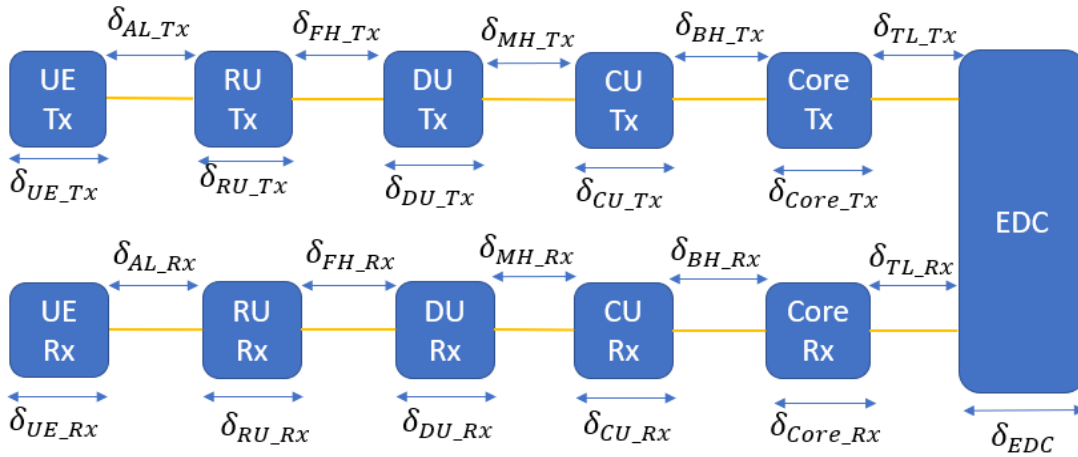


Figure 3.9 – 5G latency contributions without MEC node.

The second considered scenario is the one in which the MEC node is installed between the DU and the CU. For this specific scenario, the Total Node latency is described by:

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RU_Tx} + \delta_{DU_Tx} + \delta_{MEC} + \delta_{DU_Rx} + \delta_{RU_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.28)$$

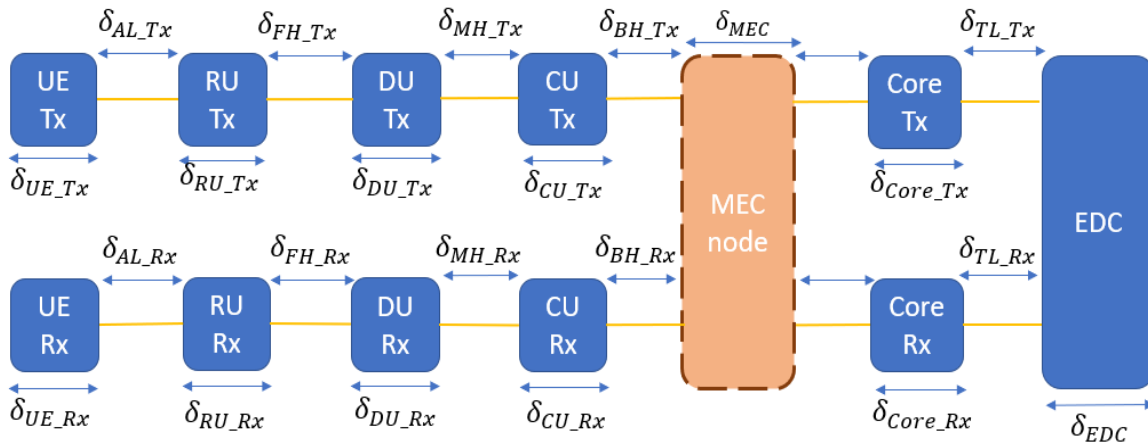


Figure 3.10 – 5G latency contributions with the MEC node between the CU and the Core.

Latency contributions in the scenario with the MEC node installed between the DU the CU are represented in Figure 3.11.

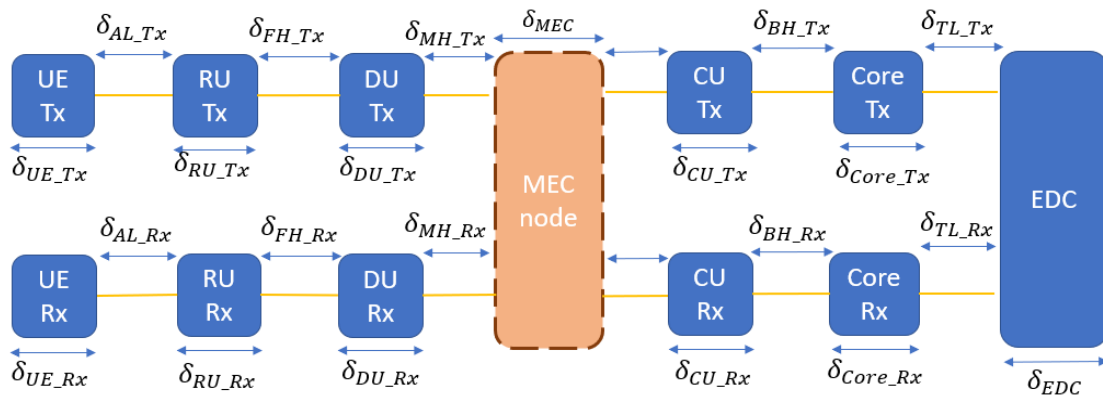


Figure 3.11 – 5G latency contributions with the MEC node between the DU and the CU.

The E2E distance is calculated in the model by splitting the E2E latency into 2 different categories: the sum of the transmission, queuing and processing latency, which consists of the total node latency, and the propagation latency, which is associated with the link length and therefore with the E2E distance.

The third considered scenario is the one in which the MEC node is installed between the RU and the DU. For this specific scenario, the Total Node latency is described by:

$$\delta_{Tot_Node} [ms] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RU_Tx} + \delta_{MEC} + \delta_{RU_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.29)$$

Latency contributions in the scenario with the MEC node between the RU the DU are represented in Figure 3.12.

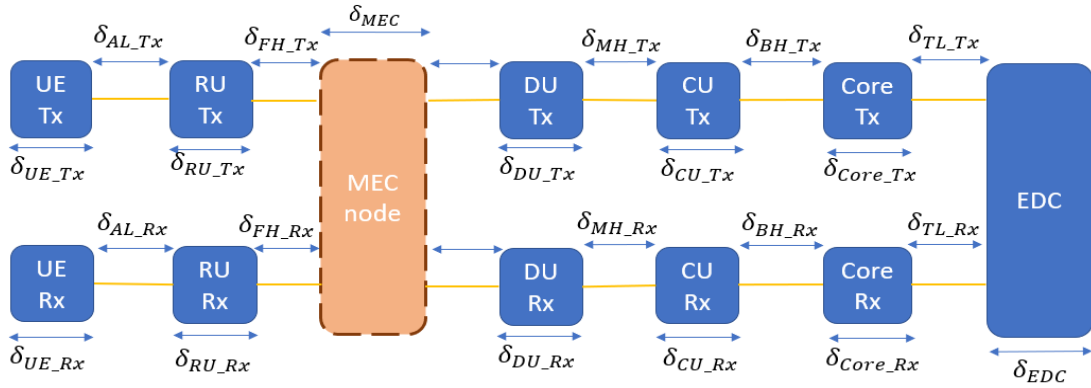


Figure 3.12 – Latency contributions in the scenario with the MEC node between the RU and the DU.

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

$$\delta_{Prop_4G} [ms] = \delta_{FH_Tx} + \delta_{BH_Tx} + \delta_{TL_Tx} + \delta_{TL_Rx} + \delta_{BH_Rx} + \delta_{FH_Rx} \quad (3.30)$$

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

$$\delta_{Prop_5G} [ms] = \delta_{FH_Tx} + \delta_{MH_Tx} + \delta_{BH_Tx} + \delta_{TL_Tx} + \delta_{TL_Rx} + \delta_{BH_Rx} + \delta_{MH_Rx} + \delta_{FH_Rx} \quad (3.31)$$

The E2E latency is calculated using:

$$\delta_{E2E} [ms] = \delta_{Tot_Node} [ms] + \delta_{Prop} [ms] \quad (3.32)$$

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

$$d_{E2E} [km] = (\delta_{App} [ms] - \delta_{Tot_Node} [ms]) \frac{v [km/s]}{2 \times 1.67} 10^{-3} \quad (3.33)$$

where:

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

The contribution of the distances in the network to the maximum E2E distance is defined by:

$$d_{E2E} [\text{km}] = d_{FH_Tx} + d_{MH_Tx} + d_{BH_Tx} + d_{TL_Tx} + d_{TL_Rx} + d_{BH_Rx} + d_{MH_Rx} + d_{FH_Rx} \quad (3.34)$$

where:

- $d_{FH_Tx}/d_{FH_Rx} [\text{km}]$ – Fronthaul link lengths.
- $d_{MH_Tx}/d_{MH_Rx} [\text{km}]$ – Middlehaul link lengths.
- $d_{BH_Tx}/d_{BH_Rx} [\text{km}]$ – Backhaul link lengths.
- $d_{TL_Tx}/d_{TL_Rx} [\text{km}]$ – Transport link lengths.

When 2 network nodes are collocated, the length of the link between them is considered negligible. In Figure 3.13, one represents the network architecture with a collocated RU and DU. For this specific scenario, the E2E latency is described by:

$$\delta_{E2E} [\text{ms}] = \delta_{UE_Tx} + \delta_{AL_Tx} + \delta_{RU_Tx} + \delta_{MEC} + \delta_{RU_Rx} + \delta_{AL_Rx} + \delta_{UE_Rx} \quad (3.35)$$

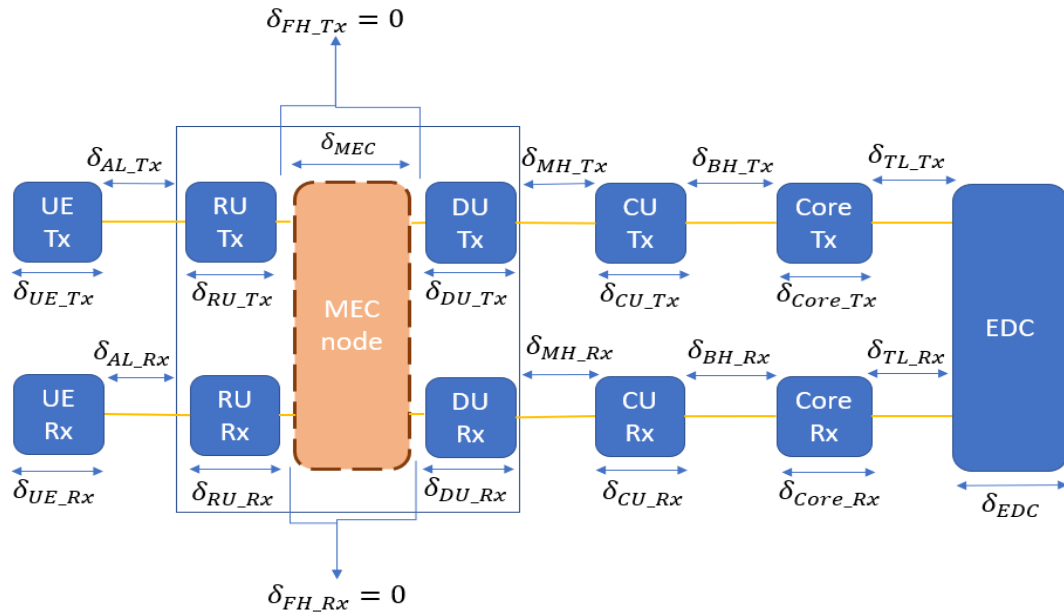


Figure 3.13 – Latency contributions with the MEC node between collocated RU and DU.

3.2.4 Link Throughput

The calculation of the maximum throughput offered by LTE is performed using a specific expression for the FDD mode. The calculation of the maximum throughput offered by NR is performed using a specific expression for the FDD mode but adapting the model to TDD with the frame structure parameter. It is important to refer that it is considered that the conditions of the link must be optimal to achieve the maximum calculated throughput.

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

$$R_{FDD} [\text{Mbits/s}] = \frac{2 \cdot 10^3 [\text{s}^{-1}] v_{Layers} N_{RB} [\text{RB}] N_s [\text{symbols/RB}] Q_m [\text{bits/symbol}] R_{code_max} (1-0)}{10^6} \quad (3.36)$$

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

$$R_{FDD} [\text{Mbits/s}] = \frac{v_{Layers} Q_m [\text{bits/symbol}] f R_{code_max} \frac{12 \cdot N_{PRB}^{BW, \mu} [\text{symbols}]}{T_s^\mu [\text{s}]} (1-O)}{10^6} \quad (3.37)$$

where:

- v_{Layers} – Number of layers (eNB/gNB transmitter streams to the UE or UE transmitter streams to the eNB/gNB), which depends on the MIMO system implemented (maximum of 8 for the DL and maximum of 4 for the UL in the 5G system).
- $N_{RB}/N_{PRB}^{BW, \mu}$ – Number of Resource Blocks that depend on the system bandwidth and the numerology (for 5G), that be taken from Table 3.7 or 3.8 (for 4G or 5G scenarios respectively).
- N_s - Number of symbols per Resource Block in 4G. This variable usually takes the value of 84, due to the existence of 12 subcarriers and 7 symbols per subcarrier per RB.
- Q_m – Average modulation order (the possible values are 2 for QPSK, 4 for 16-QAM, 6 for 64-QAM for 4G and 2 for QPSK, 4 for 16-QAM, 6 for 64-QAM, 8 for 256-QAM for 5G).
- R_{code_max} – Constant defined by 3GPP that depends on the modulation order. A higher R_{code_max} constant is correlated to a higher order of modulation and higher values of spectral efficiency, and it is determined by the CQI (Channel Quality Indicator). The R_{code_max} parameter for 4G is defined in Annex G, and the R_{code_max} parameter for 5G is defined in Annex H.
- $T_s^\mu = \frac{10^{-3}}{14 \times 2^\mu}$ – The average symbol duration in a subframe for numerology μ , for 5G.
- f - Scaling factor (1, 0.8, 0.75 or 0.4) that is defined per band or per band combination. This scaling factor has higher relevance in situations of UE high or medium mobility; due to the handover process, the required throughput of a 5G radio link may be lower or higher depending on the traffic for a certain RU, therefore this factor is important to scale the delivered throughput of 5G.
- O – The overhead for control channels is related to the bandwidth required to transmit the packet overhead, and it depends on the frequency band and the MIMO system implemented. In 4G the average value used for calculations is 0.25. In the 5G scenario, it is important to understand that 3GPP divides the frequency bands of the system in 2 different categories. The first frequency band is FR1, with a range of frequencies between 0.410 GHz and 6. GHz and the second frequency band is FR2 with frequencies above 6 GHz. In this thesis only the FR1 band is considered, because the simulated band is centred in the 3.5 GHz, and the Overheads for the FR1 band are present in Table 3.6. For the 5G system it is considered an average of the DL and UL overheads, taking into account that the URLLC services have typically transmissions in DL and UL.

The values for the overhead for control channels are represented in the Table 3.6.

Table 3.6 – Overhead for control channels in 5G (adapted from [ETSI18c]).

5G Band	UL	DL	Average Overhead
FR1	0.08	0.14	0.11

The number of available RBs (Resource Blocks) for 4G is represented in Table 3.7.

Table 3.7 – Number of Resource Blocks depending on the Bandwidth (extracted from [Corr20]).

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 5G is represented in Table 3.8.

Table 3.8 – Maximum NRBs for each transmission bandwidth and subcarrier spacing (adapted from [ETSI18b] and [3GPP18b]).

SCS [kHz]	Bandwidth [MHz]													
	5	10	15	20	25	30	40	50	60	80	90	100	200	400
15	25	52	79	106	133	160	216	270	N/A	N/A	N/A	N/A	N/A	N/A
30	11	24	38	51	65	78	106	133	162	217	245	273	N/A	N/A
60	N/A	11	18	24	31	38	51	66	79	107	121	132	264	N/A
120	N/A	N/A	N/A	N/A	N/A	N/A	N/A	32	N/A	N/A	N/A	66	132	264

For the 5G TDD mode there is an adaptation of the expression that has been defined by 3GPP. In this mode, the slot time is divided into sub-slots that can be used in UL or DL, therefore, the throughput of 5G in the TDD mode depends on the chosen format for the slot. The 5G system will also have a flexible frame structure, and therefore one considers a 57% of framing for DL and 43% for UL. The throughputs in DL and UL depend on the frame structure, and therefore, the result obtained by the expressions needs to be multiplied by the fraction of the slot used in UL and DL, which depends on the format chosen for the slot. In the 5G case, the slots are divided in UL or DL, but there is also another type of subframe called F frame, or Flexible subframe, which can be used for DL or UL, depending on the type of link under usage. The average throughput of the radio in 5G UL in TDD can be calculated using:

$$R_{TDD/UL} [\text{bits/s}] = R_{FDD} [\text{bits/s}] F_{UL} A_f \quad (3.38)$$

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 5G DL in TDD can be calculated using:

$$R_{TDD/DL} [\text{bits/s}] = R_{FDD} [\text{bits/s}] F_{DL} A_f \quad (3.39)$$

where:

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

It is important to refer that it is considered that the conditions of the link must be optimal to achieve the maximum calculated throughput, but since link conditions are not optimal, the CQI factor (for both systems), the Average Factor (for the 5G system) and the UL and DL usage ratios (for 4G) are used to account for the losses that occur due to the lack of an optimal link.

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

$$R_{FDD/DL} [\text{bits/s}] = R_{FDD} [\text{bits/s}] D_{ur} \quad (3.40)$$

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 UL in 4G can be calculated using:

$$R_{FDD/UL} [\text{bits/s}] = R_{FDD} [\text{bits/s}] U_{ur} \quad (3.41)$$

where:

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

In order to provide viable information about the radio techniques that should be used in different URLLC scenarios, it is necessary to calculate the RU/RRH required throughput. This value is calculated by summing the data rates of the services provided by the RU/RRH. The following expression is used to calculate the used throughputs of the RU/RRH in both the receiver and transmitter sides:

$$R_u [\text{Mbits/s}] = \sum_1^{N_u} R_s [\text{Mbits/s}] \quad (3.42)$$

where:

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

After being processed in the RU node and being sent to the DU through the Fronthaul, the packet is received and processed in the DU node, to be sent for the CU through the Middlehaul, and potentially to the Core of the Network through the Backhaul (in the absence of a MEC node). To define the throughputs of the nodes, the interfaces between them are defined, and this process is developed considering the functional split of the network functionalities along nodes. The Option 8 (RF/PHY) provides a link with a CPRI interface that is usually used for the Fronthaul link and allows a very high throughput. In this functional split, the BBU keeps all functionalities except the ones related to radio, working as a CU and a DU, and the RRH keeps the radio functionalities. This process creates high constant throughputs in the order of hundreds of Gbps, because the resource element mapping is executed in the BBU.

This Fronthaul link option is typical of the 4G architecture (traditional RRH-BBU split), and it generates a RU throughput in the UL and DL given by:

$$R_8 \text{ [Mbps]} = S_r[\text{megasympols/s}] N_Q[\text{bits/symbol}] N_A \quad (3.43)$$

where:

- $S_r[\text{megasympols/s}]$ – Sampling rate in samples per second.
- $N_Q[\text{bits/symbol}]$ – Bitwidth.
- N_A – Number of Antenna Ports.

The options 1, 6, 7.1, 7.2 and 7.3 are the functional splits identified by 3GPP as the possible splits that will define the Fronthaul throughputs in the NR system. In the Option 7.1 (Low PHY) the Fast Fourier Transform is applied locally in the DU, which causes data to be transmitted over the Fronthaul interface to be represented by subcarriers. In this split, the Fronthaul throughput is lower than in Option 8, but it is still constant since the resource element mapping is still executed in the CU. This Fronthaul link option generates a RU throughput in the UL and DL given by:

$$R_{7.1_DL} \text{ [Mbps]} = 2000_{[s^{-1}]} N_{SC} N_{SY} [\text{symbols}] N_Q[\text{Mbits/symbol}] N_L + M_{info} \text{ [Mbps]} \quad (3.44)$$

$$R_{7.1_UL} \text{ [Mbps]} = 2000_{[s^{-1}]} N_{SC} N_{SY} [\text{symbols}] N_Q[\text{Mbits/symbol}] N_A + M_{info} \text{ [Mbps]} \quad (3.45)$$

where:

- N_{SC} – Number of subcarriers used in the system.
- $N_{SY} [\text{symbols}]$ – Number of symbols.
- N_L – Number of layers in the system.
- M_{info} – Information in the MAC in Mbps.

In Option 7.2 (Low PHY/High PHY) the precoding and the resource element mapper are executed by the DU. This process requires a more complex DU and achieves a lower throughput in the Fronthaul. This is the first considered option where the link throughput becomes variable. This Fronthaul link option generates a RU throughput in the UL and DL given by:

$$R_{7.2} \text{ [Mbps]} = (2000_{[s^{-1}]} \cdot N_{SC} N_{SY} [\text{symbols}] N_Q[\text{Mbits/symbol}] N_A) \mu_s + M_{info} \text{ [Mbps]} \quad (3.46)$$

where:

- μ_s – Subcarrier utilisation (load).

In Option 7.3 (High PHY) the modulation and the layer mapper are included as DU functionalities, which decreases the overall throughput in the Fronthaul, and data are transmitted using codewords. This Fronthaul link option generates a RU throughput in the UL and DL given by:

$$R_{7.3} \text{ [Mbps]} = (2000_{[s^{-1}]} N_{SC} N_{SY} [\text{symbols}] N_Q[\text{Mbits/symbol}] N_L) \mu_s + M_{info} \text{ [Mbps]} \quad (3.47)$$

In Option 6 (MAC-PHY) the data link layer is separated from the physical one, and the payload is transmitted over the Fronthaul using transport blocks, which leads to a large reduction of the Fronthaul throughput. This option also assigns a higher percentage of the packet to the overhead from scheduling control, synchronisation and frame carry. This Fronthaul link option generates a RU throughput in the UL and DL given by:

$$R_6 \text{ [Mbps]} = (R_p \text{ [Mbps]} + R_c \text{ [Mbps]}) \left(\frac{B \text{ [MHz]}}{B_c \text{ [MHz]}} \right) \left(\frac{N_L}{N_{L,c}} \right) \left(\frac{\log_2 M}{\log_2 M_c} \right) \quad (3.48)$$

where:

- $R_p \text{ [Mbps]}$ – Peak rate.
- $R_c \text{ [Mbps]}$ – Signalling Rate.
- $B \text{ [MHz]}$ – System Bandwidth.
- $B_c \text{ [MHz]}$ – Bandwidth for control signals.
- $N_{L,c}$ – The number of layers for control signalling.
- M – Modulation order.
- M_c – Modulation order for control signals.

