TÉCNICO LISBOA

Analysis of the Performance of Multi-Access Edge Computing Network Slicing in 5G

Sérgio Alexandre Proença Domingues

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor: Prof. Luís Manuel de Jesus Sousa Correia

Examination Committee

Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino

Supervisor: Prof. Prof. Luís Manuel de Jesus Sousa Correia

Members of Committee: Prof. Prof. António José Castelo Branco Rodrigues

Eng. Ricardo Dinis

November 2019

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 family and friends

Acknowledgements

First of all, I would like to express my deepest appreciation to Professor Luís M. Correia for giving me the opportunity to develop my master thesis under his supervision and for all the advice throughout the year. I am thankful for all the meetings that were not only essential to the orientation of my work, but also had a strong contribution to the development of my professional and personal life, to always making me reach for the perfect results. I am grateful for the opportunity to be a part of Group of Research On Wireless (GROW).

Thank you to all GROW members, especially to my friends Magali Correia, Tomás Duarte, Mojgan Barahman, Kenan Turbic for all the support, valuable advice and time that we spent together.

To Eng. Ricardo Dinis and Eng. Luis Santo from NOS, for providing me the opportunity to work closely with the industry, and providing me with their time in helpful meetings where their support and suggestions were fundamental for my work.

To my close friends from Instituto Superior Técnico that have been with me during these five years, for all the support and good moments.

To all my family, especially my mother, Isabel Domingues, father, José Domingues, brother, Rafael Domingues, sister, Gisela Domingues, grandmother, Dina Domingues, and grandfather, José Domingues for their support, love and encouragement throughout all my life and education.

To my girlfriend, Rita Caria Mendes, for always helping me when I needed, giving me the love and support I needed in the hardest moments throughout the last five years.

To Mariana Nunes, my great friend, that helped me with my weaknesses, providing me the conditions to finish my work.

Abstract

The main purpose of this thesis is to analyse the technologies essential to support new 5G network services, with higher data rates and ultra-low latency, emphasising edge networking technology used to offload the computation tasks from the centralised network to the edge of the cloud, near users, in order to reduce latency and support more computation power near the network terminal nodes. This work studies the characteristics of the different network architectures on C-RAN in other to optimise the network for multiple services and applications from 5G. The model takes into consideration five essential parameters to support 5G services demands, including centralisation gain, network latency, node throughput, node processing power and network cost. The model is used to study the performance of the network in multiple scenarios, where one concludes that, in other to support the 1 ms latency demands, it requires the introduction of at least 5 MEC nodes and 30 MEC nodes on the Minho and Portugal scenarios, respectively. The results obtained show that it is essential to increase the splitting option used on the 4G network fronthaul, sending more processing power near users, which will compress the signal and reduce node throughput by, at least, 95%.

Keywords

5G, Cloud RAN, Radio Unit, Distributed Unit, Central Unit, MEC

Resumo

O objetivo principal desta tese é fornecer uma análise das tecnologias essenciais para suportar os novos serviços de rede 5G, com taxas de transmissão de dados mais altas e latência ultra baixa, destacando a tecnologia de *edge networking* que é utilizada para redirecionar tarefas desde a rede centralizada até ao extremo da rede, perto dos utilizadores, com o objetivo de reduzir a latência e suportar maior capacidade de computação próxima dos terminais de rede. Este trabalho estuda as características das diferentes arquiteturas de rede no C-RAN para otimizar a rede para vários serviços e aplicativos do 5G. O modelo tem em consideração cinco parâmetros essenciais para o suporte dos requisitos dos serviços 5G, incluindo o ganho de centralização, latência da rede, taxa de transferência nos nós da rede, poder de processamento do nó e custo. O modelo é utilizado para estudar o desempenho da rede em vários cenários, onde se conclui que para suportar os requisitos de latência de 1 ms, é necessário introduzir pelo menos 5 nós MEC e 30 Nós MEC nos cenários Minho e Portugal, respetivamente. Os resultados obtidos mostram que é essencial aumentar a opção de divisão usada no *fronthaul* da rede 4G, enviando mais funções da estação base próximo dos utilizadores, com o objetivo de comprimir o sinal e reduzir a taxa de transferência nos nós de agregação em pelo menos 95%.

Palavras-chave

5G, *Cloud* RAN, Unidade de Radio, Unidade Distribuída, Unidade Central, MEC

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
4G	4rd Generation
5G	5th Generation
AMC	Adaptive Modulation and Coding
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat reQuest
BBU	Baseband Unit
BS	Base Station
CA	Carrier Aggregation
CAPEX	Capital Expenditures
CCE	Congestion Control Engine
CN	Core Network
CP	Cyclic Prefix
CPRI	Common Public Radio Interface
CPU	Central Process Unit
C-RAN	Cloud Radio Access Network
CRC	Cyclic Redundancy Check
CS	Circuit-Switched
CU	Central Unit
DC	Delay Sensitive
DL	Downlink
DT	Delay Tolerant
DU	Distributed Unit
eMBB	enhanced Mobile Broadband
eMBMS	evolved Multimedia Broadcast Multicast Service
EN	Edge Nodes
eNodeB	Evolved Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
E-UTRA	Evolved Universal Terrestrial Radio Access
FDD	Frequency Division Duplex
FFT	Fast Fourier Transform
f-OFDM	filtered-OFDM
GBR	Guaranteed Bit Rate

GOPS	Giga Operations per Second
GSM	Global System for Mobile Communications
HA	Hybrid Applications
HARQ	Hybrid Automatic Repeat reQuest
HMWC	High Mobility Wireless Communication
HSR	High-Speed Railway
HSS	Home Subscription Service
ICIC	Inter-Cell Interference Coordination
IFFT	Inverse Fast Fourier Transform
IIoT	Industrial Internet of Things
IMS	IP Multimedia Core Network Subsystem
IoT	Internet of Things
IP	Internet Protocol
ITS	Intelligent Transport Systems
LSCP	Large-Scale Collaborative Processing
LTE	Long Term Evolution
MAC	Media Access Control
MC	Mobile Clones
MEC	Multi-Access Edge Computing
METIS	Mobile and wireless communications Enablers for the Twenty-twenty Information Society
MIMO	Multiple-Input Multiple-Output
MME	Mobility Management Entity
mMTC	Massive Machine-Type Communications
mm-wave	millimetre wave
NF	Network Function
NFV	Network Function Virtualisation
NR	New Radio
NRTA	Non Real-Time Applications
NSA	Non-Standalone
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal frequency-division multiple access
OPEX	Operational Expenditures
OTN	Optical Transport Network
PC	Packet-Switched
PCEF	Policy Control Enforcement Function
PCRF	Policy and Charging Rules Function
PDB	Packet Delay Budget
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network

PELR	Packet Error Loss Rate
P-GW	Packet Data Network Gateway
PHY	Physical
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAN	Radio Access Network
RE	Resource Element
RF	Radio Frequency
RLC	Radio Link Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Remote Radio Unit
RTA	Real-Time Application
RU	Radio Unit
SA	Standalone
SAP	Service Access Point
SDN	Software Defined Network
SFN	Single Frequency Network
S-GW	Serving Gateway
TDD	Time Division Duplex
TTFP	Two-Thresholds Forwarding Policy
UE	User Equipment
UHD	Ultra-High Definition
UL	Uplink
UMTS	Universal Mobile Telecommunications System
URLLC	Ultra-Reliable Low-Latency Communication
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-everything
VNF	Virtual Network Function
VR	Virtual Reality
xMBB	Extreme Mobile BroadBand

List of Symbols

α	Cost per km of the fibre
β	Cost per unit of resource for the node
Δ_{CU}	Variable cost of the CU according to the required capacity
Δ_{DU}	Variable cost of the DU according to the required capacity
Δ_{MEC}	Variable cost of the MEC according to the required capacity
Δ_{RU}	Variable cost of the RU according to the required capacity
δ_{App}	Maximum latency depending on what application is chosen
$\delta_{\mathrm{BH,C}}$	Backhaul to core transmission Latency
$\delta_{\mathrm{BH,MEC}}$	Backhaul to MEC transmission Latency
δ_{Cor}	Core processing delay
$\delta_{\mathrm{C-RAN}}$	C-RAN associated Latency
$\delta_{\mathrm{CU,DL}}$	CU DL processing delay
$\delta_{\mathrm{CU,UL}}$	CU UL processing delay
$\delta_{\mathrm{DU,DL}}$	DU DL processing delay
$\delta_{\mathrm{DU,UL}}$	DU UL processing delay
δ_{E2E}	End to End Latency
δ_{EN}	External Data centre contribution delay
δ_{FH}	Transmission delay between the RU to the DU
δ_{HARQ}	HARQ protocol requirement latency
$\delta_{\mathrm{MEC,DL}}$	MEC DL processing delay
$\delta_{\mathrm{MEC,UL}}$	MEC UL processing delay
δ_{MH}	Transmission delay between the DU to the CU
$\delta_{\mathrm{Node,que}}$	Queuing delay on the RU
δ_{Node}	Processing delay on the Node

$\delta_{\mathrm{RU,DL}}$	RU DL processing delay
$\delta_{\mathrm{RU,UL}}$	RU UL processing delay
$\delta_{Node,\mathrm{proc}}$	BS function processing delay on the RU
δ_{Tran}	Transport transmission delay from the core to the external data centres
μ_s	Subcarrier utilisation (load)
μ_{Node}	Node load
τ	Mean service duration per hour
A_{CU}	Area of a CU
A_{DU}	Area of a DU
A_{RU}	Area of a RU
A_{MEC}	Area of a MEC
a_{Link}	Boolean variable equal to 1 when there is a link between the nodes
B	Bandwidth
B_c	Bandwidth for control signals
B_{ref}	Reference bandwidth
b_{CU}	Boolean variable equal to 1 when CU is being used
b_{DU}	Boolean variable equal to 1 when DU is being used
b_{MEC}	Boolean variable equal to 1 when MEC is being used
b_{RU}	Boolean variable equal to 1 when RU is being used
C_A	Rent cost per month per square metre
C_{CAPEX}	Total CAPEX
C_{CU}	Constant cost of CUs
C_{DU}	Constant cost of DUs
C_E	Cost of energy consumed per kW/h
C_{Fibre}	Constant cost of the Fibre
$C_{Inter,i}$	Cost of interface in splitting option i
C_{Inter}	Cost per interface of the node
$C_{M,D}$	Maintenance cost of the node

$C_{M,L}$	Maintenance cost of the links
C_M	Cost related to maintenance per year
C_{MEC}	Constant cost of MECs
$C_{Microwave}$	Constant cost of the Microwave link
C_{OPEX}	Total OPEX per year
C_P	Cost related to power consumption per year
C_R	Cost related to renting per year
C_{RU}	Constant cost of RUs
$C_{t,CU}$	Total cost of CUs
$C_{t,CUMEC}$	Total cost of links between CUs and MECs
$C_{t,DU}$	Total cost of DUs
$C_{t,DUCU}$	Total cost of links between DUs and CUs
$C_{t,Fibre}$	Total cost of the Fibre link
$C_{t,MEC}$	Total cost of MECs
$C_{t,Microwave}$	Total cost of the Microwave link
$C_{t,RU}$	Total cost of RUs
$C_{t,RUDU}$	Total cost of links between RUs and DUs
C_T	Total cost of the C-RAN
d_{BH}	Backhaul distance
d_{E2E}	Maximum E2E distance
d_f	Fibre link distance
d_{FH}	Fronthaul maximum distance
d_{MH}	Middlehaul maximum distance
d_{Tran}	Distance between the core and external data centre
E	Spectral efficiency
E_{CU}	Power consumed per hour for a CU
E_{DU}	Power consumed per hour for a DU
E_{ref}	Reference spectral efficiency
E_{RU}	Power consumed per hour for a RU

E_{MEC}	Power consumed per hour for a MEC
$F_{DC,ref}$	Reference system load in the frequency-domain
F_{DC}	System load in the frequency-domain
$G_{mux,T}$	Aggregation gain
M	Modulation order
M_c	Modulation order for control signals
M_P	Processing capacity multiplier
M_{Serv}	Number of users for each service multiplier
m_{Link}	Percentage of total investment spent on maintenance of a link
m_{Node}	Percentage of total investment spent on maintenance of the nodes
MAC_{info}	Bitrate used for information to the MAC layer
$N_{A,ref}$	Reference number of antennas in the BS
N_A	Number of antennas in the BS
N_{Inter}	Number of interfaces in each node
$N_{Node,a}$	Number of aggregation nodes
$N_{Node,c}$	Number of nodes connected to the aggregation node
N_{pop}	Population in the area
N_L	Number of layers
$N_{L,c}$	Number of layers for control signalling
N_{SC}	Number of subcarriers
$N_{streams,ref}$	Reference number of transmission streams
$N_{streams}$	Number of transmission streams
N_{SY}	Number of symbols
$N_{U,máx}$	Maximum number of users in the network
$N_{U_{RU}}$	Number of users on the RU
$N_{U_{RU,Serv}}$	Number of users on the RU per service
N_y	Number of years considered for OPEX
P_{BBm}	Baseband modulation/demodulation processing component
P_{BH}	Processing power required for the backhaul interface

P_{CN}	Processing power used by the CN
P_{Code}	FEC function processing component
P_{CU}	Processing power used by the CU
P_{DU}	Processing power used by the DU
P_{MAC}	Processing power required for the MAC layer
P_{Map}	Mapping and demapping functions processing component
P_{MEC}	Processing power used by the MEC
P_{MIMO}	MIMO encoding/decoding processing component
P_{Node}	Processing power on the aggregator node
$P_{Node,Cap}$	Processing capacity assign to the aggregator node
$P_{Node,Cap,ref}$	Reference processing capacity on the node
$P_{Node,fix}$	Fixed processing power required for scheduling and signalling
P_{OFDM}	Frequency domains function for OFDM modulation processing component
P_{PDCL}	Processing power required for PDCL layer
P_{PHY}	Processing power required for the physical layer functions
P_{ref}	References scenario processing component
P_{RF}	Processing power required for RF front-end
P_{RLC}	Processing power required for RLC layer
P_{RU}	Processing power used by the RU
P_{Serv}	Service priority level
P_t	Total processing power required for each node
p_{pen}	Penetration ratio
p_u	Usage ratio
Q	Number of bits used in quantisation
Q_{ref}	Reference number of bits used in quantisation
R_1	Bitrate for splitting 1
R_2	Bitrate for splitting 2
R_6	Bitrate for splitting 6
$R_{7.1,DL}$	DL Bitrate for splitting 7.1

$R_{7.1,UL}$	UL Bitrate for splitting 7.1
$R_{7.2}$	Bitrate for splitting 7.2
$R_{7.3}$	Bitrate for splitting 7.3
R_8	Bitrate for splitting 8
R_c	Control/Schedule signalling rate
R_{FH}	Fronthaul throughput
R_p	Peak LTE data rate
R_{RU}	Throughput on RU
$R_{RU,máx}$	Maximum throughput on the RU
$R_{RU,Serv}$	Throughput on the RU for a specific service
$R_{Serv,ref}$	Reference throughput for a specific service
S	Mean file size per hour
S_{mix}	Service mix
S_r	Sampling rate
$T_{máx,4G}$	Maximum LTE traffic on the network
$T_{Node,a}$	Peak traffic generated by aggregation node
$T_{Node,c}$	Peak traffic generated on the connected node
$T_{RU,4G}$	LTE traffic on the RU
$T_{RU,Serv}$	Traffic on the RU for a specific service
v	Propagation speed in the link

List of Software

Microsoft Word 2019	Programmes Abstract
Microsoft Excel 2019	Calculation and graphical software
Matlab	Computing environment

Chapter 1

Introduction

This chapter gives a brief introduction to the thesis. In Section 1.1 one gives an overview of the state of the current mobile communications and a small introduction to the 5th Generation (5G) network. Section 1.2 starts with the main purpose of the thesis and ends with a description of the structure of its.

1.1 Overview

Mobile users and services are in constant change. Mobile networks continuously take more data and mobile industries do not only need to fulfil those requirements, but also introduce new capabilities and use cases. According to [Eric18A] the number of subscriptions grew at 4% per year, reaching 7.9 billion subscriptions in the first quarter of 2018. Not only the number of subscribers is exponentially increasing, but also the average data volume per subscription, due primarily to watching video content in higher resolutions, has also increased. Figure 1.1 illustrates the global voice and data traffic from 2013 to 2018 and the corresponding year-on-year percentage change of mobile data traffic.

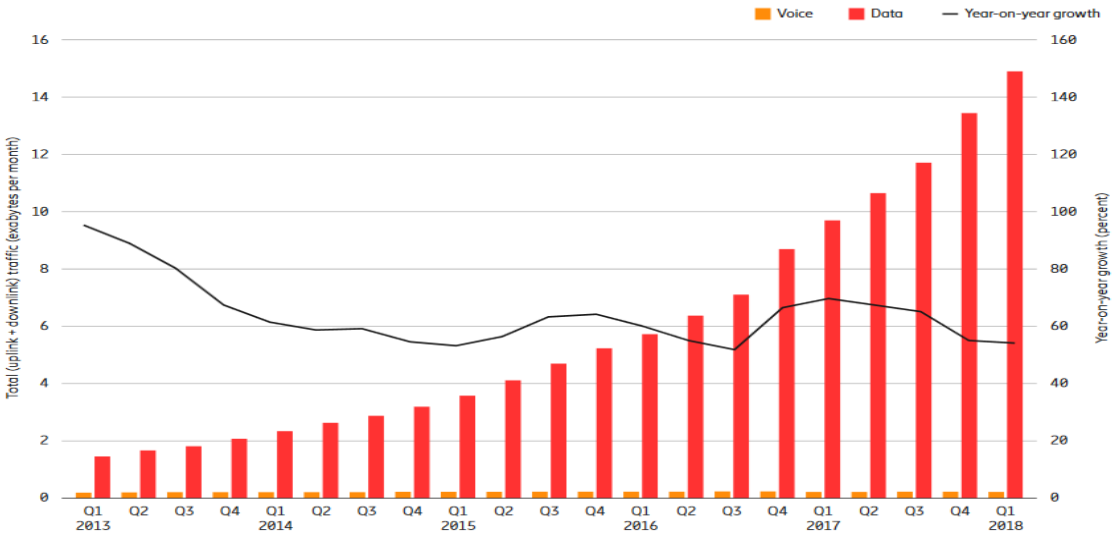


Figure 1.1. Voice and data traffic form 2013-2018 and corresponding year-on-year percentage change of mobile data traffic (extracted from [Eric18A]).

In 2016, the 3rd Generation Partnership Project (3GPP) delivery the first 5G specifications (the release 15) based on two solutions:

- 5G Non-Standalone (NSA) – It enables operators to provide 5G services with shorter time and lower cost using the existing Long Term Evolution (LTE) radio access and the core network. This configuration will likely be used for early 2019 deployments.
- 5G Standalone (SA) - All new 5G Packet Core will be introduced including a New Radio, the 5G New Radio, and 5G Core Network. This new 5G SA Packet Core architecture will include Networks Slicing, Virtualisation, Ultra-low latency, and other aspects. 5G SA is expected to be the first release in 2020.

Mobile operators have been increasing their network capacity in order to satisfy consumer demands. 5G was designed not only to secure those demands, but also to deliver connectivity to virtually every imaginable product. For example, fully-autonomous vehicles need Vehicle-to-vehicle (V2V) and Vehicle-to-everything (V2X) wireless communications technologies. Industrial Internet of Things (IIoT) needs mass connectivity, cloud computing resources, big data analytics, and artificial intelligence. These two examples are services that 5G will support, although their requirements are dramatically different. Figure 1.2 shows some use cases between people, between machines, and between people and

machines dividing those according to performance, attributes, and requirements.

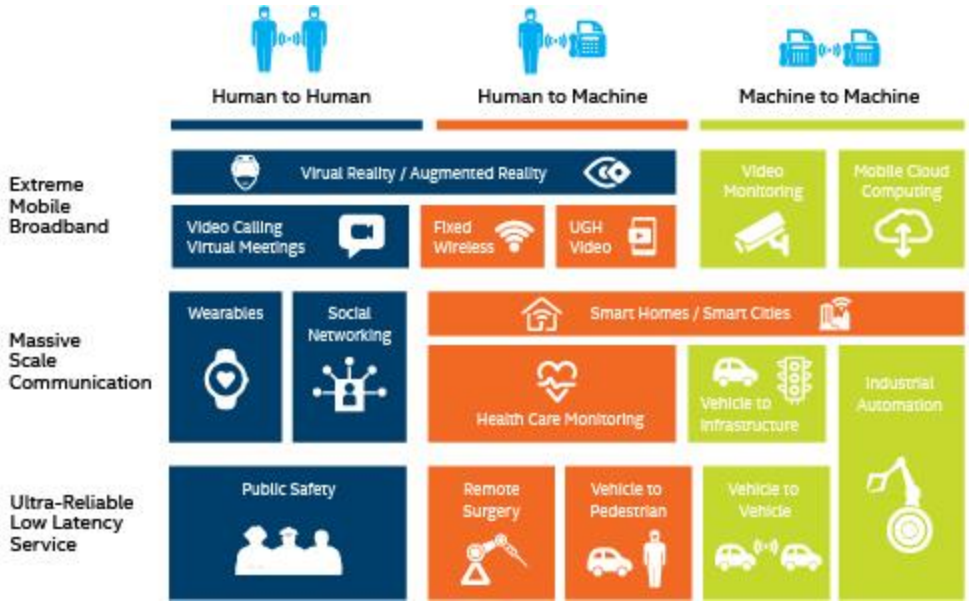


Figure 1.2. 5G use cases (extracted from [5GAm17]).

The introduction of Multi-Access Edge Computing (MEC) is an important concept used in 5G networks. MEC system brings the services close to the devices which provide computation, storage and network resources, different for each application. MEC is essential for the implementation of the services requirements, i.e., latency, scalability, and throughput. According to [Hodg18], MEC also ensures 24% Capital Expenditures (CAPEX) and 25% Operational Expenditures (OPEX) cost reduction for network operators. As illustrated in Figure 1.3, it is expected that data at the edge of the network will increase exponentially, and that by 2022 70% of the produced data will stay at the edge of the network, with just 30% being transported to the data centres of the core.

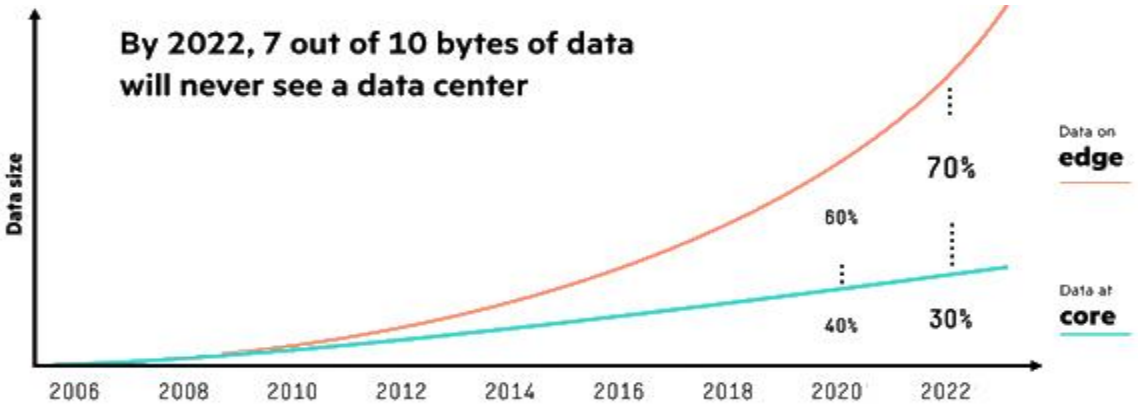


Figure 1.3. Evolution of data on the edge from 2006-2022 (extracted from [Sara18]).

In order for the MEC technology to work at its full potential, the network needs to be sliced into several isolated networks. This technology is called network slicing and is used to customise and optimise each slice of the network, to address the diverse Quality of Service (QoS) requirements for each 5G service. A combination of MEC and network slicing enables a new vertical-oriented slicing and slice management framework, which not only reduces CAPEX and OPEX for the operators but also improves user experience, new business models, reduces the time for service creation and reduces time to market.

1.2 Motivation and Contents

5G is coming, and it will have a massive impact on both consumers and multiple industries, with higher speeds and lower latency that will support new services and business models. In order to support such a variety of services, the network needs to become more flexible and autonomous than ever before. In order to support this network demands, MEC and network slicing are key technologies for the transformation of 5G into multiple industry segments. To ensure that flexible network slicing supports all 5G services, the operators need to plan carefully the position of Edge Nodes (EN) at the cell sites, in order to offload Base Station (BS) functions from the core network to the EN, and therefore, to analyse the impact of these changes on the network.

The purpose of this thesis is to develop a tool that implements a model to analyse the performance of different C-RAN architectures, according to a given deployment constraint. The model will optimise not only the position of the nodes, taking into account the optimal number of nodes and maximum distance between nodes, but also the task offloading process by shifting time sensitive BS functions to the EN. This thesis was done in collaboration with the Portuguese operator NOS.

This thesis consists of five chapters. Chapter 1, the current one, includes an introduction to the study, an overview of the problem and the motivation behind my thesis.

Chapter 2 introduces the 5G NSA architecture and the new radio interface, followed by the general aspect of the network. It begins with a brief overview of Network Virtualisation (NV), introducing the concept of Cloud Network, Software Defined Networks (SDN) and Network Function Virtualisation (NFV), which are the base technology for Network Slicing. It then presents the basic concept of Cloud Radio Access Network (C-RAN), as well as the general aspect of Edge networking and a description of the BS functions. This chapter addresses some examples of the multiple services and applications of 5G mobile network, a brief description of the performance parameters of the network, and concludes with the state of the art, presenting the latest work developments on the subject of this thesis.

Chapter 3 starts with the definition of the model used in this study, along with the model parameters that will be analysed. It presents the network architectures for the model, followed by an explanation of the different implementation layers of the model. Finally, it shows the assessment tests used to validate the model.

Chapter 4 presents the results and analysis obtained. First, it gives a description of the scenarios used. Then, each network output of the model is individually analysed for different input parameters considered in this study, and the results are compared with the reference scenario.

Chapter 5 summarises the main conclusions of the thesis, presenting some suggestions and ideas for future work.

Chapter 2

Fundamental Aspects

This chapter provides an overview of the 5G system and general aspects of the network. Section 2.1 addresses the non-standalone network architecture of 5G and a comparison between LTE and 5G Radio Interface. Section 2.2 introduces cloud network technology, SDN, and NFV, an overview of C-RAN and some basic concepts about edge computing and network slicing are also given. Section 2.3 is dedicated to an analysis of the BS functions and Section 2.4 addresses the services and use cases of 5G. Section 2.5 addresses the performance parameters of the 5G network. Section 2.6 concludes the chapter with the state of the art on the thesis subject.

2.1 5G aspects

2.1.1 Network architecture

The first wave of 5G networks is classified as Non-Standalone (NSA) based on the 3GPP Release 15, which means that 5G networks will be supported by existing 4G infrastructures. Therefore, this subsection provides an overview of LTE's network architecture based on [Alca13], [Silv16] and [Mont16].

In contrast to the Circuit-Switched (CS) model of previous cellular systems, LTE has been designed to support only Packet-Switched (PC) services in order to provide higher data rates and lower latency. The goal was to provide Internet Protocol (IP) connectivity between User Equipment (UE) and the Packet Data Network (PDN) without any anomaly to the end user.

Figure 2.1 shows the overall network architecture of LTE, covering the main components. LTE proposes an all-IP network architecture called Evolved Packet System (EPS) composed of the Evolved Universal Terrestrial Radio Access (E-UTRA), the Radio Access Network (RAN) of the LTE transport network, and the Evolved Packet Core (EPC), the core network segment of the LTE transport network. Finally, LTE EPC is connected to the external networks such as the Internet, IP Multimedia Core Network Subsystem (IMS) and roaming network.