In Option 2 (RLC-PDCP) the Packet Data Convergence Protocol (PDCP) and the Radio Link Control (RLC) are centralised while the other functionalities are performed in a local DU. This split uses an already standardised interface, F1, and is implemented in 5G as the interface option for the Middlehaul link [ETSI18e]. This Middlehaul link option generates a DU throughput in UL and DL given by:

$$R_2 \text{ [Mbps]} = R_p \text{ [Mbps]} \left(\frac{B \text{ [MHz]}}{B_c \text{ [MHz]}} \right) \left(\frac{N_L}{N_{L,c}} \right) \left(\frac{\log_2 M}{\log_2 M_c} \right) + R_c \text{ [Mbps]} \quad (3.49)$$

The throughputs of the Backhaul link (interconnecting the CU and the Core) and the Transport Network (interconnecting the Core and the External Data Centres) have typical values for the Next Generation (NG) interfaces [ITUT18]. These throughput values are represented in the Table 3.9.

Table 3.9– Typical NG Backhaul and Transport Network throughputs (extracted from [ITUT18]).

Link Type	Link Throughput [Gbps]
NG Backhaul	25
NG Transport Link	>100

3.3 Model Implementation

The model for the estimation of the E2E latency used in this thesis was developed using MATLAB. 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 architectures and MEC node deployments, and therefore these initial latency calculations are performed. After the initial calculations, the program calculates the remaining latency contributions depending on the network architecture and the system that were specified in the input parameters.

The 4G network is already installed and typically in C-RAN networks the BBU and the RRH are not collocated, and therefore the scenario of the collocated RRH and BBU is not considered in the model. If the scenario under study is 5G, there are 4 possible architectures under study: the collocated

CU/DU/RU, the collocated DU/RU, the collocated CU/DU and the independent CU, DU and RU.

Figure 3.14 represents the Model flowchart.

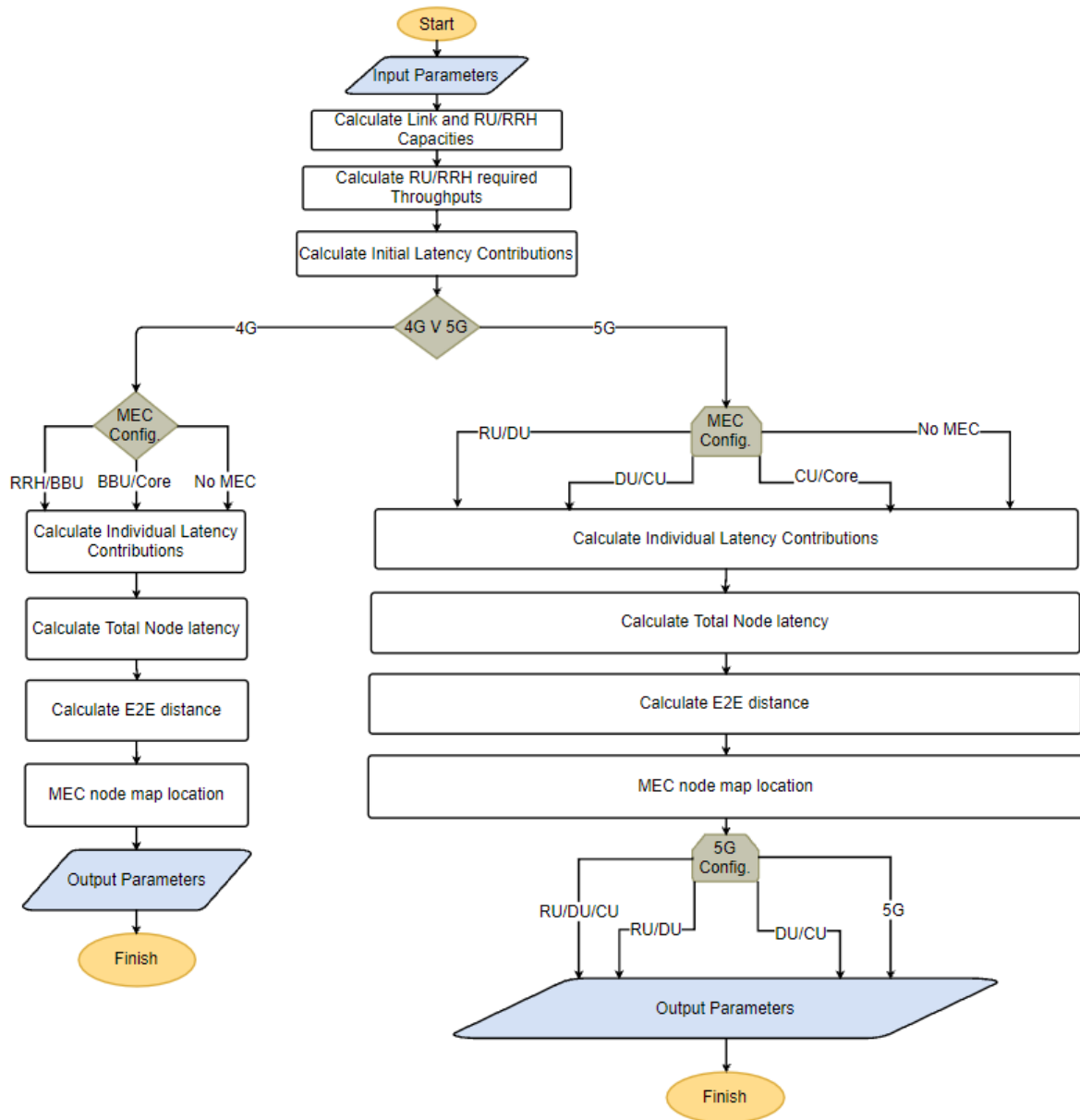


Figure 3.14 – Model Flowchart.

After defining which of the architectures is present in the studied scenario, it is important to understand if there is a MEC node installed somewhere along the network because it is considered that the packets will not be sent forward in the network in the cases where MEC nodes are installed, thereby reducing latency. The possible positioning of the MEC nodes for 5G is present in Figure 3.4, and in Figure 3.13 one represents the deployment of a MEC node between 2 collocated nodes. Figure 2.12 shows the possible implementations of the MEC node for the 4G architecture. In the 5G case, there are 3 possibilities of the MEC node deployment: Option RU/DU (between the RU and the DU), Option DU/CU (between the DU and the CU) and Option CU/Core (between the CU and the Core of the network). In the 4G case, there are 2 considered possibilities of MEC node installation: Option RRH/BBU (between

the RRH and the BBU) and Option BBU/Core (between the BBU and the EPC). After the earlier described phases, the program calculates individually each one of the remaining latency contributions. Some of these delays depend on each other, such as the processing delay of the UE, which depends on the packet transmission time to the Air Link. Also, the FH splitting option defines the distribution of functionalities between the RU and the DU, having a direct impact in the processing latency on each of the nodes and the overall E2E latency.

After calculating each of the individual latency contributions, the program computes the sum of all contributions obtaining as result the total node latency, which does not account for 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 between the nodes (E2E distance calculation) for each service group. After obtaining the 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.

All the described phases allow the program to produce output parameters, which leads to the results and allow the program user to analyse the obtained values, as well as the MEC node location in the map. Some input parameters for the algorithm to run properly are a service mix for the RU, the DU and the CU in both the transmitter and the receiver sides, which allows to calculate the traffic and the queuing delays in the nodes. The specific user under study is isolated from its priority group in the input data to be studied particularly, and to allow the calculation of the latency contributions that are not influenced by the traffic, such as the processing, transmission, and propagation delays. It is important to refer that it is considered that the sum of the data rates of the services to be provided by the RU does not exceed its throughput, since this scenario would create high delays in the node.

3.4 Model Assessment

In the model assessment phase, the validation of the developed model is performed, the input parameters, output parameters, and results of the intermediate steps of the program are checked.

It is a crucial phase, because this evaluation provides the validation that the model is well developed and implemented, and that it is ready to be used without incurring in any potential errors. The assessment of the model is performed in 8 different stages, which are shown in Table 3.11.

The first phase of the model assessment is the validation of the input file. This file contains the definition of the parameters that are needed for the program to be able to run, such as: architecture type, MEC node existence and positioning, and input parameters that are directly used to calculate latency contributions. Without the proper reading of the input file, the program is not able to be used for the calculation of latency, which creates the need for the existence of this test. The first phase of the model assessment was performed by verifying that the defined input variables of the program were correctly saved in the program variables.

Table 3.10 – Model Assessment Tests.

Test ID	Description
1	Validation of the input file, by verifying if the values stored in the program variables correspond to the ones in the input files.
2	Validation of the initial variables required to run the program properly: <ul style="list-style-type: none"> • Check if throughputs, user data and functionalities vectors are properly created and filled according to program inputs. • Check if the RU/RRH used throughputs and capacities are properly calculated and filled.
3	Validation of the initial latency contributions calculation.
4	Validation of the calculation of the individual latency contributions without MEC node: <ul style="list-style-type: none"> • Check transmission latency contributions. • Check processing latency contributions. • Check queuing latency contributions.
5	Validation of the MEC node associated latency in MEC node scenarios.
6	Validation of the maximum E2E distance and total node latency, by verifying if the sum of the latency contributions is being well performed by the program in each scenario, according to the program parameters and by checking if the sum of the link lengths is correctly calculated.
7	Validation of the MEC node positioning map, checking if the scaling of the distances is within the correct proportions and if the map generated by the program is correctly displayed.
8	Validation of the output file, by performing the verification of the values that are printed as the output file variables.

The second phase of model assessment is the validation of the initial variables required to run the program properly, which is essential for the correct calculation of all latency contributions. The throughputs vector has a direct impact on queuing and transmission delays, and the functionalities vector defines the processing delay in the nodes. The second phase of the model assessment was divided in 4 main stages: check the throughput vector in the transmitter and receiver sides, check if the functionalities vector is correctly created depending on the FH splitting option, check if the user data vector is correctly created by the program, verify if the algorithm that calculates used throughputs and capacities in the RU/RRH nodes is functioning correctly.

Both the third and fourth phases of the assessment of the model are based on the correct calculation of latency contributions. The third phase is responsible for the assessment of latency contributions that are

calculated initially, which are present in the E2E latency independently of the type of network or MEC node deployment option.

The fourth phase is responsible for the assessment of latency contributions that depend on the network type and the MEC node deployment in 4G and 5G scenarios with independent nodes and without MEC node, because these scenarios allow to verify the correct calculation of all latency contributions except the MEC node processing and transmission delays.

The third phase of the model assessment is finished by checking the correct calculation of 4 parameters: transmission latency of the transmitter UE, UE processing latency in the transmitter side and both the air link latencies in the transmitter and receiver sides. The transmission latency of the UE calculation was tested by changing the reference user service, and verifying if the results were correct, and the UE processing latency was tested by changing the system and the service type. Both tests had the expected results compared to the theoretical results generated by the expressions in this thesis. The graphical representation of the Air Link latency contribution when varying the distance between the end user and the RRH/RU took a linear shape, as expected.

The fourth phase of the model assessment is based on the testing of both 4G and 5G systems without MEC node and with the node independent architectures, because the evaluation of this specific scenario contains all latency contributions that are present in the other scenarios, except the MEC node contribution that was tested separately. For both the specific tests performed for the explained 4G and 5G scenarios, all latency contributions were correctly calculated. In both systems, transmission latency contributions were correctly scaled according to the size of the packet and the throughput/data rate. Processing latency contributions were also individually tested, and results were according to what was expected, with the increase of the variables according to the number and complexity of the functionalities in the node.

The fifth phase of model assessment was performed by checking if the MEC node processing and transmission delays are calculated correctly, since in the earlier phases the rest of the latency contributions were already verified. In this fifth phase, it was checked if the MEC node processing latency responds correctly to the increase in the size of the packet, and if the MEC node transmission latency is calculated according to the link that interconnects the previous node and the MEC node. Both calculations were verified and considered correct in the model.

The sixth phase of the assessment is performed by checking the total node latency, and if its value corresponds to the sum of the latency contributions calculated earlier, and by checking the maximum E2E distance, which depends on the total node latency. The latency accumulated in all nodes was calculated correctly in all tested scenarios after some minor corrections in the equations. The maximum E2E distance was tested in a particular scenario in which the service number 4 was used. To verify the maximum E2E distance behaviour in the model 3 scenarios were created: the RU/DU MEC node scenario, the DU/CU MEC node scenario, and the DU/CU MEC node scenario with reduced DU traffic. In the test, the links that connect the network nodes were considered optical fibre. The graphical representation of the Maximum E2E distance as a function of the E2E latency for the testing scenario took a linear shape, as expected. Figure 3.15 shows the RU queuing latency contribution when varying

the splitting option (and consequently the FH throughput), for a fixed RU traffic scenario.

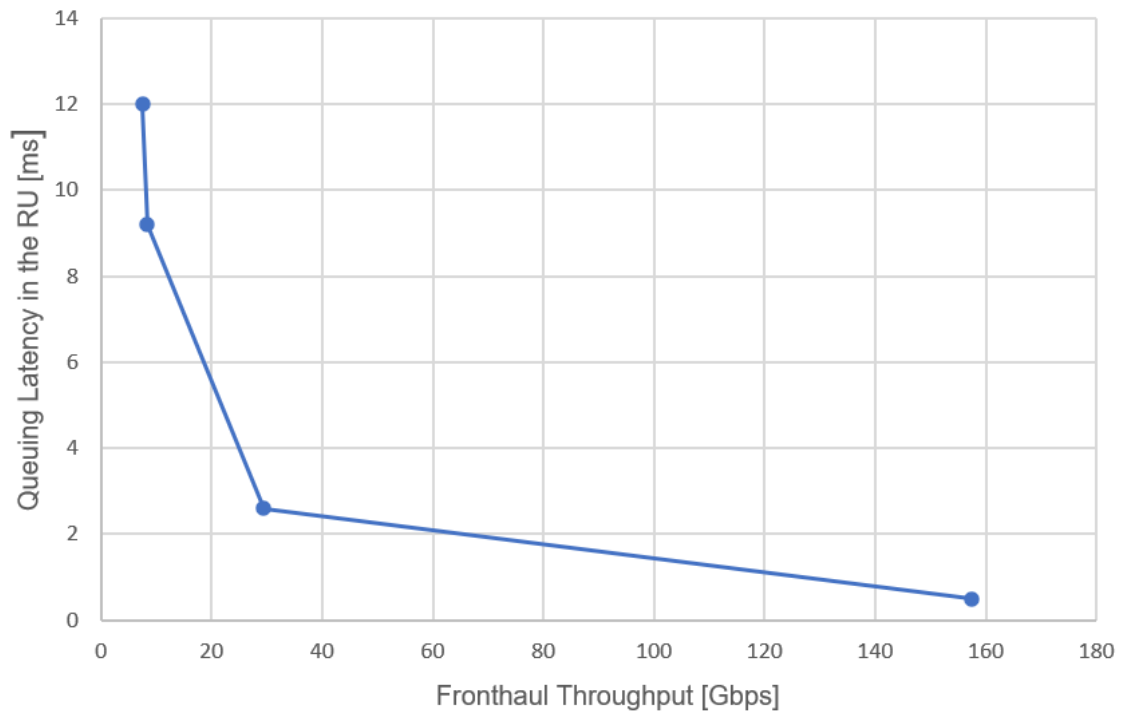


Figure 3.15 – RU queuing latency plot while varying the FH throughput.

The seventh phase of the model assessment was performed by checking the scale of the map that locates the possible location of the MEC nodes, according to the maximum E2E distance.

The last phase of the model assessment is the validation of the output file. This file contains the definition of the parameters that are needed for the program user to analyse and extract, to study the results and create the conclusions for the thesis. The list of the parameters needs to be coherent with the rest of the program phases. The correct creation of the program output file was verified several times along all previous tests, and the output file was correctly generated in all performed tests.

Chapter 4

Results Analysis

This chapter provides a description of the simulated scenarios and the results obtained after applying the latency model described in the previous chapter to the described scenarios.

4.1 Scenarios

The study in this thesis is divided into five main scenarios considering 3 types of service to be provided: Remote Surgery (inside and outside hospitals), Intelligent Transport Systems (in urban and highway scenarios) and Factory Automation. For these case studies, data provided by NOS has been taken (e.g., location of A1 highway sites, location of hospitals and latency between each site and the closest Core of the Network). Table 4.1 represents the comparison among the simulated scenarios.

Table 4.1 – Scenario Comparison.

Scenario	Simulated Service	Max. E2E latency [ms]	Data Rate [Mbps]
Santa Maria Lisbon Hospital	Internal Remote Surgery	1	10.912
Espírito Santo Évora Hospital	External Remote Surgery	3	10.912
A1 Highway ITS	Network Based Sensor Sharing	3	20
Avenida da Liberdade Urban ITS	Remote Driving	5	25
AutoEuropa Factory	Factory Automation	0.25	1

Figure 4.1 represents the Santa Maria Hospital NOS sites location.



Figure 4.1 – Santa Maria Hospital NOS sites.

The Santa Maria Hospital scenario studies the possibility of implementing internal remote surgeries inside this hospital which can be performed using Control Manipulations, High-Quality Video Streaming and Haptic Feedback. The Santa Maria Hospital has a total of 2 722 medical professionals with a total

building area of 125 504 m², in which is included the central surgery department, that has the capacity to provide 14 simultaneous remote surgeries. This hospital is located in the Lisbon region, which has:

- 2 821 697 inhabitants.
- an average populational density of 948 inh./km².
- 639 RRH nodes (in 2019, extracted from [SeDo19]).
- 55 BBU nodes (in 2019).
- 1 Network Core deployed (in 2019).

The Espírito Santo Évora Hospital scenario studies the possibility of implementing external remote surgeries in this hospital, which can be performed using Control Manipulations, High-Quality Video Streaming and Haptic Feedback. Figure 4.2 represents the NOS sites location in Évora.

In 2019, in Espírito Santo Évora Hospital were performed 18 530 surgeries and considering the possibility that 10% of these surgeries could be performed as an external surgery, outside the hospital, a plausible number of external remote surgeries per year is 1 853, which is equivalent to approximately 5 external surgeries per day (possibly simultaneously). The Évora city has:

- a total of 56 596 inhabitants
- an average populational density of 43.3 inh./km².



Figure 4.2 – Espírito Santo de Évora Hospital NOS sites.

In the Avenida da Liberdade scenario, one considered the implementation of the remote driving service in a typical urban scenario, more specifically in Avenida da Liberdade in Lisbon. In this scenario, there are 2 main services under study: remote driving and traffic information. Since the latency requirements for remote driving are in the order of 5 ms in comparison with 1 000 ms of traffic information, this is the simulated service in the program, due to its stricter latency. It is also considered that Avenida da Liberdade is one of the Lisbon avenues with the most intense traffic accumulation during the rush hour, thus having a high traffic density. Figure 4.4 represents Avenida da Liberdade NOS sites location. To generate the traffic scenario for Avenida da Liberdade one

takes into consideration the following data:

- Avenida da Liberdade in Lisbon has a total length of 1 000 m.
- 3 lanes in both ways.
- It is considered that Avenida da Liberdade is full of vehicles.
- The average car length is 4.5 m.
- All considered traffic is split among the 3 sites.
- One third of the car drivers are considered NOS clients.
- The RU/RRH capacity will need to accommodate approximately 20 cars.
- The Urban ITS scenario is studied for a traffic usage of 30%, 60% and 90%, being the above-described scenario the 60% traffic scenario.



Figure 4.3 – Avenida da Liberdade NOS sites.

The A1 Highway scenario studies the possibility of implementing advanced driving (high automation level with sensors, traffic information and intelligent routing provided by the network) along the A1 highway. It is considered that latency requirements for the highway are strict (3 ms) because the average speed in which cars are being driven is higher than the one inside cities (increasing the data rate requirements of sensors and route planning) and because sensor information is shared using the network. Figure 4.3 represents the Sacavém - São João da Talha section NOS sites location. In the highway scenario, one considered the coverage in a situation of fluent traffic flow (the worst-case scenario since data rates achieve higher requirements due to the car velocity). The A1 highway has a 303 km extension and its section with highest traffic is between Sacavém and S. João da Talha with a maximum traffic of 98 021 vehicles in Feb. 2020. The required data for the scenario specification is:

- The distance between Sacavém and S. João da Talha via A1 highway is approximately 3.5 kms,
- The highway has 3 lanes in each way.
- The average car length is 4.5 m.
- All traffic is split among the 3 sites.

- The considered safety distance between cars is 50 m on average in each of the 3 lanes.
- Only 30% of the cars are using high automation driving systems.
- One third of the car drivers are considered NOS clients.
- The RU/RRH capacity will need to accommodate approximately 10 cars.



Figure 4.4 – Sacavém - São João da Talha section NOS sites.

In the AutoEuropa Factory scenario, one considered the implementation of the factory automation services in this factory, which has critical latency requirements around 0.25 ms. Figure 4.5 represents AutoEuropa NOS sites location.



Figure 4.5 – AutoEuropa factory NOS sites location.

The required data for the scenario specification is:

- The number of AutoEuropa factory employees was 5 300 (in 2020).
- the number of AutoEuropa factory auxiliary robots was 430 (in 2015).
- Since it is known to be not viable to provide coverage for the entire factory with only one RU/RRH (due to the attenuation and high user density), one considered in the scenario that three RUs/RRHs are installed in the facilities to guarantee a good coverage and connectivity.

4.2 Radio Characteristics Analysis

This section presents an analysis of 4G and 5G radio characteristics, the throughputs that are achievable in each scenario, and also the comparison between the throughputs offered by the RRHs/RUs of each scenario and the user required data rates. This study presents the DL and UL RU/RRH capacity on both transmitter and receiver sides, as well as the required throughputs in each link for each scenario. The reference architecture is the one with the independent RU, DU and CU for the 5G scenario. Figure 4.6 illustrates the results obtained for the Santa Maria Hospital scenario, in which the radio node that transmits and receives Internal Remote Surgery data packets is the same, meaning that the Control Manipulation, Haptic Feedback and Video Streaming packets are sent in the RU/RRH UL and then received in the same RU/RRH node but in the DL.

The radio characteristics taken for all scenarios are present in Table L.1 and L.2. As it is possible to analyse from Figure 4.6, the 4G indoor RRH is not able to provide the required UL and DL capacity for the Internal Remote Surgery service, because the typical UL capacity of 4G 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. Even if the implemented radio characteristics maximised the achievable throughputs, 4G would not be able to provide the required throughputs for the scenario, although it could help to offload the 5G RU. According to the obtained results, 5G provides enough capacity in both UL and DL to allow the existence of the Internal Remote Surgery in the hospital, even without using the radio characteristics that maximise throughputs, such as the maximum Bandwidth or 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 DL and UL required throughputs of 295.18 Mbps and 197.79 Mbps.

Figure 4.7 illustrates the results obtained for the Espírito Santo Évora Hospital scenario, in which the radio nodes that transmit and receive the External Remote Surgery service data packets are different, meaning that there is a radio node to which the surgeon sends the Control Manipulation data packets, that is different from the radio node to which the robotic arm operating the patient sends the Video Streaming and Haptic Feedback data packets. The Control Manipulation packets are sent in the UL of the surgeon site and received in the DL of the patient site, and both the Video Streaming and Haptic Feedback packets are sent in the UL of the patient site and received in the DL of the surgeon site. The Espírito Santo Évora Hospital scenario RUs/RRHs are installed in environments that have different radio channel characteristics, meaning that the radio node in the surgeon site is in an indoor environment,

which typically has a better average CQI and higher offered throughputs than the outdoor environment, which characterises the radio node in the patient site.

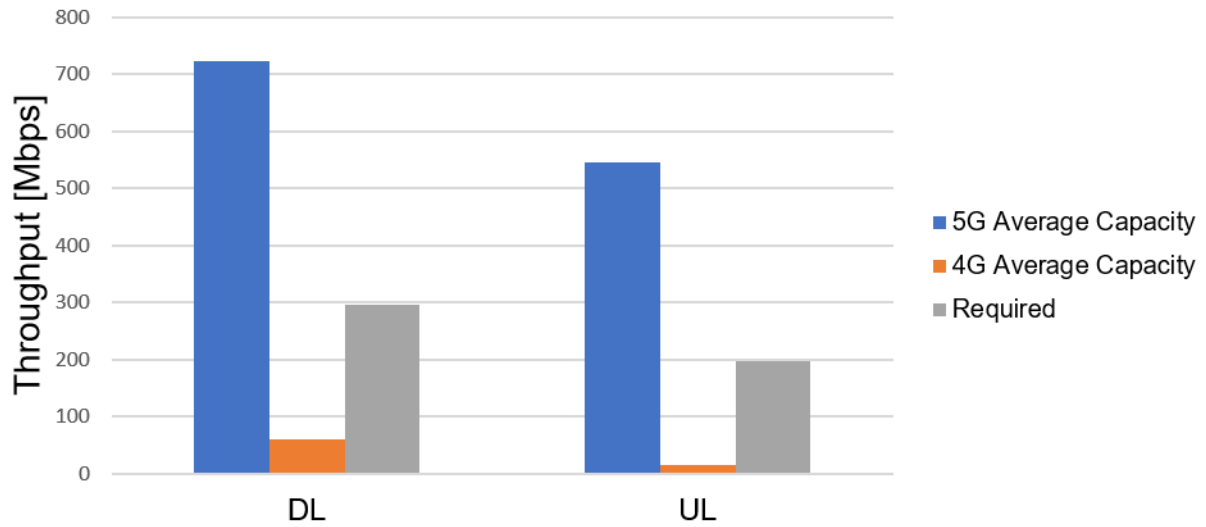


Figure 4.6 – RRH/RU throughputs for the Santa Maria Hospital scenario.

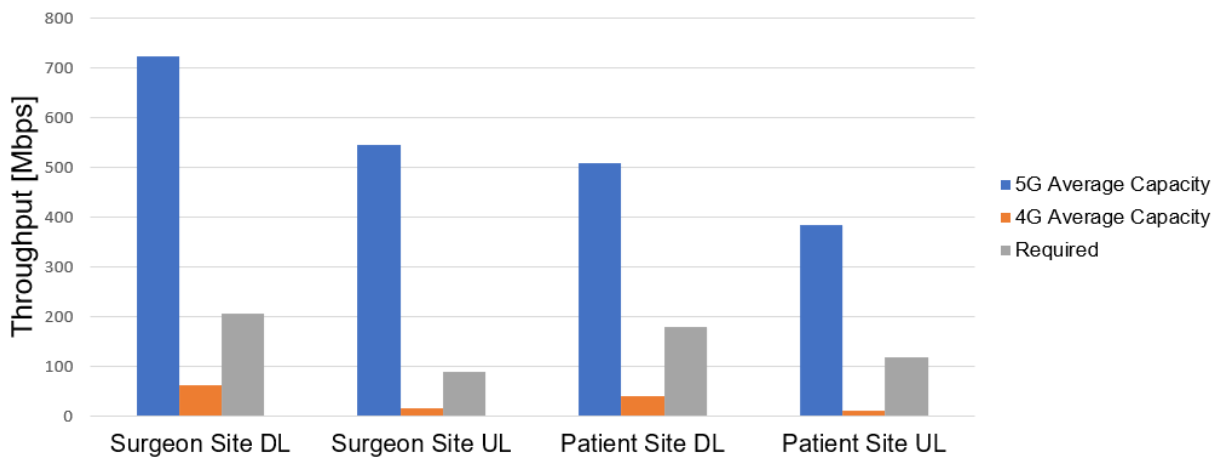


Figure 4.7 – RRH/RU throughputs for the Espírito Santo Évora Hospital scenario.

From the analysis of the graph present in Figure 4.7, it is possible to conclude that 4G does not provide enough capacity in both UL and DL to guarantee the connectivity to users. On the surgeon site for 4G, DL and UL capacities are 60.93 Mbps and 15.72 Mbps, respectively, which are below the 205.5 Mbps and 89.12 Mbps DL and UL required throughputs. On the patient site for 4G, the DL and UL capacities are 39.44 Mbps and 10.32 Mbps, respectively, which are below the 179.58 Mbps and 117.50 Mbps DL and UL required throughputs. On the other hand, 5G provides enough capacity in DL and UL in both surgeon and patient sites. On the surgeon site for 5G, the DL and UL capacities are 723.39 Mbps and 545.72 Mbps, respectively, and on the patient site the system provides 508.15 Mbps and 383.34 Mbps.

Figure 4.8 illustrates the results obtained for the Urban ITS scenario with traffic on the 60% percentile, in which the radio nodes that transmit and receive Remote Driving data packets are different, meaning that there is a radio node to which the remote driver sends the commands to the car, that is different from the radio node to which the passenger sends the Video Streaming data packets.

The commands to the car are sent in the UL of the remote driver site and received in the DL of the passenger site, and the Video Streaming and packets are sent in the UL of the passenger site and received in the DL of the remote driving site.

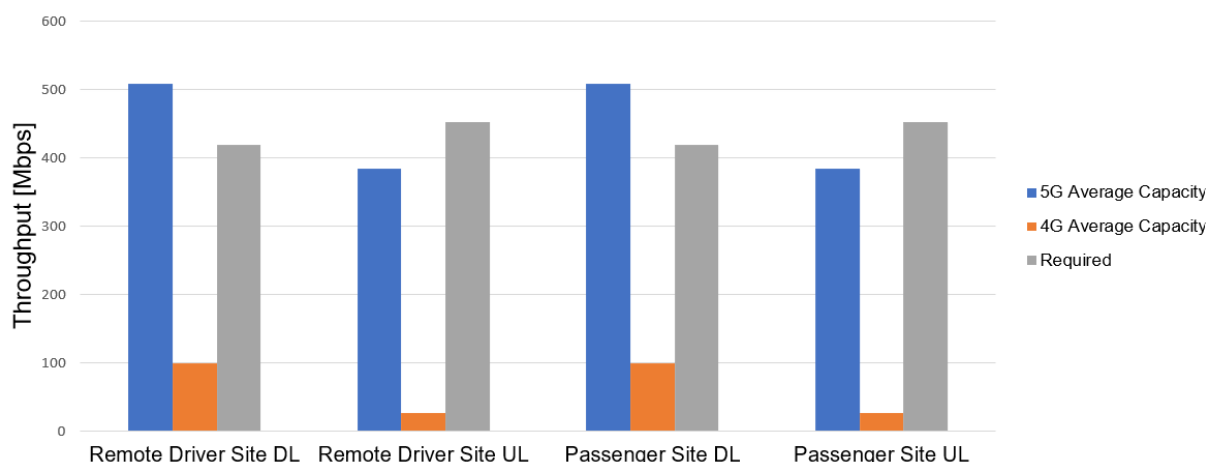


Figure 4.8 – RRH/RU throughputs for the Urban ITS scenario with 60% of the maximum traffic.

The Urban ITS scenario RUs/RRHs can be installed in environments that have different radio channel characteristics, for example, one in the city centre and other in the suburbs. In this simulation it is considered that both the radio nodes are installed in an urban environment and therefore they have approximately the same throughput requirements and capacities. From the analysis of Figure 4.8, it is possible to conclude that 4G does not provide enough capacity in both UL and DL to guarantee the connectivity of all users. On the remote driver site for 4G, DL and UL average capacities are 98.6 Mbps and 25.8 Mbps, respectively, which are below the 418.69 Mbps and 451.8 Mbps DL and UL required throughputs. These average capacities are higher than in the other scenarios because the base station that is deployed in Avenida da Liberdade uses three different frequency bands (according to the data provided by NOS): 2600 MHz, 1800 MHz, and 800 MHz. On the passenger site for 4G, both capacities and required throughputs are equal to the remote driver site because it is considered that the environments in which the sites are installed are similar. In the simulation, 5G provides enough capacity in DL, but not enough UL capacity in both the remote driver and passenger sites. These results are influenced by the fact that the data rate requirements of each considered service are at the high range of the possible interval, therefore, the required UL throughputs that would exist in a real scenario would be lower and satisfied by 5G. On both the remote driver and passenger site, DL and UL capacities are 508.15 Mbps and 383.34 Mbps, respectively. The throughput results obtained for the Urban ITS scenarios with 30% and 90% of the maximum traffic are presented in Annex M.

Figure 4.9 illustrates the results obtained for the A1 Highway scenario with traffic on the 60% percentile, in which the radio nodes that transmit and receive Network Based Sensor Sharing data packets are the same. The sender car transmits the packets in the base station UL and the receiver car receives the packets in the same base station DL.

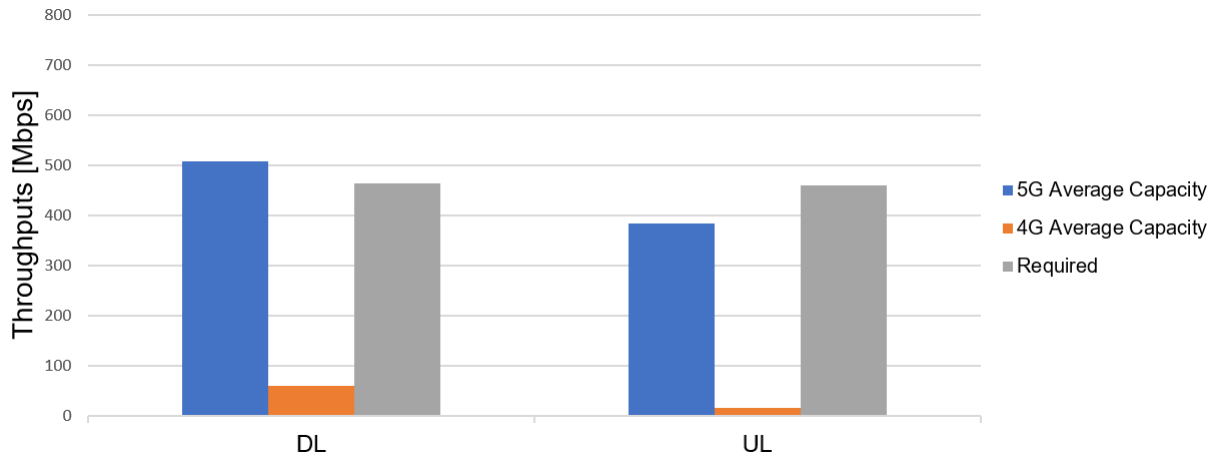


Figure 4.9 – RRH/RU throughputs for the A1 Highway scenario with 60% of the maximum traffic.

From the analysis of the graph present in Figure 4.9, it is possible to conclude that 4G does not provide enough capacity in both UL and DL to guarantee the connectivity of all users. For 4G, 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. These average capacities are higher than in the other outdoor scenarios because the base stations that are deployed in the A1 highway use two different frequency bands (according to the data provided by NOS): 1800 MHz and 800 MHz. In the simulation, 5G provides enough capacity in 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 5G. The throughput results obtained for the A1 Highway scenarios with 30% and 90% of the maximum traffic are presented in Annex M.

Figure 4.10 illustrates the results obtained for the AutoEuropa factory scenario, in which the radio node that transmits and receives Factory Automation service data packets is the same, meaning that the Packaging Machine, Machine Tools and Printing Machine packets are sent in the RU/RRH UL and then received in the same RU/RRH node but in the DL.

As it is possible to analyse from Figure 4.10, 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. If the implemented radio characteristics maximised the achievable throughputs, 4G would not be able to provide the required throughputs for the scenario, because the peak UL capacity is still below the required one. According to the obtained results, 5G provides enough capacity in both UL and DL to allow the existence of the factory automation services in the AutoEuropa factory, even without using the radio characteristics that maximise 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.

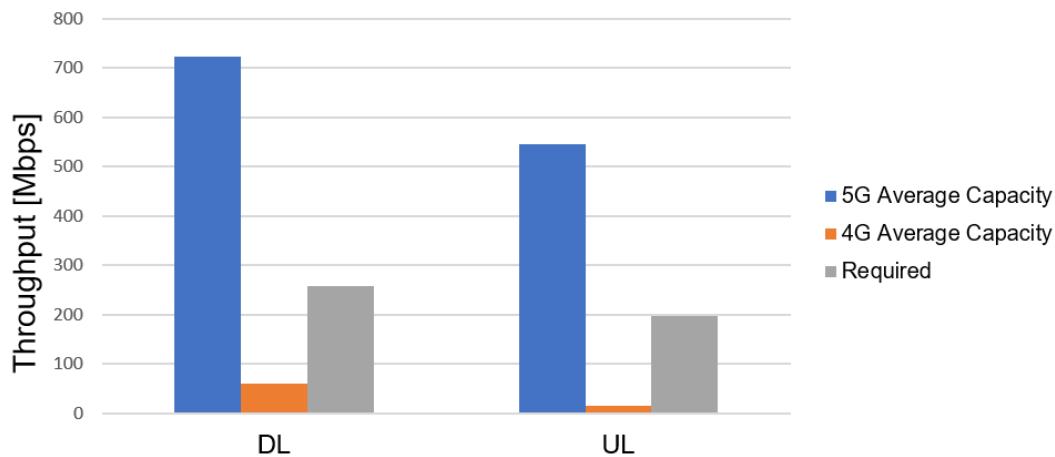


Figure 4.10 – RRH/RU throughputs for the AutoEuropa Factory scenario.

4.3 Total Node Latency and MEC Node Deployment Analysis

This section presents an analysis of 4G and 5G E2E latency and the MEC node deployment option chosen in each scenario to guarantee that the maximum E2E latency is not exceeded. The reference architecture is the one with the independent RU, DU and CU for 5G. This study is performed taking into account the several splitting options, even knowing that the splitting option that will be implemented in the 5G network is option 7.2. It is also important to refer that the viability of the MEC node deployments presented in the previous chapters is tested for each of the simulated applications.

Figure 4.11 illustrates the total node latency results obtained for the Santa Maria Hospital scenario, in which the maximum allowed latency value is 1 ms. There are several experiments of remote surgeries that were performed with higher E2E latency values, but the objective is to guarantee a high availability of the service and to make sure that the network keeps the delays in the packet transmission, queuing and processing as low as possible. For safety and insurance purposes, one considered a margin of 10% regarding the maximum E2E latency, meaning that any value that is higher than 90% of the maximum E2E latency is considered above the ideal value.

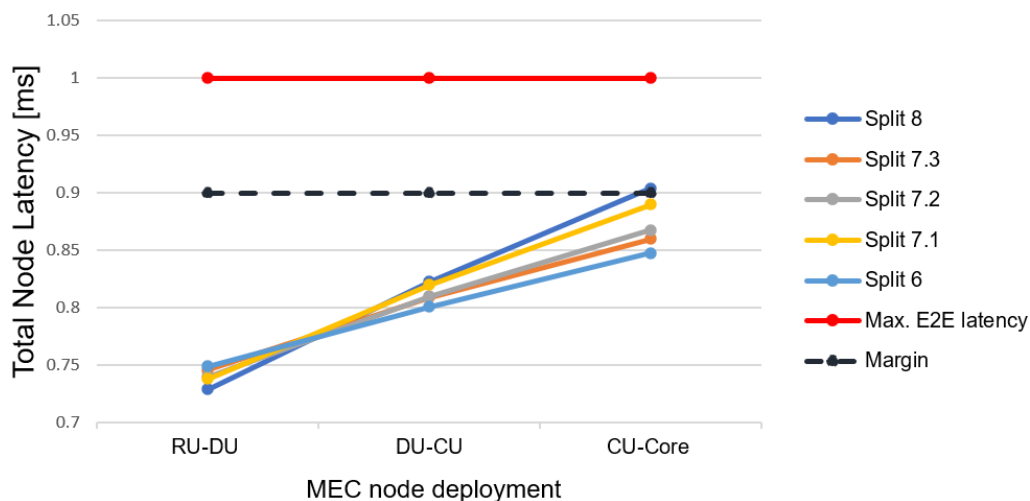


Figure 4.11 – Total node latencies for the Santa Maria Hospital scenario.

As Figure 4.11 shows, 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 QoS. 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 between the RU and DU. 4G does not provide sufficiently low latency values for the service implementation, since the minimum E2E latencies of the system are around 8 ms. The different curves 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 4.12 represents the individual latency contributions for each splitting option for the RU-DU MEC node deployment. The queuing latency is almost inexistant, because the remote surgery has a very high priority in the network. The processing latency is around 0.1 ms for every splitting option, because with the RU-DU MEC node deployment the only nodes that are required to process data are the RU and the MEC node. The represented transmission latency (0.625 ms) corresponds to the air transmission delay because the FH link capacity is very high, resulting in very low latency values for the optical fibre link transmission. The maximum propagation latency is calculated using the margin between the total node latency and the maximum E2E one that the service requires to be performed in good conditions.

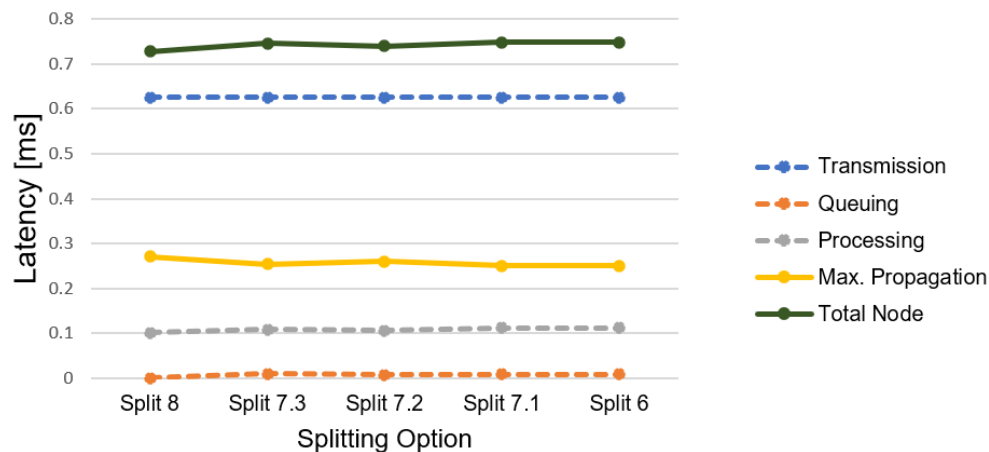


Figure 4.12 – Individual latency contributions for the RU-DU MEC node deployment in Santa Maria.

Figure 4.13 illustrates the total node latency results obtained for the Espírito Santo Évora Hospital scenario, in which the maximum allowed latency value is 3 ms. For safety and insurance purposes, one considered a margin of 10% regarding the maximum E2E latency. The E2E latency is kept below the margin in the CU-Core MEC node deployment option, with values below the 1 ms, which are significantly below the required margin. According to this information, the RU-DU and DU-CU MEC node positioning is not needed to achieve the latency requirements of the service, but if the Internal Remote surgery service is installed in the hospital, like in the performed simulation, the MEC node that processes the Internal Remote Surgeries may also be used for the External Remote Surgeries, which reduces the E2E latency significantly. The minimum 4G E2E latency is around 8 ms which is still high in comparison with the 3 ms requirement. The curves are also near each other independently of the MEC node deployment option, which happens because the queuing latency added to the E2E one is only due to the remote

surgery packets, since this is a high priority service.

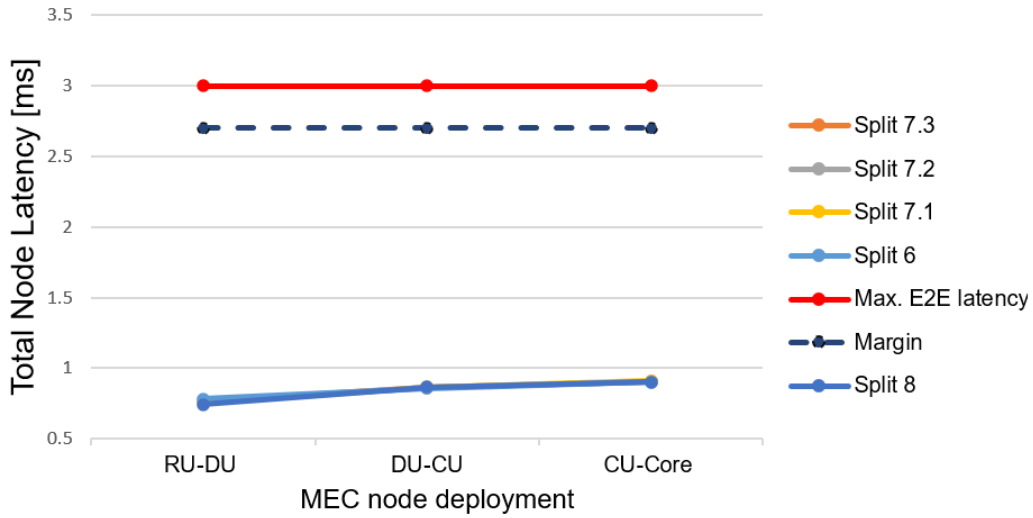


Figure 4.13 – Total node latencies for the Espírito Santo Évora scenario.

Figure 4.14 represents the individual latency contributions for each splitting option for the CU-Core MEC node deployment.