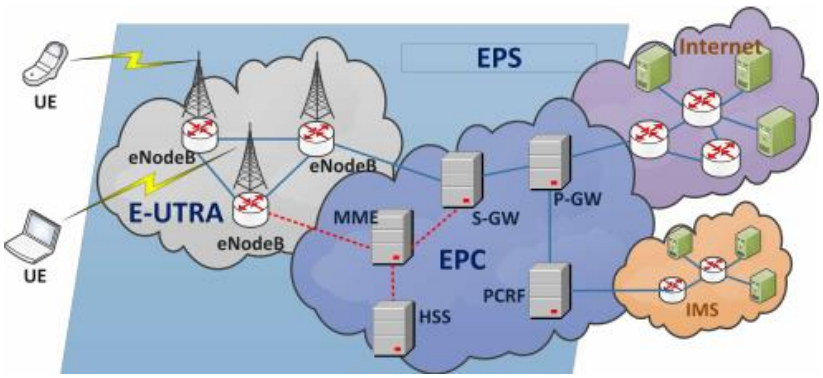


Figure 2.1. Network architecture of LTE (extracted from [LAYG15]).

The E-UTRAN of LTE is the evolution of the Universal Mobile Telecommunications System (UMTS) radio access, and is responsible for the complete radio management in LTE. E-UTRAN consists of Evolved Node B (eNodeB) interconnected with each other by X2, an interface used to support UE mobility and multi-cell functions, as Inter-Cell Interference Coordination (ICIC), defining a new distributed control system. When the UE is connected to the network, the eNodeB is responsible for Radio Resource Management (RRM), which provides radio bearer control, radio admission control, allocation of uplink and downlink to the UE.

When an eNodeB receives a packet from the UE, the eNodeB compresses the IP header and encrypts the data stream. It is also responsible for encrypting all data and sending it to the Serving Gateway (S-GW). Furthermore, eNodeB is also responsible for choosing the Mobility Management Entity (MME) using MME selection function, and is used to achieve QoS demands.

EPC is responsible for the overall control of the UE and the establishment of bearers. The main nodes of the EPC are the Mobility Management Entity (MME), Serving Gateway (S-GW), Packet Data Network Gateway (P-GW), Policy and Charging Rules Function (PCRF), and Home Subscription Service (HSS). These nodes are shown in Figure 2.1 and discussed in more detail below:

- MME - The control entity that is responsible for all control plane operation functions related to signalling from UE to EPC. MME is also responsible for tracking area list management, selection of P-GW/S-GW, and selection of other MME during handovers.
- S-GW - Used to transmit all IP packets, for each UE there is a single S-GW associated with EPS at a given point of time. S-GW is used for inter eNodeB handovers and inter-working with GSM (Global System for Mobile Communications) and UMTS. S-GW also performs administrative functions like data for charging and legal issues.
- P-GW - Terminates the interface between the EPS and the external packet data networks. It is responsible for all IP packet-based operations such as deep packet inspection, UE IP address allocations, transport level packet marking in uplink and downlink, among others. P-GW also contacts the PCEF to assurance QoS demands.
- PCRF- Responsible for policy control decision-making, and controlling the flow-based charging functionalities in the Policy Control Enforcement Function (PCEF) in P-GW. This ensures that data flow is treated accordingly to each traffic profile.
- HSS - A central database that contains user-related and subscription-related information. HSS is responsible for mobility management, call establishment support, user authentication, and access authorisation.

2.1.2 Radio interface

This subsection explains some technologies of the 5G New Radio (NR) interface, comparing them with the LTE existing one. This subsection is based on [ZBAF17], [Eric18A], [Qual16], [JPYK18] and [GAMZ10].

5G NR builds on LTE radio interface, making LTE an integral part of the 5G radio access solution. However, NR has a radio structure better prepared for future technology, with higher spectral efficiency and traffic capacity, and shorter user plane latency.

LTE has been designed to accommodate both paired spectrum for Frequency Division Duplex (FDD) and unpaired one for Time Division Duplex (TDD). In 5G NR, the frame structure supports both full-duplex FDD and TDD, which support transmitted and received data simultaneously and on the same channel, respectively. These full-duplex technologies can double the capacity of wireless networks.

According to [Eric18A], 5G spectrum will be composed of low-, mid-, and high-bands. In general, all the current 3GPP lower bands (600 MHz, 700 MHz, 800 MHz, 850 MHz, and 900 MHz) and mid-bands (1.5 GHz, 1.7 GHz, 1.8 GHz, 1.9 GHz, 2.1 GHz, 2.3 GHz and 2.6 GHz) will be considered in 5G network. Beyond these existing bands, a new band will be created in 600 MHz, providing coverage both in remote areas and buildings. In the mid-band spectrum, new bands will appear in the 3.5 GHz to 6 GHz

range in order to support the terrestrial 5G access network. 5G NR will also support millimetre wave (mm-wave) frequencies, first in the 28 GHz band, and then in the 26 GHz, 37 GHz and 39 GHz ones. The mm-waves will be restricted to direct line-of-sight links, since these bands have a very high free space attenuation and they are easily blocked by buildings. Mm-waves will enable high traffic capacity and enhanced Mobile Broadband (eMBB). However, mm-waves have a much shorter coverage area that can be compensated with a multi-antenna transmission, beamforming and lower frequency transmission.

5G NR will use Orthogonal Frequency-Division Multiple Access (OFDMA) as the multiple access technique but, unlike LTE, which only use a sub-carrier spacing of 15 kHz, NR supports flexible Orthogonal Frequency Division Multiplexing (OFDM) symbols, named filtered-OFDM (f-OFDM). F-OFDM enables flexibility of wave-forms depending on the services demands, the subcarrier spacing is scalable according to 15×2^nkHz, where n is an integer. The choice of n depends on the services requirements (latency, reliability, and throughput), hardware, mobility, and implementation complexity. The wider subcarrier spacing can be used in latency critical services (e.g. vehicle-to-vehicle communication), small coverage areas, and high carrier frequencies. Conversely, narrower subcarrier can be used in large coverage areas, lower carrier frequencies, narrowband devices (e.g. Internet of Things (IoT)), and evolved Multimedia Broadcast Multicast Services (eMBMSs). Figure 2.2 illustrates the scalability of the subcarriers depending on the working frequency.

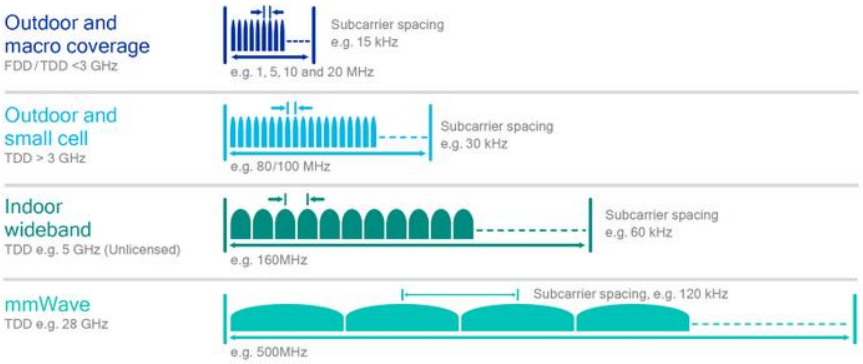


Figure 2.2. 5G NR scales from cellular to mm-wave frequency bands (extracted from [Qual16]).

LTE uses Adaptive Modulation and Coding (AMC) for both Uplink (UL) and Downlink (DL) to improve system throughput by ensuring more reliable transmissions. LTE uses Quadrature Phase Shift Keying (QPSK) and Quadrature Amplitude Modulation (QAM), 16 QAM and 64 QAM.

Turbo coding is used with a different coding rate in order to provide a higher data rate, possible to users according to their specific needs. 5G NR uses the same modulation scheme as LTE, but with the Released 12, it is possible to implement a higher order modulation, up to 256 QAM in some networks.

Multiple-Input Multiple-Output (MIMO) is also used in 4G systems to increase quality performance and data rates. Essentially, MIMO employs multiple antennas on the receiver and transmitter - normally 2x2 or 4x4. The 5G NR interface will support massive MIMO, which can support a very large number of antennas at the base station to serve many independent terminals simultaneously. Massive MIMO will offer excellent spectral efficiency, achieved by spatial multiplexing of many terminals in the same time-

frequency resource that greatly increases the achievable throughput at the cell edge, and superior energy efficiency derivative from the antennas array gain that allows a reduction of radiated power.

In 5G NR, the physical layer needs to support Ultra-Reliable Low-Latency Communication (URLLC) [JPYK18], a wide range of frequency band, and various services categories, so it is important to include a packet structure that minimises the latency and a flexible frame structure that supports different services demands. The URLLC packet has a non-square form, stretched in the frequency axis in order to reduce the transmission latency. Furthermore, the three packet components (pilot, control, and data part) are grouped together in order to reduce the processing latency time. A comparison between 5G and 4G resource block is illustrated in Figure 2.3.

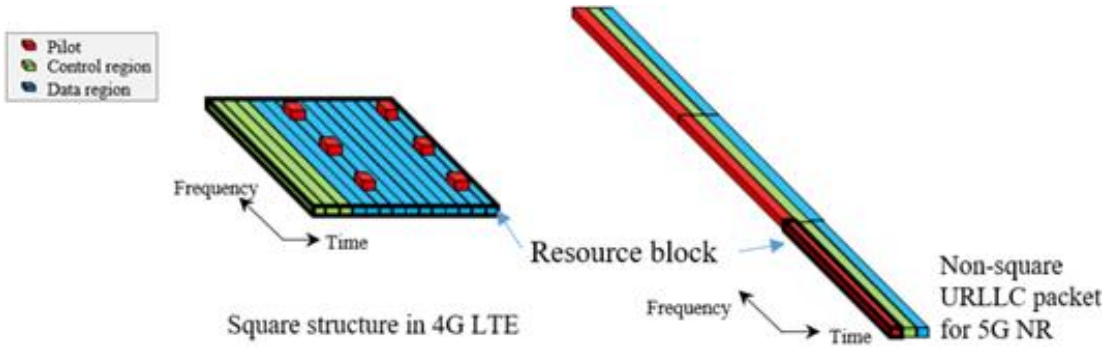


Figure 2.3. Example of Resource Block allocation in LTE and a URLLC 5G NR (extracted from [JPYK18]).

To efficiently support various QoS classes of services, 5G and LTE adopt a hierarchical channel structure. There are three different channel types, each associated with a Service Access Point (SAP) between different layers: logical, transport, and physical channels.

- Logical channels - They are used by the Media Access Control (MAC) to provide services to the Radio Link Control (RLC). In order to support the low-latency requirements from 5G, the NR redesigns MAC and RLC protocols to enable processing data without knowing the amount of data to transmit. The NR RLC does not support in-sequence delivery in order to reduce the delay from the packets streaming. Both 5G and LTE have two categories of logical channels depending on the service they provide: control and traffic channels.
- Transport Channels - A transport channel is basically characterised by how and with what characteristics data is transferred over the radio interface - that is, the channel coding scheme, the modulation scheme, and antenna mapping. The transport channel connects the MAC and the physical layers.
- Physical channels - Each physical channel corresponds to a set of resource elements in the time-frequency grid that carry information from higher layers. The basic entities that make a physical channel are resource elements and resource blocks. A resource element is a single subcarrier over one OFDM symbol, which could typically carry one modulated symbol (or two, with spatial multiplexing). A resource block is a collection of resource elements.

2.2 Network Aspects

2.2.1 Network Slicing and Virtualisation

Cloud networking, Software Defined Network (SDN), and Network Function Virtualisation (NFV) combine various features that create the desired environment for flexible virtualised network functions, which allow the introduction of slicing technology in the network. This subsection gives an overview of each technology and their contribution to the implementation of network slicing, and is based on [Amaz17], [AHGZ16], [RoSh17], [MSGB15], [Eric18B] and [5Gam16].

Cloud networking is the technology to distribute data processing, where scalable resources and capacities are provided to deliver a service to multiple customers through a cloud network environment. Unlike traditional hardware-based solutions, cloud networks enable resource sharing, flexibility, and resource pooling, enabling companies to benefit from economies of scale, while providing levels of centralised control and network visibility.

SDN introduces the control plane on the transport networks and decouples the network control and data planes, enabling the network to become directly programmable. The network infrastructure is abstracted for applications and services. SDN technology is essential to enable real-time software-controlled configuration to a specific service. SDN introduces flexibility within the cloud infrastructure.

NFV provides the ability to process Network Function (NF) in real time at any location within the operator cloud platform. NFV is essential to optimise resources and increase operational efficiency gains. It enables the speed, scalability, and efficiency to support the new 5G business cases.

Each technology is distinct and non-dependent on each other, and the advantages of agility, cost reduction, dynamism, and resource scaling are similar to each of them. Each of these technologies is an abstraction of different resources: the SDN is an abstraction of the network, the NFV an abstraction of the functions, and the Cloud network is an abstraction of computation and storage.

Network slicing is the technology that divides the network into slices, providing virtual networks dedicated to each service or customer, using the same physical infrastructure. The NFV and SDN technologies allow traditional structures to be broken down into customisable elements, and the goal of network slicing is to chain together these elements to provide just the right level of connectivity, where each element can run on the different architectures.

Since performance requirements placed on the network (i.e. data rate, latency, QoS, security, and availability) vary from one service to another, network slicing can be used to balance cost-optimised and performance-optimised views, crucial to profit in each service. In 5G, network slicing can be used to slice a single physical network into multiple virtual networks that can support multiples RANs. For example, a device could use multiple access slices that connect to multiple core networks. Figure 2.4 illustrates device allocation of network slices architecture.

This architecture contains radio access and fixed access slices, Core Network (CN) slices, and a slice pairing function that connects the slices between the access network and the CN. Each CN uses a set

of Network Functions - some of them can be used across multiple slices, while others are used just for a specific slice. The pairing between access slices and CN slices can be static or semi-dynamic in order to achieve the required network function and demands.

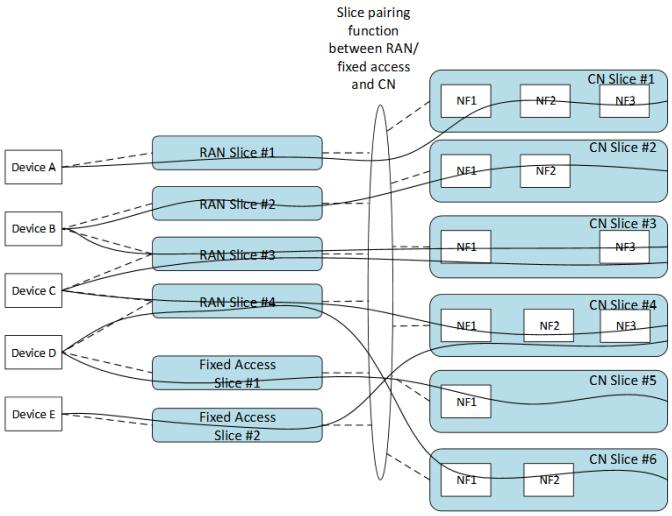


Figure 2.4. Devices connection with the network slices (extracted from [5Gam16]).

To conclude, both SDN and NFV, aggregated with the cloud, provide the foundation for network slicing technology. With these technologies it is possible to efficiently increase the management of network traffic, providing significant OPEX and CAPEX savings, and automated management of increasingly complex network architectures.

2.2.2 Cloud Radio Access Network

This subsection provides an extensive introduction to Cloud Radio Access Network (C-RAN), outlining the fundamental aspects and advantages, based on [CCYS14], [RWNL17] and [WZHW15].

C-RAN is the architecture that can address a number of challenges that operators face while trying to support the growing needs of end-users and is, therefore, a major technological foundation for 5G networks. C-RAN is a network architecture where baseband resources are pooled so that they can be shared between base stations. Figure 2.5 gives an overview of the overall C-RAN architecture.

The base station is separated into a radio unit and a signal processing one. The radio unit is called a Remote Radio Header (RRH), or Remote Radio Unit (RRU). RRH is responsible for power amplification, frequency conversion, and analogue and digital conversion. In C-RAN, the baseband signal processing part is called a Baseband Unit (BBU). This technology performs BS functions like baseband and packet processing. To achieve optimal BBU utilisation in base stations, the BBUs are centralised and shared amongst sites in a virtualised BBU Pool. This technology is prepared to adapt to non-uniform traffic, and new BBUs can be added and upgraded easily, improving scalability and easing network maintenance. A virtualised BBU Pool can be shared by different network operators, allowing them to rent RAN as a cloud service. As BBUs from many sites are co-located in one pool, they can interact with lower delays.

The link between RRH and BBUs is called fronthaul, and the transmission of information is done using

Common Public Radio Interface (CPRI). The fronthaul provides high capacity, low delay, and low jitter for a large number of cells in an efficient way, in terms of both cost and energy. Amongst many wired and wireless technologies, the Optical Transport Network (OTN) is considered to be the best candidate for the 5G fronthaul link due to high capacity.

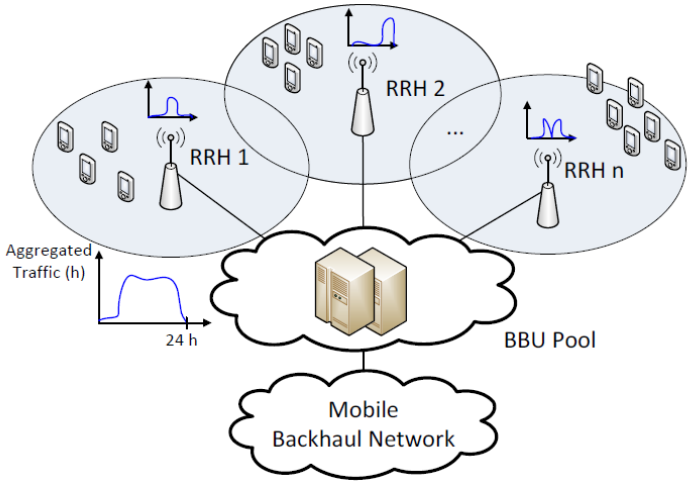


Figure 2.5. C-RAN architecture (adapted from [CCYS14]).

The 5G fronthaul network has a huge capacity required to support targeted data rates and latency. According to [RWNL17], the bandwidth requirement of a CPRI-based 5G fronthaul network to achieve the targeted data rate of 1 Gbps with 8×8 MIMO antennas from 3 sectors is a 147.5 Gbps of optical fibre link. The capacity requirement of CPRI links can be reduced by about 50% with compression techniques, but it still requires a significant capacity by these CPRI links.

To conclude the study of C-RAN networks, the main advantages related to this technology are:

- Adaptability to non-uniform traffic and scalability - Since in C-RAN baseband processing of multiple cells is carried out in the centralised BBU pool, the overall utilisation rate can be improved. The required baseband processing capacity of the pool is expected to be smaller than the sum of capacities of single base stations.
- Energy and cost saving - A centralised BBU pool reduces the power consumption mostly related to air-conditioning and processing equipment. Civil work on remote sites can also be reduced by gathering equipment in a central room.
- The increase in throughput, decrease in delays - The co-location of multiple BBUs in a pool facilitates advanced cooperative techniques, reducing processing needs and communications delays compared to traditional architecture. Since handovers can be done inside the BBU pool instead of between eNodeBs, the time needed to perform handovers is reduced.
- Ease in network upgrades and maintenance - Co-locating BBUs in a BBU Pool enables more frequent Central Process Unit (CPU) updates, making it possible to benefit from the technology improvements in CPU technology.

2.2.3 Edge Networking

This subsection starts with an explanation of the basic aspects of 5G Edge Networking and an introduction to Multi-access Edge Computing (MEC), followed by a presentation of the 5G NR solution of C-RAN. This section is based on [Saty17], [TSMF17], [LiHW18] and [ITUT18].

The heart of a connected network is the cloud. Cloud processes send out information to devices, including all the software to run all applications, and act as a central management point for the whole network. In recent years, new concepts were conceived to optimise cloud computing, which brings processing units closer to the devices as mediators at the physical edge of the network, known as Edge Networking. Edge nodes include many varieties, such as gateways connected to sensors or radio towers. In the 5G context, edge computing is usually discussed as MEC and offers applications and content providers cloud-computing capabilities at the RAN edge, in close proximity to end users. Figure 2.6 illustrates an architecture combining MEC into C-RAN. In this case, MEC is introduced in the path between the end service and the cloud server.

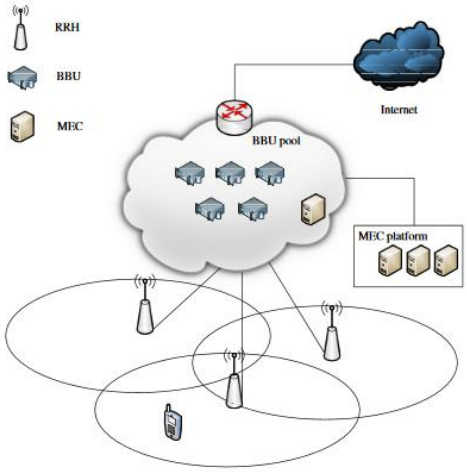


Figure 2.6. C-RAN and MEC system combination architecture (adapted from [LiHW18]).

MEC can be used as a tool to handle offloading tasks by shifting time sensitive BS functions to the Edge Nodes (EN), which in an LTE network are considered to be the BBU pools. The BBU decrypts the packets before computing and decides if the task is addressed on the MEC or is routed for the core cloud. This decision depends on the services demand, and since the path to the cloud is much longer than the path to the EN, the maximum latency and corresponding maximum distance needs to be taken into account, as well as the computing demands of the task, since the computing capability in the cloud would be more powerful than in MEC.

To summarise, the benefits of Edge Computing are:

- Reducing network loads - Network edge nodes connect numerous devices and communicate with the cloud through a single central point, substantially reducing network loads and enhancing data transfer and processing speed.
- Increased security - The edge nodes process data from devices at a physical point closer to users, which allows for a more secure transfer of data and considerably diminishes security

threats. Data coming from the edge node can be further encrypted, with security measures implemented at the edge node level (before the data is sent).

- Decreased latency - By optimising the information shared with the cloud, the network speed is increased and latencies are shortened. Time-sensitive information is processed first, substantially raising user experience.

However, C-RAN traditional implementation has a major problem of capacity demands on the fronthaul links - according to [LCCh18], in a 20 MHz LTE with 2 antennas the bitrate of the link reaches 2.5 Gbps for one connection between an RRH and BBU. Therefore, with the increase of demand of the 5G NR, it is important to reduce the bitrates on the fronthaul link and maintain the many benefits of centralisation from the traditional C-RAN implementation. Instead of the common splitting between the RRH and BBU, in 5G NR the functions from the 3GPP protocol stack will be split into a Distributed Unit (DU) and a Centralised Unit (CU). This implementation provides the possibility for more functions to be processed locally in the DU closer to the user before being transmitted to the CU, where the processing power is higher and can benefit from processing centralisation. The connection between the DU and CU is called Middlehaul (MH). Figure 2.7 illustrates the difference between the 4G C-RAN implementation and 5G NR C-RAN implementation. The functions on the RRH and BBU will be split between the Radio Unit (RU), DU, and CU. Section 2.3 provides an overview of the BS function along with the different splitting option for the implementation of the new 5G C-RAN.

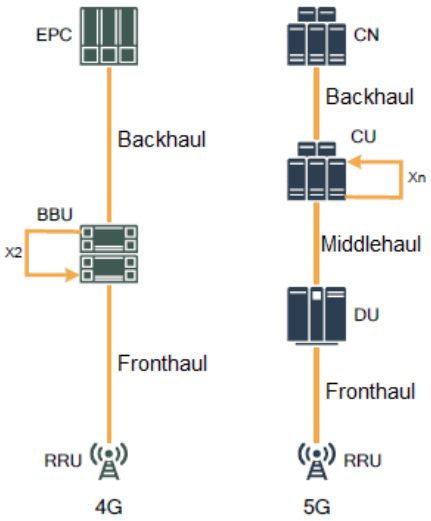


Figure 2.7. 4G vs 5G C-RAN implementation (adapted from [ITUT18]).

2.3 Base Station's Functionalities

This subsection presents the functionalities of the LTE user plane protocol, starting with the downlink functions to the LTE eNodeB and, after, the multiple functional procedures in the uplink. This subsection is based on [Bara17] and [LCCh18].

The first wave of 5G networks will be classified as NSA and supported by existing 4G infrastructures, so it is essential to do an overview of the existing function in the LTE user plane protocol. The multiple functions of the BS are associated with the different sublayers of the user plane protocol stack of LTE composed of the Packet Data Convergence Protocol (PDCP), RLC, MAC, and the Physical (PHY) layer.

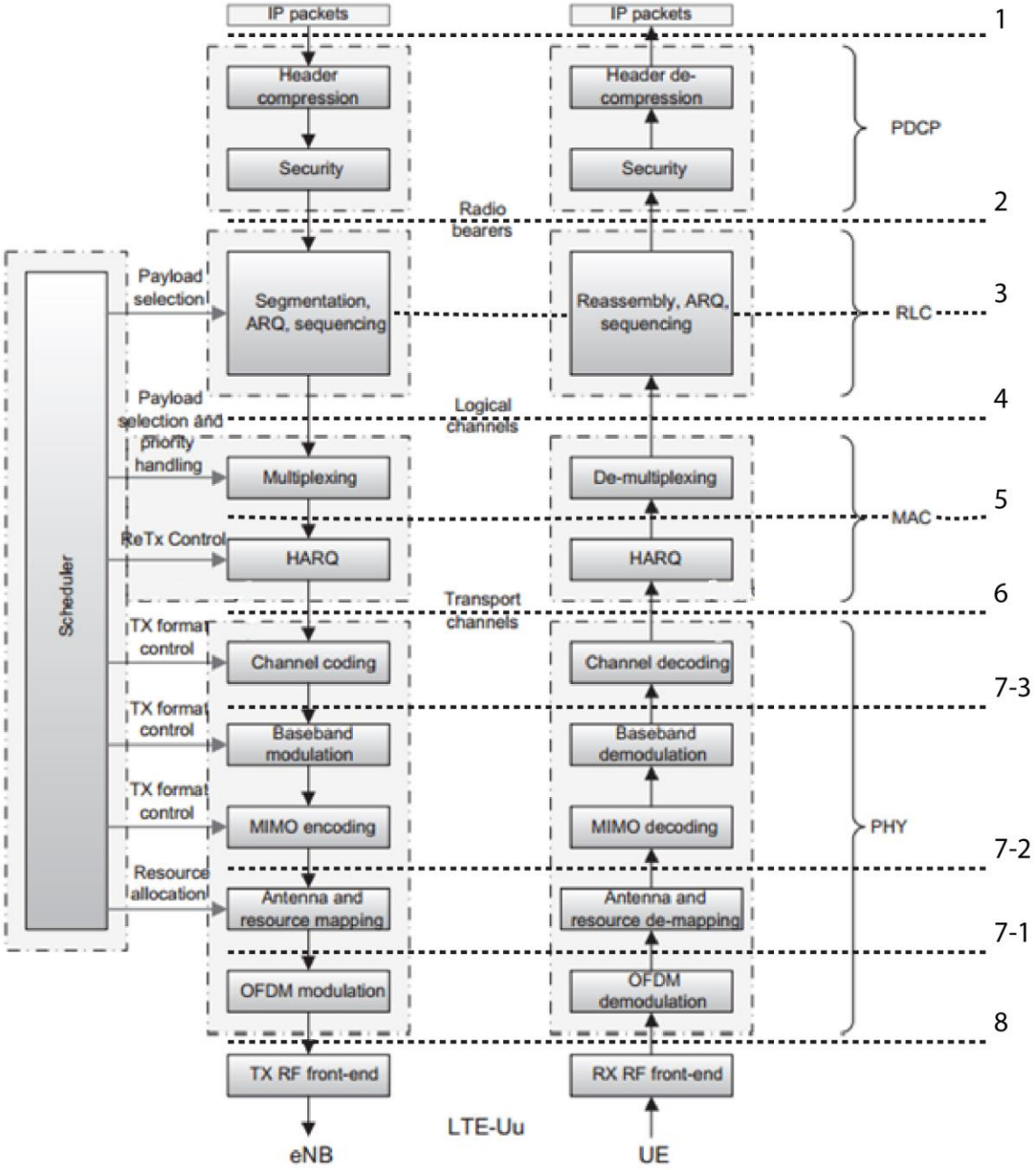


Figure 2.8. Overall downlink and uplink BS functions (extracted from [Alme13]).