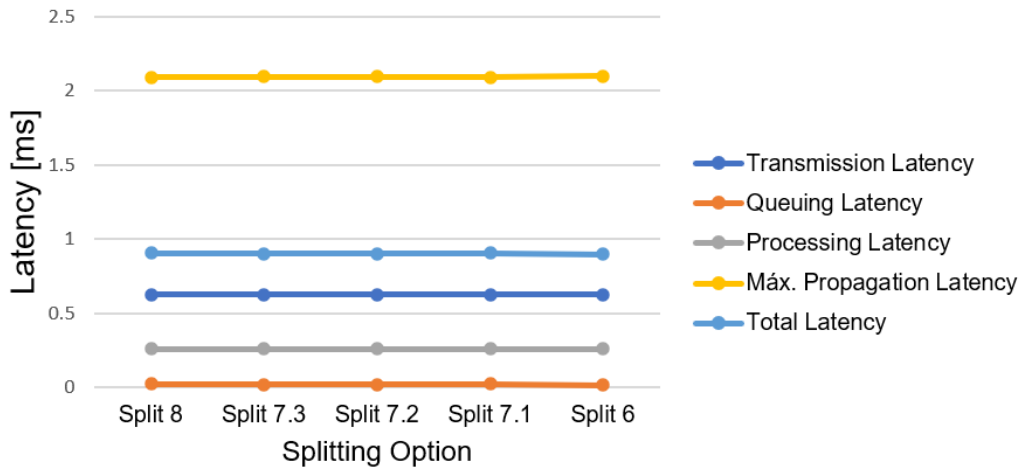


Figure 4.14 – Individual latency contributions for the CU-Core MEC node deployment in Espírito Santo Évora.

The queuing latency is almost inexistent, because the external remote surgery has a very high priority in the network. The processing latency is around 0.26 ms for every splitting option, because with the CU-Core MEC node deployment there is a fixed number of nodes that process the packets, resulting in a stable processing delay. The represented transmission latency (0.625 ms) corresponds to the air transmission delay because the FH link capacity is very high, resulting in very low latency values for the optical fibre link transmission. The maximum propagation latency is calculated using the margin between the calculated Total Node latency and the maximum E2E latency that the service requires to be performed with good conditions.

Figure 4.15 illustrates the E2E latency results obtained for the Urban ITS with 60% of the maximum traffic scenario, in which the maximum allowed latency value is 5 ms.

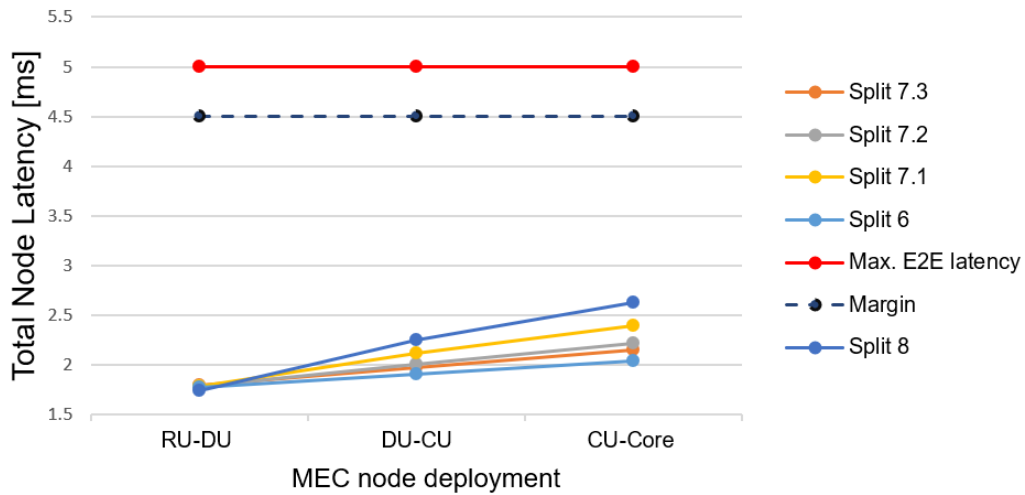


Figure 4.15 – Total node latencies for the Urban ITS scenario with 60% of the maximum traffic.

The studied service in this scenario is Remote Driving, in which the network is fully responsible for the correct and efficient transmission and reception of packets, because this service is classified as a pure V2N service, in contrast with Sensor Sharing that can be performed between the vehicles. For safety and insurance purposes, one considered a margin of 10% regarding the maximum E2E latency.

It is possible to conclude that the MEC node deployment between the CU and the Core in 5G is the most adequate deployment to keep the E2E latency below the margin, since it reduces the complexity and processing capacity of the MEC nodes, and simultaneously keep the latency considerably below the 4.5 ms margin. 4G has a minimum air transmission latency of 8 ms which keeps the E2E latency above the 5 ms threshold and makes the system not suitable to provide this application.

Figure 4.16 represents the individual latency contributions for each splitting option for the CU-Core MEC node deployment.

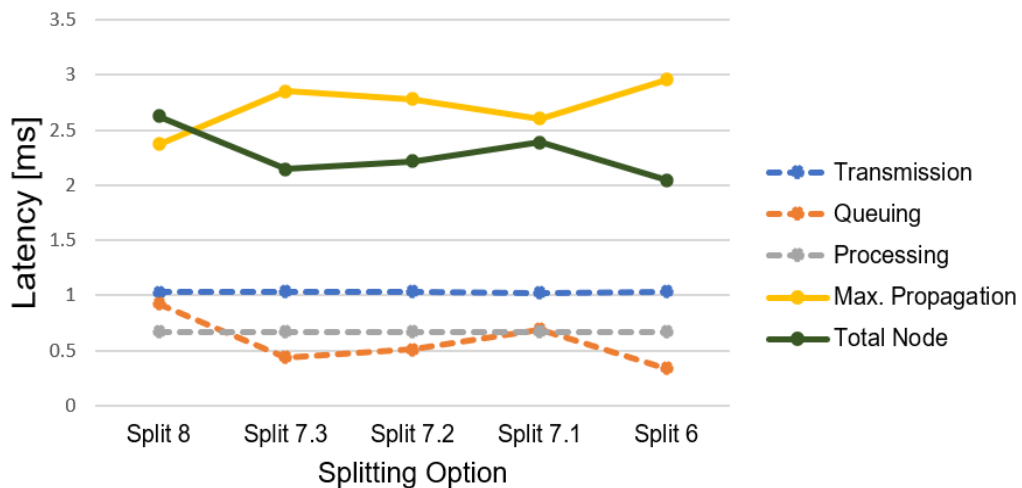


Figure 4.16 – Individual latency contributions for the CU-Core MEC node deployment in the Urban ITS scenario for 60% traffic usage.

Queuing latency oscillates between 1 ms and 0.3 ms, because traffic aggregation depends on the chosen splitting option. The processing latency is around 0.671 ms for every splitting option, because with the CU-Core MEC node deployment there is a fixed number of nodes that process packets,

resulting in a stable processing delay. The represented transmission latency (1.03 ms) corresponds to the air transmission delay accumulated along the network nodes.

Figure 4.17 presents the total node latency results for 30%, 60% and 90% of the maximum traffic load for the splitting option 7.2. The CU-Core MEC node deployment option keeps the E2E latency below the margin even for intense traffic loads, which makes that option of installation a viable solution to the Remote Driving service.

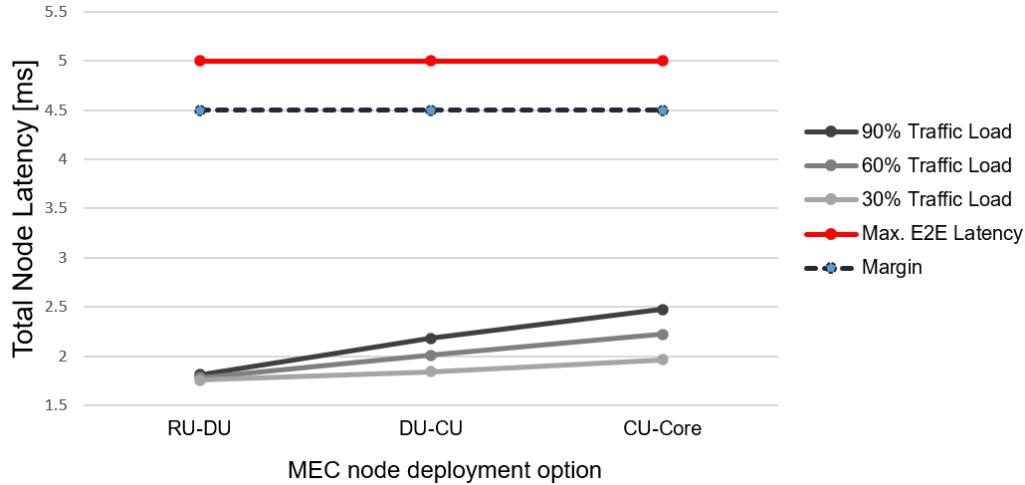


Figure 4.17 – Total node latencies for the Urban ITS scenario with increasing percentages of the maximum traffic for the splitting option 7.2.

Figure 4.18 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 ITS. For safety and insurance purposes, one considered a margin of 10% regarding the maximum E2E latency.

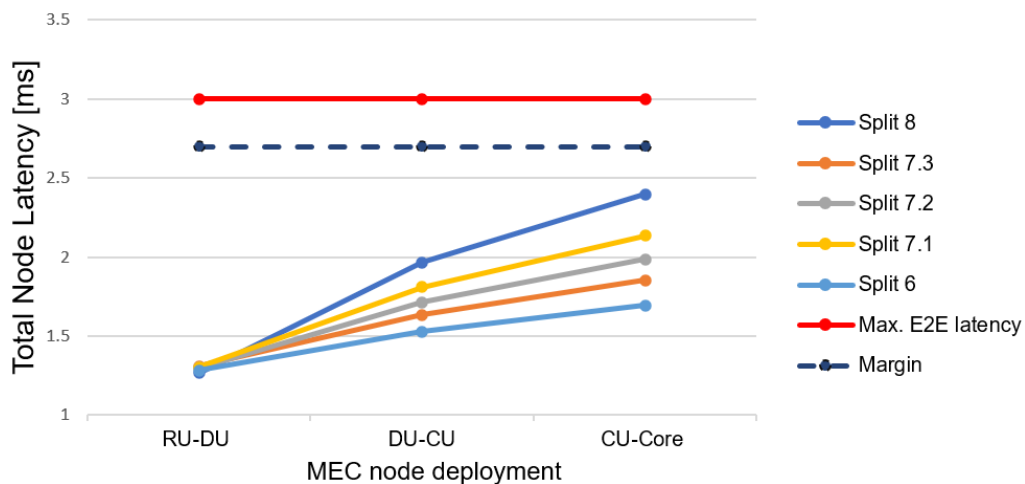


Figure 4.18 – Total node latencies for the A1 Highway scenario with 60% of the maximum traffic.

Figure 4.19 represents the individual latency contributions for each splitting option for the DU-CU MEC node deployment. The queuing latency oscillates between 0.3 ms and 0.8 ms because the added delay that occurs due to the queuing of the packets depends on the traffic aggregation, which is correlated to

the splitting options. The processing latency is around 0.36 ms for every splitting option, because with the DU-CU MEC node deployment there is a fixed number of nodes that process the packets, resulting in a stable processing delay. The represented transmission latency of 0.8 ms exists mainly due to the air transmission, because the link transmission adds very low latency values.

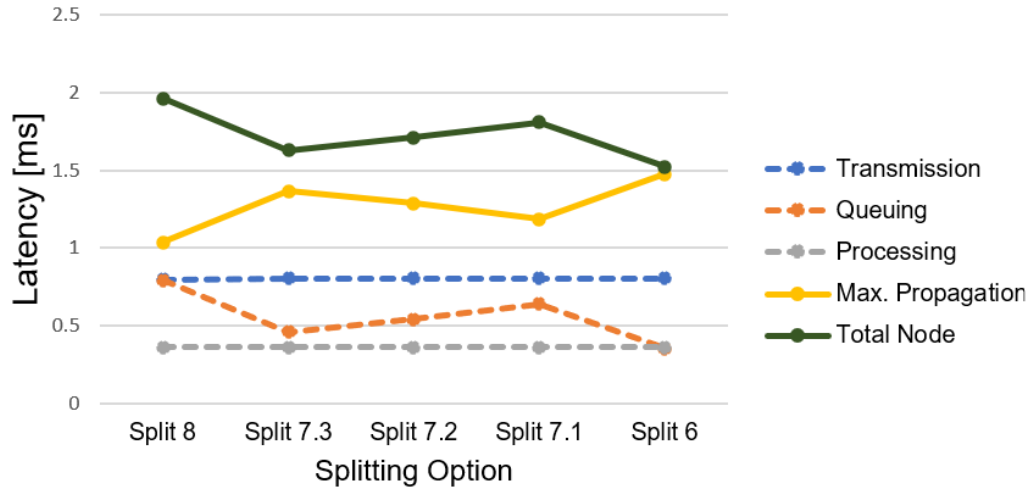


Figure 4.19 – Individual latency contributions for the DU-CU MEC node deployment in the Highway scenario for the 60% traffic usage.

Figure 4.20 presents the E2E latency values for 30%, 60% and 90% of the maximum traffic load for the splitting option 7.2. It is possible to conclude that the MEC node deployment between the DU and the CU (in 5G) is the most adequate installation to keep the E2E latency below the margin. The installation of the MEC node between the CU and the Core keeps the latency too close to the margin for the 60% percentile of the maximum traffic.

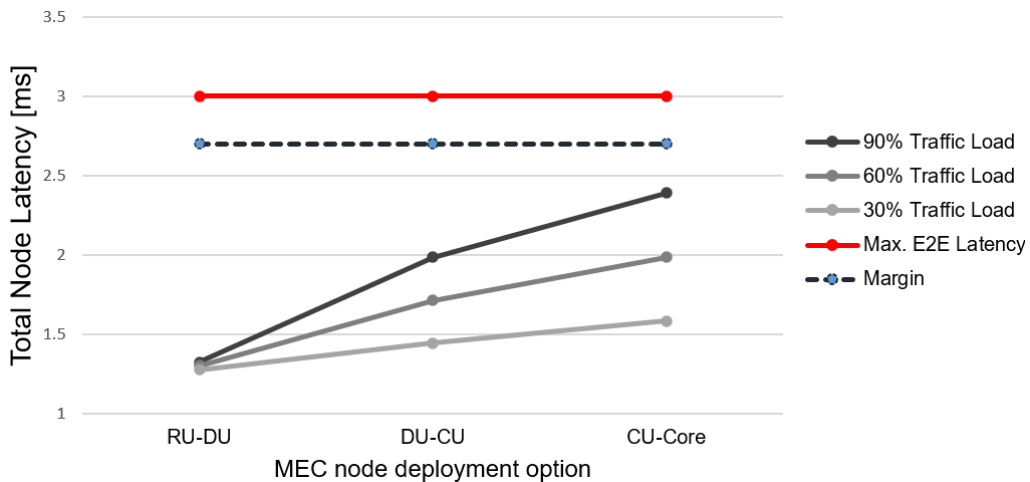


Figure 4.20 – Total node latencies for the A1 Highway scenario with increasing percentages of the maximum traffic for the splitting option 7.2.

The CU-Core MEC node deployment option is not enough to keep the E2E latency significantly below the margin for intense traffic loads, which makes that option of installation an inviable solution for the Network Based Sensor Sharing. On the other hand, the DU-CU deployment option responds well in terms of keeping the latency below the margin threshold even for intense traffic scenarios, which makes

this deployment option the most viable solution of the graph, since there is no need to increase the MEC node load by using the RU-DU MEC node deployment.

Figure 4.21 illustrates the total node latency results obtained for the AutoEuropa factory scenario, in which the maximum allowed latency value is 0.25 ms. This value generates very strict delay demands on the network, and accounting for the latency margin of 90%, the E2E latency should be below 0.225 ms. 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 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 exceed the margin. 4G 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.

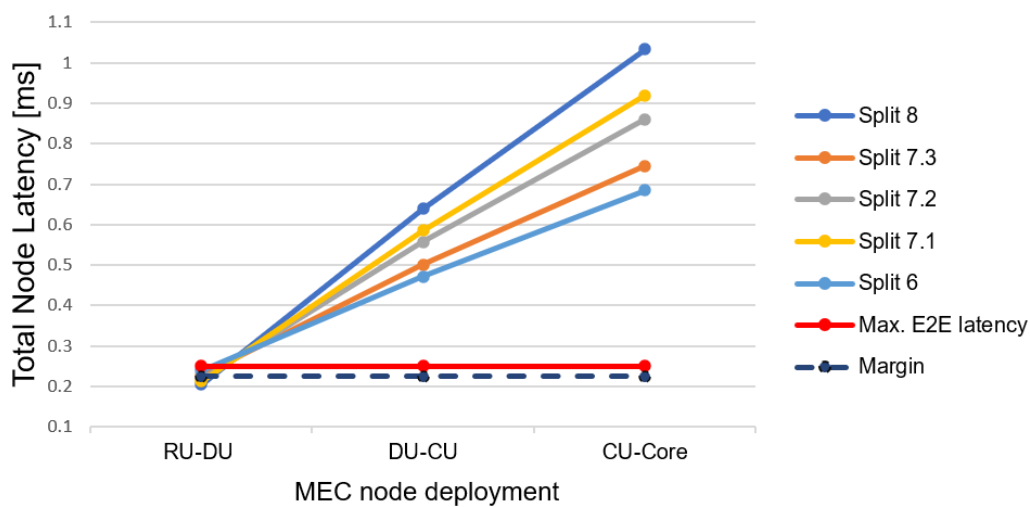


Figure 4.21 – Total node latencies for the AutoEuropa Factory scenario.

Figure 4.22 represents the individual latency contributions for each splitting option for the RU-DU MEC node deployment in the AutoEuropa factory scenario.

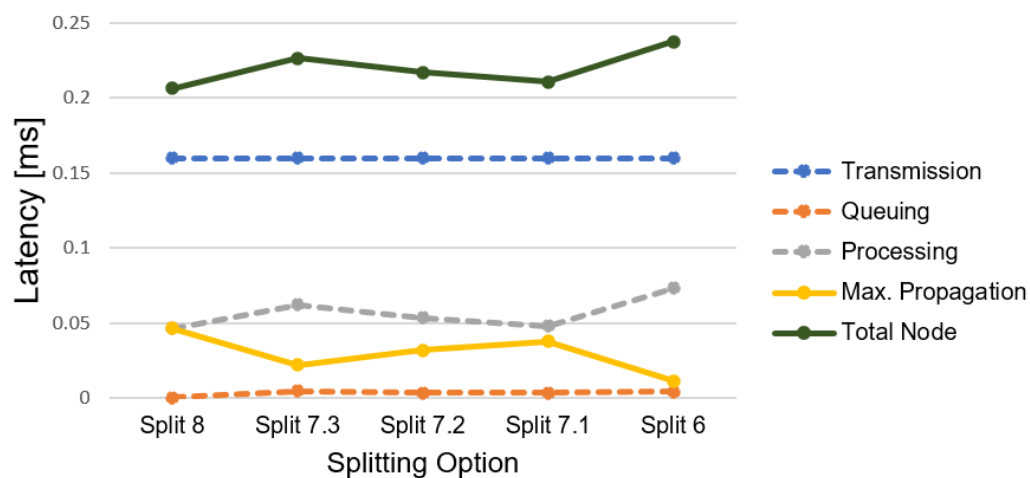


Figure 4.22 – Individual latency contributions for the RU-DU MEC node deployment in AutoEuropa.

The queuing latency takes low values, because the RU-DU MEC node deployment only adds queuing delay in the RU and the MEC node, since the Edge technology is installed between the RU and the DU.

The processing latency takes values between 0.05 ms and 0.08 ms, because packets for the simulated factory automation service only have 10 bytes. The represented transmission latency of 0.16 ms exists mainly due to the air transmission, because the link transmission adds very low packet delays, since the links provide high throughputs.

4.4 E2E Distance Analysis

The E2E distance is calculated in the model, representing a result that indicates the maximum distance from the user transmitter, in which the MEC node can be installed to keep the E2E latency below the maximum value allowed for a certain service. In this section, results are presented in a graphical way in a map, and in a numerical way through tables that show the maximum E2E distances.

Figure 4.23 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 Santa Maria Hospital.

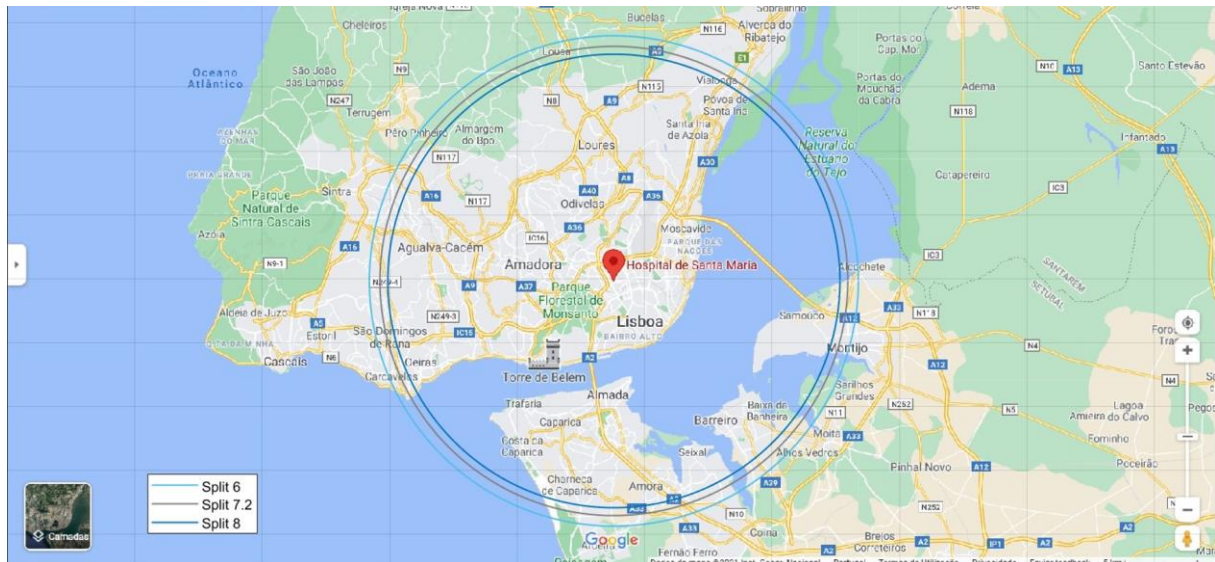


Figure 4.23 – Santa Maria Hospital scenario maximum E2E distances.

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 is 15.0 km 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. These results are obtained due to the low maximum E2E latency limit of 1 ms, and also as consequence of the size of the packet, since the air transmission corresponds to 62.5% of the maximum E2E latency.

Table 4.2 represents the maximum E2E distances for each MEC node deployment and splitting option for Santa Maria Hospital, showing that when the MEC node is placed closer to the Core of the Network, the maximum E2E distance is reduced, since the accumulated latency in the nodes due to the processing, queuing and transmission takes higher values. The results shown in the map of the

Figure 4.23 distance results are present in the second line of Table 4.2.

Table 4.2 – Santa Maria Hospital maximum E2E distances for each MEC node Deployment and splitting option.

	Maximum E2E distance [km]				
MEC Node Deployment Option	Split 8	Split 7.3	Split 7.2	Split 7.1	Split 6
RU-DU	16.2	15.2	15.6	15.7	15.0
DU-CU	10.6	11.5	11.4	10.8	11.9
CU-Core	5.7	8.4	7.9	6.6	9.1

Figure 4.24 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 Espírito Santo Évora Hospital scenario. For this scenario only the splitting option 7.2 is represented, which is the one that will be implemented in 5G, because the map circles for the other splitting options are very close to the represented one, since the E2E distance values are close to each other independently of the chosen option.

The maximum E2E distance shown in the map corresponds to 125.4 km, which exists due to the fact that the maximum E2E latency for the External Remote Surgery is 3 ms, and the packet size defined for the service is the same as the Santa Maria Hospital scenario, which keeps the transmission and processing latencies low in comparison with the maximum allowed E2E latency.

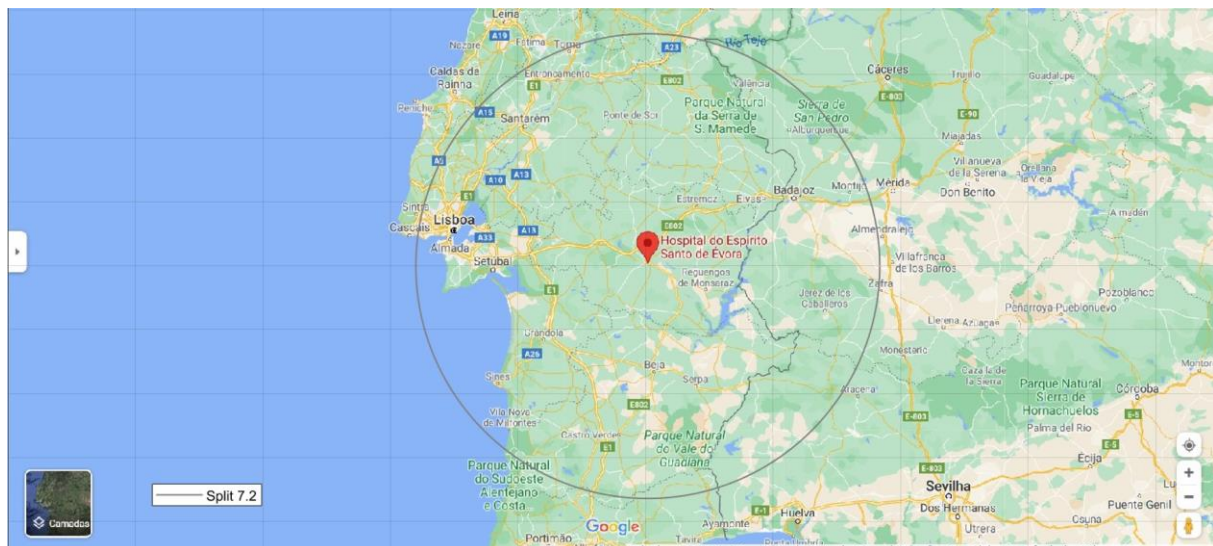


Figure 4.24 – Espírito Santo Évora Hospital scenario maximum E2E distance.

Table 4.3 represents the maximum E2E distances for each MEC node deployment and splitting option for the Espírito Santo Évora Hospital scenario. As it is possible to understand from the analysis of the table, the E2E distances increase around 10 km comparing the CU-Core with the RU-DU MEC Node

Deployment Option, which happens due to the high priority of the service, that keeps queuing delays at almost 0 ms and due to the small packet size, which results in a reduced packet processing time.

Table 4.3 – Espírito Santo Évora Hospital scenario maximum E2E distances for each MEC node Deployment and splitting option.

MEC Node Deployment Option	Maximum E2E distance [km]				
	Split 8	Split 7.3	Split 7.2	Split 7.1	Split 6
RU-DU	135.2	133.6	134.2	134.6	132.8
DU-CU	127.9	128.0	127.9	127.8	128.2
CU-Core	125.2	125.5	125.4	125.2	125.8

Figure 4.25 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 Urban ITS scenario with 60% of the maximum traffic. The circles present in the map correspond to the splitting options 6, 7.2 and 8, and the Split 6 and Split 8 correspond to the circles with the longest and shortest radius length, respectively. The maximum E2E latency allowed for this service is 3 ms, which in the 90% of the maximum traffic scenario results in shorter maximum E2E distances in comparison with the results shown in Figure 4.25.

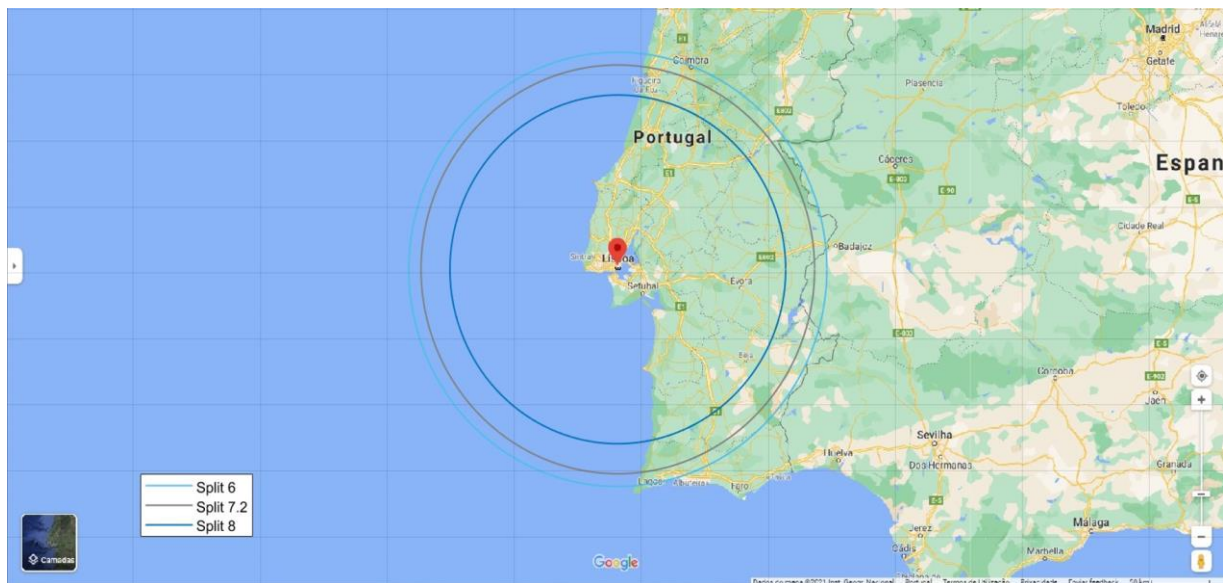


Figure 4.25 – Urban ITS scenario (with 60% of the maximum traffic) maximum E2E distances.

Table 4.4 represents the maximum E2E distances for each MEC node deployment and splitting option for the Urban ITS scenario (with 60% of the maximum traffic). The E2E distance gain, when comparing the different MEC node deployment options, oscillates between 50 km (in Splitting Option 8) and 16 km (in Splitting Option 6) because Split 8 is related to higher numbers of aggregated nodes, which generates more queuing delay, in comparison with the other splitting options.

Table 4.4 – Urban ITS scenario (with 60% of the maximum traffic) maximum E2E distances for each MEC node Deployment and splitting option.

MEC Node Deployment Option	Maximum E2E distance [km]				
	Split 8	Split 7.3	Split 7.2	Split 7.1	Split 6
RU-DU	195.3	192.0	192.7	192.7	193.1
DU-CU	164.7	181.4	179.1	172.4	185.4
CU-Core	142.1	170.9	166.6	155.8	177.0

Figure 4.26 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 between 62.0 km and 88.3 km 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.

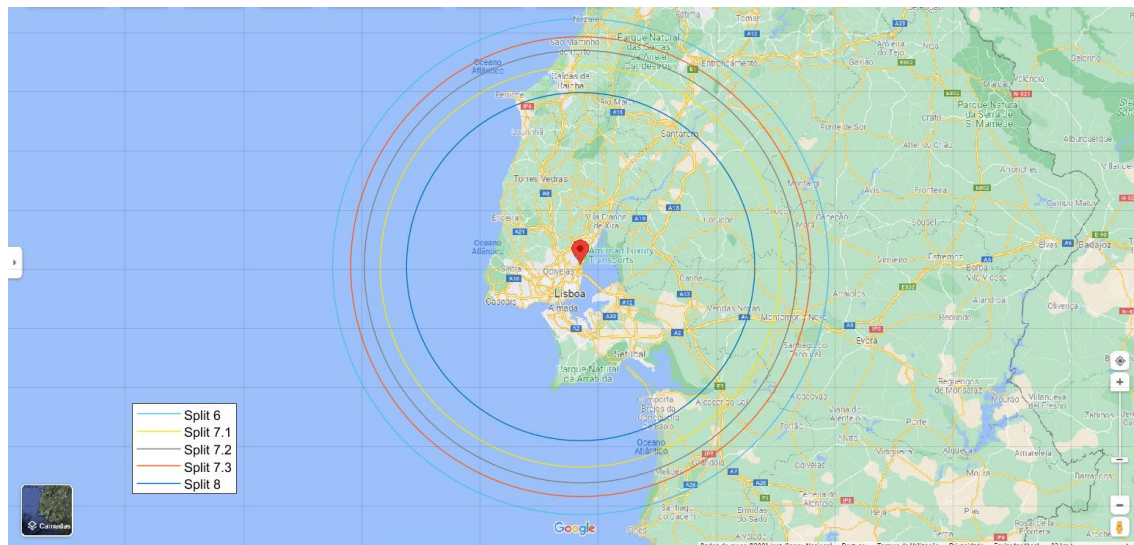


Figure 4.26 – A1 Highway scenario (with 60% of the maximum traffic) maximum E2E distances.

Table 4.5 represents the maximum E2E distances for each MEC node deployment and splitting option for the A1 Highway scenario (with 60% of the maximum traffic). The different MEC Node deployment options have significant maximum E2E distance differences due to the increase in the queuing and processing time of the packets in the nodes. The differences in maximum E2E distances when comparing the different splitting option exist due to the aggregation factor and the difference in the number of functionalities processed in the nodes. The CU-Core deployment has an average maximum E2E distance of approximately 60 km, which in a scenario with 90% of the maximum traffic can result in a E2E latency that exceeds the maximum allowed value. After this observation, and the analysis of the

latency results, the DU-CU deployment seems to be the best option in terms of achieving a balance between the latency reduction and the MEC node complexity reduction.

Table 4.5 – A1 Highway scenario (with 60% of the maximum traffic) maximum E2E distances for each MEC node Deployment and splitting option.

MEC Node Deployment Option	Maximum E2E distance [km]				
	Split 8	Split 7.3	Split 7.2	Split 7.1	Split 6
RU-DU	103.9	101.2	101.8	101.7	102.4
DU-CU	62.0	81.9	77.0	71.3	88.3
CU-Core	36.2	68.7	60.7	51.9	78.3

Figure 4.27 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.

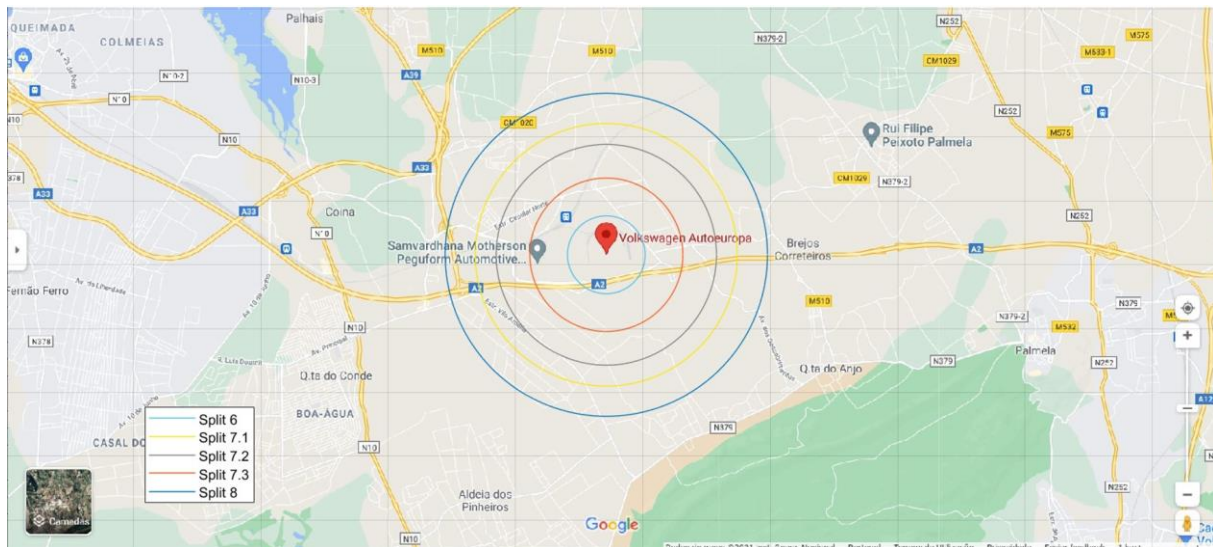


Figure 4.27 – AutoEuropa Factory scenario maximum E2E distances.

Table 4.6 represents the maximum E2E distances for each MEC node deployment and splitting option for the AutoEuropa Factory scenario. The only MEC node Deployment option that can fulfil the latency requirements of the factory automation services is the RU-DU deployment. This fact occurs mainly because of the packet transmission, which corresponds to 64% of the maximum E2E latency, fact that can be observed in Figure 4.22.

Table 4.6 – AutoEuropa Factory scenario maximum E2E distances for each MEC node Deployment and splitting option.

MEC Node Deployment Option	Maximum E2E distance [km]				
	Split 8	Split 7.3	Split 7.2	Split 7.1	Split 6
RU-DU	2.8	1.3	1.9	2.3	0.7
DU-CU	Not Possible	Not Possible	Not Possible	Not Possible	Not Possible
CU-Core	Not Possible	Not Possible	Not Possible	Not Possible	Not Possible

The obtained results show that 5G is able to provide enough capacity to guarantee URLLC services. On the other hand, 4G results show that the system does not provide enough capacity to guarantee URLLC services data rates, although it can be used to offload 5G in certain conditions.

To sum up the E2E distance and latency results, the best MEC node deployment option for the Santa Maria Hospital scenario is between the RU and DU nodes, inside the hospital facilities, to reduce the probability that the latency exceeds the maximum allowed values. The same results are shared by the AutoEuropa factory, because the maximum E2E latency takes very low values, and therefore the MEC node should be installed between the RU and DU inside the factory facilities.

To guarantee the Network Based Sensor Sharing service in the A1 Highway scenario, one concludes that the MEC node should be installed between the DU and the CU nodes, as close as possible to the DU.

The Remote Driving in Avenida da Liberdade and External Remote Surgery in Espírito Santo Évora Hospital services should be provided with latency values below the maximum allowed E2E latency if the MEC nodes are deployed between the CU and the Core. Simulations show that this deployment is enough to keep latency with values considerably below the margin in both scenarios.

Chapter 5

Conclusions

This chapter finalises the thesis, compiling the main conclusions of the study.

The main objective of this thesis was to analyse the different strategies to minimise the E2E latency in 4G and 5G networks. The strategies studied to achieve the Thesis objectives were MEC node deployment options, different splitting options, different network architectures, and different available radio techniques. In addition to the delays that packets from the different services suffer along the network, it was also performed a study the throughputs that both 4G and 5G can offer to users, because these capacities can also have an influence on E2E latency.

The first chapter contains a brief introduction to the work, by establishing a comparison between 4G and 5G in terms of capacity, availability and latency. This chapter also describes the importance of Edge Computing technology and contains explanations about how this technology will be crucial to achieve lower latencies than in the previous generations of mobile communication systems. After the previously described overview, the motivation and the contents of the thesis are presented.

The second chapter of the thesis describes the fundamental aspects that are important to understand the context in which the model was developed. It starts by differentiating the Non-standalone and Standalone versions of 5G, because the first phase of 5G deployment focuses on eMBB, which can be achieved using the 5G non-Standalone version, but URLLC services can only be provided to users with the Standalone version. The radio interface is studied by giving a brief explanation of the TDD and FDD modes and their impacts on latency, and afterwards, presenting the impacts of 5G numerologies and frequency bands. Network Slicing and Virtualisation is a process that simultaneously reduces the physical network resources and guarantee that the available network resources are used in a more efficient way, which ultimately is an important tool for the evolution of the mobile communication systems. Cloud Networks are described in order to understand the centralisation process and how the 4G network is installed with the RRH and the BBU located independently, and also how most of the network functionalities are processed by the BBU, in a centralisation process. The Edge Network is a useful technology in terms of latency reduction, since it moves network resources closer to the end user, preventing further accumulated delays, and the several MEC node deployments along the 4G network (which are used to create the mathematical model) are present in this chapter. After the Edge Networking description, there is a characterisation of the NR system directions and services. The fifth section describes latency in the network, containing a description of queuing, transmission, processing and propagation latencies, and the comparison between 4G and 5G Transport Network architectures. The state of the art is present in the end of the chapter, summarising the most relevant information related to the thesis.

The third chapter of the thesis provides a detailed description of the developed model, namely its mathematical formulation and its implementation. The assessment of the model is provided in the end of the chapter. In the beginning, the input and output parameters of the program are identified, as well as intermediate calculations and steps that the program performs.

The 5G network can be deployed following different architectures in which RU, DU and CU can be collocated or independently deployed, and the MEC node can be deployed in different positions along the network. After analysing the different network topologies, the latency is analysed from an individual viewpoint, meaning that the text is developed to explain how each latency contribution is calculated in

the model, and how the different contributions are aggregated to calculate E2E latency. After this process, the document contains an explanation of the process that the program performs to calculate the total node for each MEC node deployment option, by summing the latency contributions that are accumulated along the network nodes and links in both 4G and 5G.

The model development also allows the radio and link throughput study, meaning that the different 4G and 5G radio techniques and splitting options are described in this chapter and used in the simulations to acquire realistic results. The studied 4G system frequency bands (2.6 GHz, 1.8 GHz, and 0.8 GHz) use the FDD mode, but the studied 5G system band (3.5 GHz) uses TDD mode, which adds some diversity to the expressions and the performed calculations.

A detailed explanation of the model implementation is given, using flowcharts to better illustrate the program calculations and procedures, and the model assessment is performed, checking if each phase of the program is functioning in the expected way. The model starts by reading as input parameters, the network, the service and the user specifications and it stores the important variables to perform further calculations. All radio throughputs and required throughputs, as well as intermediate calculations required to obtain the final throughput and latency results are performed in the beginning of the program.

The analysed services, present in the service specifications, belong to the URRLC 5G direction in which E2E latency requirements are stricter than in other provided services. These are the services that can only be implemented by analysing different strategies that minimise the latency between the two extremities of the network.

After calculating some initial latency contributions that are common to all network and MEC node deployment options and architectures, the program proceeds to calculate all latency contributions (depending on the MEC node deployment option and the simulated system) and after this phase it sums all individual latency contributions (except the propagation latency), resulting in the total node latency. The E2E distance is calculated using the Total Node Latency and the maximum E2E latency for a certain service. This length corresponds to the maximum distance in which the MEC node can be installed (using the transmitter site as a reference) and it is represented in a geographical map as a program output, as well as the other variables calculated in the algorithm, such as the Total Node latency, link throughputs, used throughputs, amongst others.

The fourth chapter presents the analysis of the results obtained as program outputs. Firstly, it defines the scenarios under analysis, considering the RRHs location, traffic and capacity provided by NOS. The considered scenarios are the implementation of Internal Remote Surgery inside the Santa Maria Hospital in Lisbon, the implementation of the External Remote Surgery service for the Espírito Santo Hospital in Évora, the implementation of Remote Driving in Avenida da Liberdade 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.

All the available splitting options and MEC node deployment options were tested for each scenario in both 4G and 5G, to have a better understanding and analysis of the impact of each strategy in network latency reduction, and to understand which techniques can be used to guarantee that each of the

URLLC services can be implemented. The differences among each scenario are the traffic generated in the RU/RRH and the aggregation factor, because a higher traffic density is expected in the areas with higher mobile data usage. It is also important to mention that there are scenarios in which some of the radio nodes are installed inside the facilities, like in the hospital and factory automation scenarios. In the Urban ITS and A1 Highway scenario, the radio nodes are installed in an outdoor environment, which corresponds to a lower average CQI value, and consequently a lower 5G and 4G capacity, compared with indoor radio nodes.

For indoor radio nodes, the average 5G capacity obtained in simulations is 723.39 Mbps in DL and 545.72 Mbps in UL, while in outdoor ones the average capacity is 508.15 Mbps and 383.34 Mbps, respectively. The considered frame structure for 5G divides 57% of the frame for DL and the rest for UL, which is important to guarantee enough capacity for the services. It is important to refer that this is not the usual frame structure, but according to NOS, NR will have a flexible frame structure regime, which generates the possibility of having this frame split.

For the indoor capacity, the program calculates an average DL throughput of 60.93 Mbps and an UL one of 15.72 Mbps. The outdoor capacity provided by LTE depends on the frequency bands with which the site works; for example, the Avenida da Liberdade site works with the 2600 MHz, 1800 MHz and 800 MHz frequency bands, while the A1 Highway site works with the 1800 MHz and 800 MHz bands.

On the other hand, one considers that the other outdoor sites work only with the 2600 MHz band, which represents a DL capacity of 39.44 Mbps and a UL one of 10.32 Mbps. The calculated LTE capacity in the Urban ITS scenario is 98.6 Mbps and 25.8 Mbps, for DL and UL respectively, while in the A1 Highway scenario it is 59.16 Mbps and 15.48 Mbps, for DL and UL respectively.

The obtained results show that 5G is able to provide enough capacity to guarantee URLLC services. On the other hand, 4G results show that the system does not provide enough capacity to guarantee URLLC services data rates, although it can be used to offload 5G.

For the hospital and factory scenarios one considered a fixed network traffic, unlike the Urban ITS and A1 Highway scenarios, in which three different usage scenarios (having also into account the increase in the URLLC services usage along the years) were considered: 30%, 60 and 90% of the usage traffic. For the Santa Maria Hospital scenario, the required throughputs in DL and UL are 295.18 Mbps and 197.79 Mbps, respectively, which are below the radio node capacity.