An LTE eNodeB in DL has the existing functions:

- Functions in PDCP sublayer
    - IP header compression - Used to reduce the number of bits transmitted over the radio interface.
    - Ciphering and Security - Provides security for both the user plane and control plane by ciphering.

- Function in RLC sublayer
    - Segmentation and concatenation - Minimise protocol overheads and it is responsible for retransmission handling, duplicated detection and delivery to higher layers.
- Function in MAC sublayer
    - Multiplexing - Multiplexes data from several logical channels into one transport channel.
    - Hybrid Automatic Repeat request (HARQ) - Used for retransmission of data in case of error detection.
- Function in PHY layer.
    - Channel coding - Provides error detection, error correction, rate matching and scrambling.
        - Error detection - Using a Cyclic Redundancy Check (CRC) method, the receiver can detect an error. A certain number of check bits are sent with the message for the receiver to determine if the check bits agree with the data.
        - Error correction - It uses a turbo coder for error correcting with different coding rates.
        - Rate matching - This algorithm is capable of producing different coding rates.
        - Scrambling code - Allows the eNodeB to separate signals coming simultaneously from many different UEs, and also the ability for UE to separate signals coming simultaneously from different eNodeBs.
    - Baseband modulation - Used for mapping the input bits with a modulation scheme.
    - MIMO encoding - Used for mapping the input symbols from the modulation phase to be transmitted over multiple antennas.
    - Antenna and resource mapping - The previously modulated symbols are mapped into subcarriers in OFDMA symbols.
    - OFDM modulation - The OFDM Modulator converts the symbols from all the subcarriers into time domain using Inverse Fast Fourier Transform (IFFT), and in the end, the Cyclic Prefix (CP) is inserted and the data is transmitted.

Meanwhile, these are the existing functions of the UL:

- Functions in PHY layer
    - OFDM demodulation - First the CP is removed, and then the received subcarrier values are recovered using Fast Fourier Transform (FFT) per symbol.
    - Antenna and resource demapping - Used to demap Resource Elements (REs) to the physical channels.
    - MIMO decoding - Separates and detects the received symbols from MIMO antennas.
    - Baseband demodulation - Aims to recover the information from the modulated carrier signal.
    - Channel decoding - Provides descrambling, decoding, rate dematching, CRC check and HARQ.
- Function in MAC sublayer
    - Demultiplexing - It aims to demultiplex the data from the different logical channels for

one transport channel.

- Function in RLC sublayer.
    - Segmentation, Automatic Repeat request (ARQ) and concatenation - Used to reverse the segmentation, concatenation and ARQ in the BS side.
- Function in PDCP sublayer.
    - Security - Performs the reverse of the security operations and is responsible for duplicate detection, user plane, and integrity protection for the control plane.
    - IP header decompression - Used to perform the reverse procedure described in the BS.

The splitting decision in the implementation of the 5G NR, a BS function, is an important factor in order to reduce bitrate demand on the fronthaul, including more functions locally at the RU, and processing the signal before it is transmitted, reducing the fronthaul capacity, but also reducing the centralisation gains of the classical C-RAN implementation. Lower splitting option (Option 8) provides more centralisation gains but a very heavy link interface, so using a higher splitting option will reduce the bit rate required on the link. This being said, it is important to analyse the pros and cons of the different splitting options based on [LCCh18].

- Option 8 - This option splits the RF front-end to the RU when all other BS functions are processed in the CU, and it is the solution used in LTE C-RAN implementation, which leads to cheap RU nodes, high levels of centralisation, and constant bitrate, but with the disadvantages of leading to a very heavy fronthaul link scaling with the number of antennas. This option can be used in scenarios with high capacity fibres and a real time communication requirement.
- Option 7.1 - This option sends the FFT to the RU. The data in the interface is represented by a constant flux of subcarriers which, by removing the cycle prefix and transforming the signal to the frequency-domain, can reduce the bitrate. The use cases for this option are similar to option 8: a high capacity fibre and real time communication, but with less extreme demands.
- Option 7.2 - In this option, the resource element mapping is included in the RU and the data is transported on subcarrier symbols. The data symbols are only exchanged when data is available, so the transport capacity is reduced and scaled with the cell load.
- Option 7.3 - In this option, the signal is modulated in the RU, reducing significantly the bitrate while keeping the close relation between the FEC and the MAC layer. This scenario requires a protocol in the RU for the Modulation and MIMO that demands a link distance less than 10 km.
- Option 6 - This split divides the data link layer from the physical layer, and the data is transported over a resource block, which reduces bandwidth requirements. The HARQ is processed in the RU, so the network latency is directly proportional to the applications latency. This option can be used for delay tolerant application and when centralised scheduling is wanted in the network.
- Option 5 - In this case, the RU handles time critical processing, reducing the latency requirements, so the distance of the link can be longer and the bitrate requirements smaller.
- Option 4 - This option splits the RLC layer and the MAC layer - the network transport RLC Protocol Data Units (PDUs) in the DL and MAC Service Data Units (DDUs) in the UL. This option has no benefits for LTE and it is not applicable for 5G shorter subframe sizes.

- Option 3 - This option splits the RLC layer, just the PDCP and asynchronous RLC is processed in the CU. This option allows multiple MAC entities to be associated with a common RLC entity. This architecture provides robust adaptability to non-ideal transmission conditions.
- Option 2 - Only the PDCP and the Radio Resource Control (RRC) are processed in the CU. This split uses an already standardised interface called F1. The traffic is divided into multiple flows directly to different access nodes, so a re-sequencing buffer in nodes is required. This option provides higher security and resiliency and can be used for low link bitrate like wireless.
- Option 1 - In this case, all User Plane (UP) functions are allocated in the RU, close to the user. This scenario has lower bitrate requirements on the link, but has higher RU complexity. This option can be used for few cells aggregated in the CU-pool and applications like caching.

2.4 Services and Applications

This section first presents an overview of the most important 5G services and applications, then lists the QoS Class Identifier (QCI) characteristics for 4G and new 5G services, based on [EFSZ16], [MMBT15], [IEEE17], [Bara17] and [Corr18].

The expected launch of 5G services in the next few years will deliver much faster mobile data connections. For consumers, that means faster web browsing, quicker download content and stream high-quality video without buffering. But 5G will not only bring increased speed, but will also enable all new services and new business models. 5G can be divided into three main services type:

- Enhanced Mobile BroadBand (eMBB) or extreme Mobile BroadBand (xMBB) - Require extremely high data rates, low-latency and reliable broadband access over a large coverage area with high user density. The network needs to support more than 1 Gbps/user.
- Massive Machine-Type Communications (mMTC) - This requires wireless connectivity for millions of devices worldwide. In this type of service, scalable connectivity is essential for an increasing number of devices, wide coverage area, deep indoor penetration, high energy efficiency, and low cost. This service requires a very large number of device coverage, low data rate, and high energy efficiency (in some cases more than 10 years of battery lifetime).
- Ultra-reliable Low Latency Communications (URLLC) - This service uses ultra-reliable low-latency and resilient communication links primary between machines. This service required latency lower than 10 ms, reliability of 99.999%, and more than 100 Mbps/user (e.g. Intelligent Transportation Systems (ITS).

To measure the overall performance of each service from the user viewpoint, it is essential to quantify QoS characteristics and functionalities for QCI. The tables on Annex B summarise this information.

One can divide QCI into two resource types: Guaranteed Bit Rate (GBR) bearers, used for real-time services which guarantee minimum bit rate for the services, and the Non-GBR bearers, which do not offer minimum bitrate guarantee, so they are suitable for non-real-time services. Each bearer is

characterised by the Allocation and Retention Priority (ARP), which indicates the order of priority of a certain service, the Packet Delay Budget (PDB), which is associated with latency constraints of the application and is directly proportional to the value of the priority level, and the maximum Packet Error Loss Rate (PELR), which is related to the reliability of the service.

The implementation of different splitting options from the BS functions through the C-RAN nodes depends on the 5G service type chosen. The latency and data rate values for the most popular use cases of a 5G network can be seen in Table 2.1, the latency specification corresponds to an End to End (E2E) latency requirement from the user viewpoint. The use cases are divided into three main 5G services types.

Table 2.1. Use cases specifications (adapted from [PRRG18]).

Service type	Use Case	Latency [ms]	Data rate [Mbps]	Remarks
URLLC	Factory Automation	0.25	1	• Small data rates for motion and remote control • An application like Machine tools operations require low latency (0.25 ms)
URLLC	Telepresence	1	100	• Remote control with require 1 ms latency • Synchronous visual-haptic feedback requires 100 Mbps
URLLC	Health Care	1	100	• Tele-diagnosis, tele-surgery and tele-rehabilitation may require latency of 1 ms
URLLC	ITS	10-100	10 to700	• Road safety requires 10 ms latency • An application like virtual mirrors require 700 Mbps
eMBB	Virtual Reality	1	1000	• Hight resolution 360° VR
eMBB	Real-time Gaming	1	1000	• High resolution and high performance for immersive entertainment and interaction
eMBB	Education and Culture	5	1000	• Human-machine interface may require latency of 5 ms • High resolution 360° VR requires a data rate of 1 Gbps
mMTC	Smart Grid	1-20	0.01 to 1.5	• Dynamic activation and deactivation in smart grid require 1 ms of latency • Wide area situational awareness requires 1.5 Mbps

Regarding QoS, one can divide the type of services into four different classes based on [3GPP02].

- Conversation Class - This class is characterised by real-time conversation services between end-users (e.g. telephone speech, voice over IP, and video conferencing). The required characteristics of low latency for QoS are given by human perception which makes the acceptable transfer delay very strict. The fundamental characteristics for QoS are to preserve time relation between information entities of the stream, and low delay conversation patterns.

- Streaming Class - Real-time data flow (e.g. video and audio) is characterised by a low latency, but it can support a small delay variation because it is not limited by human sensory perception. The fundamental characteristic of QoS is to preserve time relation between information entities of the stream.
- Interactive Class - The interaction between user and remote equipment (e.g. web browsing, database retrieval, and server access). This class is characterised by the request-response pattern of end-users. The fundamental characteristics for QoS are the request-response pattern and preserve payload content.
- Background Class - The end-user is usually a computer, which sends or receives data in the background (e.g. delivery E-mails, download of databases, and reception of measurement records). In this case, traffic is not time sensitive and should be transparently transferred with a low bit error rate. The fundamental characteristics for QoS are the destination, which is not expecting the data within a certain time, and preserve payload content.

2.5 Performance Parameters

This section enumerates some of the 5G performance parameter demanded by the different services of the network, Table 2.2, based on [EFSZ16], [EmFS18] and [Silv16].

Table 2.2. Parameters in 5G identified by ITU-R IMT 2020 (adapted from [EFSZ16]).

Parameters	Values
Area traffic capacity	10 Mbps/ m^2
Peak data rate	20 Gbps
User experienced data rate	100 Mbps
Latency	1 ms
Spectrum Efficiency	2/3/5 times greater than 4G
Connection density	$10^6/\mathrm{km}^2$
Network energy efficiency	100 times greater than 4G
Mobility	500 km/h

Each application of 5G has key parameters that the network needs to achieve based on ITU-R IMT-2020 (International Mobile Telecommunications 2020 in the International Telecommunication Union Radiocommunication Sector). These performance parameters are essential to provide users with the expected Quality of Experience (QoE):

- Link capacity – The mobile networks traffic capacity is one of the key parameters demanded by the user. In order to support the data traffic exponential growth, 5G is expected to support a traffic capacity of 10 Mbps/m^2, a peak throughput of 20 Gbps, and a user experienced data rate from 100 Mbps in an urban/suburban environment up to 1 Gbps in a hotspot.

- Processing power – In order to create a balanced load in the network nodes, the processing power needs to be quantified by measuring the computation frequency (CPU cycles per second) on Giga *operations per second* (GOPS). The goal is to offload the cloud core to the edge cloud using MEC technology, using a balancing algorithm among nodes.
- Latency – Latency is one of the most important key parameters in 5G. In order to support ultra-low latency services, it is expected that latency achieves less than 1 ms using edge computing technology. Latency can be divided into 3 parts: the Link Latency describes the time of DL and UP to transmit the signal, including the travel time on the C-RAN and the time for a packet to be routed through the backhaul network (MEC is not affected by backhaul latency because data do not interact with the core network), Processing delay, which represents the time it takes to process data on the nodes depending on the functions assigned to the node and the processing power of the network, and finally the Queuing delay, which depends on the traffic arriving at the node and the throughput supported by the node.
- Spectrum efficiency – In order to accommodate the high capacity throughput in the minimal channel bandwidth possible, the network needs to improve the spectrum efficiency. 5G spectrum efficiency will increase 2/3/5 times compared to 4G networks, and this improvement is achieved, in the most part, by massive MIMO technology.
- Connection density – 5G not only supports person-to-person communication, but also person-to-machine and machine-to-machine communications, so, as expected, the network will need to support an area with 100 times more devices than LTE, reaching 10^6per km^2.
- Network energy efficiency – Since multiple devices in the new 5G application (e.g., smart cities and autonomous industries) are self-powered using energy harvesting and need to have an autonomy of months, energy efficiency is a key parameter especially in the massive machine type communication, so the energy efficiency of 5G is expected to be 100 times greater than the one of 4G. Energy reduction also plays an important part in reducing OPEX in a cell site.
- Mobility – An important challenge of URLLC is the high mobility demands, so the network is expected to support velocities up to 500 km/h, especially in High Mobility Wireless Communication (HMWC) systems like High-speed Railway (HSR) ones.

2.6 State of the Art

In this subsection, an overview is given on the relevant research regarding the subject of the thesis, the main focus of which is the implementation of MEC at the cell sites in order to ensure a flexible network slicing resource on a 5G network.

[HKPS18] presents a novel edge computing architecture that customises network resources at the edge cloud, the closest to users as possible in order to minimise network signalling overhead on IoT services. In this case, a novel architecture of MEC enables slicing technologies to increase the level of automation, flexibility, and programmability for IoT application. Depending on services requirements,

dedicated slices are created for different IoT services, and different virtualised network functions needs to be placed on different parts of the slice (Edge, Core, and Service Cloud) in order to support the multiple 5G application requirements.

Flexibility is an important aspect of a 5G network. In [SKRA18], the author discusses the implementation of a MEC-enabled 5G architecture that supports the flexibility of the network and Virtual Network Functions (VNFs). The MEC architecture is divided into two tiers of computation capabilities - the core tier cloud, that has a lot of computing resources and can host application VNFs and network VNFs, and the edge tier cloud, which has limited resources allocated to the MEC entity and should be placed closer to the user for specific services requirements. The VNFs are divided into Real-Time Applications (RTAs), Non-Real-Time Applications (NRTAs), and Hybrid Applications (HAs). A real implementation of a MEC with VNFs in an LTE network is used to demonstrate the potential of the architecture. For example, depending on the available resources, the network decides to redirect the needed resources of an NRTA from the edge to the core in order to release resources to possible RTAs.

[NHHS18] proposes a congestion control mechanism for reducing RAN congestion in a MEC environment. The mechanism aims to make a real-time decision for buffering traffic in order to improve QoS based on SDN technology. The basic idea of the mechanism is to intentionally delay Delay Tolerant (DT) services, buffering these services data in an intermediate cloud server. This process uses the Congestion Control Engine (CCE), which evaluates the congestion of the RAN and makes the decision of offloading traffic. The Congestion Control Mechanism can be divided into three steps. Firstly, the packets are inspected in order to identify DT content, and are assigned to a deadline constraint. Secondly, the CCE monitors RAN and, in case of network congestion, the SDN redirects the DT content to be stored in the MEC. Finally, the content is transmitted from storage when the CCE identifies that the deadline of a DT content is approaching.

QoE is an important part of the 5G network, so it is essential to study a resource management mechanism that optimises the QoE of 5G services. [LHWW18] presents the Min-Fit algorithm that provides a flexible slicing solution to choose a server that has the maximum resources available to satisfy users' demands. This algorithm is used to minimise the delay gap tolerance, which is defined as the difference between the delay tolerance and the actual time delay to achieve the expected QoE. There are three experimental cases presented in this article in which the Min-Fit algorithm was tested and compared with three other algorithms in the literature. The first case scenario was the increasing number of users, the second the increasing of the number of edge servers, and the last the increasing number of core servers.

In order to achieve the 5G requirements of latency and data traffic, [LLHC18] divides the MEC architecture into three tiers: core, edge, and devices. The authors use a two-phase interactive optimisation method to optimise capacity and traffic allocation in a MEC-based architecture. The paper uses a latency percentage constraint metric that calculates the percentage of the latency that satisfies the latency constraint threshold. The latency percentage constraint is calculated according to three different traffic types, depending on the services and application demands, the DC-Type Traffic, served by the device and the core, the EC-Type Traffic, which is served by the edge and the core, and the EE-

Type Traffic, the edge to edge traffic. First, the algorithm adjusts the traffic distribution based on the currently allocated capacity to satisfy the latency percentage constraint, then, the capacity allocation is adjusted based on current allocated traffic in order to minimise the total capacity.

MEC is essential to offload tasks from the centralised BBU to an edge cloud, which can reduce offloading latency. [LMWX17] focused on the matching problem in the hybrid offloading architecture of C-RANs with MEC. The authors designed an efficient offloading control to minimise the refusal ratio of offloading request (the portion of offloading tasks that are not able to meet their demands). The authors use a multi-stage duplex matching, dividing the cross-layer optimisation problem into three stages: matching between RRH and UEs, matching between BBUs and UEs, and matching between Mobile Clones (MC) and UEs. The paper compares the results of using optimal baseline (optimal solution using brute-force searching), Linear programming relaxation (solution without using the multi-stage duplex matching proposed by the authors), and the solution using Multi-stage Duplex matching.

[Silv16] studies the pros and cons of different solution designs for the C-RAN fronthaul on an LTE network. The study focuses on analysing the connection between the RRHs and BBU Pools based on traffic profile, positioning and delay characteristics. The study uses a model divided into three layers in order to analyse the parameters of the network. The first layer is the physical layer used to compute the maximum distance between RRHs and BBUs, which is directly related to the maximum latency of the fronthaul link. The second layer, or technical one, aims to identify the best connection between RRHs and BBU Pools, taking the demands of the different networks into account. The third is the cost layer and deals with OPEX and CAPEX of the network.

[Mont16] addressed the implementation of C-RAN in small cells. The goal was to study the assignment of RRHs to BBU pools using different algorithms in order to study the different performance parameters of C-RAN. The author uses a proliferation algorithm in order to forecast the growth of RRHs and traffic demands in the future, introducing a scale factor to the architecture. To achieve the best results for each traffic profile, this thesis studies multiple algorithms - the Minimise Delay Algorithm, the Load Balancing Algorithm, the Minimise Number of Pools Algorithm, and the Maximise Multiplexing Gain Algorithm.

[LiHW18] presents the Two-Thresholds Forwarding Policy (TTFP) algorithm, which is used to control data that go either to the MEC platform or to the cloud server. This study considers two types of applications: Delay Tolerant (DT) and the Delay Sensitive (DS). The algorithm evaluates the state of the system CPU utilisation for a specific input of the system busy threshold and the traffic intensity threshold of each application, and then decides the traffic routing path. Simulations results show that the TTFP algorithm is essential to control the DT application arrival on the MEC BBU data computing part to avoid CPU utilisation overflow, in order to maintain the waiting time of DS application as low as possible.

Chapter 3

Model and Simulator Description

This chapter provides a description of the model used in this thesis, providing an overview of the metrics used by the model and a detailed explanation of the model implementation. At the end of the chapter, the model assessment is presented.

3.1 Model Overview

The purpose of this thesis is to optimise the BS functions splitting among the RU, DU, CU, CN, and MEC for the different use cases considered in this study, concerning the performance parameters assigned to network demands. Figure 3.1 represents an overview of the model under study considering the relation between Input and Output parameters.

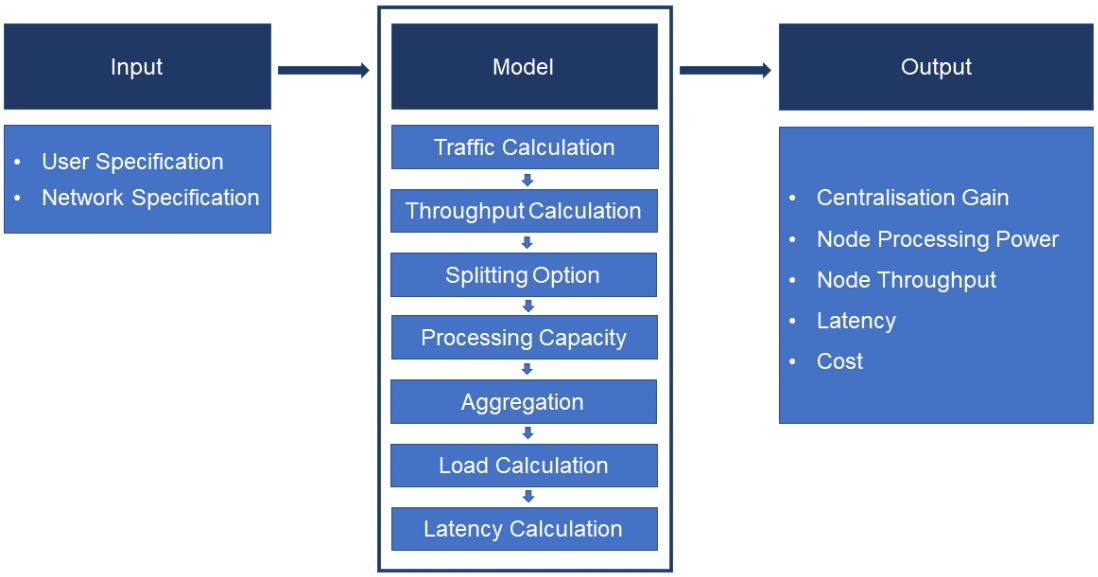


Figure 3.1. Model overview.

The inputs are divided into two classes. First, presented in Table 3.1, User Specification, which addresses the specific parameters of the use cases, and the required parameters necessary to establish the number of users in each BS. Second, presented in Table 3.2, Network Specification, which takes the input specification of the architecture used by the model into account, the information on link distances and the different cell parameters.

Table 3.1. Input User specification.

Use case	• Latency – Maximum E2E latency admissible on the use case. • Data rate – Required throughput per user to support the use case. • Service duration – Average usage duration on a real time use case. • File size – Average file size on a non-real time use case. • Service mix - Percentage of the total users on the cell that are assigned to each use case.
Usage ratio	Percentage of the usage ratio of the network.
Penetration ratio	Percentage of the penetration ratio of the network.
Service user's multiplier	Parameter used to adjust the number of users in a specific 5G service. (i.e. eMBB, URLLC, mMTC)

The first step consists of the computation of the traffic and the throughput arriving at the RU node, which represents the BS at the cell site. The calculation of these parameters is calculated based on the number of users assigned to it and use cases specifications. The next step is the splitting options of the BS network functions throughout the RU, DU, CU, and MEC in order to analyse the impact of different

architectures of the network on the different use cases considered in this study. The splitting options of the nodes are based on the FH (i.e. Option 8, 7.1, 7.2, 7.3, 6), MH (i.e. Option 2), and BH (Option 1). To choose a splitting option, the required processing capacity in the nodes is computed, depending on the assigned computation requirement of each BS function. The aggregation process of the nodes starts with the ones closest to users. First, the available RUs are aggregated onto the DUs or the CUs, if the DU connection is not possible. Then, the DUs are aggregated onto the CUs depending on the distance between the nodes, and the same is done for the connection between CU and the Core, or, if necessary, the MEC. Finally, the model computes the latency on the different parts of the network and the total E2E latency depending on the distance between nodes, the processing delay from the assigned BS function that is called GOPS delay, and the Queuing delay from the traffic arriving at the node. The two parameters of the processing delay (i.e. GOPS and Queuing delays) are computed based on the load of the node that depends on the assign processing power due to data throughput and the BS functions. The required processing capacity is defined so the load of the nodes does not exceed a certain threshold. The purpose of the model is to evaluate the different network architectures performance, and the behaviour of these architectures with each type of service.

Table 3.2. Input Network specification.

Network architecture	Deployment scenario used by the simulator, depending on the number and position of the nodes.
FH splitting Option	FH splitting Option – Splitting option of the FH, used to define the BS functions assigned to the RU node.
Maximum Link distance	Specification of the maximum link distance, considering the FH, MH, and BH links.
RU specification	• Number of RU – The number of RU nodes on the network is equal to the number of BSs in a traditional 4G network. • Location – Location of the BS. • Processing capacity - Specification of the maximum capacity allowed in the node. The Processing power capacity measured in GOPS or Throughput capacity measured Mbps.
DU specification	• Percentage of RU nodes converted to DU nodes – Since the DU node do not have an assign location, this parameter is used to define the number of DU nodes on the network. • Processing capacity.
CU specification	• Number of CU - The number of CU nodes on the network is equal to the number of BBU pools on a 4G C-RAN architecture. • Location – Location of the BBU pools. • Processing capacity
CN specification	• Number of CN - The number of Core nodes on the network. • Location – Location of the Core node. • Processing capacity.
MEC specification	• Percentage of CU nodes converted to MEC nodes • Processing capacity.
Cell specification	• 4G Traffic – Specification of the 4G traffic profile. • Operating bandwidth – Bandwidth of the RU nodes, depending on the node density of the network.

For each considered architecture, the model evaluates the network performance considering the following output parameters:

- Centralisation Gain - Achieved centralisation gains for each architecture implementation in order to evaluate the performance of the network for different levels of centralisation.
- Processing Capacity - Processing Capacity required on the nodes depending on the assign capacity demands of each BS function measure in GOPS
- Node throughput - Input and output throughputs in the nodes, which are directly related to the data rates created by the user in the nodes and the signal compression level.
- Latency - Total latency of the network depending on the processing delays and link latency. The processing delay has two different components: the queuing delay from the throughput of the node and the processing one from the assigned functions on the node.
- Cost - The cost of the C-RAN implementation for each architecture proposed by the model.

3.2 Architecture scenarios

C-RAN implementations in a 5G network have multiple scenarios, depending on the position of RUs, CUs, and DUs. Operators may use different deployments scenarios on the network to address the different applications of 5G, so it is important to address all possible solutions.

The different implementation approaches correspond to a different mapping of BS functions in the nodes. Figure 3.2 illustrates the scenarios for the splitting of BS functions for the different C-RAN implementations, based on [ITUT18]:

- RU-DU-CU (Independent RU, DU, and CU locations) - In this scenario, the distance between RU and DU can go up to 10 km, while the distance between DU and CU can range from 20 km to 40 km.
- RU-DU+CU (Independent RU and Co-located CU and DU) - In this scenario, there is no middlehaul and, in consequence, the CPRI interface between the nodes is heavier.
- RU+DU-CU (Independent CU and Co-located RU and DU) - In this case, the distance between RU and the DU is very small (i.e. in the same building) and, in this case, there is no fronthaul.
- RU+DU+CU (Collocated RU, DU, and CU) - In this scenario, the network only has backhaul, and the processing on the network is all done in the RU node.

In order to support low latency communication, a MEC implementation scenario to reduce the core transmission and processing delay is considered. In this case, the information is routed to the nearest MEC node to be processed instead of going to the CN, which can be hundreds of kilometres away. The information stays on the edge of the network being process on the MEC, closer to the user to support the very low latency requirements of several 5G applications. Figure 3.3 illustrates the general architecture of a 5G network considering all the independent nodes architecture with the implementation of MEC nodes.