For the Espírito Santo Hospital scenario, the required throughputs in DL and UL are 205.5 Mbps and 89.12 Mbps, respectively, for the site inside the hospital, and for the site outside the hospital DL and UL required throughputs are 179.58 Mbps and 117.50 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 remote surgery service in different scenarios.

For the Urban ITS scenario with 60% of the maximum traffic, the required throughputs in DL and UL are 418.69 Mbps and 451.8 Mbps, respectively, for both the remote driver and passenger sites since the environments in which the RU/RRHs are installed are similar.

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 Remote Driving and the Network Based Sensor Sharing services in the scenarios, the considered required data rates for the services were in the highest range of the possible values, meaning that on average the required throughputs will be lower, and therefore 5G sites are expected to be able to provide these services. For the AutoEuropa Factory scenario, the required throughputs in DL and UL are 256.92 Mbps and 197.28 Mbps, respectively.

After analysing the previous results, one can conclude that NR can provide capacity for the factory automation services in different scenarios. The analysis of the radio capacities and requirements for both systems is followed by latency studies. An E2E latency study is presented with a consequent E2E distance analysis and a latency contribution study of the queuing, processing, transmission, and maximum propagation latencies. It is important to refer that these studies are only performed for the best considered MEC node deployment option, after the total node latency analysis.

The obtained results mixed with the information available in papers leads to values in the 8 ms to 20 ms range for the RRH-BBU MEC node deployment and values in the 10 ms 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 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 between the calculated Total Node latency and the maximum E2E latency that generates a maximum E2E distance of approximately 15 km, it is advised that the MEC node is installed inside the hospital, to prevent the latency increase under certain circumstances. With this MEC node deployment, the impact of the splitting options choice is reduced, because the traffic that generates the queuing delay is only present in the RU/RRH node.

Since the service has a high priority in the network the queuing delays are almost inexistant, and since the considered packets are very small, the obtained processing delays are close to 0.1 ms. The transmission delays are the ones that add more latency to the packets.

For the Espírito Santo Hospital scenario it is possible to conclude that the best MEC node deployment option is between the CU and the Network Core, which still presents a considerable latency margin because the calculated Total Node latency of about 0.9 ms is lower than the maximum E2E latency of 3 ms, which results in a high E2E distance length of approximately 125 km for all splitting options. Considering that the service has a high priority in the network, queuing delays are almost inexistant, and since the considered packets are very small, the obtained processing delays are close to 0.25 ms. The transmission delays are the ones that add more latency to packets.

For the Urban ITS scenario, one can conclude that the best MEC node deployment option is also between the CU and the Network Core, because even for intense network usage scenarios (90% of the maximum traffic) the Total Node latency is considerably below the maximum allowed Remote Driving E2E latency of 5 ms.

The processing delays are approximately 0.67 ms and the queuing delays are in the 0.34 ms to 0.93 ms

range. Since the maximum allowed E2E latency is 5 ms, and the Total node latencies for the most favourable MEC node deployment option are between 2.04 ms and 2.63 ms the maximum E2E distances are between 142.1 km and 177.0 km, meaning that the MEC nodes can be installed in other city and still provide coverage to Lisbon.

For the A1 Highway scenario, one can conclude that the best MEC node deployment option is 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 between the CU and the Core. After analysing the transmission, processing, and queuing latencies, it is possible to conclude that the transmission latency is responsible for most of the packet delay, followed by the queuing latency with values between 0.80 ms and 0.35 ms, and the processing latency with delay values in the order of 0.36 ms.

For the AutoEuropa factory scenario, one can conclude that the best and only possible MEC node deployment option is between the RU and the DU, since the other options exceed the maximum allowed E2E latency. After analysing the transmission, processing and queuing latencies, it is possible to conclude that the transmission latency is responsible for most of the packet delay with fixed values of 0.16 ms, followed by the processing latency with values between 0.05 ms and 0.07 ms, and the queuing latency with almost inexistent delays since the only node in which the packet can be queued is the RU, and the factory automation service has also a high priority service. The E2E distance calculated for the best considered MEC node deployment are low, 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 traffic profiles from the future URLLC services, which is unpredictable based on the currently available information. The used model is a packet-based one, in which the E2E latency studied for 5G is calculated based on a reference packet instead of time frames. The processing delays were calculated based on the number of functionalities in the nodes and the packet size, which is an approximation that can 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 possible network deployments.

Annex A

User's Manual

This Annex presents the detailed instructions on how to run a simulation and configure the parameters.

A.1 Run the Simulator

The simulator was developed in a Matlab environment, it was used an 2019a version with the most common Matlab toolboxes.

To run the model simulator, first it is necessary to configure the Network_specification.xlsx, User_specification.xlsx and Service_specification.xlsx files which contain definition of the input parameters to run the simulator.

After configuring the specification parameters, one should be able to run the simulator by running the script “main.m” in the Latency_Program folder.

When the simulator finish, the output variables will be in the Output_File in the Latency_Program folder.

A.2 Simulator Configuration

The input parameters allow the program user to change the characteristics of the network and the nodes for the simulations. In this annex it is explained how these parameters can be configured to simulate specific scenarios. Table A.1 explains the configuration of the network specification parameters of the excel files.

Table A.1 – Network Specifications input parameters definition.

NA (Network Architecture)	0	4G architecture with the independent RRH and BBU.
	1	5G architecture with independent RU, DU and CU.
	2	5G architecture with collocated RU and DU.
	3	5G architecture with collocated DU and CU.
	4	5G architecture with collocated RU, DU and CU.
MEC (MEC node Deployment Option)	0	No MEC for both 4G and 5G systems.
	1	MEC node between the BBU and the Core or between the CU and the Core for the 4G and the 5G systems, respectively.
	2	MEC node between the RRH and the BBU or between the DU and the CU for the 4G and the 5G systems, respectively.
	3	MEC node between the RU and the DU for the 5G system.
NL (Network Link Type)	0	Optical Fibre links.
	1	High-Capacity Microwave links.

FH1 /FH2 (Splitting options in the transmitter and receiver sides)	8	Splitting Option 8.
	7.3	Splitting Option 7.3.
	7.2	Splitting Option 7.2.
	7.1	Splitting Option 7.1.
	6	Splitting Option 6.
MH1/MH2	2	Splitting Option 2 in the MH.
BH1/BH2	Backhaul link capacity (BH1 is the capacity on the transmitter side and BH2 is the capacity on the receiver side). The used capacity in the simulations is 25 Gbps.	
TL1/TL2	Transport link capacity (TL1 is the capacity on the transmitter side and TL2 is the capacity on the receiver side). The used capacity in the simulations is 100 Gbps.	
MM1/MM2	Number of MIMO layers in the transmitter and receiver sides.	
NP1/NP2	Numerology used in the transmitter and receiver sides. The used value in the simulations is 0 for the 4G system and 1 for the 5G system.	
BW1/BW2	System bandwidths in the transmitter and receiver sides.	
CQ1/CQ2	Average channel quality indicators in the transmitter and receiver sides (between 1 and 15).	
SF1/SF2	Scaling factors for the 5G system in the transmitter and receiver sides. The used value in the simulations is q.	
FS/DUR1 or FS/DUR2	Frame structure (for the 5G TDD band) or Downlink usage ratio (for the 4G FDD band).	
AF1/UUR1 or AF2/UUR2	Average Factor (for the 5G TDD band) or Uplink usage ratio (for the 4G FDD band).	
SV List	List of the simulated services with the packet size, priority, latency adaptation parameter and link type.	

Table A.2 explains the configuration of the user specification parameters of the excel files.

Table A.2 – User specification input parameters definition.

DT1/DT2	Distances between the UEs and the radio node in the transmitter and receiver sides.
US	Service to be simulated (between 1 and 17).
NRUU_Tx/NRUU_Rx	Number of users actively connected to the RU/RRH in the

	transmitter and receiver side.
List of RU_Tx/RU_Rx	List of the percentages of the users actively connected to the RU/RRH in the transmitter and receiver sides.
NDUU_Tx/NDUU_Rx	Number of users actively connected to the DU in the transmitter and receiver sides.
List of DU_Tx/DU_Rx	List of the percentages of the users actively connected to the DU in the transmitter and receiver sides.
NCUU_Tx/NCUU_Rx	Number of users actively connected to the CU/BBU in the transmitter side in the transmitter and receiver sides.
List of CU_Tx/CU_Rx	List of the percentages of the users actively connected to the CU/BBU in the transmitter and receiver sides.

Table A.3 explains the configuration of the service specification parameters of the excel files.

Table A.3 – Service specification input parameters definition.

List of Services	List of services that contains the required data rates and the maximum allowed E2E latencies for each service in the list.
-------------------------	--

Annex B

4G and 5G services

This Annex presents the requirements that 4G and 5G systems need to fulfil in order to provide the URLLC services. This Annex presents the detailed instructions on how to run a simulation and configure the parameters

Table B.1 – Service requirements.

Service Group	Service	File Size [MB]	Duration [min]	Service Class	Latency [ms]	Data Rate [Mbps]	Packet Size [B]	Priority	References
Rem. Surg.	Int. Cont. Manip.	-	30	Conv. URLLC	1	0.512	20	1	[JMjL02] [IAIH18] [QJGZ18]
	Vid. Str.	-	30		100	10	188		
	Hap. Feed.	-	30		3	0.400	20		
	Ext. Cont. Manip	-	30		3	0.512	20		
ITS	Traffic Info.	0.36	-	Conv. URLLC eMBB	100	2	300	2	[ALT120] [3GPP18a]
	Rem. Driv.	0.36	-		5	25	1600		
	Net. Bas. Sen. Shar.	0.36	-		3	20	1000		
Fact. Auto.	Mach. Tools	0.18	-	Conv. URLLC	0.25	1	10	2	[PSMM17] [IAIH18]
	Print. Mach.	0.18	-		1	1	30		
	Pack. Mach.	0.18	-		2.5	1	15		
-	Aug. Real.	-	30	Stream. URLLC	15	600	650	4	[Qual18]
-	Voice	-	2.0	Conv.	100	0.032	218	3	[SeDo19]
-	Video Conf.	-	30	Conv.	150	2	800	5	[SeDo19]
-	Web Brow.	3.0	-	Interac.	300	0.5	512	9	[SeDo19]
-	Email	0.5	-	Back.	300	0.512	128	12	[SeDo19]
-	Soc. Net.	30	-	Interac.	300	2	1000	8	[SeDo19]
-	File Transfer	5.0	-	Interac.	300	1	4096	10	[SeDo19]

Annex C

Latency Contributions Description

This Annex presents a description of the latency contributions accumulated along the network.

Table C.1 – Latency contributions description.

Latency Parameter	Description
$\delta_{UE_Tx}, \delta_{UE_Rx}$	The δ_{UE_Tx} represents the accumulated latency in the UE in the transmission, which accounts for the processing and transmission delays to the Air link. The δ_{UE_Rx} represents the accumulated latency in the UE in the reception, which accounts for the processing in the terminal.
$\delta_{AL_Tx}, \delta_{AL_Rx}$	The δ_{AL_Tx} accounts for the air link propagation latency in the transmitter side, and the δ_{AL_Rx} accounts for the air link propagation latency in the receiver side.
$\delta_{RU_Tx}, \delta_{RU_Rx}$ or $\delta_{RRH_Tx}, \delta_{RRH_Rx}$	The $\delta_{RU_Tx}/\delta_{RRH_Tx}$ represents the accumulated latency in the RU/RRH in the transmitter side, which accounts for the processing, queuing and transmission to the FH link delays. The $\delta_{RU_Rx}/\delta_{RRH_Rx}$ represents the accumulated latency in the RU/RRH in the receiver side, which accounts for the processing, queuing, and transmission to the Air Link delays.
$\delta_{FH_Tx}, \delta_{FH_Rx}$	The δ_{FH_Tx} accounts for the propagation latency in the FH at the transmitter side, and the δ_{FH_Rx} accounts for the propagation latency in the FH at the receiver side. Both the δ_{FH_Tx} and δ_{FH_Rx} can be from the RU to the MEC node, or from the RU to the DU and vice versa, depending on the considered side.
$\delta_{DU_Tx}, \delta_{DU_Rx}$	The δ_{DU_Tx} represents the accumulated latency in the DU in the transmitter side, which accounts for the processing, queuing, and transmission to the MH link delays. The δ_{DU_Rx} represents the accumulated latency in the DU in the receiver side, which accounts for the processing, queuing, and transmission to the FH Link delays.
$\delta_{MH_Tx}, \delta_{MH_Rx}$	The δ_{MH_Tx} accounts for the propagation latency in the MH at the transmitter side, and the δ_{MH_Rx} accounts for the propagation latency in the MH at the receiver side. Both the δ_{MH_Tx} and δ_{MH_Rx} can be from the DU to the MEC node, or from the DU to the CU and vice versa, depending on the considered side.
$\delta_{CU_Tx}, \delta_{CU_Rx}$ or $\delta_{BBU_Tx}, \delta_{BBU_Rx}$	The $\delta_{CU_Tx}/\delta_{BBU_Tx}$ represents the accumulated latency in the CU/BBU in the transmitter side, which accounts for the processing, queuing, and transmission to the BH link delays. The $\delta_{CU_Rx}/\delta_{BBU_Rx}$ represents the accumulated latency in the CU/BBU in the receiver side, which accounts for the processing, queuing, and transmission to the MH Link delays in the CU case, and to the FH Link in the BBU case.

$\delta_{BH_Tx}, \delta_{BH_Rx}$	The δ_{BH_Tx} accounts for the propagation latency in the BH at the transmitter side, and the δ_{BH_Rx} accounts for the propagation latency in the BH at the receiver side. Both the δ_{BH_Tx} and δ_{BH_Rx} can be from the CU to the MEC node, or from the CU to the Core and vice versa, depending on the considered side.
$\delta_{Core_Tx}, \delta_{Core_Rx}$	The δ_{Core_Tx} represents the accumulated latency in the Core in the transmitter side, which accounts for the processing and transmission to the Transport link delays. The δ_{Core_Rx} represents the accumulated latency in the Core in the receiver side, which accounts for the processing and transmission to the BH Link delays.
$\delta_{TL_Tx}, \delta_{TL_Rx}$	The δ_{TL_Tx} accounts for the propagation latency in the Transport Link at the transmitter side, and the δ_{TL_Rx} accounts for the propagation latency in the Transport Link at the receiver side.
δ_{EDC}	The δ_{EDC} accounts for the processing delay in the EDC and the transmission delay to the Transport Link on the receiver side.
δ_{MEC}	The δ_{MEC} accounts for the processing in the MEC node and the transmission delay to the Link on the receiver side, depending on the MEC node positioning.

Annex D

Data Centre and MEC node processing latency

This Annex presents a graphic with the considered data centre and MEC node processing delays for a single functionality.

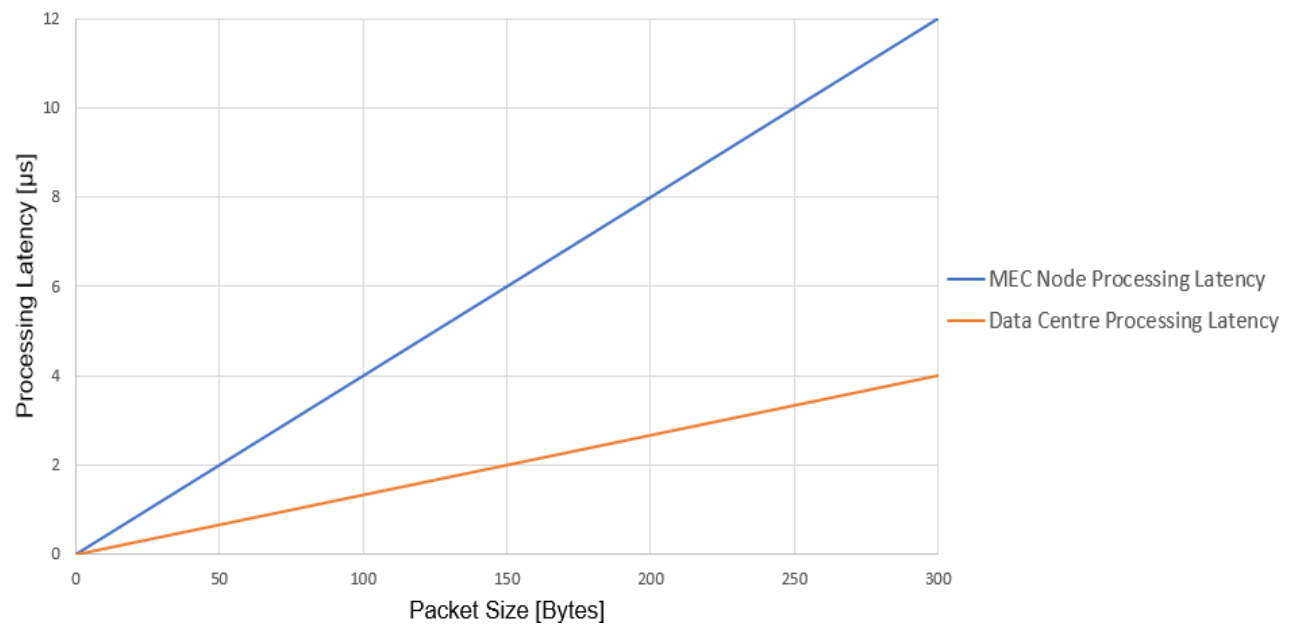


Figure D.1 – Data Centre and MEC node processing latency for a single functionality.

Annex E

Latency Adaptation Parameter

This Annex presents a table with the considered latency adaptation parameter for each service.

Table E.1 – Latency adaptation parameter.

Service	Service Number	ρ_{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

Annex F

4G and 5G Link Throughputs

This Annex presents the 4G and 5G link throughputs for the reference scenarios depending on the splitting option.

Table F.1 – 4G and 5G link throughputs depending on the splitting options.

Splitting Option		Link Throughput	Reference Values
Option 8 – 4G (30 kHz of Subcarrier spacing as reference)		DL: 157.3 Gbps	S_r : 30.72 N_Q : 32 N_A : 32
		UL: 157.3 Gbps	S_r : 30.72 N_Q : 32 N_A : 32
Option 8 – 5G (30 kHz of Subcarrier spacing as reference)		DL: 157.3 Gbps	S_r : 30.72 N_Q : 32 N_A : 32
		UL: 157.3 Gbps	S_r : 30.72 N_Q : 32 N_A : 32
Option 7 (30 kHz of Subcarrier spacing as reference)	7.3	DL: 5.9 Gbps	N_{SC} : 3276 (N_{RB} * 12 subcarriers) N_{SY} : 14 N_Q : 8 N_L : 7 MAC_{info} : 713.9
		UL: 7.5 Gbps	N_{SC} : 3276 N_{SY} : 14 N_Q : 8 N_L : 10 MAC_{info} : 120
	7.2	DL: 5.3 Gbps	N_{SC} : 3276 (N_{RB} * 12 subcarriers) N_{SY} : 14 N_Q : 8 N_L : 7 MAC_{info} : 121
		UL: 29.4 Gbps	N_{SC} : 3276 (N_{RB} * 12 subcarriers) N_{SY} : 14 N_Q : 32 N_A : 10 MAC_{info} : 80

	7.1	DL: Same as option 7.3	Same as option 7.3
		UL: Same as option 7.2	Same as option 7.2
Option 6 (250 Mbps of throughput as reference)		DL: 6.8 Gbps	R_p : 250 R_c : 5 B: 100 B_c : 20 N_L : 8 $N_{L,C}$: 2 M: 256 M_c : 64
		UL: 8.4 Gbps	R_p : 96 R_c : 44 B: 100 B_c : 20 N_L : 8 $N_{L,C}$: 1 M: 64 M_c : 16
Option 2 (250 Mbps of throughput as reference)		DL: 6.7 Gbps	R_p : 250 B: 100 B_c : 20 N_L : 8 $N_{L,C}$: 2 M: 256 M_c : 64
		UL: 5.0 Gbps	R_p : 83 B: 100 B_c : 20 N_L : 8 $N_{L,C}$: 1 M: 64 M_c : 16

Annex G

Target Code Rate R for the 4G system

This Annex presents the target code rate for the 4G system.

Table G.1 – R parameter for 4G radio throughput (extracted from [3GPP17]).

CQI Index	Modulation Order	R parameter	Spectral Efficiency
0	Out of Range		
1	2	0.076	0.152
2	2	0.117	0.234
3	2	0.188	0.377
4	2	0.301	0.602
5	2	0.438	0.877
6	2	0.588	1.176
7	4	0.369	1.477
8	4	0.479	1.914
9	4	0.602	2.406
10	6	0.455	2.731
11	6	0.554	3.322
12	6	0.650	3.902
13	6	0.754	4.523
14	6	0.853	5.115
15	6	0.926	5.555

Annex H

Target Code Rate R for the 5G system

This Annex presents the target code rate for the 5G system.

Table H.1 – R parameter for 5G radio throughput (extracted from [ETSI18d]).

CQI Index	Modulation Order	R parameter	Spectral Efficiency
0	Out of Range		
1	2	0.076	0.152
2	2	0.188	0.377
3	2	0.438	0.877
4	4	0.369	1.477
5	4	0.479	1.914
6	4	0.602	2.406
7	6	0.455	2.731
8	6	0.554	3.322
9	6	0.650	3.902
10	6	0.754	4.523
11	6	0.853	5.115
12	8	0.694	5.555
13	8	0.778	6.227
14	8	0.864	6.914
15	8	0.926	7.406

Annex I

5G TDD slot formats

This Annex presents the 5G system TDD slot formats including the configurations with the Flexible frame F.

Table I.1 – 5G TDD slot formats.

Format	Symbol Number in a Slot														Slot in the UL	Slot in the DL
	0	1	2	3	4	5	6	7	8	9	10	11	12	13		
0	D	D	D	D	D	D	D	D	D	D	D	D	D	D	0	1
1	U	U	U	U	U	U	U	U	U	U	U	U	U	U	1	0
2	F	F	F	F	F	F	F	F	F	F	F	F	F	F	1	1
3	D	D	D	D	D	D	D	D	D	D	D	D	D	F	0.07	1
4	D	D	D	D	D	D	D	D	D	D	D	D	F	F	0.14	1
5	D	D	D	D	D	D	D	D	D	D	D	F	F	F	0.21	1
6	D	D	D	D	D	D	D	D	D	D	F	F	F	F	0.29	1
7	D	D	D	D	D	D	D	D	D	F	F	F	F	F	0.36	1
8	F	F	F	F	F	F	F	F	F	F	F	F	F	U	1	0.93
9	F	F	F	F	F	F	F	F	F	F	F	F	U	U	1	0.86
10	F	U	U	U	U	U	U	U	U	U	U	U	U	U	1	0.07
11	F	F	U	U	U	U	U	U	U	U	U	U	U	U	1	0.14
12	F	F	F	U	U	U	U	U	U	U	U	U	U	U	1	0.21
13	F	F	F	F	U	U	U	U	U	U	U	U	U	U	1	0.29
14	F	F	F	F	F	U	U	U	U	U	U	U	U	U	1	0.36
15	F	F	F	F	F	F	U	U	U	U	U	U	U	U	1	0.43
16	D	F	F	F	F	F	F	F	F	F	F	F	F	F	0.93	1
17	D	D	F	F	F	F	F	F	F	F	F	F	F	F	0.86	1
18	D	D	D	F	F	F	F	F	F	F	F	F	F	F	0.79	1
19	D	F	F	F	F	F	F	F	F	F	F	F	F	U	0.93	0.93
20	D	D	F	F	F	F	F	F	F	F	F	F	F	U	0.86	0.93
21	D	D	D	F	F	F	F	F	F	F	F	F	F	U	0.79	0.93
22	D	F	F	F	F	F	F	F	F	F	F	F	U	U	0.93	0.86
23	D	D	F	F	F	F	F	F	F	F	F	F	U	U	0.86	0.86
24	D	D	D	F	F	F	F	F	F	F	F	F	U	U	0.79	0.86
25	D	F	F	F	F	F	F	F	F	F	F	U	U	U	0.93	0.79
26	D	D	F	F	F	F	F	F	F	F	F	U	U	U	0.86	0.79
27	D	D	D	F	F	F	F	F	F	F	F	U	U	U	0.79	0.79
28	D	D	D	D	D	D	D	D	D	D	D	D	F	U	0.14	0.93
29	D	D	D	D	D	D	D	D	D	D	D	F	F	U	0.21	0.93
30	D	D	D	D	D	D	D	D	D	D	D	F	F	F	0.29	0.93
31	D	D	D	D	D	D	D	D	D	D	D	F	U	U	0.21	0.86
32	D	D	D	D	D	D	D	D	D	D	D	F	F	U	0.29	0.86
33	D	D	D	D	D	D	D	D	D	D	F	F	F	U	0.36	0.86
34	D	F	U	U	U	U	U	U	U	U	U	U	U	U	0.93	0.14
35	D	D	F	U	U	U	U	U	U	U	U	U	U	U	0.86	0.21
36	D	D	D	F	U	U	U	U	U	U	U	U	U	U	0.79	0.29
37	D	F	F	U	U	U	U	U	U	U	U	U	U	U	0.93	0.21
38	D	D	F	F	U	U	U	U	U	U	U	U	U	U	0.86	0.28
39	D	D	D	F	F	U	U	U	U	U	U	U	U	U	0.79	0.36
40	D	F	F	F	U	U	U	U	U	U	U	U	U	U	0.93	0.29
41	D	D	F	F	F	U	U	U	U	U	U	U	U	U	0.86	0.36
42	D	D	D	F	F	F	U	U	U	U	U	U	U	U	0.79	0.43
43	D	D	D	D	D	D	D	D	D	F	F	F	F	U	0.36	0.93
44	D	D	D	D	D	D	F	F	F	F	F	F	U	U	0.57	0.86
45	D	D	D	D	D	D	F	F	U	U	U	U	U	U	0.57	0.57
46	D	D	D	D	D	F	U	D	D	D	D	D	F	U	0.29	0.86
47	D	D	F	U	U	U	U	D	D	F	U	U	U	U	0.71	0.43
48	D	F	U	U	U	U	U	D	F	U	U	U	U	U	0.86	0.29
49	D	D	D	D	F	F	U	D	D	D	D	F	F	U	0.43	0.86
50	D	D	F	F	U	U	U	D	D	F	F	U	U	U	0.71	0.57
51	D	F	F	U	U	U	U	D	F	F	U	U	U	U	0.86	0.43
52	D	F	F	F	F	F	U	D	F	F	F	F	F	U	0.86	0.86

Annex J

Relevant Data for the Scenarios

This Annex presents the relevant data to generate the input traffic for each studied scenario. This Annex presents the simulated traffic for each studied scenario.

Table J.1 – Link type and UE distances for the simulated scenarios.

Scenario	Link Type	Tx UE distance [m]	Rx UE distance [m]
Sta. Maria Hospital	Optical Fibre	200	200
Espírito Santo de Évora Hospital	Optical Fibre	200	800
A1 Highway ITS	Optical Fibre	700	700
Urban ITS	Optical Fibre	300	300
AutoEuropa Factory	Optical Fibre	200	200

Annex K

Reference Scenario Configuration

This Annex presents the simulated traffic for each studied scenario. This Annex presents the simulated radio techniques for each studied scenario.

Table K.1 – Average number of users per node (both receiver and transmitter sides) on the Santa Maria Hospital scenario.

Splitting Option	RU/RRH	DU	CU	BBU
8	200	1200	6000	7000
7.3	200	700	3500	-
7.2	200	800	4000	-
7.1	200	1000	5000	-
6	200	600	3000	-

Table K.2 – Santa Maria Hospital scenario receiver and transmitter RU/RRH service mix.

Service	I. C.M	V. SM	H. FB	E. C.M	T. IF	R. DV	NB. S. S	M. TL	PT. MC	PK. MC	A. RT	VC	V. CF	W. BW	EM	S. NW	F. TF
RU Service mix [%]	8	10	10	2	0	0	0	0	0	0	0	31	3	15	7	11	3
DU Service mix [%]	6	8	8	2	12	2	5	0	0	0	0	30	2	10	6	7	2
CU/BBU Service mix [%]	5	8	8	3	10	1	4	1	1	1	0	32	2	5	4	14	1

Table K.3 – Average number of users per node on the Espírito Santo de Évora Hospital scenario.

Splitting Option	Rec. RU/RRH	Trans. RU/RRH	Rec. DU	Trans.DU	Rec. CU	Trans. CU	Rec. BBU	Trans. BBU
8	200	200	800	1000	3000	3000	4000	4000
7.3	200	200	550	600	1800	1800	-	-
7.2	200	200	600	650	2000	2000	-	-
7.1	200	200	700	800	2400	2400	-	-
6	200	200	350	400	1200	1200	-	-

Table K.4 – Espírito Santo de Évora Hospital scenario transmitter RU/RRH service mix.

Service	I. C.M	V. SM	H. FB	E. C.M	T. IF	R. DV	NB. S. S	M. TL	PT. MC	PK. MC	A. RT	VC	V. CF	W. BW	EM	S. NW	F. TF
Transmitter RU/RRH Service mix [%]	3	5	5	2	0	0	0	0	0	0	0	38	3	20	7	15	2
Receiver RU/RRH Service mix [%]	0	1	1	1	10	0	1	0	0	0	0	40	2	20	6	16	2
Transmitter DU Service mix [%]	3	4	4	1	14	1	5	0	0	0	2	30	5	15	5	10	1
Receiver DU Service mix [%]	0	1	1	1	10	2	6	0	0	0	0	39	5	13	5	12	5
CU/BBU Service mix [%]	1	2	2	1	12	2	6	2	2	2	1	33	2	17	3	10	2

Table K.5 – Average number of users per node on the Urban ITS scenario for 30 %, 60 % and 90% of network usage.

Splitting Option	RU/RRH	DU	CU	BBU
8	100	1000	3000	3500
	200	2000	6000	7000
	300	3000	9000	10500
7.3	100	400	1200	-
	200	800	2400	-
	300	1200	3600	-
7.2	100	500	1500	-
	200	1000	3000	-
	300	1500	4500	-
7.1	100	700	2100	-
	200	1400	4200	-
	300	2100	6300	-
6	100	300	900	-
	200	600	1800	-
	300	900	2700	-

Table K.6 – Urban ITS scenario RU (both transmitter and receiver sides) service mix.

Service	I. C.M	V. SM	H. FB	E. C.M	T. IF	R. DV	NB. S. S	M. TL	PT. MC	PK. MC	A. RT	VC	V. CF	W. BW	EM	S. NW	F. TF
RU Service mix [%]	0	0	0	0	15	2	7	0	0	0	0	42	1	12	5	14	2
DU Service mix [%]	4	8	8	4	10	1	4	0	0	0	0	34	2	9	4	10	2
CU/BBU Service mix [%]	3	6	6	3	6	1	4	1	1	1	0	44	5	5	3	10	1

Table K.7 – Average number of users per node on the A1 Highway for 30 %, 60 % and 90 % of the network usage.

Splitting Option	RU/RRH	DU	CU	BBU
8	100	800	2400	3000
	200	1600	4800	6000
	300	2400	7200	9000
7.3	100	400	1200	-
	200	800	2400	-
	300	1200	3600	-
7.2	100	500	1500	-
	200	1000	3000	-
	300	1500	4500	-
7.1	100	600	1800	-
	200	1200	3600	-
	300	1800	5400	-
6	100	300	900	-
	200	600	1800	-
	300	900	2700	-

Table K.8 – A1 Highway scenario RU (transmitter and receiver) service mix.

Service	I. C.M	V. SM	H. FB	E. C.M	T. IF	R. DV	NB. S. S	M. TL	PT. MC	PK. MC	A. RT	VC	V. CF	W. BW	EM	S. NW	F. TF
RRH/RU Service mix [%]	0	0	0	0	9	1	9	0	0	0	0	40	0	17	11	11	2
DU Service mix [%]	0	1	1	1	11	2	11	0	0	0	0	41	2	11	8	9	2
CU/BBU Service mix [%]	2	4	4	2	17	1	6	2	2	2	1	35	2	7	5	7	1

Table K.9 – Average number of users per node on the AutoEuropa factory scenario.

Splitting Option	RU/RRH	DU	CU	BBU
8	300	1200	3600	4000
7.3	300	700	2100	-
7.2	300	900	2700	-
7.1	300	1000	3000	-
6	300	600	1800	-

Table K.10 – AutoEuropa factory scenario RU (on both the transmitter and receiver side) service mix.

Service	I. C.M	V. SM	H. FB	E. C.M	T. IF	R. DV	NB. S. S	M. TL	PT. MC	PK. MC	A. RT	VC	V. CF	W. BW	EM	S. NW	F. TF
RU Service mix [%]	0	0	0	0	0	0	0	20	10	10	0	20	10	10	10	10	0
DU Service mix [%]	0	0	0	0	10	1	5	22	8	8	0	20	7	6	6	6	1
CU/BBU Service mix [%]	2	4	4	2	14	2	7	20	6	6	0	20	3	3	3	3	1

Annex L

Radio Characteristics Configuration

This Annex presents the simulated radio techniques for each studied scenario.

Table L.1 – 4G Radio Characteristics for each scenario.

4G scenario	MIMO layers	Numerology	Bandwidth [MHz]	Average CQI	DL Usage Ratio	UL Usage Ratio
Santa Maria Hospital	4	0	20	12	0.31	0.08
Aveiro Outdoor	2	0	20	9	0.65	0.17
Espírito Santo de Aveiro Hospital	4	0	20	12	0.31	0.08
A1 Highway (LTE 800)	2	0	10	9	0.65	0.17
A1 Highway (LTE 1800)	2	0	20	9	0.65	0.17
Avenida da Liberdade (LTE 800)	2	0	10	9	0.65	0.17
Avenida da Liberdade (LTE 1800)	2	0	20	9	0.65	0.17
Avenida da Liberdade (LTE 2600)	2	0	20	9	0.65	0.17
AutoEuropa factory	4	0	20	12	0.31	0.08

Table L.2 – 5G Radio Characteristics for each scenario.

Scenario	MIMO layers	Numerology	Bandwidth [MHz]	CQI	Scaling Factor	DL Frame Structure	Average Factor
Santa Maria Hospital	4	1	100	12	1	0.57	0.7
Espírito Santo de Aveiro Hospital	4	1	100	12	1	0.57	0.7
Aveiro Outdoor	4	1	100	9	1	0.57	0.7
A1 Highway	4	1	100	9	1	0.57	0.7
Avenida da Liberdade	4	1	100	9	1	0.57	0.7
AutoEuropa factory	4	1	100	12	1	0.57	0.7

Annex M

Urban ITS and A1 Highway Scenarios

This Annex presents the results obtained for the Urban ITS and the A1 Highway scenarios with 30% and 90% of the maximum traffic.

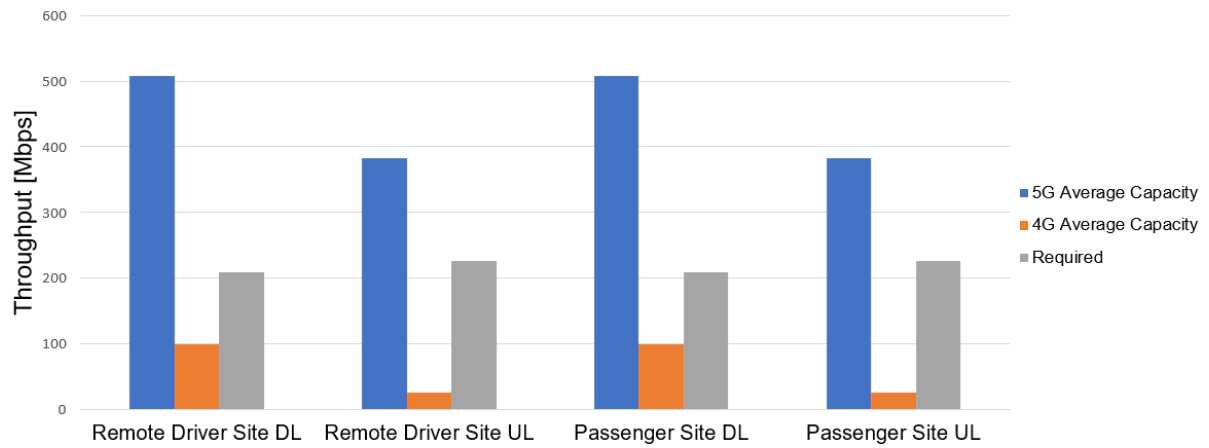


Figure M.1 – RRH/RU throughputs for the Urban ITS scenario with 30% of the maximum traffic.

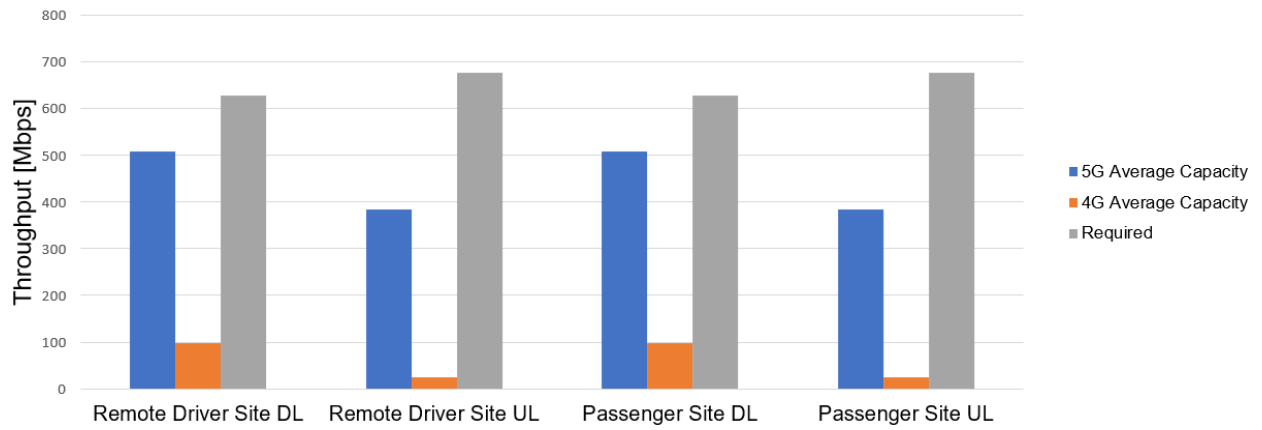


Figure M.2 – RRH/RU throughputs for the Urban ITS scenario with 90% of the maximum traffic.

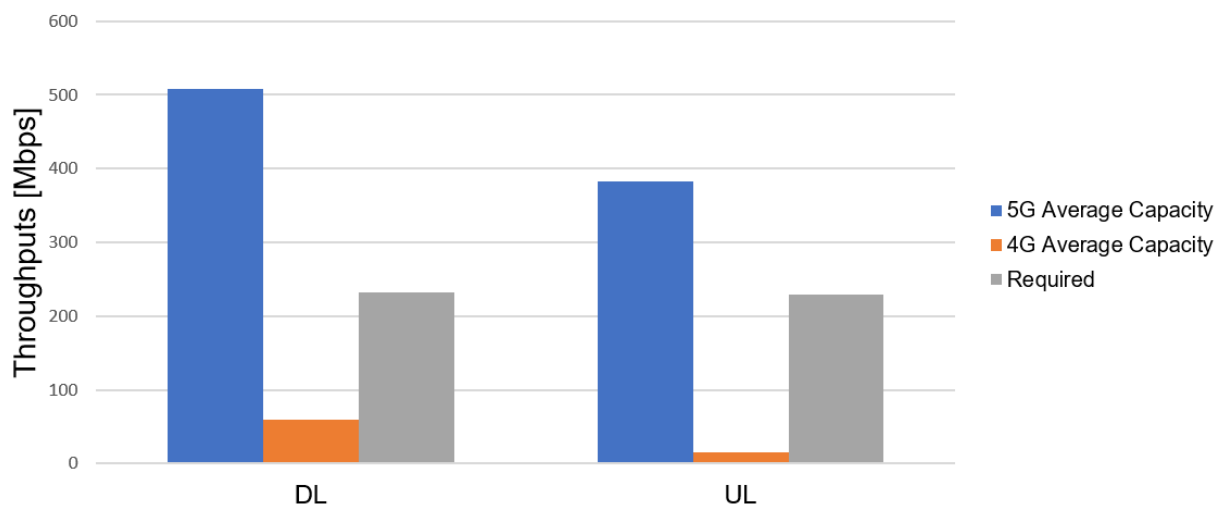


Figure M.3 – RRH/RU throughputs for the A1 Highway scenario with 30% of the maximum traffic.

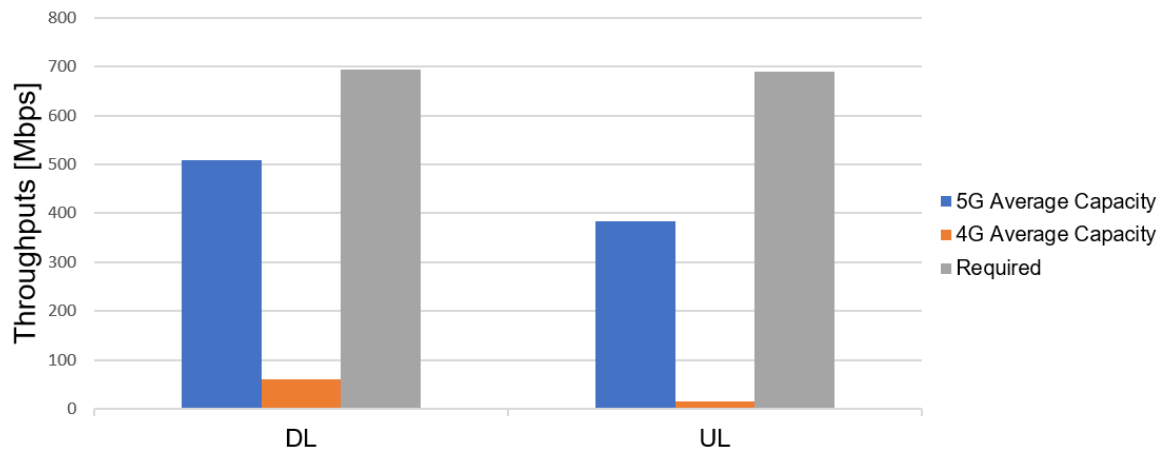


Figure M.4 – RRH/RU throughputs for the A1 Highway scenario with 90% of the maximum traffic.

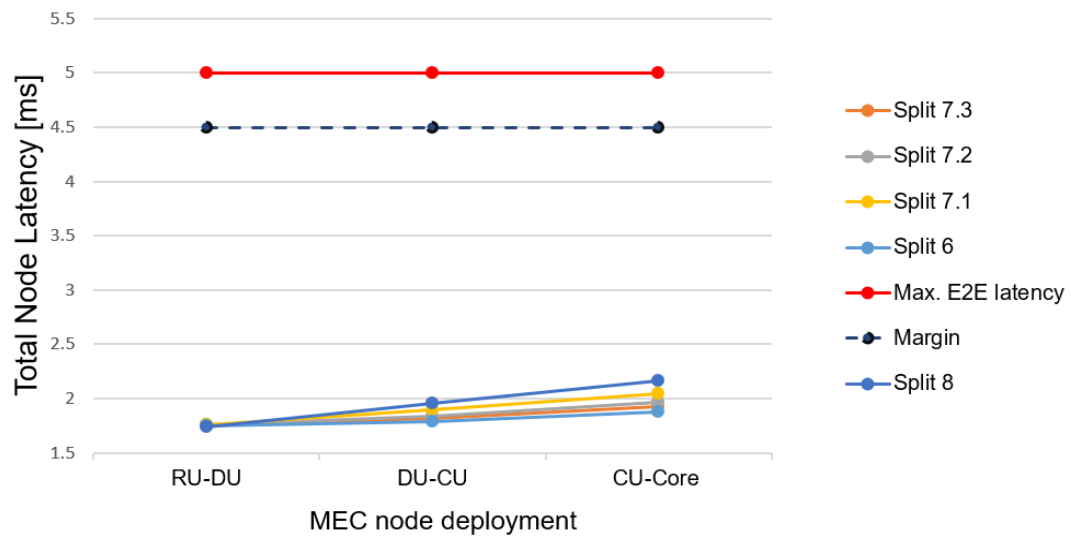


Figure M.5 – Latency results for the Urban ITS scenario with 30% of the maximum traffic.