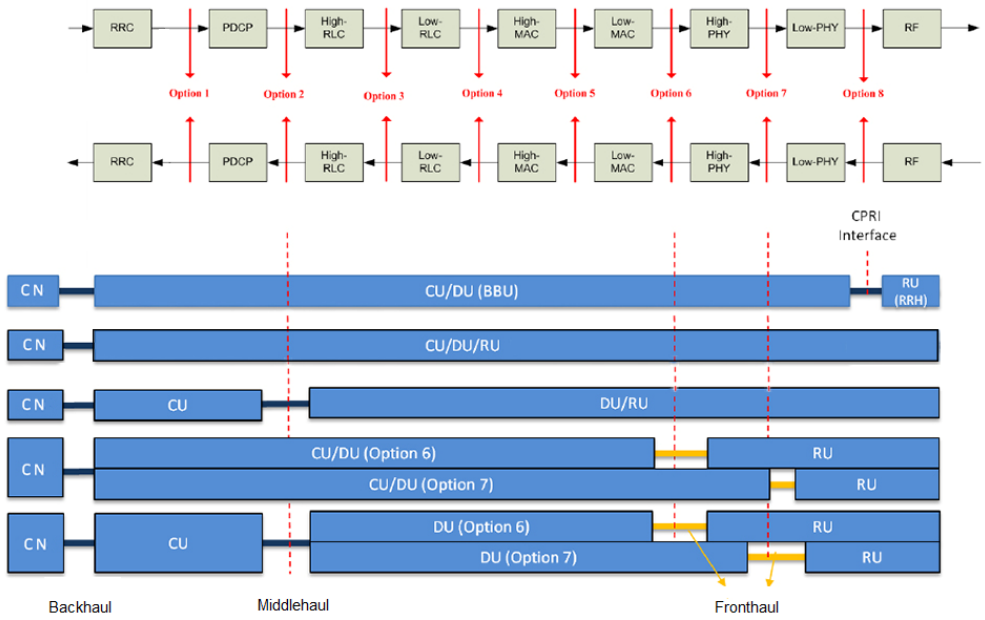


Figure 3.2. Mapping of CU, DU and RU functions according to the split points (adapted from [ITUT18]).

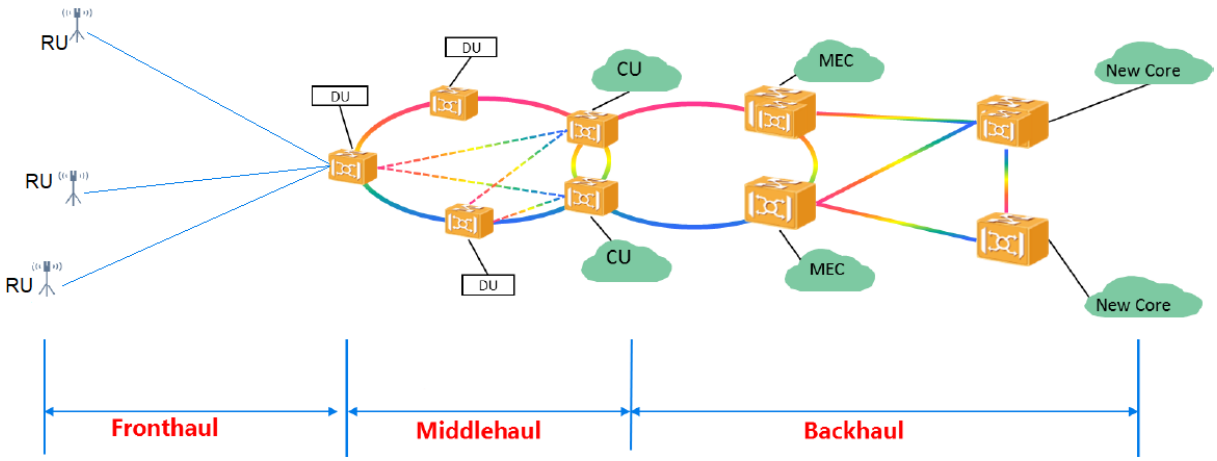


Figure 3.3. General network architecture (adapted from [ITUT18]).

3.3 Traffic Computation

The network traffic demand on each node is one of the input parameters necessary for the computation of the link capacity on the links and the load of the different nodes that are directly related to the queuing delay of the nodes.

The traffic assigned to each RU node depends, firstly, on the number of users assigned to each RU node and, secondly, on the different services characteristics. In this study, one addresses the main network services already supported by the network and the new 5G use cases that will be supported by the 5G network. The reference characteristics of the services are described in Table 3.3:

Table 3.3. Reference services characteristics (adapted from [Rouz19], [Mart17] and [Khat16]).

Service name	Service Class	Duration [min]	File size [MB]	Data rate [Mbps]	Latency [ms]	Priority
Voice	Conversational	2.0	-	0.032	100	3
Video conference	Conversational	1.5	-	2	150	5
Video streaming	Streaming	30	-	5.12	300	6
Music streaming	Streaming	3.5	-	0.128	300	7
Web browsing	Interactive	-	3.0	0.5	300	9
Social networking	Interactive	-	30	2	300	8
File sharing	Interactive	-	5.0	1	300	10
Email	Background	-	0.5	0.512	300	12
Virtual reality	Streaming	10	-	1000	1	4
Realtime gaming	Streaming	30	-	1000	1	4
Smart Meters	Background	-	0.1	0.1	300	11
Factory automation	Conversational	-	0.18	1	0.25	2
Road safety ITS	Conversational	-	0.36	10	10	2
Remote surgery	Conversational	-	9.0	100	1	1

The calculation of the number of users per cell is based on the traffic profile of the 4G network presented in [Silva16] and the traffic mix from the different use cases. First, one computes the maximum number of users $N_{U,máx}$, based on the total population in the area of the cell, the penetration ratio, and the usage ratio,

$$N_{U,máx} = N_{pop} p_{pen} p_u \tag{3.1}$$

where:

- N_{pop} – Population in the area.
- p_{pen} – Penetration ratio.
- p_u – Usage ratio.

Next, the population is distributed by each RU depending on the traffic load:

$$N_{U_{RU}} = \frac{T_{RU,4G[\mathrm{GB/h}]}}{T_{máx,4G[\mathrm{GB/h}]}} N_{U,máx} \tag{3.2}$$

where:

- $N_{U_{RU}}$ - Number of users on the RU.
- $T_{RU,4G}$ - 4G traffic on the RU.
- $T_{máx,4G}$ - Maximum 4G traffic on the network.
- $N_{U,máx}$ - Maximum number of users in the network.

The number of users for each service $N_{U_{RU,Serv}}$ is express in (3.3) taking the number of users in the site and the percentage of users using each application into account, being considered a multiplier factor in the users to simulate the increase of connected devices in the network, from the new 5G use cases for future years:

$$N_{U_{RU,Serv}} = N_{U_{RU,Serv}} S_{mix[\%]} M_{Serv} \tag{3.3}$$

where:

- $N_{U_{RU,Serv}}$ - Number of users in the RU per service.
- S_{mix} - Service mix.
- M_{Serv} – Number of users for each service multiplier.

The throughput in the RU node is given by:

$$R_{RU,Serv[\text{Mbps}]} = N_{U_{RU,Serv}} R_{Serv,ref[\text{Mbps}]} \tag{3.4}$$

where:

- $R_{RU,Serv}$ - Throughput in the RU for a specific service.
- $R_{Serv,ref}$ - Reference throughput for a specific service.

The throughput in the node is an important parameter to compute the queuing delay in the nodes (which is addressed in Subsection 3.4.1).

The traffic in the RU is measured in [GB/h], being computed in (3.5) and (3.6). If the service is a real time one, the traffic is based on the service duration, but if it is a non-real time one, it is based on the data size created per hour by each user.

$$T_{RU,Serv[\text{GB/h}]} = N_{U_{RU,Serv}} \tau_{[\text{s}]} R_{Serv,ref[\text{Mbps}]} \frac{1}{8} 10^{-3} \tag{3.5}$$

$$T_{RU,Serv[\text{GB/h}]} = N_{U_{RU,Serv}} S_{[\text{MB}]} 10^{-3} \tag{3.6}$$

where:

- $T_{RU,Serv}$ - Traffic in the RU for a specific service.
- τ - Mean service duration per hour.
- S - Mean file size per hour.

The traffic in the node is an important parameter to compute the traffic in the different links of the network (which is addressed in Subsection 3.4.3).

3.4 Output Parameters

3.4.1 Latency

Based on Table 3.3, the latency critical services in 5G can require an E2E latency from 1 ms to 300 ms. The E2E latency is based on the delay of packet transmission through the network. Two scenarios are considered: one without the implementation of MEC that takes C-RAN, Core backhaul, core network, and external data centre delays into account, whose delay contribution to the network is presented in (3.7), and another, with the implementation of MEC. In this second case, there are two possibilities: the first considers that information does not go to the CN and takes just the C-RAN, MEC backhaul, and the MEC processing delays into account, whose delay contribution in the network is presented in (3.8); in the second, traffic can also be routed to the core if delay network demands allow extra network latency:

$$\delta_{E2E[\mathrm{ms}]} = \delta_{C-RAN[\mathrm{ms}]} + 2\delta_{BH,C[\mathrm{ms}]} + \delta_{Cor[\mathrm{ms}]} + \delta_{Tran[\mathrm{ms}]} + \delta_{EN[\mathrm{ms}]} \quad (3.7)$$

$$\delta_{E2E[\mathrm{ms}]} = \delta_{C-RAN[\mathrm{ms}]} + 2\delta_{BH,MEC[\mathrm{ms}]} + \delta_{MEC,UL/DL[\mathrm{ms}]} \quad (3.8)$$

where:

- δ_{E2E} - End to End Latency.
- δ_{C-RAN} - C-RAN associated Latency.
- $\delta_{BH,C}$ - Backhaul to core transmission Latency.
- $\delta_{BH,MEC}$ - Backhaul to MEC transmission Latency.
- δ_{Cor} - Core processing delay.
- δ_{Tran} - Transport transmission delay from the core to the Internet data centres.
- δ_{EN} - External Data centre contribution delay.
- $\delta_{MEC,UL/DL}$ - MEC processing delay.

Application delay requirements are presented in Table 3.3.

The C-RAN delay represents the latency contribution from the network edge, delay contributions coming from the RU, DU, and CU processing delays and the transmissions ones from FH and MH.

$$\delta_{C-RAN[\mathrm{ms}]} = \delta_{RU,UL[\mathrm{ms}]} + \delta_{RU,DL[\mathrm{ms}]} + 2\delta_{FH[\mathrm{ms}]} + 2\delta_{MH[\mathrm{ms}]} + \delta_{DU,UL[\mathrm{ms}]} + \delta_{CU,UL[\mathrm{ms}]} + \delta_{DU,DL[\mathrm{ms}]} + \delta_{CU,DL[\mathrm{ms}]} \quad (3.9)$$

where:

- $\delta_{RU,UL/DL}$ - RU processing delay on UL and DL.
- δ_{FH} - Transmission delay between the RU to the DU.
- $\delta_{DU,UL/DL}$ - DU processing delay on UL and DL.
- δ_{MH} - Transmission delay between the DU to the CU.
- $\delta_{CU,UL/DL}$ - CU processing delay on UL and DL.

The processing delay, (3.10), in the nodes depends on two factors: first, the delay from the process of the BS function, which is directly related to the amount of functions that are addressed in the node; second, the queuing delay from the input traffic.

$$\delta_{Node,UL/DL[\text{ms}]} = \delta_{Node,proc[\text{ms}]} + \delta_{Node,que[\text{ms}]} \tag{3.10}$$

where:

- δ_{Node} - Processing delay in the node.
- $\delta_{Node,proc}$ - BS function processing delay in the node.
- $\delta_{Node,que}$ - Queuing delay in the node.

The queuing delay depends on the maximum throughput in the devices, the throughput assigned by each service in the network, and the priority level of each service described in Table 2.1.

$$\delta_{RU,que[\text{ms}]} = \sum_{P=1}^{P<P_{Serv,j}} \frac{R_{RU,Serv,j[\text{Mbps}]}}{R_{RU,máx}} \tag{3.11}$$

where:

- $R_{RU,Serv,j}$ - Throughput in the RU for a specific service *j*.
- $R_{RU,máx}$ - Maximum throughput on the RU.
- $P_{Serv,j}$ - Priority level of service *j*.

Figure 3.4 illustrates the delay contributions from the different nodes and links of the network.

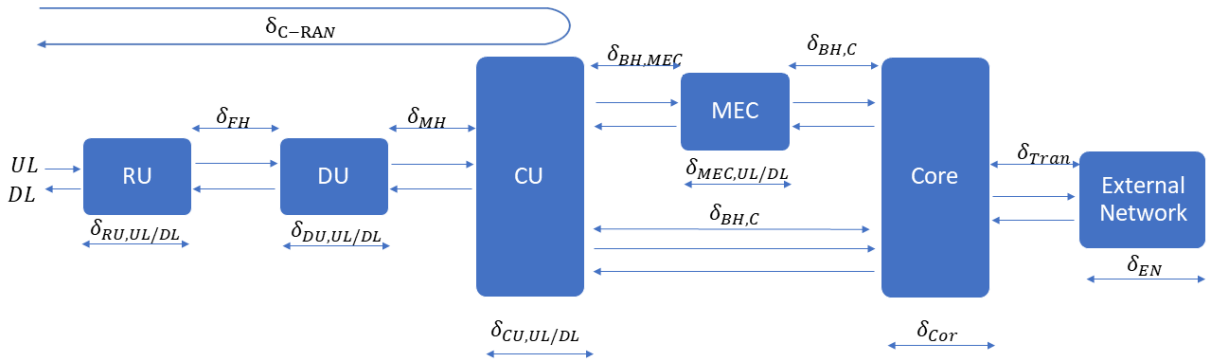


Figure 3.4. Delay contributions on the network.

The C-RAN latency is limited by two factors: application latency and HARQ protocol requirements. For the splitting option between 8 and 6, HARQ is implemented in the RU so the retransmission process restriction needs to be taken into account, considering a retransmission maximum latency of 3 ms,

$$\begin{cases} \delta_{C-RAN[\text{ms}]} < \delta_{HARQ[\text{ms}]}, & if\ \delta_{HARQ[\text{ms}]} < \delta_{App[\text{ms}]} \\ \delta_{C-RAN[\text{ms}]} < \delta_{App[\text{ms}]}, & if\ \delta_{HARQ[\text{ms}]} > \delta_{App[\text{ms}]} \end{cases} \tag{3.12}$$

where:

- δ_{HARQ} - HARQ protocol requirement latency.

- δ_{App} - Maximum latency depending on what application is chosen.

Assuming splitting options higher than 6, HARQ is sent to the DU so the network delay is only affected by the applications one.

The latency of the network is essential to determine the length of the links in the network. The distance of an E2E communication is determined by the time between application delay requirements and network delay, expressed by

$$d_{E2E\,[\mathrm{km}]} = \left(\delta_{App[\mathrm{ms}]} - \delta_{E2E[\mathrm{ms}]}\right)\frac{v_{[\mathrm{km/ms}]}}{2} \tag{3.13}$$

where:

- d_{E2E} - Maximum E2E distance.
- v - Propagation speed in the link.

The total distance in the network is divided into four parts:

$$d_{[\mathrm{km}]} = d_{FH[\mathrm{km}]} + d_{MH[\mathrm{km}]} + d_{BH[\mathrm{km}]} + d_{Tran[\mathrm{km}]} \tag{3.14}$$

where:

- d_{FH} - Fronthaul maximum distance.
- d_{MH} - Middlehaul maximum distance.
- d_{BH} - Backhaul distance.
- d_{Tran} - Distance between the core and external data centre.

3.4.2 Node Processing Power

In order to achieve the maximum network performance, it is important to balance the processing capacity among RU, DU and CU specific for each use case requirements. The processing power in the node is one of the two parameters that define the node processing capacity, and processing power requirements are directly correlated with the splitting option of the BS function, so it is important to analyse the processing required for each BS function, measured in GOPS, being based on [DDLo15].

The model presented in [DDLo15] estimates the processing power used in each node instance (i.e. RU, DU, CU, MEC, and CN) for DL and UL, considering the multiple physical layer functions processing power, the processing power associated with the data flow management and system control of the MAC and RLC layers, the processing of the PDCL, and the processing power used for the transmission to the core network:

$$P_{t\,[\mathrm{GOPS}]} = P_{RF\,[\mathrm{GOPS}]} + P_{PHY\,[\mathrm{GOPS}]} + P_{MAC\,[\mathrm{GOPS}]} + P_{RLC\,[\mathrm{GOPS}]} + P_{PDCL\,[\mathrm{GOPS}]} + P_{BH\,[\mathrm{GOPS}]} \tag{3.15}$$

where:

- P_t - Total processing power required for each node.

- P_{RF} - Processing power required for the RF front-end.
- P_{PHY} - Processing power required for the physical layer functions.
- P_{MAC} - Processing power required for the MAC layer.
- P_{RLC} - Processing power required for the RLC layer.
- P_{PDCL} - Processing power required for the PDCL layer.
- P_{BH} - Processing power required for the backhaul interface depending on the data rate.

The processing power of the physical layer depends on the complexity of the multiple digital processing components:

$$P_{PHY[\mathrm{GOPS}]} = P_{OFDM[\mathrm{GOPS}]} + P_{MAP[\mathrm{GOPS}]} + P_{MIMO[\mathrm{GOPS}]} + P_{BBm[\mathrm{GOPS}]} + P_{Code[\mathrm{GOPS}]} \quad (3.16)$$

where:

- P_{OFDM} - Frequency domains function for OFDM modulation processing component including FFT and IFFT.
- P_{Map} - Mapping and demapping functions processing component.
- P_{MIMO} - MIMO encoding/decoding processing component.
- $\mathrm{P_{BBm}}$ – Baseband modulation/demodulation processing component.
- P_{Code} – FEC function processing component.

The processing power associated with each component can be calculated by:

$$P = P_{ref}\left(\frac{B_{[\mathrm{MHz}]}}{B_{ref[\mathrm{MHz}]}}\right)^{e1}\left(\frac{E_{[\mathrm{bps/Hz}]}}{E_{ref[\mathrm{bps/Hz}]}}\right)^{e2}\left(\frac{N_A}{N_{A,ref}}\right)^{e3}\left(\frac{F_{DC[\%]}}{F_{DC,ref[\%]}}\right)^{e4}\left(\frac{N_{streams}}{N_{streams,ref}}\right)^{e5}\left(\frac{N_{Q_{[\mathrm{bits}]}}}{N_{Q_{ref[\mathrm{bits}]}}}\right)^{e6} \quad (3.17)$$

where:

- P_{ref} - Complexity associated with each function, measured in GOPS, based on the reference scenarios presented in Annex C.
- B - Bandwidth used in the BS.
- B_{ref} - Reference bandwidth used in the BS.
- E - Spectral efficiency dependent on the modulation and coding rate used.
- E_{ref} - Reference spectral efficiency dependent on the modulation and coding rate used.
- N_A - Number of antennas in the BS.
- $N_{A,ref}$ - Reference number of antennas in the BS.
- F_{DC} - System load in the frequency-domain.
- $F_{DC,ref}$ - Reference system load in the frequency-domain.
- $N_{streams}$ - Number of transmission streams, up to the number of antennas.
- $N_{streams,ref}$ - Reference number of transmission streams, up to the number of antennas.
- N_Q - Number of bits used in quantisation.
- $N_{Q_{ref}}$ - Reference number of bits used in quantisation.

Depending on the chosen splitting option, the different power component is assigned to the RU, DU, or CU, assigning the corresponding functions to the different splitting points. The processing capacity in each node can be computed from:

$$P_{RU\,[\mathrm{GOPS}]} = \sum P_{i[\mathrm{GOPS}]} \tag{3.18}$$

$$P_{DU\,[\mathrm{GOPS}]} = \sum^{N_{RU}\,Connected} \sum P_{i[\mathrm{GOPS}]} \tag{3.19}$$

$$P_{CU\,[\mathrm{GOPS}]} = \sum^{N_{DU}\,connected} \quad \sum^{N_{RU}\,Connected} \sum P_{i[\mathrm{GOPS}]} \tag{3.20}$$

$$P_{MEC/CN\,[\mathrm{GOPS}]} = \sum^{N_{CU}\,connected} \sum^{N_{DU}\,connected} \quad \sum^{N_{RU}\,Connected} \sum P_{i[\mathrm{GOPS}]} \tag{3.21}$$

where:

- P_{RU} - Processing power used by the RU.
- P_{DU} - Processing power used by the DU.
- P_{CU} - Processing power used by the CU.
- $P_{MEC/CN}$ - Processing power used by the MEC or CN.
- P_i - Function *i* assign to the node.

One considers that the total processing power is always divided among nodes, without existing any additional process required:

$$P_{t\,[\mathrm{GOPS}]} = P_{RU\,[\mathrm{GOPS}]} + P_{DU\,[\mathrm{GOPS}]} + P_{CU\,[\mathrm{GOPS}]} + P_{MEC/CN\,[\mathrm{GOPS}]} \tag{3.22}$$

The calculation of the load of the aggregation node is based on the functions processing power assigned by each connected node and a fixed component independent of the number of connected nodes required for scheduling and signalling:

$$\mu_{Node} = \frac{P_{fix,Node[\mathrm{GOPS}]} + P_{Node\,[\mathrm{GOPS}]}}{P_{Node,Cap\,[\mathrm{GOPS}]}} \tag{3.23}$$

where:

- μ_{Node} - Node load.
- $P_{Node,fix}$ - Fixed processing power required for scheduling and signalling, independent of the number of connected nodes.
- P_{Node} - Processing power on the aggregation node assigned by the connected nodes.
- $P_{Node,Cap}$ - Processing capacity assigned to the aggregation node.

In order to analyse the impact of the load of the node on network performance, one considers a multiplier factor of the processing capacity assign to the nodes:

$$P_{Node,Cap_{[\mathrm{GOPS}]}} = P_{Node,Cap,ref_{[\mathrm{GOPS}]}} M_P \tag{3.24}$$

where:

- $P_{Node,Cap,ref}$ – Reference processing capacity on the node.

- M_P – Processing capacity multiplier.

3.4.3 Node Throughput

The node throughput is the second parameter that defines the processing capacity in the node, the throughput is an important factor to choose the best splitting option of the network architecture, since a higher throughput in the nodes leads to more expensive nodes and interfaces. There are two different splitting options that need to be studied. The lower splitting options, which correspond to the splitting between the RU and the DU, and are directly related to the FH link capacity, and a high splitting, which divides the function from the DU to the CU and has a direct impact on the bitrate of the middlehaul. It is considered that the splitting option in the BH is always splitting option 1. To calculate the throughput in the nodes, it is necessary to do an overview of the different link capacities for each functional split option proposed in the model, since it is considered that the signal compression factor is proportional to the link capacity for the different splitting options. This study is based on [LCCh18] and [3GPP16].

Option 8 (RF/PHY) provides a constant bitrate in the link, being considered a widely used CPRI interface in this case. Since the bitrate is very high and scales with the number of antennas, the DL and UP fronthaul bitrate for split 8 is defined as:

$$R_{8[\text{Mbps}]} = S_{r[\text{sample/s}]}\, N_{Q[\text{bits}]} N_A \;\; 5 \tag{3.25}$$

where:

- S_r - Sampling rate.

Option 7.1 (Low PHY) continues to provides a constant bitrate in the link and, in this case, the data transmitted in the interface is represented by subcarriers by removing the cyclic prefix and transforming the received signal to frequency-domain using FFT. The guard subcarriers can be removed in the RU, which reduces the bitrate. The DL and UP bitrate for splitting 7.1 is defined in by:

$$R_{7.1,DL[\text{Mbps}]} = N_{SC} N_{SY}\, N_{Q[\text{bits}]} N_L \;\; 2 \times 1000 + MAC_{info[\text{Mbps}]} \tag{3.26}$$

$$R_{7.1,UL[\text{Mbps}]} = N_{SC} N_{SY}\, N_{Q[\text{bits}]} N_A \;\; 2 \times 1000 + MAC_{info[\text{Mbps}]} \tag{3.27}$$

where:

- N_{SC} - Number of subcarriers.
- N_{SY} - Number of symbols.
- N_L - Number of layers.
- MAC_{info} - Bitrate used for information to the MAC layer.

Option 7.2 (Low PHY/High PHY) began to provide a variable bitrate in the link depending on the network load since the FFT and the resource elements mapper are included in the RU, so the data transported in the link are subcarrier symbols. The DL and UP bitrates for splitting 7.2 is defined by:

$$R_{7.2[\text{Mbps}]} = \left(N_{SC} N_{SY}\, N_{Q[\text{bits}]} N_A\ \ 2 \times 1000\right)\mu_s + MAC_{info[\text{Mbps}]} \tag{3.28}$$

where:

- μ_s - Subcarrier utilisation (load).

Option 7.3 (High PHY) achieves a reduce bitrate in DL since the modulation is included in the RU, being considered only for DL, and data is transmitted using code words.

$$R_{7.3[\text{Mbps}]} = (N_{SC} N_{SY}\, N_{Q[\text{bits}]} N_L\ \ 2 \times 1000)\mu_s\ + MAC_{info[\text{Mbps}]} \tag{3.29}$$

Option 6 (MAC/PHY) splits the data link layer from the physical layer and, in this case, the payload transmitted over the middlehaul are transported blocks that lead to a large reduction in the link bandwidth. In this option, data will have extra overhead from scheduling control, synchronisation and frame carry.

$$R_{6[\text{Mbps}]} = (R_{p[\text{Mbps}]} + R_{c[\text{Mbps}]}) \left(\frac{B_{[\text{MHz}]}}{B_{c[\text{MHz}]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{log_2 M}{log_2 M_c}\right) \tag{3.30}$$

where:

- R_p - Peak LTE data rate.
- R_c - Control/Schedule signalling rate.
- B_c - Bandwidth for control signals.
- $N_{L,c}$ - Number of layers for control signalling.
- M - Modulation order.
- M_c - Modulation order for control signals.

Option 2 (RLC/PDCP) uses an already standardised interface (F1), which makes the inter-operation simpler, and the centralisation of the PDCP offers header compression protocols.

$$R_{2[\text{Mbps}]} = R_{p[\text{Mbps}]} \left(\frac{B_{[\text{MHz}]}}{B_{c[\text{MHz}]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{log_2 M}{log_2 M_c}\right) + signaling_{[\text{Mbps}]} \tag{3.31}$$

In option 1 (PDCP/RRC), the entire User Plane (UP) is located in the RU and DU, which gives the lowest link capacity on the middlehaul, but on the other hand achieves small centralisation gains, the bit rate for spiting 1 being calculated from:

$$R_{1[\text{Mbps}]} = R_{p[\text{Mbps}]} \left(\frac{B_{[\text{MHz}]}}{B_{c[\text{MHz}]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{log_2 M}{log_2 M_c}\right) \tag{3.32}$$

The values for all options link capacities are in Annex B.

Finally, it is possible to compute the input and output throughputs in the nodes. Since the input throughput in the RU does not suffer any compression, the data rate that arrives at the RU is precisely the one generated by the user connected to the RU and can be calculated from (3.4). Next, the data are

compressed in the node, depending on the splitting option and the node output throughput, given by:

$$R_{RU/DU/CU/MEC/CN,out\,[\mathrm{Mbps}]} = R_{RU/DU/CU/MEC/CN,in\,[\mathrm{Mbps}]} \frac{R_{in,split,[\mathrm{Mbps}]}}{R_{out,split[\mathrm{Mbps}]}} \quad (3.33)$$

where:

- $R_{RU/DU/CU/MEC/CN,out}$ – RU/DU/CU/MEC/CN output throughput.
- $R_{RU/DU/CU/MEC/CN,in}$ – RU/DU/CU/MEC/CN input Throughput.
- $R_{i,split}$ – Bit rate of the input splitting option.
- $R_{out,split}$ – Bit rate of the output splitting option.