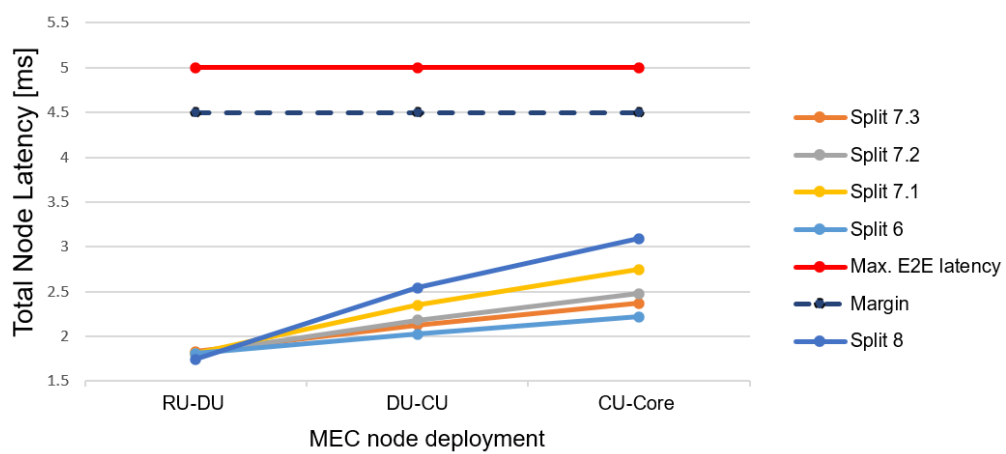


Figure M.6 – Latency results for the Urban ITS scenario with 90% of the maximum traffic.

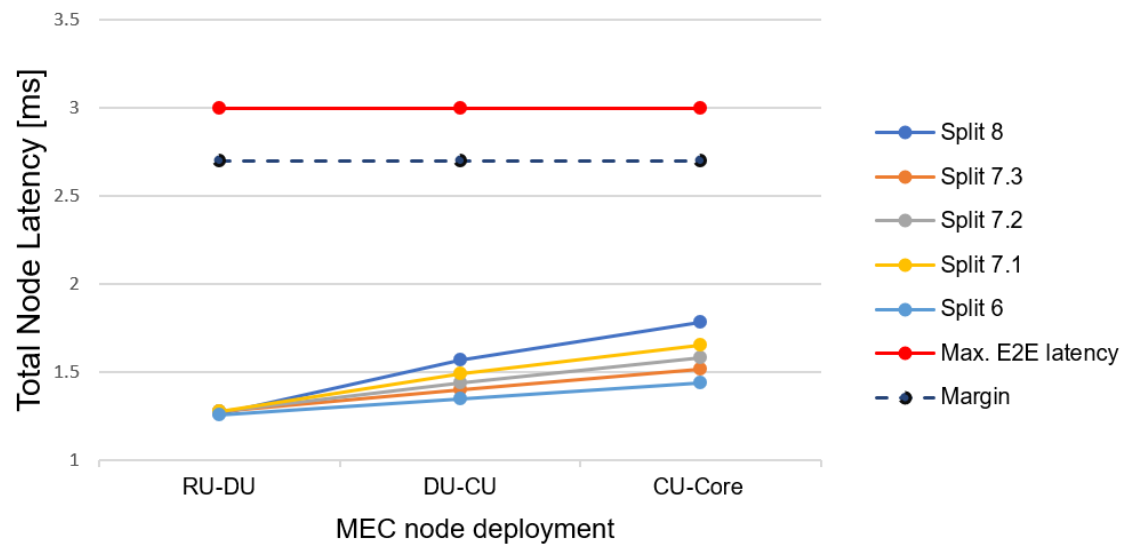


Figure M.7 – Latency results for the A1 Highway scenario with 30% of the maximum traffic.

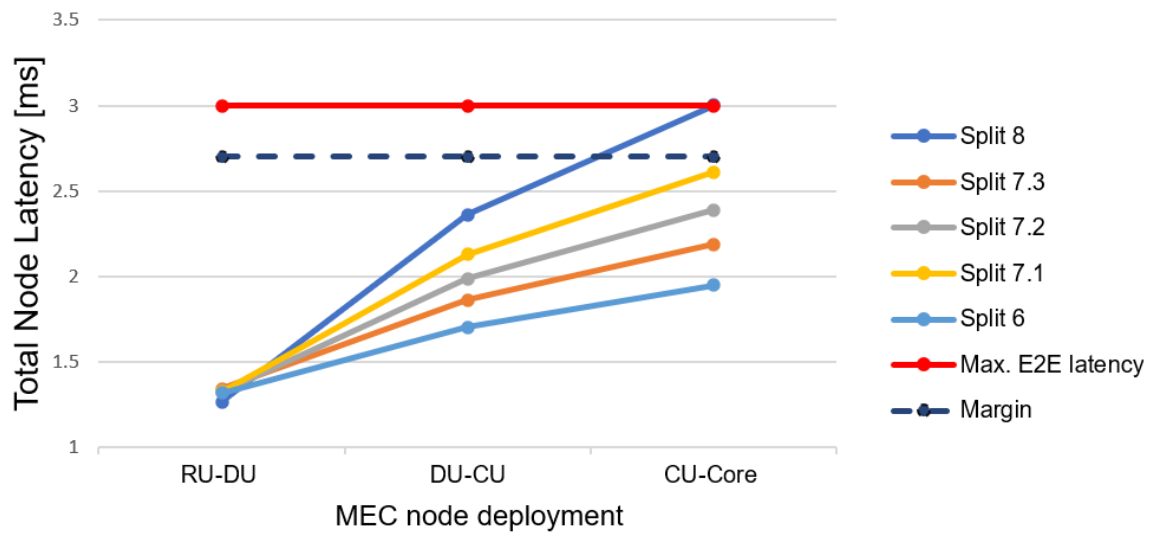


Figure M.8 – Latency results for the A1 Highway scenario with 90% of the maximum traffic.

Annex N

MEC node Coverage

This Annex presents the coverage provided to other hospitals by the MEC node installed in the Santa Maria Hospital and the coverage provided by the MEC nodes installed along the A1 Highway.

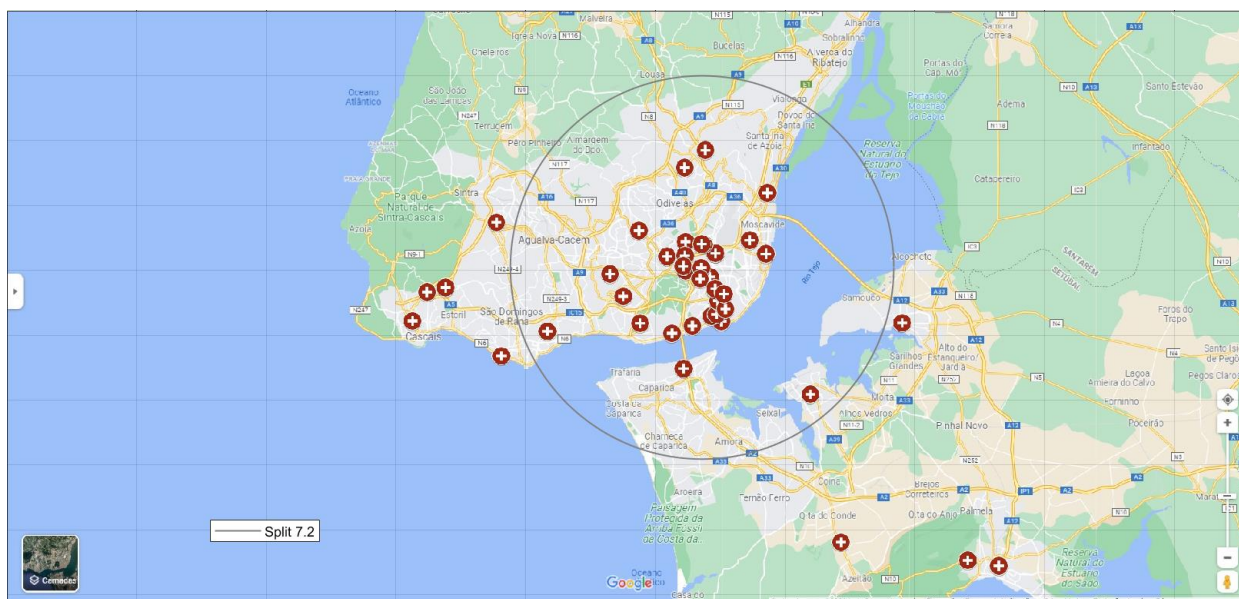


Figure N.1 – Santa Maria scenario MEC node coverage.

Table N.1 – Santa Maria scenario MEC node coverage hospitals list.

Hospital List
Trofa Saúde Loures
Hospital Beatriz Ângelo
Hospital do Mar
Hospital SAMS
Hospital CUF Descobertas
Centro Hospitalar Psiquiátrico de Lisboa
Hospital Pulido Valente
Hospital das Forças Armadas
Trofa Saúde Amadora
Hospital Curry Cabral
Hospital da Luz Oeiras
Hospital Professor Dr. Fernando Fonseca
Instituto de Oncologia Francisco Gentil
Hospital da Cruz Vermelha
Maternidade Alfredo da Costa
Hospital Dona Estefânia
Hospital Santo António dos Capuchos
Hospital São José
Hospital St. Louis
CUF Infante Santo
Hospital São Francisco Xavier
Hospital Santa Cruz
Hospital Egas Moniz
Hospital da Ordem Terceira Chiado
Hospital da Nossa Senhora do Rosário
Hospital Garcia da Orta
Hospital de Jesus

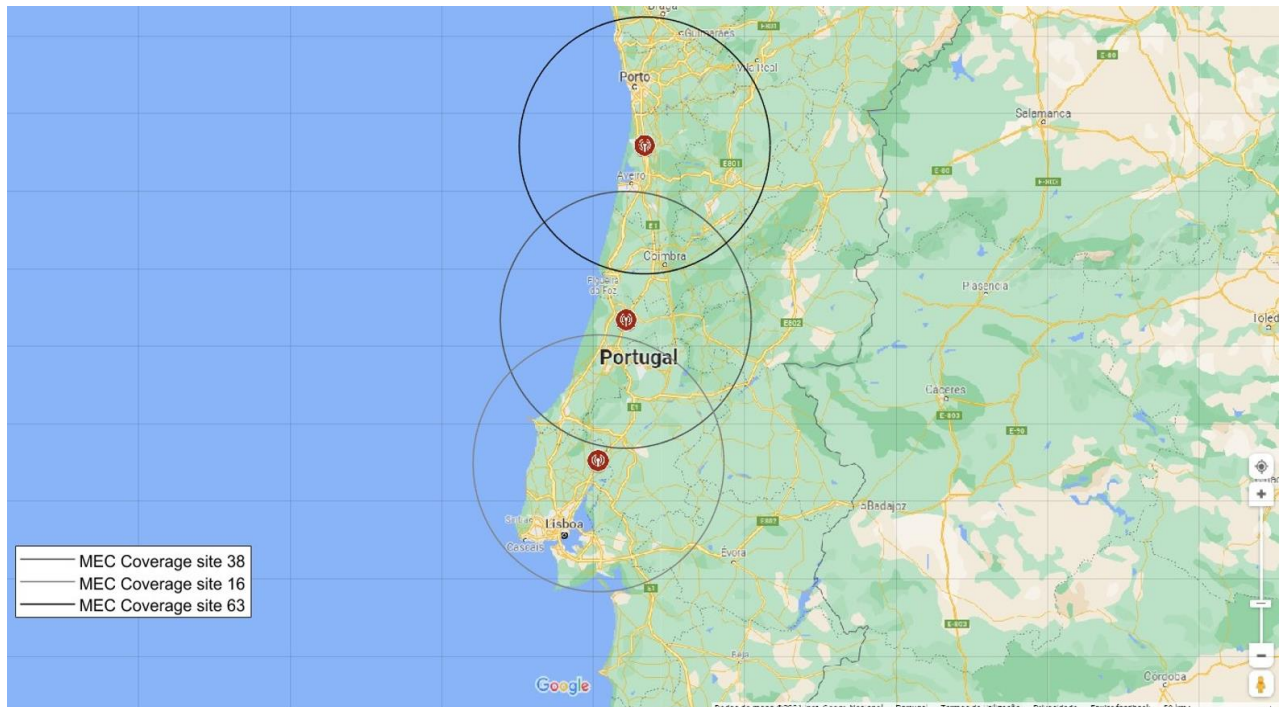


Figure N.2 – A1 Highway MEC node coverage.

References

- [3GPP19a] 3GPP, *System Architecture for the 5G System (Release 15)*, Report TS 23.501, Stage 2, V15.3.0 Sep.2019.
- [3GPP19b] 3GPP, *System Architecture for the 5G System (Release 15)*, Report TS 38.211, V15.4.0, Apr.2019.
- [3GPP19c] 3GPP, *Latency needs to support example use cases from vertical industries (Release 15)*, Report TS 22.261, V15.7.0, Mar.2019
- [3GPP18a] 3GPP, *Technical Specification Group Services and System Aspects: Study on enhancement of 3GPP Support for 5G V2X Services (Release 16)*, Report TR 22.886, V16.2.0, Dec.2018.
- [3GPP18b] 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.
- [3GPP17] 3GPP, *LTE Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures; (Release 14)*, Report TS 36.213, V14.2.0, Apr.2017.
- [5GOB20] European 5G observatory, *National 5G Spectrum Assignment*, Available: <http://5gobservatory.eu/5g-spectrum/national-5g-spectrum-assignment/#1533307441058-1f1bbc1b-307c>.
- [ALT120] Altice Labs, *5G Intelligent Communications for V2X ecosystems Whitepaper*, Jul. 2020 Available: https://www.alticelabs.com/content/WP_5G_Intelligent_Communications.pdf.
- [ASCC18] Alejandro Santoyo-González, Cristina Cervelló-Pastor, "Edge Nodes Infrastructure Placement Parameters for 5G Networks", in *Proc. of 2018 IEEE Conference on Standards for Communications and Networking (CSCN)*, Paris, France, Oct.2018. Available: <https://ieeexplore.ieee.org/document/8581749>.
- [ARTM19] Amine El Rhayour, Tomader Mazri, "5G Architecture: Deployment scenarios and options", in *Proc. of 2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Rome, Italy, Nov.2019. Available: <https://ieeexplore.ieee.org/document/9069723>.
- [AZRB17] Ali Zaidi and Robert Baldemair, *5G NR Physical Layer Design*, Internal Report, Stockholm, Sweden, Jun.2017. Available: <https://www.ericsson.com/49e9d0/assets/local/reports-papers/ericsson-technology-review/docs/2017/designing-for-the-future---the-5g-nr-physical-layer.pdf>.
- [Corr20] Luís Manuel Correia, *Notes from Mobile Communication Systems course*, Instituto Superior Técnico, Lisbon, Portugal, 2020.
- [DMAG18] Daniel Maaz, Ana Galindo-Serrano, Salah Eddine Elayoubi, "URLLC User Plane Latency

- Performance in New Radio", in *Proc. of IEEE 2018 25th International Conference on Telecommunications (ICT)*, Saint-Malo, France, Sep.2018. Available: <https://ieeexplore.ieee.org/document/8464912>.
- [Eric20] Ericsson, *Ericsson Mobility Report*, Jun. 2020. Available: <https://www.ericsson.com/49da93/assets/local/mobility-report/documents/2020/june2020-ericsson-mobility-report.pdf>.
- [ETSI18a] ETSI, *MEC in 5G Networks*, White Paper, Jun.2018, Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf.
- [ETSI18b] 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.
- [ETSI18c] 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.
- [ETSI18d] ETSI, *5G NR; Physical layer procedures for data (Release 15)*, Internal Report TS 38.214, V15.3.0, Oct.2018. Available: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/15.03.00_60/ts_138214v150300p.pdf?fbclid=IwAR3DDIbC3WKuD_cFsM3rdL_Di8j5Mzplau9QQot5BIX8m5jjCleBfQML-Ow.
- [ETSI18e] ETSI, *5G NG-RAN; F1 Application Protocol (F1AP) Release 15*, Internal Report TS 38.473, V15.2.1, Jul.2018. Available: https://www.etsi.org/deliver/etsi_ts/138400_138499/138473/15.02.01_60/ts_138473v150201p.pdf?fbclid=IwAR0Cux-OzDho1KmS3pl-8rmCGKwR1drFR-sFDkOWeFc4f3OS4jOBCI3bAo4.
- [ETSI13] ETSI, *Network Functions Virtualisation (NFV) Architectural Framework*, Group Specification GS NFV.002, V1.1.1, Oct.2013. Available: https://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.01.01_60/gs_NFV002v010101p.pdf.
- [GLJW18] Guangshun Li, Jiping Wang, Junhua Wu, Jianrong Song, "Data Processing Delay Optimisation in Mobile Edge Computing", *Wireless Communications and Mobile Computing Journal*, Vol. 2018, No. 1, Feb.2018, pp. 1-9. Available: <https://downloads.hindawi.com/journals/wcmc/2018/6897523.pdf>.
- [GMFF20] Gordana Barb, Marius Ottesteanu, Florin Alexa, Flavius Danuti, "OFDM Multiple-Numerology for Future 5G New Radio Communication Systems," in *Proc. of IEEE 2020 International Conference on Software, Telecommunications and Computer Networks*, Split,

Hvar, Croatia Sep.2020. Available:
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9238308>.

- [HMRS16] Hugo Silva, *Design of the C-RAN Fronthaul for Existing LTE Networks*, M.SC thesis, Instituto Superior Técnico, Lisbon, Portugal, Nov.2016. Available:
https://grow.tecnico.ulisboa.pt/wp-content/uploads/2016/12/ThesisFinal_HugoS.pdf.
- [HUAW20] Huawei, *5G Spectrum - Public Policy Position*, Internal Report, Feb.2020. ,Available:
https://www-file.huawei.com/-/media/corporate/pdf/public-policy/public_policy_position_5g_spectrum_2020_v2.pdf?la=en.
- [IAIH18] 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>.
- [ITUT19] ITU, *CoE Training on Traffic engineering and advanced wireless network planning*, 4G to 5G Network and Standard Releases, Bangkok, Thailand, Oct.2019 Available:
https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/SiteAssets/Pages/Events/2019/ITU-ASP-CoE-Training-on-3GPP_4G%20to%205G%20networks%20evolution%20and%20releases.pdf.
- [ITUT18] 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.
- [ITUT17] Denis Andreev, *Overview of ITU activities on 5G*, Saint Petersburg, Russia, June. 2017 Available: https://www.itu.int/en/ITU-D/Regional-Presence/CIS/Documents/Events/2017/06_Saint_Petersburg/Presentations/ITU%20Workshop%2019.06%20-%20Denis%20Andreev%202.pdf.
- [JMIL02] Jacques Marescaux, Joel Leroy, Francesco Rubino and Michelle Smith, "Transcontinental Robot-Assisted Remote Telesurgery: Feasibility and Potential Applications", *Annals of Surgery*, Vol. 235, No. 4, Apr.2002, pp 487–492. Available:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422462/pdf/20020400s00005p487.pdf>.
- [KMNT18] Kmar Thaalbi, Mohamed Taher Missaoui, Nabil Tabbane, "Short Survey on Clustering Techniques for RRH in 5G networks" in *Proc. of IEEE 2018 Seventh International Conference on Communications and Networking (ComNet)*, Hammamet, Tunisia, Nov.2018. Available : <https://ieeexplore.ieee.org/document/8622300>.
- [MIW15] Mikio Iwamura, "NGMN View on 5G Architecture", in *Proc. of 2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, Glasgow, United Kingdom, May.2015. Available:
<https://ieeexplore.ieee.org/document/7145953>.
- [NGOF18] Next Generation Optical Transport Network, *5G Oriented OTN technology White Paper*, Mar.2018. Available: <http://www.ngof.net/download/5G.pdf>.

- [NHKA19] Najmul Hassan, Kok-Lim Alvin Yau and Celimuge Wu, "Edge Computing in 5G: A Review", *IEEE Access Journal*, Vol. 7, May.2015, pp 127276 –127289. Available: <https://ieeexplore.ieee.org/document/8821283>.
- [OAGW18] Osama Al-Saadeh, Gustav Wikstrom, Joachim Sachs, "End-to-End Latency and reliability Performance of 5G in London", in *Proc. of IEEE 2018 Global Communications Conference*, Abu Dhabi, United Arab Emirates, Feb.2019. Available: <https://ieeexplore.ieee.org/document/8647379>.
- [PaMa17] Pavel Mach and Zdenek Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," in *IEEE Communications Surveys & Tutorials*, Vol. 19, No. 3, Aug.2017, pp. 1628 -1656. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7879258>.
- [PSMM17] Philipp Schulz, Maximilian Matth  , Henrik Klessig and Meryem Simsek, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture", in *IEEE Communications Magazine*, Vol. 55, No. 2, Feb.2017, pp. 70 – 78. Available: <https://ieeexplore.ieee.org/document/7842415>.
- [QJGZ18] Qi Zhang, Jianhui Liu and Guodong Zhao, *Towards 5G Enabled Tactile Robotic Surgery*, Aarhus University, Aarhus, Denmark, Mar.2018. Available: <https://arxiv.org/pdf/1803.03586.pdf>.
- [Qual16] Qualcomm, *Making 5G NR a reality*, Internal Report, Dec.2016. Available: <https://www.qualcomm.com/media/documents/files/whitepaper-making-5g-nr-a-reality.pdf>.
- [Qual18] Qualcomm, *VR and AR pushing the connectivity limits*, Internal Report, Oct.2018. Available: <https://www.qualcomm.com/media/documents/files/vr-and-ar-pushing-connectivity-limits.pdf>.
- [SBDC18] Steinar Bjornstad, David Chen and Raimena Veisllari, "Handling Delay in 5G Ethernet Mobile Fronthaul Networks" in *Proc. of IEEE 2018 European Conference on Networks and Communications*, Ljubljana, Slovenia, Jun.2018. Available: <https://ieeexplore.ieee.org/document/8442755>.
- [SeDo19] S  rgio Domingues, *Analysis of the Performance of Multi-Access Edge Computing Network Slicing in 5G*, M.SC thesis, Instituto Superior T  cnico, Lisbon, Portugal, Nov.2019. Available: https://grow.tecnico.ulisboa.pt/wp-content/uploads/2020/07/Thesis_SergioD_vPublic.pdf.
- [Thal20] Thales, *Introducing 5G technology and Network*, Dec.2020. Available: <https://www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/inspired/5G>.
- [VGMS12] Vijay Gurbani, Michael Scharf, Lakshman and Volker Hilt, "Abstracting network state in Software Defined Networks (SDN) for rendezvous services", in *Proc. of 2012 IEEE International Conference on Communications (ICC)*, Ottawa, Canada, Jun.2012. Available:

<https://ieeexplore.ieee.org/abstract/document/6364858>.

- [VODA18] 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>.
- [YCNP18] Young-il Choi and Noik Park, "Support for Edge Computing in 5G networks," in *Proc. of IEEE Tenth International Conference on Ubiquitous and Future Networks*, Prague, Czech Republic, Jul.2018. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8436806>.
- [YZWW19] Yongli Zhao and Wei Wang, "Edge Computing and Networking: A Survey on Infrastructures and Applications", *IEEE Access*, Vol.7, Jul.2019, pp. 101213 - 101230. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8758431>.