The input throughput on the other nodes of the network depends on the number of connected nodes and can be calculated from:

$$R_{DU/CU/MEC/CN,in\,[\mathrm{Mbps}]} = \sum_{i=1}^{N_{RU/DU/CU}} (R_{RU/DU/CU/out\,[\mathrm{Mbps}]}) \quad (3.34)$$

3.4.4 Centralisation Gain

In order to evaluate the performance of the different C-RAN architectures, one must consider the centralisation gain achieved by centralising the BS functions into the aggregation node, depending on the splitting option. This study is based on [Mont16] in order to quantify the different gains parameters.

First, (3.35) characterises the aggregation gain, comparing the traffic peaks of each node with the traffic peak in the aggregation node, depending on the number of nodes that are aggregated to it. In an FH link the aggregation node is a DU and the connected nodes are the RU, in an MH link the aggregation node is the CU and the connected nodes correspond to the DU, and in the BH link the connected nodes are the CUs and the aggregation is the MEC or the CN.

$$G_{mux,T} = \frac{\sum_{i=1}^{N_{Node,c}} T_{Node,c,i[\mathrm{GB/h}]}}{\sum_{j=1}^{N_{Node,a}} T_{node,a,j[\mathrm{GB/h}]}} \quad (3.35)$$

where:

- $G_{mux,T}$ - Aggregation gain.
- $N_{Node,c}$ - Number of nodes connected to the j^{th} aggregation node.
- $T_{Node,c,i}$ - Peak traffic generated in the i^{th} connected node.
- $N_{Node,a}$ - Number of aggregation nodes.
- $T_{Node,a,j}$ - Peak traffic generated by the j^{th} aggregation node.

The gain achieved by the multiplexing of nodes can also be characterised by the processing capacity of the nodes measure in GOPS, characterised in (3.36) and by the throughput in the node in (3.37):

$$G_{\mathrm{mux,P}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}_{Node,c}} P_{Node,c[\mathrm{GOPS}]}}{\sum_{\mathrm{j=1}}^{\mathrm{N}_{Node,a}} P_{node,a,\mathrm{j}[\mathrm{GOPS}]}} \quad (3.36)$$

$$G_{mux,P} = \frac{\sum_{i=1}^{N_{Node,c}} R_{Node,c,i[\mathrm{Mbps}]}}{\sum_{j=1}^{N_{Node,a}} R_{node,a,j[\mathrm{Mbps}]}} \quad (3.37)$$

The centralisation gain can also be based on the distribution of the BS functions. Increasing the number of function processes in the aggregation node increases the centralisation gain, i.e., the processing gain:

$$G_{proc} = \frac{\sum_{j=1}^{N_{Node,a}} P_{node,a,j[\mathrm{GOPS}]}}{\sum_{i=1}^{N_{Node,c}} P_{Node,c[\mathrm{GOPS}]} + \sum_{j=1}^{N_{Node,a}} P_{node,a,j[\mathrm{GOPS}]}} \quad (3.38)$$

3.4.5 Cost Model

Cost is an important part of the implementation strategy of the network either in CAPEX or OPEX. This study is based on [ATNO18], which presents a model to estimate the network implementation cost, and also on [Mont16] and [Silva16], for the OPEX cost.

The study under analysis only takes C-RAN cost into account, since it is assumed that the operator already has implemented the backhaul and core networks, and that it will not suffer any additional cost from the new C-RAN implementation. The total cost is divided into two parts, CAPEX and OPEX:

$$C_{T[€]} = C_{CAPEX[€]} + N_y C_{OPEX[€]} \quad (3.39)$$

where:

- C_T - Total cost of the C-RAN.
- C_{CAPEX} - Total CAPEX.
- C_{OPEX} - Total OPEX per year.
- N_y - Number of years considered for OPEX.

Since the model aims to optimise the splitting options of the BS functions, one proposes a cost function model in order to compare the different options considering the different data rates in the link connections and the different nodes processing power requirements (i.e. RU, DU, CU or MEC). The CAPEX of the C-RAN implementation accounts for hardware, licences and civil work costs, (3.40) describing the implementation cost of C-RAN based on [ATNO18]:

$$C_{CAPEX[€]} = C_{t,Link[€]} + C_{t,RU/DU/CU/MEC[€]} \quad (3.40)$$

where:

- $C_{t,Link}$ - Total cost of the link.
- $C_{t,RU/DU/CU/MEC}$ - Total cost of RUs, DUs, CUs and MECs.

Regarding the links among nodes, one assumes that they can be fibre or microwave links. First, one presents the cost of the fibre link, where the different terms of (3.41) are based on the number of RUs, DUs, and CUs, having constant and variable terms, depending on network requirements:

$$C_{t,Fibre[€]} = \sum_{i=1}^{N_{Node,c}} \sum_{j=1}^{N_{Node,a}} (C_{Fibre[€]}(i,j) + d_{f\,[\mathrm{km}]}(i,j)\alpha_{[€/\mathrm{km}]})a_{Link}(i,j) \quad (3.41)$$

where:

- $C_{Fibre}(i,j)$ - Constant cost of the Fibre.
- $a_{Link}(i,j)$ - Boolean variable equal to 1 when there is a link between the nodes and 0 otherwise.
- d_f - Fibre link distance.
- α - Cost per km of the fibre.

The cost of the microwave link does not depend on a variable term, since it is considered that the cost of the link does not depend on the network distance:

$$C_{t,Microwave[€]} = \sum_{i=1}^{N_{Node,c}} \sum_{j=1}^{N_{Node,a}} (C_{Microwave[€]}(i,j))a_{Link}(i,j) \quad (3.42)$$

where:

- $C_{t,Microwave}$ - Total cost of the microwave link.
- $C_{Microwave}(i,j)$ - Constant cost of the microwave link.

Considering the cost of the network nodes, the model leads to:

$$C_{t,RU/DU/CU/MEC[€]} = \sum_{i=1}^{N_{RU/DU/CU/MEC}} (C_{RU/DU/CU/MEC[€]}(i) + \Delta_{RU/DU/CU/MEC[€]}(i))b_{RU/DU/CU/MEC}(i) \quad (3.43)$$

where:

- $C_{RU/DU/CU/MEC}(\mathrm{i})$ - Constant cost of the specific node *i*.
- $\Delta_{\mathrm{RU/DU/CU/MEC}}(\mathrm{i})$ - Variable cost of a node *i* according to the required capacity.
- $b_{RU/DU/CU/MEC}(i)$ - Boolean variable equal to 1 when node *i* is being used and 0 otherwise.

The model considers constant and variable values from the C-RAN implementation, and (3.44) provides a cost model depending on the power capacity of the node implementation in order to emphasise the different processing power depending on the different splitting options proposed in the model.

$$\Delta_{[€]}(i,j) = P_{[GOPS]}(i,j)\beta_{[€/GOPS]} + N_{Inter}C_{Inter\,[€/Interface]} \quad (3.44)$$

where:

- β - Cost per unit of resource for the node.

- N_{Inter} - Number of interfaces in each node.
- C_{Inter} - Cost per interface of the node.

The costs of the interfaces on the nodes are proportional to the chosen splitting option in the link, which has a direct impact on the throughput in the node, being considered that:

$$C_{Inter,i_{[€/\text{Interface}]}} = C_{Inter\,[€/\text{Interface}]} \frac{R_{i[\text{Mbps}]}}{R_{8[\text{Mbps}]}} \tag{3.45}$$

where:

- $C_{Inter,i}$ - Cost of the interface in splitting option *i*.

Regarding OPEX, one considers three main factors, based on [Mont16] and [Silva16]:

$$C_{OPEX[€]} = C_{P[€]} + C_{R[€]} + C_{M[€]} \tag{3.46}$$

where:

- C_P - Cost related to power consumption per year.
- C_R - Cost related to renting per year.
- C_M - Cost related to maintenance per year.

The energy consumption of this model only takes the required digital processing of each node into account, and the energy consumption per unit of processing is considered to be the same in every node,

$$\begin{aligned} C_{P[€]} = 24_{[\text{h}]} \times 365 \times (&E_{RU[\text{kW}]} N_{RU} + E_{DU[\text{kW}]} N_{DU}\left(1 - G_{proc,FH}\right) \\ &+ E_{CU[\text{kW}]} N_{CU}(1 - G_{proc,MH}) + E_{MEC[\text{kW}]} N_{MEC}(1 - G_{proc,BH}))C_{E[€/\text{kW/h}]} \end{aligned} \tag{3.47}$$

where:

- E_{RU} - Power consumed per hour for a RU.
- E_{DU} - Power consumed per hour for a DU.
- E_{CU} - Power consumed per hour for a CU.
- E_{MEC} - Power consumed per hour for a MEC.
- C_E - Cost of energy consumed per kW/h.

Since the different splitting options do not change the area of the node and consequently the renting cost of the area, one considers a constant cost for renting each RU, DU, CU and MEC area:

$$C_{R[€]} = 12C_{A[€/\text{m}^2]}(N_{RU}A_{RU[\text{m}^2]} + N_{DU}A_{DU[\text{m}^2]} + N_{CU}A_{CU[\text{m}^2]} + N_{MEC}A_{MEC[\text{m}^2]}) \tag{3.48}$$

where:

- C_A - Rent cost per month per square metre.
- A_{RU} - Area of a RU.
- A_{DU} - Area of a DU.

- A_{CU} - Area of a CU.
- A_{MEC} - Area of a MEC.

Regarding the network maintenance, it is considered the maintenance of the optical fibre and the maintenance of the nodes as a fraction of the total investment:

$$C_{M[€]} = C_{t,Link[€]} m_{Link} + C_{t,Node[€]} m_{Node} \quad (3.49)$$

where:

- m_{Link} - Percentage of total investment spent on maintenance of a link.
- m_{Node} - Percentage of total investment spent on maintenance of the nodes.

3.5 Model Implementation

In this section, one explains the different steps of model implementation. First, one gives a general overview of the implementation process, followed by a deeper analysis on the aggregation process of the different nodes in the network, and the implementation process of the DU for an all independent location for the edge network nodes and the implementation process of MEC nodes.

3.5.1 Model Workflow

The main point of the model is to analyse the different architecture implementation scenarios of the network for the different input parameters, depending on the chosen use case and network specification. In the C-RAN side, the model considers the different locations and aggregations scenarios of RUs, DUs and CUs to study the fronthaul, middlehaul connections in the network. It considers the possibility of a MEC implementation in order to support the ultra-low latency services in 5G. Figure 3.5 illustrates a detailed implementation perspective of the general model.

The model receives the input parameters, computes the traffic and starts to identify the implementation scenario chosen for the simulation. Then, data is loaded and the BS functions are split into the nodes. These data will be used to calculate the different model parameters of nodes processing power, link capacity, and latency impact, which will be necessary in the nodes' aggregation process.

The model works as a sequence of multiple nodes aggregation processes, starting with the RU, which corresponds to the BS, then the DU to CU connection and at last the CU to CN or the CU to MEC connection. This implementation structure is used to interconnect the different possible collocated nodes scenarios, instead of creating a separate, independent and less efficient workflow for the model.

The model under development is based on a 4G C-RAN network, adapted from [Silva16] and [Mont16], so the input parameters only have information on the RRH and BBU locations, in order to convert the RRH and BBU implementation to an RU, DU and CU one. It considers that the RUs locations correspond

to RRHs and the CUs locations correspond to the BBUs, and in the DU implementation scenarios the model is used to efficiently convert RRHs locations to a collocated RU and DU locations. The same principle is used in the MEC implementation process. In this case, some CU possible locations are converted to a collocated MEC and CU node.

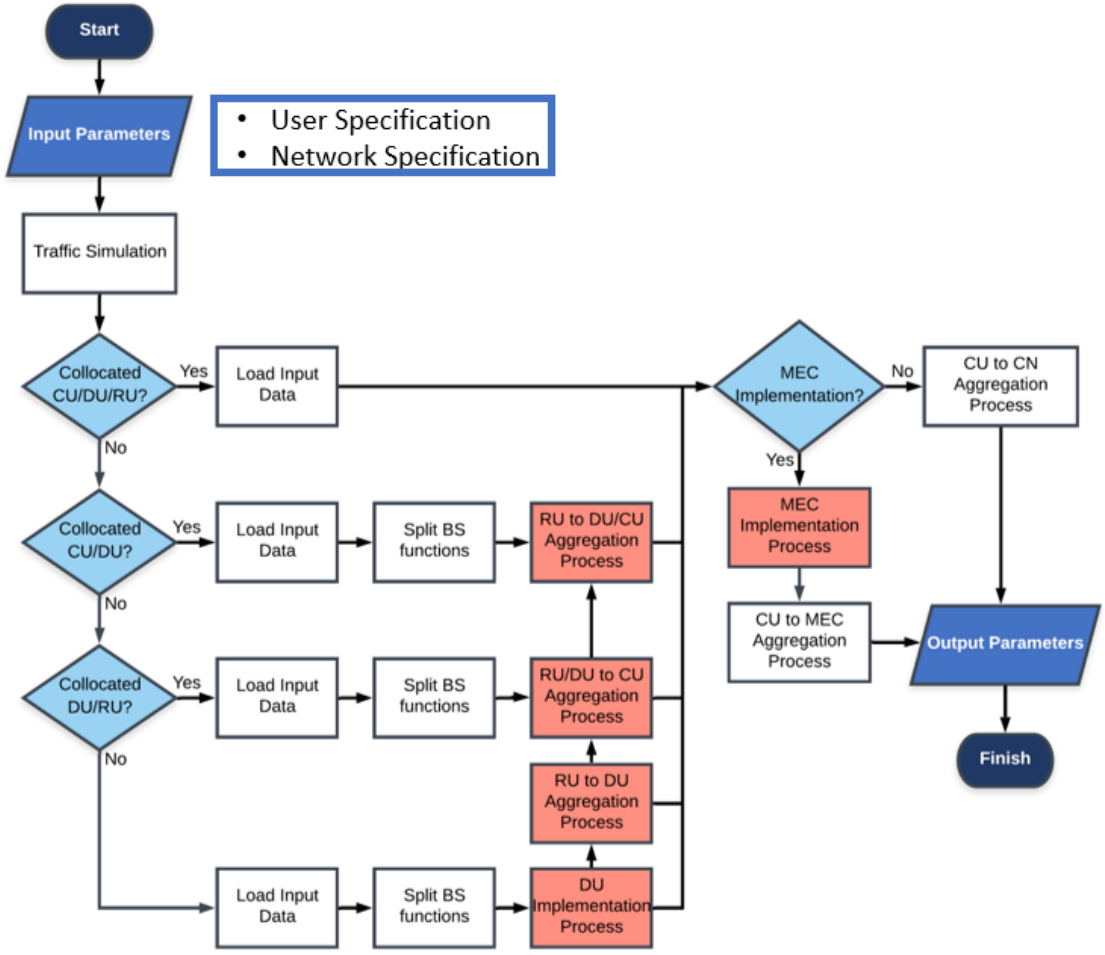


Figure 3.5. Model Flowchart

3.5.2 Aggregation Process

The aggregation process is responsible for efficiently interconnect the possible locations of the different nodes to different scenarios. This process is based on the work of [Silva16] and [Mont16].

In this study, there are multiple different possible aggregations, and each considered aggregation has different input parameters, but the algorithm is the same. Therefore, in order to easily explain the process, one considers an example of a RU to DU aggregation process that is illustrated in Annex E.

The first step of the process is to compute the model parameters, starting with the processing power in the RU, which depends on the BS functions served by the node and the throughput in the node due to the traffic in the network. Next, with the input distances between FH links, the program computes the network latency that depends on the link distance and the RU processing time.

The model starts a loop to process all the RUs, in order to evaluate the RUs possible connections to the

DU. To connect the two nodes there are two network requirements - firstly is the maximum FH distance, and secondly the maximum capacity of the DU. So, in order to evaluate these two requirements, the model starts to compute the number of neighbours of the RU. There are three possibilities, and if there is more than one possible aggregation the model checks the conditions based on the algorithm chosen in the input parameters. These algorithms were developed by [Silva16] and are further explained in this section. After that, the model checks the available DU processing power to verify first if the RU can be connected to the chosen DU, and, if there is only one possible connection, the model will skip the analysis of the algorithms and directly analyses the processing power requirement. Finally, if there is no possible connection or the DU does not have enough processing power to support the RU, the RU will be marked as a standalone node, which means that the RU will be converted to an RU+DU node.

If the two requirements are fulfilled, the RU is connected to the DU, the RU is marked as processed, the DU parameters are updated and the model advances to the following RU.

The model provides different aggregation algorithms developed by [Silva16]. This study considers two different algorithms. In the first connection of the network - the connection closest to the user – one considers the Balance Number of Connections Algorithm in order to balance the processing capacity, traffic, and connections of the aggregation node, and also to avoid overload on specific parts of the network. In the second, and in some scenarios, third aggregation process, one uses the Minimise Delay Algorithm in order to efficiently reduce the delay contribution of the link distances of the nodes, despite losing some balance of the network utilisation. This algorithm was chosen because this study focuses on the evaluation and reduction of the network latency to support the new 5G use cases. Since the number of connections nodes was balanced in the first connection, the network will still be balanced.

The aggregation algorithms used are explained based on [Silva16]:

- Minimise Delay Algorithm - This algorithm is used to aggregate the closest nodes with the required processing capacity. In this case, the model analyses all the possible connection distances and choose the smallest one. This algorithm aims to reduce link distances, reducing the delay of the network.
- Balance Number of Connections Algorithm - This algorithm aims to balance the number of aggregations for each node. The model checks the maximum capacity of the node with fewer connections until it is capable to aggregate new nodes. The number of aggregations for each node is always updated and available for the next step of the aggregation process evaluation.

3.5.3 New Node Implementation Process

Since the model is a continuation of the work develop in [Mont16] and [Silva 16], it is necessary to convert the RRH and BBU locations provided by NOS to an RU, DU, CU, and MEC implementation.

There are two scenarios where the location problem needs to be addressed. One is on RU, DU, and CU independent location scenarios, and the other is on a MEC implementation scenario. In order to solve the problem, the model aims to efficiently add to the location of the existing node the new node that is being implemented in the network. Considering the first scenario, the model evaluates the density

of RUs depending on the FH distance. A threshold level is taken into account in order to consider the possibility to only implement DU nodes in dense traffic areas, like urban and dense urban scenarios, if desired. In order to compute the best DU location, one uses the K-means algorithm, which first receives the RU locations, dividing the RUs into N_{DU} clusters, and computes the centroid of each cluster that corresponds to the best DU theoretical location.

After that, the model analyses each centroid and finds the closest RU to the best theoretical location of the DU. This RU location is converted to a DU, converting the RU node to a RU+DU collocated node. Finally, data from the processing power and traffic are updated.

The K-means algorithm is the one used to find the best theoretical location of the DU. K-means is an unsupervised machine learning algorithm, a self-organised learning algorithm that finds patterns in data without a pre-existing label. The process starts to randomly select N_{DU} points on the map and creates N_{DU} clusters of the points that correspond to the closest RU to each point. Next, the centroid is computed for each cluster, and the original random points are updated. The process continues until all clusters stay the same, which means that it has found the centroid for each cluster that leads to the minimum error, or minimum distances to the RUs assign to the cluster. The K-means algorithm process is illustrated in detail in Figure E.3 of Annex E.

The algorithm runs multiple times with different initial values and returns the result with the minimum error of clusters. As explained before, this algorithm returns the best theoretical location of the DU, and since the DU location can only be allocated to an existing RU one, the optimal possible DU location is addressed to the closest RU node one.

3.6 Model Assessment

The model assessment aims to validate the model in the development stage, and uses a set of empirical tests in which the outcome of the results is already expected in order to verify if the model follows the theoretical results. Table 3.4 described the structure of the empirical tests used to validate the model, and the results are illustrated in Annex F.

In the assessment of the model, several tests are considered to validate the model output parameters. The validation test starts focusing on the three main model parameters that are used to compute network latency, which is the network link distances, the nodes processing capacity, and the traffic in the links. One also validates the central gain analysing the process gain between the RU and CU nodes, and at the end the model for the cost of the network is taken into assessment by analysing the cost of the RU node. These four parameters are subjected to a test for a set of different splitting options of the network architecture. Minho is used as a scenario, with 374 RU and 42 CU locations to evaluate the model, assuming an RU-DU+CU network architecture, without the implementation of new nodes.

Table 3.4 List of model assessment tests.

Test ID	Description
1	Validation of the input file read, by verifying if the size and type of inputs values stores in the different variables are the same as in the input files.
2	Scattering the position of the RUs, CUs and CN positions in the Matlab plot over a Google Maps to inspect the node placement on the scenario.
3	In case of implementation of new nodes: • Scattering the position of the new DU or MEC positions, plotting the original nodes and centroids positions in the Matlab plot over a Google Maps to inspect the node placement on the map. • Check if the computational and link capacity values are updated. • Check if the connection is correctly stored.
4	Validation of the maximum distances constraints by checking if there are no connections that do not respect the constraints.
5	Validation of the aggregation process: • Check if the computational and link capacity values are update. • Check if the connection is correctly stored. • Check if the node is marked as served and not assigned again.
6	Validation of the output files, by checking if they are correctly printed and plotting the output results.

In the first test, the percentage of nodes successfully connected to the aggregation node is tested and then compared to the results of a FH scenario where the maximum link distance is 10 km and with an MH scenario with the maximum link distance varying from 20 km to 40 km. The results from the first test are shown in Figure 3.6, and, as expected with a fronthaul splitting scenario, since the maximum distance is lower than a middlehaul network, the percentage of RUs connected to the aggregation node is lower.

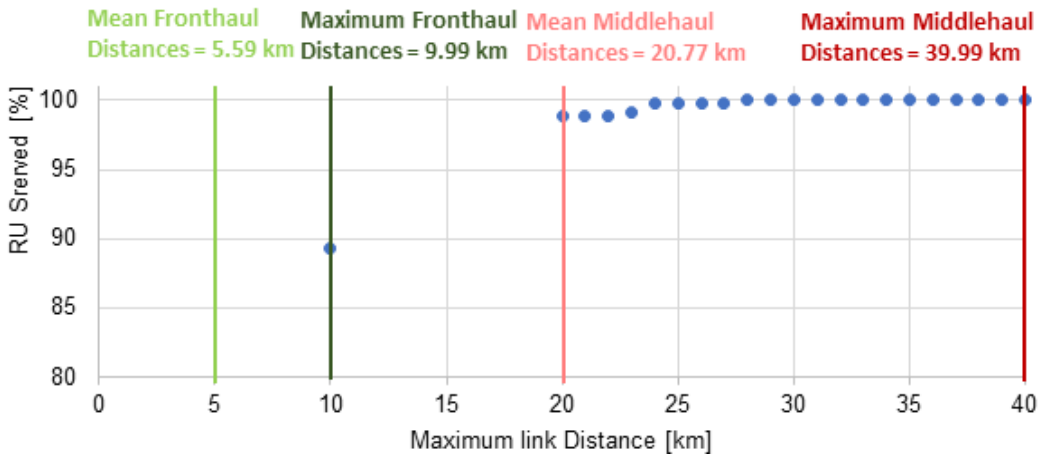


Figure 3.6. Served function of RUs with the maximum link distance.

The second test evaluates the processing power in the nodes. One can verify that when the splitting option is higher, the processing capacity on the RUs is higher, since more functions are addressed to

the RU. The DU+CU processing capacity is reduced and the CN processing capacity stays the same, since the functions of the BS are distributed among the edge nodes and the functions addressed by each node linked to the core is always the same.

The third test evaluates the capacity in the links. As expected from the theoretical viewpoint, the throughput on the FH, between the RU, and the DU+CU is proportional to the link data rate described in the table of Annex B, and the throughput on the spitting option 6 is lower than the lower splitting option. The throughput from the splitting option of the physical layer is similar in the DL, but as a high impact from splitting option 7.2 to 7.3 because the signal is modulated in the RU.

Regarding the assessment of the central gain output, the results validate the model since, as expected, increasing the number of functions in the RU node (i.e. higher splitting option) decreases the process gain since the function is not centralised in the CU node.

Chapter 4

Results Analysis

This chapter starts to provide a description of the scenarios considered, and it then presents the results and their respective analysis.

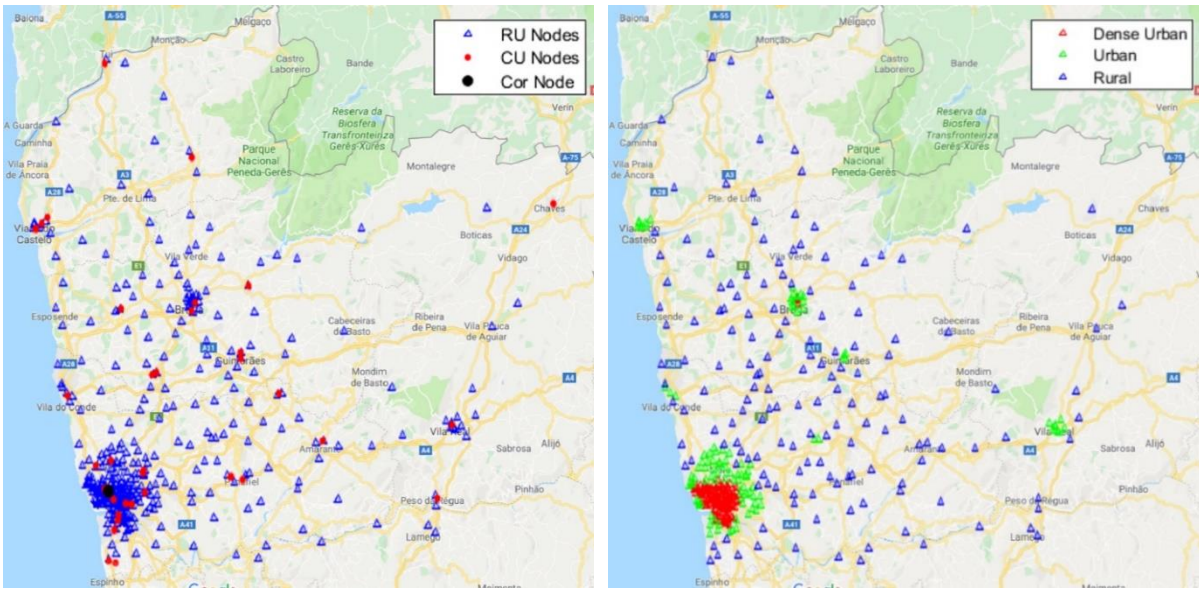
4.1 Scenarios

The study under analysis is based on three different scenarios, considering the data provided by NOS in the Minho and Portugal scenarios required in order to run the model simulator. One also analysed a real case scenario, which includes latency information data from NOS network that was applied to a separate simulator specifically to analyse the latency parameters on the network.

4.1.1 Minho Scenario

The area of Minho is located in the north-west of Portugal, where the main regions into consideration are Porto, Braga, Viana do Castelo, and Vila Real. This scenario has around 3.4 million inhabitants in around 11 600 km^2 of area. The Minho scenario has a population density of 290 inh./km^2, and the majority of the population in Minho are in the Porto metropolitan area, which has a population density of 843 inh./km^2.

This thesis analyses the multiple architecture scenarios of C-RAN in a 5G network. From NOS' nodes data locations, one considers the location of the cell sites as the location of the nodes closest to the user. In an RU and DU independent location architecture, the cell sites are the RU location, in an RU and DU collocation architecture, the cell sites are the RU+DU node, and finally in an RU, DU, and CU collocated node, the cell sites are the RU+DU+CU locations. The aggregation points' location is also available, which in this case can be considered as the CU location or the DU+CU collocated location. In an architecture of all independent nodes, the model computes the best DU locations and, in this scenario, the implemented DU are collocated at the best RU locations defined by the model. Table 4.1 summarises the number of nodes in the Minho scenario, and considers the RU nodes as the cell sites, the CU nodes as the aggregation ones, and also the core node.



(a) Node Location (b) RU Density

Figure 4.1. Node location and RU density type distribution on Minho.

Table 4.1. The number of RU nodes, CU nodes and Core.

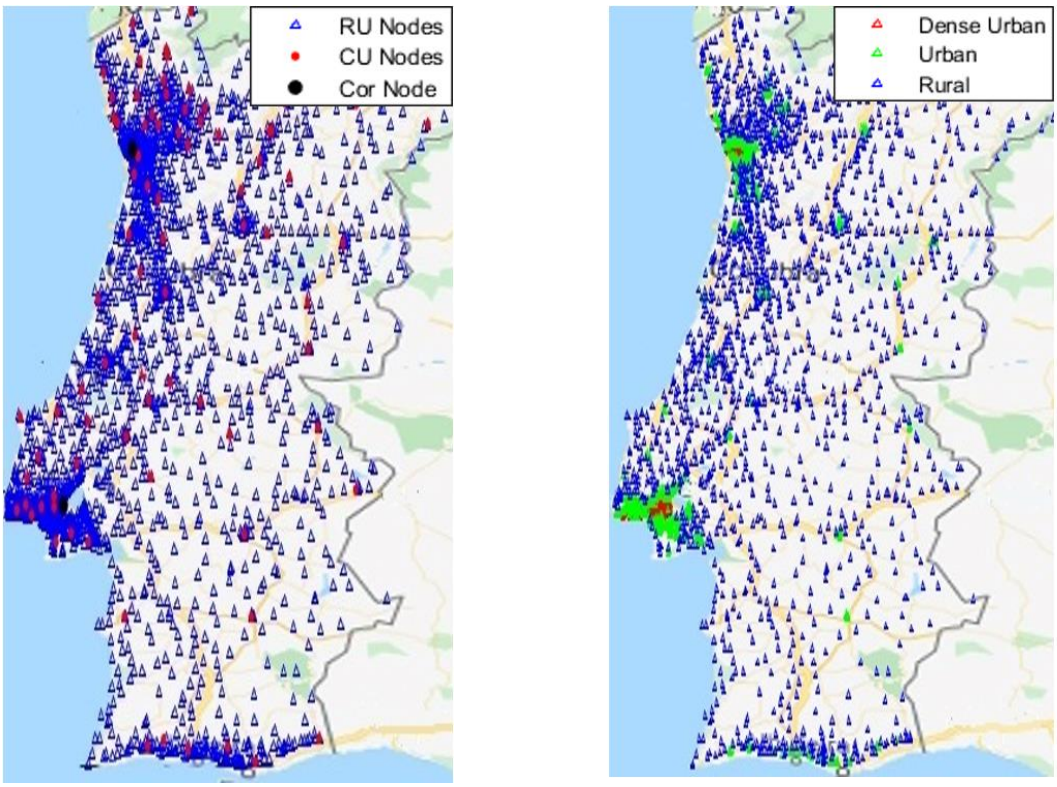
Number of RU Nodes	Number of CU Nodes	Number of CN Nodes
374	42	1

The RUs on the map are analysed based on the density of nodes. Three environments are considered for the classification of RU density types: dense urban, urban, and rural. This classification is essential for the analyses of the nodes processing power, as it considers three different bandwidths for the nodes: 100 MHz for dense urban RUs, 50 MHz for urban RUs, and 20 MHz for rural RUs. Figure 4.1 shows the different RU types in the Minho scenario, where there are around 35% of dense urban RUs, 30% of urban, and 35% of rural RU nodes.

4.1.2 Portugal Scenario

Portugal has around 10.5 million inhabitants in around 92 090 km^2, corresponding to a population density of approximately 115 inh./ km^2, and the majority of the population in Portugal is concentrated in coastal areas - almost half of the total population lives in Lisbon and Porto´s Metropolitan Area.

The mobile network in Portugal follows the same behaviour as the population density in the country, so as expected the majority of the RU nodes are located in the coastal areas, which one can observe in Figure 4.2.



(a) Node Location (b) RU Density

Figure 4.2. Node location and RU density type distribution on Portugal.

The information related to the number of nodes in Portugal is summarised in Table 4.2.

Table 4.2. The number of RU nodes, CU nodes and Cor.

Number of RU Nodes	Number of CU Nodes	Number of Core Nodes
2755	86	2

Following the same principles as used in the Minho scenario, Figure 4.2 illustrates the different types of the density of the RU nodes, where one can verify that the city of Lisbon, Porto, and Setubal are the ones with higher RU nodes density.

4.1.3 Scenarios with Input Network Latency

This scenario takes the latency values of the existing mobile network of NOS into consideration, based on the input data provided by NOS.

The network latency information provided by NOS is divided into four different regions in Portugal, including:

- North of Portugal – Defined, approximately, as the regions north of Santarém.
- Lisbon – The region of Lisbon has an average density of 948 inh./ km^2, which represents the higher population density in Portugal, with is approximately 2 808 000 inhabitants.
- Madeira Archipelago - Composed of the Madeira and Porto Santo islands, with an area of 800 km^2, and an average density of 334 inh./ km^2, the Madeira archipelago is approximately 1 000 km away from Portugal continent.
- Azores Archipelago - Composed of nine islands, with a total area of 2330 km^2, and an average density of 106 inh./ km^2, the average distance between Azores and Portugal is around 1500 km.

Table 4.3. The number of RU nodes and CU nodes.

Region	Number of RUs	Number of CUs
North of Portugal	1 405	134
Lisbon	639	55
Madeira	77	17
Azores	70	11

In this scenario, the nodes in the network are already connected according to the existing network configuration of NOS, so it is not considered any algorithm to reduce network distance between the nodes or a balancing algorithm to level the number of connections in the CU nodes.

The input data describes an overall latency information between cell sites to the aggregation node, including the processing delay in the node, and the latency information between aggregation nodes until the arrival to the Core. Since the processing delay in the CN is not considered in the input data, one assumed a 150 µs processing delay based on [Inte19]. It is considered that network links are all fibre ones, taking into consideration a distance multiplier factor to compensate for the fact that fibre is not implemented on a straight line. The cell site is not always directly connected to the aggregation node, having in some cases multiple points connecting wireless and fibre links that are not specified in the input parameters of the scenario. The following figures illustrate the cell sites (i.e. RU nodes), the aggregation nodes (i.e. CU nodes), and CN location on the scenario.

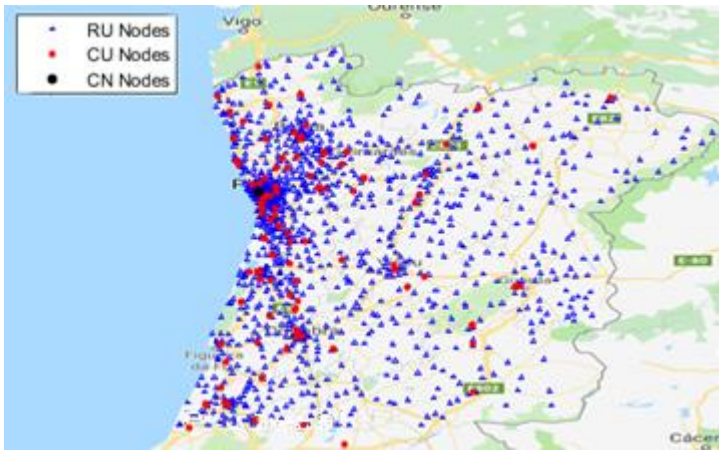


Figure 4.3. North of Portugal map with nodes location.

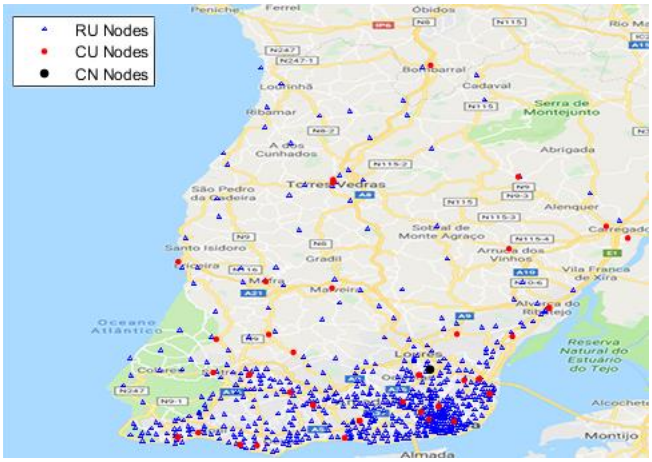


Figure 4.4. Lisbon map with nodes location.

4.1.4 Reference Scenario Configuration

In order to analyse the different output parameters for the model, one considered a reference scenario that works as a means of comparison for the study to analyse the evolution of a specific output, changing an input parameter of the reference values, and comparing them with the reference one.

First, it is presented the reference values for the calculation of the number of users of each service. Considering the assumption on the user population on the RU nodes, the reference scenario considers:

- Penetration ratio: 30%,
- Usage ratio: 10%.

It is assigning a service mix for each use case; the service mix represents the percentage of the total users in the cell that are assigned to each use case depending on the scenario under study. The reference values of the service mix for each use case are presented in Annex G. In order to simulate the increase of devices connected to the network of the new 5G use cases, a multiplier factor of the 5G services was taken into consideration, which in the reference scenario is set to one.

The calculation of the devices connected to the network is based on the traffic mix of the use case and

the traffic profile from the generated 4G traffic in the network. One considered a reference traffic on DL and UL extracted from [Silva16].

The reference values on the cell specification are also taken into account, including the input 4G traffic profile, which will have a direct impact on the computation of the number of users in the cell, and the RU bandwidth that is assigned based on the node density on the map, which will have a direct impact on the processing capacity of the node.

Next, regarding the network reference values specification, it is considered that the rural RUs use a wireless connection to the DUs, and that the urban and dense urban RUs use a fibre connection. The wireless connection is only implemented on the first connection because of the link capacity and link distance restriction. One uses a 2×10^8m/s propagation speed in fibre, and a 3×10^8m/s one in the wireless links. A 1.67 factor increase is considered on the fibre links distance, which takes into account that the fibre is not implemented in a straight line. Regarding the maximum distance in the links, the reference maximum link distances are 10 km on the FH link and 40 km on the MH link, while there is no maximum link distance defined on the BH link in order to ensure that nodes are connected into the BH.

An independent RU and a collocated CU+DU for the architecture reference scenario are assumed since this implementation is the closest to a 4G network used in [Silva16] and [Mont16]. It is assumed a 7.1 splitting option between the nodes of the FH since the majority of the function are in the CU node to guaranty high levels of centralisation, with a significant lower data rate on the FH link (i.e. 90% reduction on the DL and 60% reduction on the UL) when comparing to an 8 splitting option, which corresponds to a 4G network architecture.

When considering the RU-DU-CU architecture, 50 DU nodes are implemented in the Minho scenario and 300 DU nodes in the Portugal scenario, distributed throughout the dense urban and urban areas, establishing the reference scenario since the implementation of DU nodes is focused on the offload of traffic and processing demands on the CU node, so the DUs are implemented on the more users density areas. The analysis of the implementation of DU nodes is presented in Annex L.

In the reference scenario, the restriction on the maximum processing power on the node for the aggregation process is not defined as the load varies during the day, considering the normalisation against the peak power of the node (corresponding to 80% of load) required to achieve the 1 ms of processing delay. The same principle is applied for the maximum throughput on the node, where one assumes that the normalisation against the peak throughput of the node corresponds to 80% of load required to achieve the 1 ms of queuing delay on the nodes. In this scenario, it is not considered the transition latency between the CN and the external data centre (i.e. $\delta_{Tran} = 0$) and no processing delay on the external data centre (i.e. $\delta_{\mathrm{EN}} = 0$).

Finally, the last assumptions for the reference scenario are related to the cost. The reference values for the CAPEX parameters are divided into the constant cost, which does not depend on the output parameters of the model, that represent the initial implementation of the links and the nodes, independent of distance and processing power, with values for the RUs based on [Silva16]. The other nodes have an increasing implementation cost with the maximum processing power required in the

node. These values are very similar, since the nodes are virtual nodes, so the implementation of the different nodes is made using different virtual machines with different processing capacities. The variable costs depend on the output parameters of the model, which take into account the required processing capacity on each node in order to establish a linear increment of the costs of the nodes directly related to the processing power needed, and the link distance of the fibre depending on the chosen architecture. It is considered the cost per interface in the nodes which is directly related to the number of devices connected to the different aggregation nodes.

Regarding the reference cost for the OPEX, first it is assumed that the microwave links on the network do not need maintenance, and that the fibre and node maintenance cost is established on a small percentage of the CAPEX cost of the equipment, based on [Silva16]. The parameter C_A represents the mean renting cost per square meter based on statistical information from housing companies in Portugal. The area of the nodes is based on the size of the cabinet used in the work of [Mont16]. Finally, the C_E value is based on the energy fee available for low voltage tariffs with EDP, and the energy consumption on the nodes is directly proportional to the maximum processing power on the node, using a conversion of GOPS per *Watt* with a factor of C_{GOPS}, based on [DDLo15]. Annex G summarises the reference scenario configuration presented in this section.

4.1.5 Reference Scenario Variation

In order to study the impact of the input parameters on the performance of the network, one must analyse the output parameters applying variations on the reference scenario configuration values.

First, one analyses the impact of the chosen architecture on the network. The reference scenario considers an RU-DU+CU architecture, since this is the one used in a 4G network so, in this analysis, one considers a variation with all the possible architectures presented in Section 3.2. The model analyses the impact of the different FH splitting options (i.e. 8, 7.1, 7.2, 7.3, 6) on the output parameters of the network. When considering the different architectures, one measures the impact of the implementation of the new node locations on the network (i.e. DU and MEC ones). In the case of the implementation of DUs, these nodes can only be located at the RU location, so one considers the percentage of RU nodes converted to RU+DU ones. In the case of the implementation of MECs, these nodes can only be located at the CU locations. The variation on the network architecture has a direct impact on the network distance and nodes processing power.

Secondly, one verifies the impact of the assigned processing power to the nodes on the overall performance of the network to process the BS functions assigned to the nodes. When considering the reference scenario, it is established that the network is implemented to verify an 80% maximum load on the nodes, and it is taken into account an input factor multiplying on the processing power of the nodes, changing the load of the nodes, and measuring the latency impact on the overall network. The variation on the nodes processing power has a direct impact on the processing delay due to the BS functions assigned to the node called GOPS delay.

Finally, it is considered the impact that the number of users connected to the network has on the

performance and coverage of the different use cases. The variation of the usage and penetration ratio is taken into account, so the variation of the number of users is uniform for all services under analysis. The variation of users on the three 5G services type (i.e. eMBB, mMTC and URLLC) was considered, and in this case the user distribution does not follow the same proportion as in the reference scenario, in other to focus the analysis on the impact that the new 5G services bring to network performance and the consequence and requirements that need to be taken into consideration, Table 4.4.

Table 4.4. Specification on the variation of the reference scenario.

Architecture	{RU-DU-CU; RU-DU+CU; RU+DU-CU; RU+DU+CU}
Splitting Option	{8; 7.1; 7.2; 7.3; 6}
RU nodes converted to DU nodes [%]	[0; 100]
CU nodes converted to MEC nodes [%]	[0; 100]
Node Processing Capacity multiplier	$[10^{-3}; 10^{4}]$
Usage and Penetration ratio [%]	[10; 100]
eMBB users [%]	[0; 100]
mMTC users [%]	[0; 100]
URLLC users [%]	[0; 100]
eMBB user multiplier	$[1; 10^{3}]$
mMTC user multiplier	$[1; 10^{7}]$
URLLC user multiplier	$[1; 10^{4}]$

4.2 Centralisation Gain Analysis

This section features the results obtained from the analysis of the multiplexing gain. The purpose of this section is to compare the values of the different gain metrics used in this study for the different architecture splitting options. The results are illustrated in Table 4.5.

The aggregation gain is used to compare the impact of the peak processing capacity on the nodes with the peak processing capacity of the aggregation node. In the FH scenario, for example, this means that higher values of aggregation gain lead to more processing power on the sum of RUs compared to the CU or DU that the nodes are aggregated. The aggregation gain is measured in 3 different metrics: the traffic aggregation gain evaluates the traffic on the network measure in GB/h, the throughput aggregation gain measures the data rate on the nodes in Mbps and, finally, the GOPS aggregation gain analyses the impact of the BS functions on the nodes measured in GOPS. It is also analysed the process gain that aims to evaluate the BS functions distribution on the nodes, since a higher process gain means

more functions are implemented in the CU or DU compared with the functions on the RU, which leads to more benefits from data centralisation. The traffic and throughput aggregation gain are similar since the metric is proportional. The throughput and traffic gain are proportional to the number of functions processed in the RU node since a higher number of BS function on the RU provides a higher compression of the output signal.

Table 4.5. DL Centralisation Gain for RU-DU+CU architecture.

RU-DU+CU-CN		Throughput/Traffic Aggregation Gain	GOPS Aggregation Gain	Process Gain
FH	8	1.115	0.410	0.709
	7.1	17.90	1.264	0.442
	7.2	19.07	1.413	0.414
	7.3	17.90	7.601	0.116
	6	31.18	9.847	0.092
BH	8	1.060	4.951	0.168
	7.1		3.085	0.244
	7.2		2.906	0.256
	7.3		0.853	0.540
	6		0.679	0.596

The first architecture into analysis is the RU-DU+CU-CN scenario, which works as a reference scenario for further analysis in this section. Option 8 of this architecture matches the 4G C-RAN architecture with the RRH and BBU nodes and, in this case, the aggregation gain in the FH is the lowest since the signal processing is all done in the DU+CU node, so the traffic in the FH is not pre-processed and does not achieve gains from signal processing. On the other hand, the process gain in option 8 achieves the highest values, reaching more centralisation benefits but with a heavy FH link. In this architecture, the splitting of the physical layer functions (i.e. 7.1, 7.2, and 7.3) leads to similar traffic gains on the node, but it is worth noting that the option 7.3 achieves the most GOPS aggregation gain, so the CU device has lower processing requirements. Option 6 is the one with the most signal processing in the RU, so the FH links have higher aggregation gains, reaching 31.2 traffic aggregation gain on the FH, but does not benefit from centralisation (i.e. 0.1 process gain). In the BH link, the traffic and throughput gain is much lower compared with the FH gain - this is expected, since the majority of the signal processing is done on the lower splitting options, and the variation from the different splitting options is more significant on the FH splitting option. In the BH, the splitting option on the architecture goes from an option 2 to an option 1, which leads to a 0.4% variation on the throughput per aggregated node. The process gain on BH is inversely proportional to the process gain on the FH because in all options the number of BS functions on the CN is the same.

Considering the network gain in all independent nodes, the more important differences are the GOPS aggregation gain since there are more nodes in the network, since the BS functions on the different nodes are balanced, achieving higher GOPS aggregation gains. One should notice that, for splitting option 8, this architecture distributes the heavy load on the 4G BBU node between the DU and CU node, so one concludes that implementing DU nodes in the network can be an alternative to offload CU nodes, instead of sending more BS functions to the RU nodes, which requires a higher investment.

In a high splitting option, like in a RU+DU-CU architecture, the signal is highly processed in the RU node, and the FH link achieves a 40% increase on the traffic aggregation gain and around 150% on the GOPS aggregation gain compared with option 6 of the reference scenario. The high aggregation gain means that almost all network functions are in the edge of the network, providing high link data rate reduction but low benefits from centralisation, since the RU+DU+CU architecture has an 82% reduction compared with option 8 on the reference scenario.

Concerning UL results, the first conclusion is that the traffic in the CU is higher than in DL, which leads to a reduction in the traffic and throughput aggregation gain for all splitting options. In the FH, one notices a more significant improvement from the physical layer splitting 7.3 (i.e. 10 times higher traffic gain compared with option 8). The values for GOPS aggregation gain on UL are similar to DL since just a small percentage of the processing power in the node is traffic loaded dependent. The process gain in UL follows similar results as DL, as explained before, and the BS functions processing power in the nodes are low traffic load dependent. As a general overview of the results on the comparison of DL and UL, it is worth emphasising that UL achieves more significant improvements on network gains with the 7.3 splitting option, since options 7.1 and 7.2 are not beneficial in a UL network environment.

4.3 Processing Capacity Analysis

This section presents the information regarding the processing capacity required in the different nodes on the network for the different architecture scenarios considered in this study. The processing capacity in the node is divided into two measurements. Subsection 4.3.1 analyses the data throughput on the nodes, which is related to the traffic arriving at the nodes. Subsection 4.3.2 analyses the processing power, measured in GOPS, required to support the BS functions assigned to each node.

The two subsections follow the same structure, analysing the processing capacity on the nodes for the different architectures.

4.3.1 Throughput Analysis

This section analyses the input and output throughputs in the nodes of the network. Throughput depends on the traffic generated by each cell site, depending on the use cases specifications presented in Table 3.3 and Table G.1, and the number of users for each service in the cell site.

The RU-DU+CU scenario is analysed as the reference architecture. In this case, the input throughput in the nodes of the network is measured which, in this case, are the RUs, the DU+CUs, and the CN. Figure 4.5 illustrates the results obtained for the reference architecture.

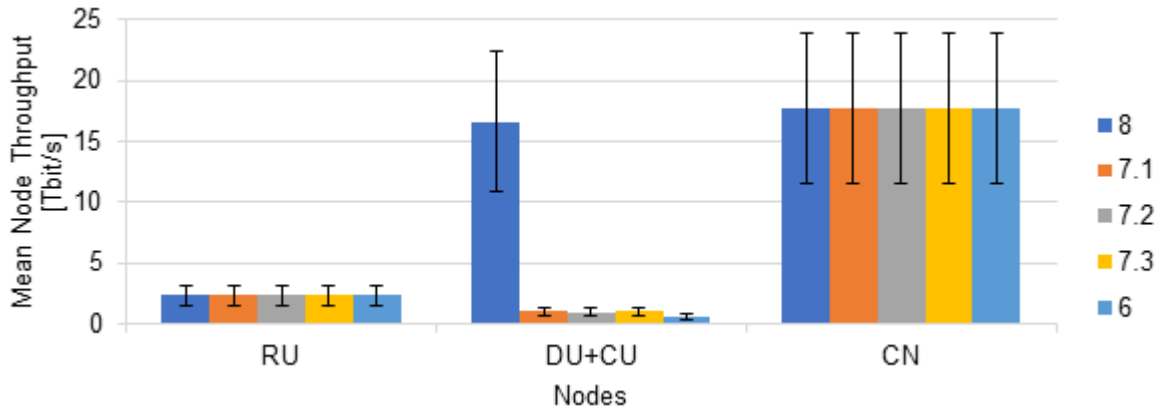


Figure 4.5. Mean input throughput on the network nodes in different splitting options for RU-DU+CU architecture on DL.

The interval established by the vertical black lines represents the confidence intervals of the mean output values of the throughput on the network nodes.

It is worth noticing that the input throughput on the RU does not change with the chosen splitting option since the signal was not processed in the node yet, so the 2.4 Tbps of the RU input throughput is the mean value of the traffic created in the cells. The throughput in the CN is also identical in all splitting options since the BH connection assumes a BS splitting option 1 that does not depend on the chosen splitting option on the FH, so the traffic on the C-RAN is divided between the RU and the DU+CU. The input throughput on the DU+CU node is proportional to the amount of signal processing in the RU, and for the lower splitting option (i.e. Option 8) the signal is not compressed in the RU since all the BS functions are assigned to the CU node, so the throughput arriving in the DU+CU node per RU is the same as the input throughput on the RU. Higher splitting options provide higher signal compression on the RU, which substantially reduces the throughput on the DU+CU. From Option 8 to Option 7.1 the node achieves a 94% reduction, and this variation occurres since the RU has the capacity to modulate the signal with FFT and the data starts to be transported on the FH as subcarriers.

The output throughput on all the different RU splitting options was analysed, since the chosen link capacity on the link is related to the output throughput in the node. As expected, the most noticeable reduction on the throughput is from splitting option 8, which corresponds to a 4G architecture, and the splitting option 7.1, where it achieves a 93.7% reduction. When considering the splitting option between PHY and MAC (i.e. Option 6), there is a 42.6% throughput reduction from option 7.3 to option 6. From the independent RU splitting option 6 to a collocated RU and DU, that corresponds to an MH splitting option 2, the output node throughput decreases 28.6%, and the throughput from the RU+DU node to the RU+DU+CU node only reduces 0.4%. From these results, one can conclude that the 5G C-RAN architecture should change the FH splitting option 8 to a higher splitting option on the splitting of the physical layer in order to support the high throughputs arriving on the CU nodes from the new 5G use cases. Figure 4.6 summarises the results of the output throughput on the RU node.

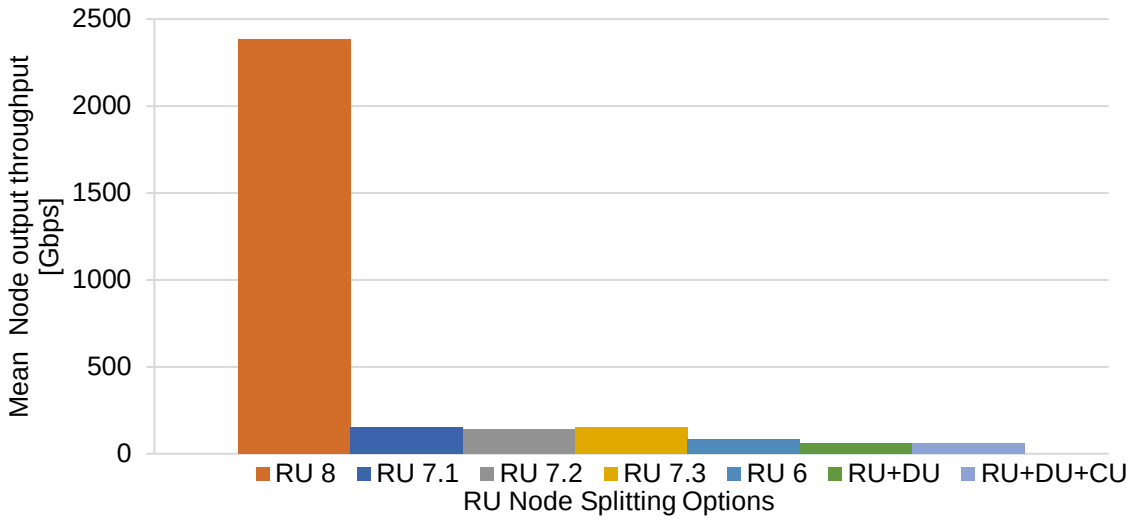


Figure 4.6. Mean output throughput on the network RU nodes in different splitting options on DL.

4.3.2 Processing Power Analysis

This subsection analyses the processing capacity measure in GOPS, required in the nodes of the network for the different architecture scenarios. The processing power depends on the BS functions assigned to the node so it is necessary to analyse the impact of the splitting option on the distribution of the processing capacity throughout the network.

First, the reference scenario was analysed in Figure 4.7 and, as expected, higher splitting options assign more BS function to the RU, which results in higher processing power needed in the RU and lower processing power assigned to the DU+CU node. It is worth noticing that the options with higher impact on the processing power on the nodes are the transition from an option 8 to an option 7.1 and the transition from an option 7.2 to an option 7.3. In the transition from an option 8 to an option 7.1, there is a 43.3% variation on the processing power on the nodes. The difference between these options is the FFT that modulates the signal, and it is the BS function with higher GOPS assigned to it.

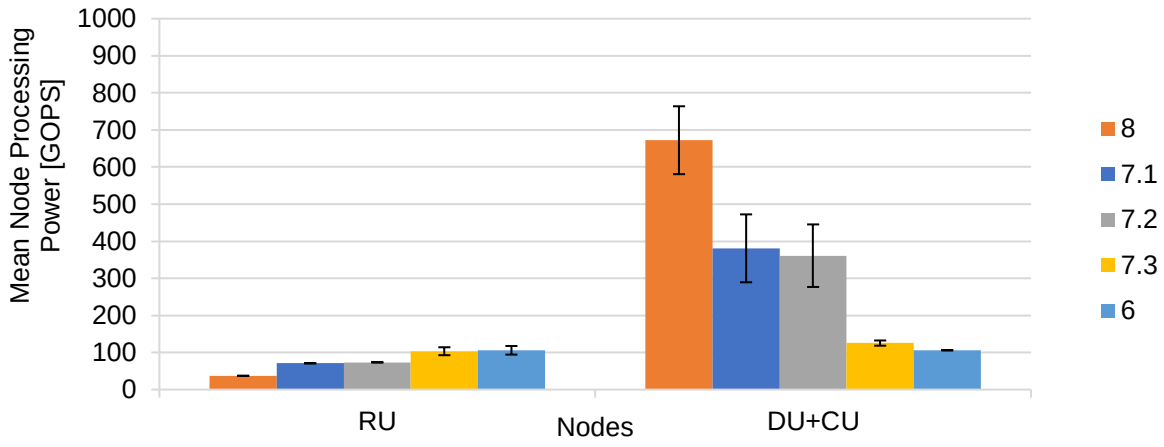


Figure 4.7. Mean processing power on the network nodes in different splitting options RU-DU+CU architecture on DL.

The transition from options 7.2 to 7.3 has a 65.2% variation, and this high variation between options appears because in this transition there is a variation of two BS functions - the MIMO coding/decoding and the baseband modulation/demodulation -, when in all the other splitting option transitions only one BS function is sent from one node to the other. Finally, one verifies that the confidence interval changes with the splitting option, which is explained by the fact that the BS functions of MIMO encoder, baseband modulation, and channel coding depend on the load of the node, so when these functions are assigned to the RU (i.e. option 7.3 and 6) the processing power on the node changes throughout the day, and when these functions are in the DU+CU node (i.e. option 8, 7.1 and 7.2), the standard deviation of the processing power increases on the DU+CU node.

Regarding the RU-DU-CU architecture, the CU node has low processing power requirements since the only BS function assigned to the CU is the PDCP BS function. The splitting option between the DU and CU for all FH splitting options is option 2, so the variation on the CU processing power that is illustrated in Figure 4.8 happens because, if an RU node does not have the capability to connect to a DU node, the RU is connected to the closest CU node. This architecture achieves a 28.5% reduction of processing power between the DU+CU node to the DU one and 83.9% reduction compared with the DU+CU and the CU processing power. This architecture can be an alternative to offload CU processing requirements, instead of increasing the number of BS functions in the RU nodes.

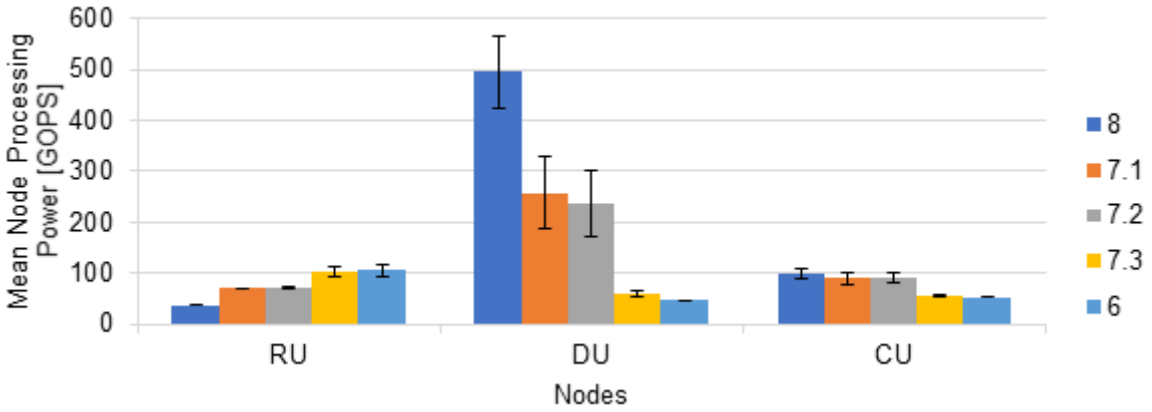


Figure 4.8. Mean processing power on the network nodes in different splitting options RU-DU-CU architecture on DL.

Figure 4.9 presents the processing power required on the RU for all the splitting option possibilities, from where one can conclude that 89.1% of the processing power on the nodes is assigned on the lower splitting option (i.e. physical layer). As explained before, the more significant increments on the processing power of the RU occur on the transition from option 8 to option 7.1, with a 47.6% increase, and the transition from options 7.2 to 7.3, with a 28.9% increase, so one concludes that, regarding the processing power on the node, the best splitting options to be implemented in the FH network are 8, 7.1, and 7.2, but since option 8 assigns an input throughput on the CU node significantly higher than the other options - 94% variation between 8 and 7.1 -, it should not be considered for the implementation of the splitting option in the nodes of the network.

In the UL scenario, the results for the processing power on the nodes are similar to the DL one. The RU node has a mean variation of 0.67% increase on the processing power, which is not noticeable on the

overall performance of the network, and the CU node has a 10.4% increase on the 7.2 and 7.3 splitting option, but a reduction of 14% of the processing power on splitting option 6.

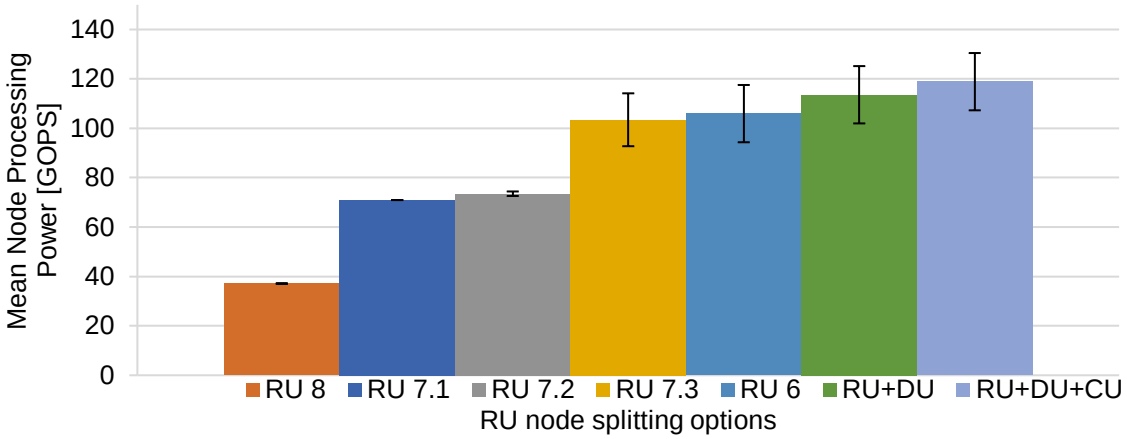


Figure 4.9. Mean processing power on the network RU nodes in different splitting options on DL.

4.4 Latency Analysis

In the following section, one analyses the latency impact on the network, being divided into five subsections in order to analyse the output parameters related to latency for several different variations of the input parameters. First, the impact of the chosen network architecture on the distance of the network is studied, since the distance is directly related to the link latency of the network. Next, the latency impact with the difference splitting architectures is analysed, considering a constant processing capacity on the nodes. Subsection 4.4.3 evaluates the output latency of the network for different latency requirements on the input of the model, using an input multiplier of the processing capacity of the nodes. Finally, the last subsection studies the response of the network latency to the variation of users on the cell sites, looking into the impact of the usage and penetration ratio and the impact of the 5G services users, such as eMBB, mMTC, and URLLC users.

4.4.1 Distance Analysis

This subsection studies the network distance for the different architecture scenarios. The network distance analysis is important since the link latency of the network represents the minimum physical latency of the network architectures, or the latency of the network without the processing delay due to traffic queuing and processing delays.

In the reference scenario, the average total network distance is 88 km. In this case, there is an FH with a maximum distance of 10 km between RU and CU, and a BH to the CN. The RU-DU-CU architecture has an additional MH with a maximum distance of 40 km, but it is worth noticing that the total network distance is reduced 10.8% than in the reference Minho scenario, this appears since, with the new DU

nodes, the FH distance is substantially reduced and, in the MH link connections, the model follows an algorithm to minimise link distance instead of a balancing algorithm used on the FH connections, so even though there are 3 links in this architecture, the total network distance is lower. It is important to remember that the first link connection of the model uses a balancing algorithm in order to balance the traffic on the network, and the second or third link connections use a minimise delay algorithm.

The highest network distance is in the RU+DU-CU architecture. In this case, there is an MH with 40 km maximum distance and, since the MH is the first connection of the network, it is using a balancing algorithm. There is an increment of 48.1% total RU to CN distance compared with the reference scenario. Finally, all collocated nodes achieve a 5.2% reduction compared with the reference scenario, because, in this case, there is only the BH, so the RU node is directly connected to the CN.

Figure 4.10 depicts the expected results that the most of RU+DU+CU architecture links are below 30 km mainly because of the direct connection between the cell site and the CN near Porto. The RU-DU+CU and RU-DU-CU architectures have a more balance distance between the nodes since the RU on the network are first connected to the aggregation node, increasing the minimum network distance between the site and the core. Finally, since the RU+DU-CU architecture has an MH instead of an FH, the network distance increases due to the increment on the allowed link distance from 10 km to 40 km.

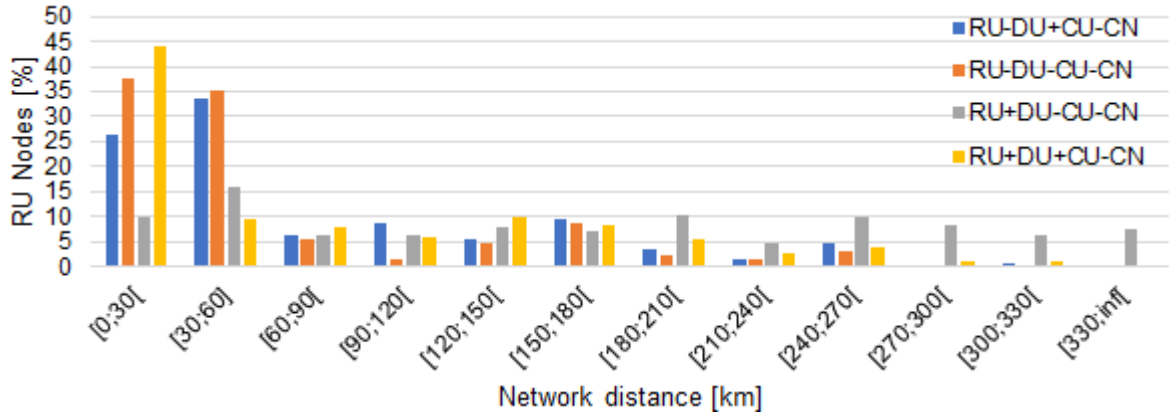


Figure 4.10. Network distance variation for all architecture options.

4.4.2 Impact of the architecture

This subsection analyses the reference scenario latency impact considering a constant processing power on the nodes for all splitting options. The processing power values considered are from splitting option 7.1, while this analysis focuses on the processing delay parameters, i.e., GOPS delay, which is associated with the processing of the BS functions, and queuing delay, which is proportional to the input traffic in the nodes. Figure 4.11 illustrates the results, where one can verify that, for a constant processing capacity, option 8 has 45.4% higher total latency than option 7.1. This variation is mainly due to the queuing delay increase in this option, since the throughput on the CU node in option 8 is 94% higher than in option 7.1. For the higher splitting options, the queuing delay is nearly constant and processing delay reduction is achieved due to the BS functions distribution on the nodes. The transition from the splitting options 7.2 to 7.1 has a 58.1% reduction in the mean GOPS delay.

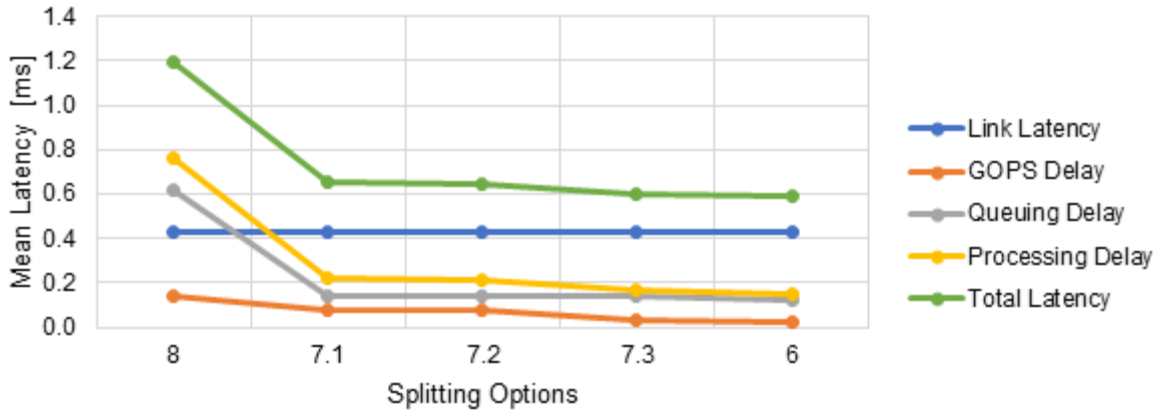


Figure 4.11. Mean network latency for RU-DU+CU architecture with fix processing power.

The standard variation of the total network latency in the network increases with higher splitting options since the resources on the network are assigned closer to the user. The network performance becomes more load dependent: on splitting option 8 the standard deviation is 47% of the mean values, and in the option 6 it achieves 70% of the mean latency value, Figure 4.12.

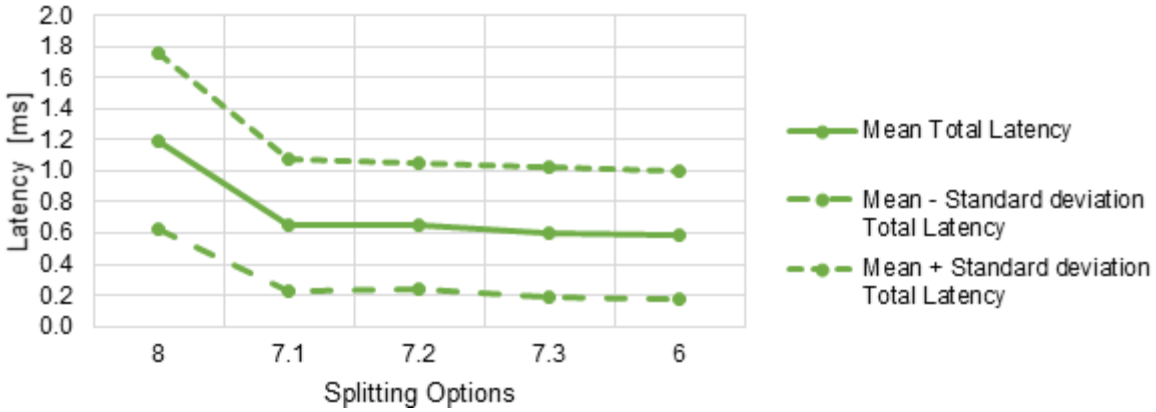


Figure 4.12. Total network latency for RU-DU+CU architecture with fix processing power.

4.4.3 Impact of maximum latency

The influence of the maximum latency allowed in the network is an important perspective to analyse in order to design the network to deal with all latency demands from the different use cases. In this study it is considered a scenario using a 7.1 spiting option on a RU-DU+CU architecture.

Figure 4.13 shows the mean network latency in the network for different maximum latency demands. In order to support lower maximum latencies, the network is forced to use MEC nodes in order to reduce the network distance. It is important to know that only 5G use cases can be processed in the MEC nodes without going to the CN node.

One can clearly observe that the network reaches a maximum total latency around 2 ms, and this happens when the network no longer needs MEC nodes to support the latency demands. The minimum value for the maximum latency value that the network can support is 0.72 ms, due to the restriction of the minimum processing delays when the network is heavily loaded, like in the evenings, so in order to reduce that processing latency it is necessary to provide more node capacity in the network.

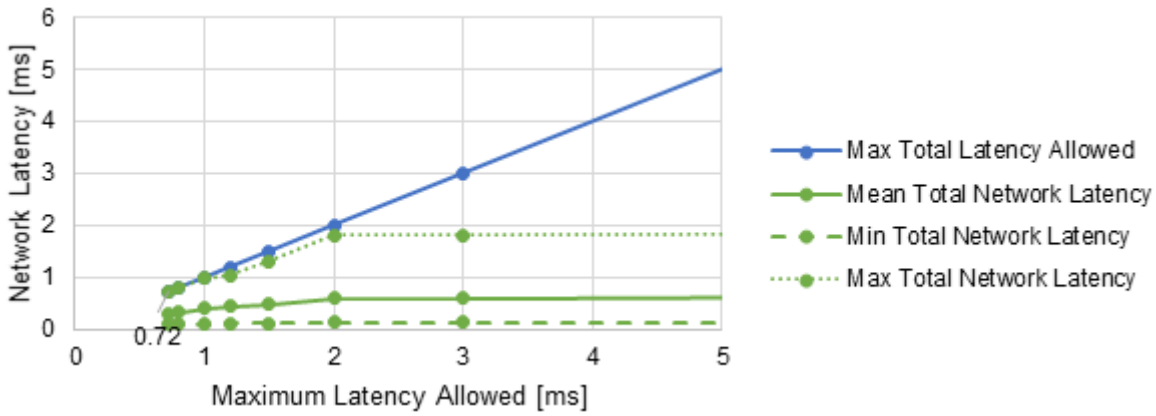


Figure 4.13. Network Latency variation with maximum network latency.

4.4.4 Impact of the node processing capacity

The nodes processing power is an important factor in the latency analysis. Higher processing capacities provide lower network latency, but, on the other hand, the cost of the network nodes increases, so it is important to balance the processing capacity of the BS functions with the maximum throughput in the nodes in order to find a middle point between network performance and its cost.

This subsection studies the network latency performance using the RU-DU+CU architecture and splitting option 7.1, changing the input processing power in the nodes, so the study focuses on the processing delay in the nodes, Figure 4.14.

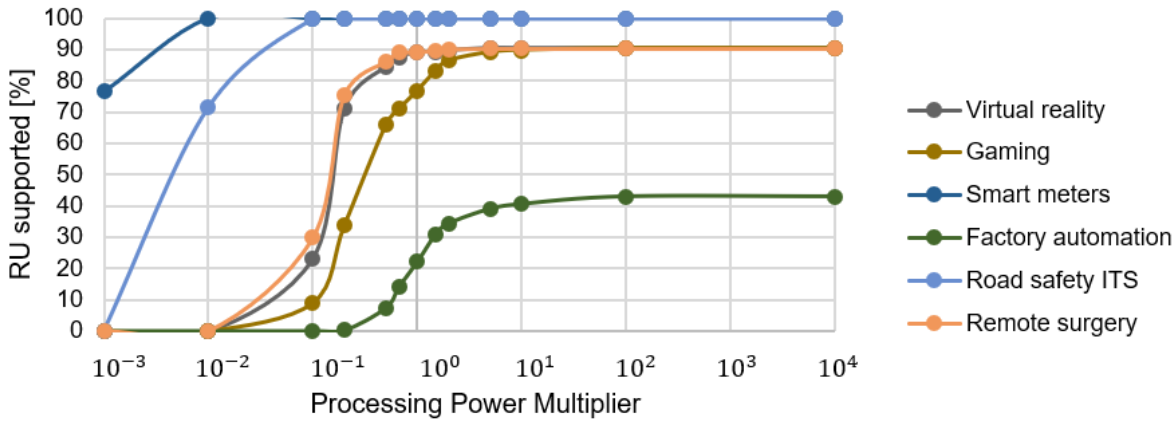


Figure 4.14. RU use cases coverage for RU-DU+CU architecture with variable processing power.

When analysing the percentage of RU coverage on the 5G use cases, one can conclude that the road safety ITS use cases have low processing power requirements, achieving 100% coverage using 10% of the reference scenario processing power. For the reference scenario, the remote surgery and virtual reality applications coverage stabilise on 90%, which means to increase these use cases cover the network needs to apply a network with MECs. In the real-time gaming use cases, the processing power on the nodes should be 10 times higher than the reference scenario to achieve the same coverage probability as the other eMBB use cases, since the gaming use case has lower priority level than the VR use case, so if the network nodes do have lower processing power the VR use case overloads the

network and the performance of the gaming use case is drastically affected due to its high resource requirements.

4.4.5 Impact of the users

This subsection analyses the impact of the number of users on the latency of the network. First, it studies the behaviour of the network as a function as the usage ratio and the penetration ratio - in this case, the number of users is uniformly changed for all use cases. Next, one analyses the impact that increasing the number of users of 5G specific services brings to the overall network: first, the impact of the eMBB users, then, the number of mMTC devices and, finally, the impact of URLLC connected devices on the cell site. Additional information related to this section is presented in Annex K.

The number of users in the site is directly proportional to the traffic produced in the network so this subsection focuses on the queuing delay parameter that is related to the throughput in the nodes. The 7.1 splitting option is considered in the reference scenario, so the link latency and the GOPS processing delay are constant in this subsection analysis.

First, as the usage ratio and penetration ratio variation on the input parameters on the network is analysed, it is worth remembering that the reference scenario considers a usage ratio of 10% and the penetration ratio 30%. Figure 4.15 illustrates the results of the simulation presenting the queuing and processing delays, and total latency.

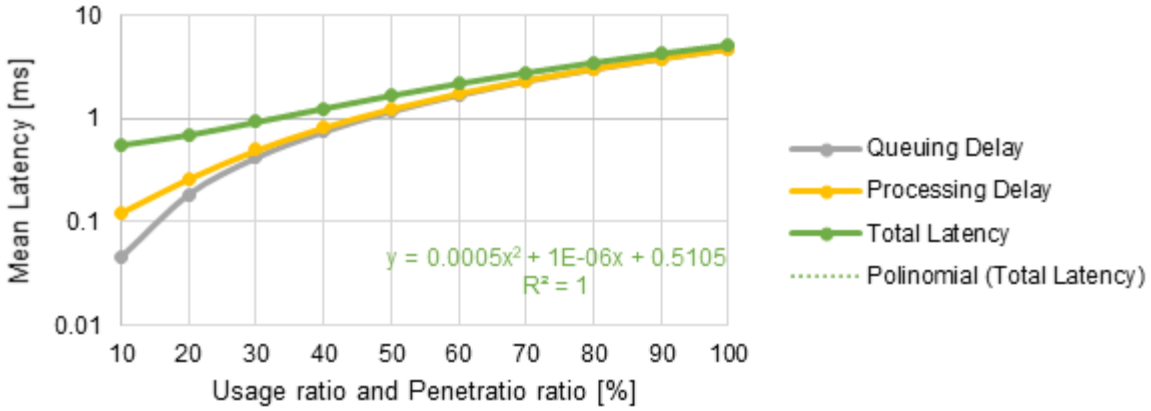


Figure 4.15. Mean network latency with variable usage and penetration ratio.

In the reference scenario, the average queuing delay is already 64.5% of the total processing time, which means that normally the majority of the processing time is due to queuing delays. For a scenario with 30% usage and penetration ratio, the average network latency is almost 1 ms and, as verified before, for an average network time of 1 ms the network cannot support eMBB services like VR and real-time gaming. This leads to conclude that increasing the number of users on the cell has a strong impact on network latency, reaching an average 25% increase on the network latency, for a 10% variation on the usage and penetration ratio.

Figure 4.16 illustrates the coverage of the 5G use cases when changing the number of eMBB users. Since the eMBB services use a lot of network resources, eMBB is very affected by the increase of users.

Doubling the number of eMBB users reduces by 11% the RUs that support real-time gaming. The virtual reality use case is more robust to the increment of users, since the priority level of this use case is higher than gaming. It is worth noticing that there is no impact on the URLLC services, since the priority level of these use cases is higher than the eMBB one.

Regarding the number of RUs that support the 5G use cases, the coverage does not change with the number of mMTC users, since the priority level of the service is very low. When increasing the number of devices more than 1 million times, the reference scenario starts to reduce the coverage of the RU, but as foreseen with the implementation of smart cities in the near future, this number of devices connected to the network will not be achieved.

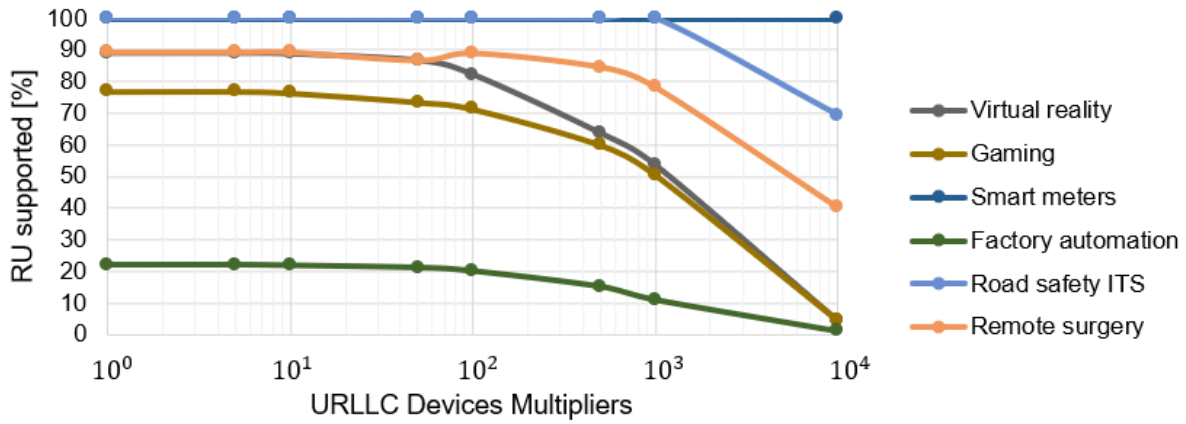


Figure 4.16. RU use cases coverage for RU-DU+CU architecture with variable eMBB users.

The final simulation on the network latency analysis is the impact of URLLC users connected to the network. URLLC services are characterised by a very low latency requirement (i.e. 1 ms to 10 ms). In this case, the reliability of the service is extremely important to the QoS of the use cases. An application like the road safety ITS requires high reliability from the network, so it is important that the service is supported on all the nodes of the network with high reliability.

When considering the coverage of the user cases throughout the network, one concludes that the impact of URLLC users is uniform for all use cases. Increasing the number of devices 100 times leads to a decrease on the performance of eMBB services and the factory automation service, but it is worth noticing that Figure 4.17 shows that the remote surgery use cases support only starts to decrease on 500 times the number of users. This can be explained, firstly, because the reference number of users of remote surgery is the lowest of all use cases and, secondly, because the remote surgery priority level is very high so the resources of the network prioritise this use case. Road safety ITS is an important use case of URLLC services, since it has a network latency requirement of 10 ms and a very high priority level. The network supports one 1 000 the reference value of users before coverage starts to drop, but it is noticeable that these values do not account for a safety margin, which reduces the network capacity to support these use cases.

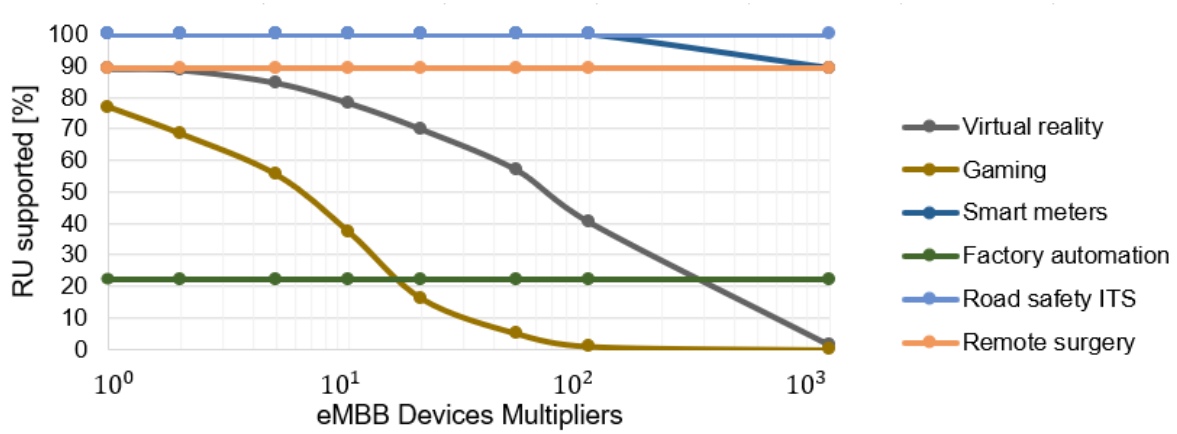


Figure 4.17. RU use cases coverage for RU-DU+CU architecture with variable URLLC users.

4.5 Analysis of the Implementation of MEC

This section analyses the MEC impact on the overall network performance, using the reference scenario (i.e. splitting option 7.1, RU-DU+CU architecture). In this architecture, new nodes called MEC have been implemented. The main purpose of the MEC nodes is to reduce network distance allowing the network to achieve lower network latencies to support the new 5G use cases. Instead of the network signal being routed to the CN, which can be very far from the user, MEC nodes are implemented to reduce network distance, where the information is processed in the MEC without needing to go to the CN node. It is considered that only 5G use cases are routed to the MEC nodes since 4G use cases are already routed to the CN network and, in some cases, it is required to access an external network. MEC nodes are essential to support the new 5G use cases, since, as concluded in Subsection 4.4.4, reducing the processing delay in the nodes is not enough to achieve a full coverage of the use cases.

This section analyses the different output parameters of the network for different MEC nodes architectures, and it considers that the MEC can only be implemented on a CU position, so MEC nodes can vary from 0% to 100% of CU nodes converted to MEC ones.

Regarding the centralisation gain, this section analysis the output on the BH link changing the number of MEC nodes in the C-RAN. It is possible to conclude that the implementation of MEC nodes does not provide a significant variation on the gain of the network compared with the CN scenario, since the splitting option between the CU and the CN are the same as the CU to MEC.

The next analysis measures the impact of MEC nodes on the network distance. Figure 4.18 illustrates the results, from where one can conclude, as mentioned before, that by implementing more than 5 MEC nods in the Minho scenario (i.e. around 10% of CU nodes), the total network latency impact does not compensate for the implementation costs of MEC. Considering the 5 MEC nodes scenario, the network latency is reduced by 45.3% compared with an architecture with no MEC nodes.

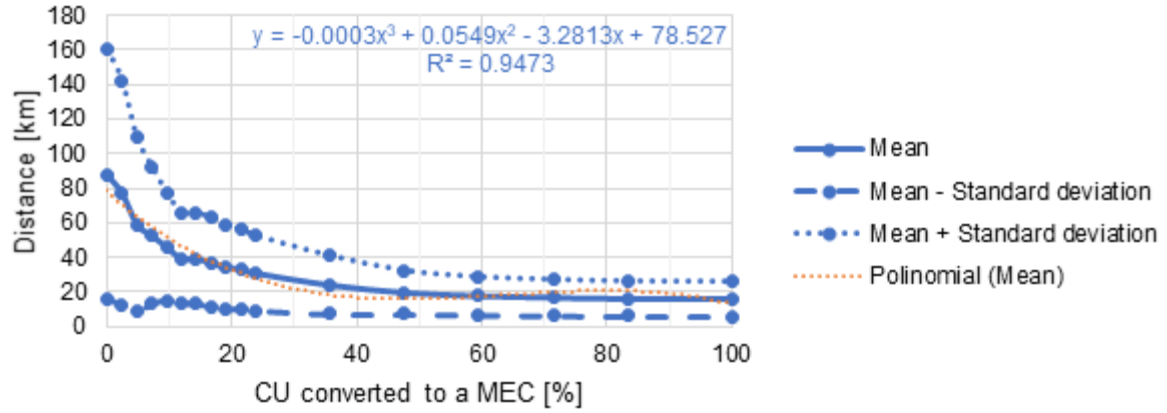


Figure 4.18. Network distance for RU-DU+CU architecture with a variable number of MEC nodes.

Regarding the use cases coverage in the network, one presents its dependence with the implementation of MEC nodes, since the 5G use cases have different priority levels, having the same latency requirements like the eMBB services, which does not mean that the use cases (i.e. Virtual Reality and Real-Time Gaming) will need the same network architecture. The results are presented in Figure 4.19. First, one verifies the previous conclusion that, for 5 MECs on the network, without considering the factory automation cases that have an extremely lower latency demand, all the network use cases considered are supported on more than 98% the RU on the network, which is extremely important since VR and, especially, real-time gaming experience are services that need to be available for a wide area on the map. For factory automation, since it is more region specific, it is not relevant that the network supports this use cases for all RUs on the network, but it is worth noticing that, from a zero MEC nodes scenario to a 5 MEC nodes one, the coverage of this use case increases 51.5%, covering 45.6% of the RUs on the network.

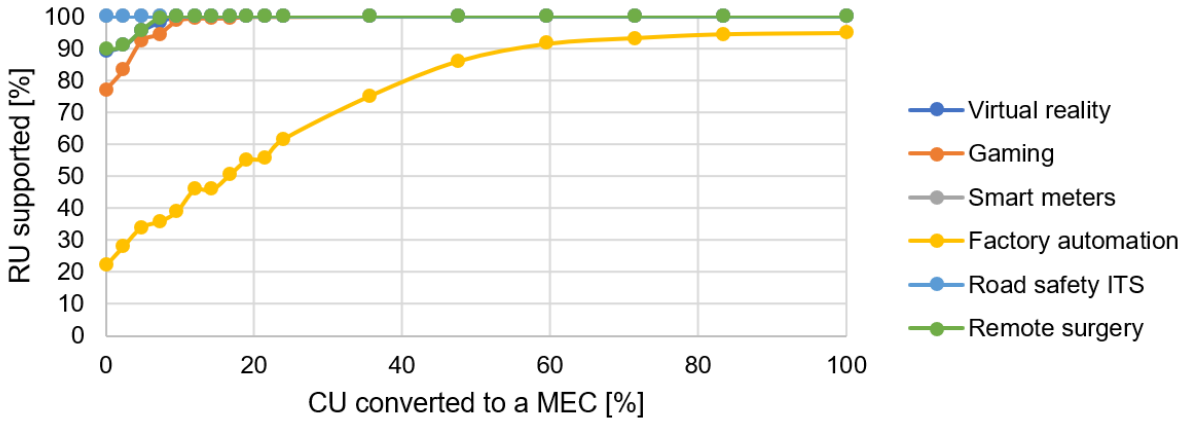


Figure 4.19. RU use cases coverage for RU-DU+CU architecture with a variable number of MECs.

In the previous analysis, the network cannot achieve full coverage using an RU-DU+CU architecture for the factory automation use case, so one considers an independent study using a RU+DU+CU architecture, since this architecture has the smallest network distance due to only having a BH connection and a lower node processing delay due to the existence of only two nodes. In this case, Table 4.6, the RU nodes are directly connected to the MEC ones. It is important to notice that in the RU+DU+CU architecture there are 374 CU nodes that correspond to RU ones, and for all the other

architectures there are the normal 42 CU nodes in this scenario.

Table 4.6. Required MEC nodes on the network to achieve full coverage for different use cases for different architectures.

CU converted to a MEC [%]	RU-DU+CU	RU-DU-CU	RU+DU-CU	RU+DU+CU
Virtual Reality	9.52	9.52	40.48	1.07
Gaming	38.10	Impossible	Impossible	4.28
Factory Automation	Impossible	Impossible	Impossible	12.30
ITS	0.00	0.00	0.00	0.00
Remote Surgery	9.52	9.52	42.86	1.07

Due to the 0.25 ms latency requirements of factory automation, assuming the reference scenario processing power, the MEC node needs to be located at a maximum of 11.5 km from the RU node that is covering the factory, on a fibre link connection, and at a maximum of 19 km on a microwave link.

4.6 Cost Analysis

This section analyses the cost of the different architecture options of the network, since CAPEX and OPEC are one of the most important parameters when a network is being deployed.

One considers relative values of CAPEX and OPEX concerning the reference scenario. Regarding CAPEX, Figure 4.20, it is taken into account the cost of implementation of the links between nodes and of the C-RAN nodes. Concerning OPEX, Figure 4.21, it is taken into account the cost of energy, rent, and maintenance of the network per year. This analysis considers the scenario where the processing capacity of the nodes is adjusted to the demands of the different splitting options.

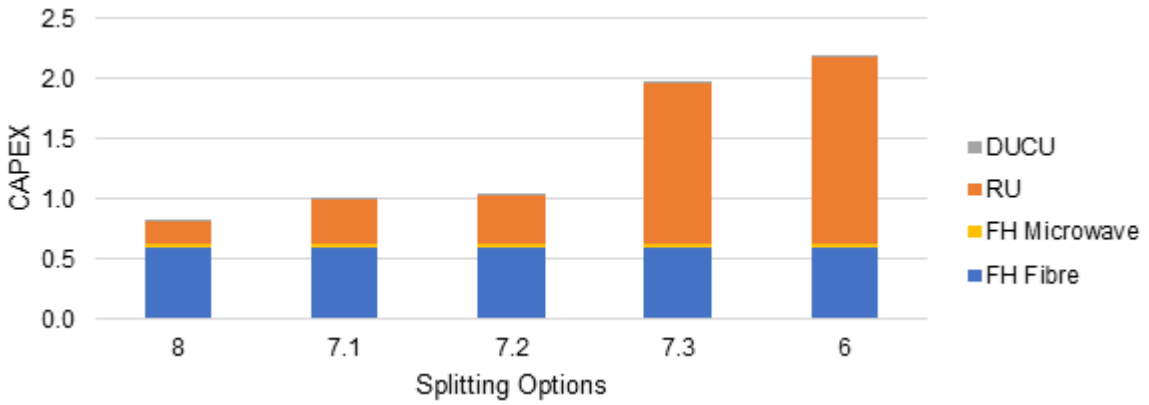


Figure 4.20. CAPEX for RU-DU+CU architecture with different splitting options.

When analysing the RU-DU+CU architecture, one can verify that increasing the number of BS functions in RU nodes increases CAPEX, since the network does not benefit from centralisation. From splitting option 8 to option 7.1, CAPEX increases 21%, and it is worth remembering that, from splitting options 8 to 7.1, the network throughput in the CU nodes is drastically reduced. Changing the splitting options from 7.2 to 7.3 increases CAPEX in 47% due to the high increment of the required processing capacity

on the RU nodes. Regarding OPEX, the results follow the same variation with the different splitting options, mainly due to the increase of maintenance cost of the RU nodes. It is noticeable that the rent does not change with the splitting options since for all splitting options the nodes have the same area.

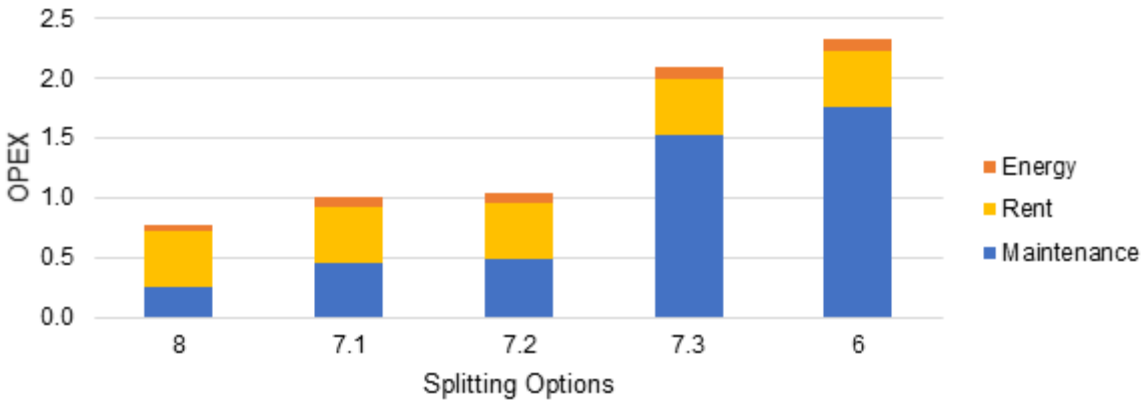


Figure 4.21. OPEX for the RU-DU+CU architecture with different splitting options.

4.7 Analysis of Portugal Scenario

The following section analyses the output parameters of the model applied to the Portugal scenario. First, compared with the Minho scenario, there are 7.3 times more RU nodes in the scenario. But in Portugal 53% of RUs are in rural areas, compared with 42% when considering Minho. The differences in the node density between the two scenarios consequently impact the output parameters of the model.

First, the output values of the central gain from the Portugal scenario are illustrated in Table 4.7.

Table 4.7. DL Centralisation Gain for all network architectures.

Central Gain		Throughput/Traffic Aggregation Gain	GOPS Aggregation Gain	Process Gain
FH	RU-DU-CU	16.68	2.26	0.307
	RU-DU+CU	16.74	1.94	0.340
MH	RU-DU-CU	2.14	4.13	0.195
	RU+DU-CU	40.40	19.84	0.048
BH	RU-DU-CU	1.00	0.80	0.556
	RU-DU+CU	2.45	3.83	0.207
	RU+DU-CU	1.00	0.53	0.653
	RU+DU+CU	40.55	11.07	0.083

When analysing the centralisation gain on the network, it is noticeable that there is an overall reduction on the aggregation gain, which is expected. Since the node density is lower, the benefits achieved by aggregating more RU nodes to the aggregator node are smaller. The load in the nodes is, on average, reduced since the traffic generated on rural nodes is lower than the traffic in DU nodes. Therefore, one verifies that the process gain in the network is reduced compared with the Minho scenario. Nevertheless, the trend amongst the different architectures remains with the same behaviour as in the Minho scenario.

Another interesting analysis is the impact of the network architecture on the network delay, illustrated in Figure 4.22. Firstly, it is important to notice that the processing latency is almost constant throughout the different architectures, since an adaptative processing capacity is considered for the different architectures, having a small increase of 16.8% on the processing time of the RU-DU-CU, mainly because the information on that architecture needs to pass through three nodes. The major impact on the total network latency is the latency in the links. The RU+DU-CU architecture continues to have the largest average network latency, but one notices that the advantages of using an RU-DU-CU architecture are less noticeable in the Portugal scenario mainly because the south and interior of Portugal have much lower nodes density so the benefits of offload traffic in the urban scenario are undermined by an average latency that is highly increased by the south and interior of Portugal. All the network architectures have a mean latency higher than 1 ms, so one can conclude that to support 5G use cases it is absolutely necessary to implement new MEC nodes in the network.

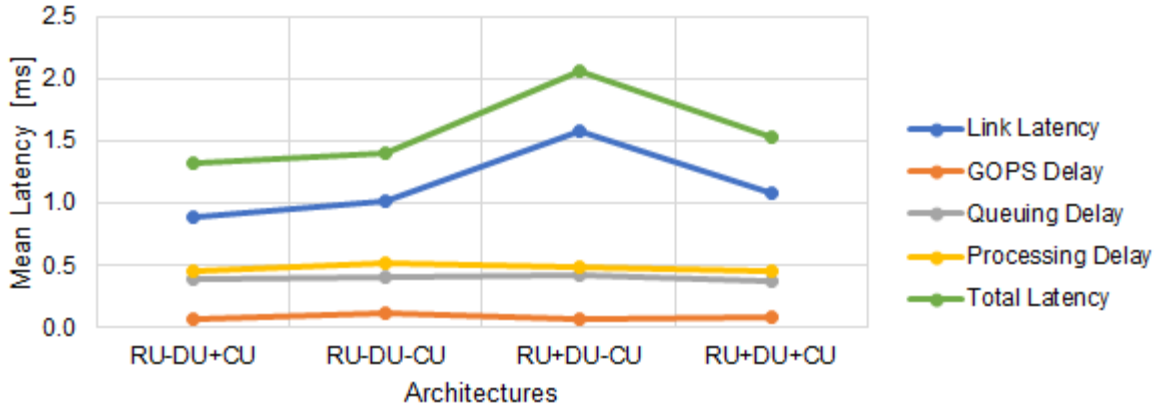


Figure 4.22. Mean Latency comparison between all network options with adjustable processing power.

Taken into account the possibility of implementing MEC nodes in the network, Figure 4.23 presents the impact on the total network latency of the network considering different restrictions on the maximum latency allowed in the network.

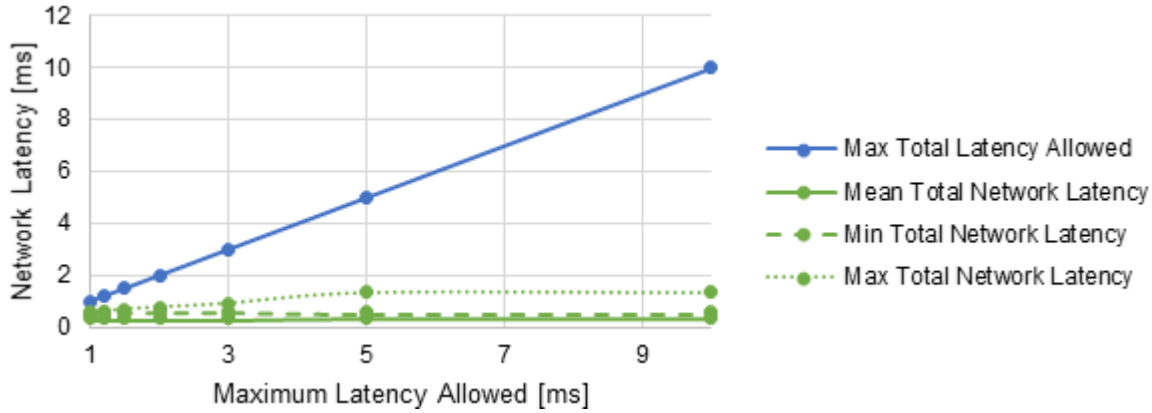


Figure 4.23. Network Latency variation with maximum network latency on Portugal scenario.

To support a 1 ms maximum latency in this architecture the network needs 96.5% of CUs converted to MEC nodes. Increasing the allowed maximum latency from 1 ms to 1.2 ms decreases 50% of the required MEC nodes to support the network, and one can clearly see that the average network latency

does not suffer a major impact when one changes the maximum latency, because the maximum latency on the network occurs due to the low node density in the interior areas of Portugal. This being said, with a high distance between RU and CU nodes, the utilisation of a RU+DU+CU architecture should be considered for the nodes with low node density.

After analysing the impact of the network architecture on the latency of the network, it is important to understand the investment that needs to be done considering the different network architectures in the Portugal scenario, and comparing the results with the Minho scenario presented in Section 4.6.

The first conclusion is that the percentage of CAPEX allocated to the implementation of network links is much higher than in the Minho scenario, which is expected since the node density is lower, so the average network distance is higher. The increment of 6 and 14 times the CAPEX on the RU+DU-CU and RU+DU+CU architectures, respectively, is explained, since those architectures do not have FH, so it is not possible to implement microwave links in the rural nodes due to the link distance restrictions and characteristics of those links. Since the Portugal scenario has a percentage of rural nodes higher than Minho, the implementation of higher splitting option architectures has a bigger impact on CAPEX. On the other hand, if it is considered that network links are already implemented, and consequently considering the link implementation cost equal to zero, one verifies that assuming an RU-DU-CU architecture has less than 1% impact on the total node cost, while the RU+DU-CU architecture brings a 26.5% increase on the node cost, and the RU+DU+CU a 30% one. Figure 4.24 illustrates CAPEX results considering a relative value with the RU-DU+CU architecture as a reference value.

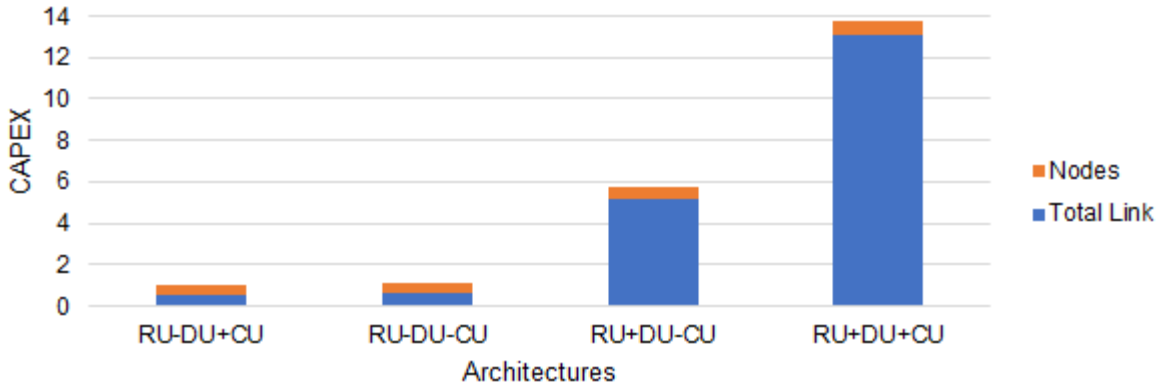


Figure 4.24. CAPEX comparison between all architectures.

The main purpose of the implementation of MEC nodes is to reduce link latency and not the processing delay, since the CN has higher processing capacities than the MEC node. Converting 10% of CU nodes to MEC nodes leads to a 53.5% reduction on the mean network distance, which consequently reduces 36.2% of the total network delay. It is important to mention that the first MEC implemented on the network reduces the confidence interval of the network distance and consequently the network's delay. This occurs since the maximum distance of the network, due to the south and interior of Portugal, are primarily reduced on the implementation of the first MEC nodes.

Even with the implementation of MEC nodes, in the gaming use case, due its high usage of network resources and the 1 ms of latency requirement, and in the factory automation use case, due to its 0.25 ms latency demand, full coverage on the Portugal scenario cannot be achieved using the RU-

DU+CU architecture, as one can observe in Figure 4.25.

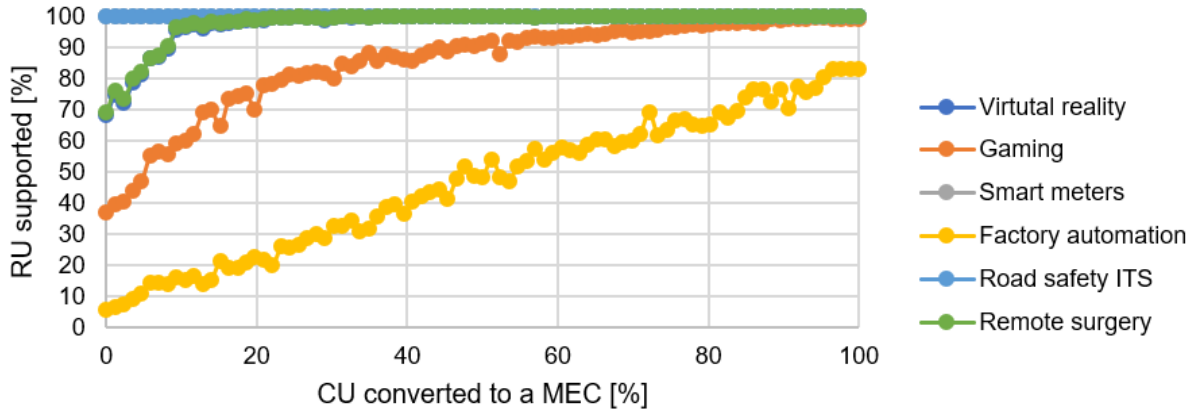


Figure 4.25. RU use cases coverage for RU-DU+CU architecture with a variable number of MEC nodes on Portugal scenario.

To understand how to support those services, one must analyse the required percentage of CU nodes that need to be converted into MEC nodes in other to support all 5G use cases, and the results are presented in Table 4.8.

Table 4.8. Required MEC nodes on the network to achieve full coverage for different use cases for different architectures.

CU converted to a MEC [%]	RU-DU+CU	RU-DU-CU	RU+DU-CU	RU+DU+CU
Virtual Reality	38.37	56.98	94.19	1.71
Gaming	Impossible	Impossible	Impossible	45.52
Factory Automation	Impossible	Impossible	Impossible	36.52
ITS	0	0	0	0
Remote Surgery	30.23	56.98	86.05	1.63

As expected, the only architecture capable to support all use cases is the architecture of all collocated nodes, since the BS is directly connected to the MEC node. It is interesting to notice that even though gaming can allow a higher network latency, more MEC nodes are needed to support the gaming use case than the factory automation one, because the factory automation priority level is much higher than the gaming one, so the queuing delay on the service is lower, and consequently the network delay of the factory automation is lower than for gaming. The gaming use case has a much more difficult use case to support than the VR one, since the priority level for gaming is lower than for VR and, since the VR service is an eMBB service, it uses a lot of resources of the network, drastically increasing the queuing delay on the gaming use case, which is the use case considered in this study that uses the most resources from the network. Therefore, it is safe to conclude that to support the eMBB services it is necessary to increase node processing capacity.

4.8 Analysis of the scenario with Input Network Latency

This study only focus on the network latency of the network since the real network latency is the only input data available in this simulation, where the main purpose is to analyse the performance of the network to cover 5G use cases with the implementation of MEC nodes, without creating any new network links between nodes.

The scenario is tested for the 5 main latency demands of the network use cases. The 100 ms maximum E2E latency represents the 4G latency requirement, where the latency demand is supported without any MEC node. When considering the ITS use case, with a 10 ms maximum latency, the cell sites in the Madeira and Azores required MEC nodes in the aggregation nodes due to the high propagation delay to the CN nodes located in Lisbon and Porto. The average E2E network latency on the sites of the island is 21.5 ms, since the fibre distance to the core is around 2 000 km, and the signal propagation speed is 200 km/ms.

When considering the low latency demands use cases, it is clearly observed in Figure 4.26 that the network cannot support those use cases assuming the processing delay currently produced by the nodes, and reducing the allowed network latency from 10ms in the north region of Portugal drastically increases the number of MEC nodes required in the network and the percentage of RUs covered by the network drops, since the main point of the MEC nodes is reducing propagation delay by reducing network distance. If the majority of the network distance is due to the processing delay on the nodes, the introduction of MEC nodes does not bring a strong improvement on the performance of the network.

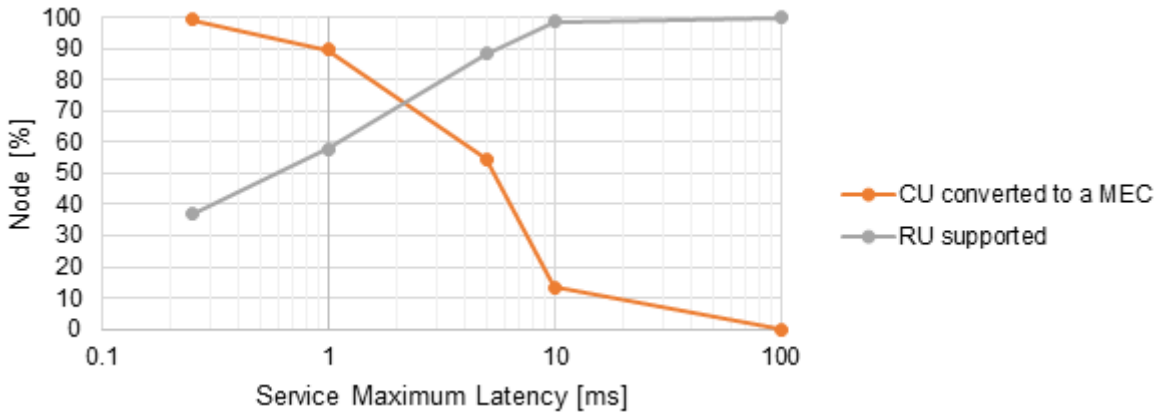


Figure 4.26. RU coverage and number of MEC nodes variation with maximum network latency on the north of Portugal.

From Figure 4.27, one verifies that, in the current state of the network, the processing delay in the nodes is 75.5% of the total E2E delay. It is important to mention that the processing delay in the network is calculated by subtracting the link delay based on the distance between the nodes and the total network latency from the input parameters, so the high values of the processing delay can be explained since, in some cases, the input latency values were not acquired from a direct connection between the cell site and the aggregation node, including multiple unidentified links until the arrival at the node.

The high processing delay on the network leads to conclude that to support the new 5G use cases it is

necessary to increase the processing capacity on the nodes, reducing the processing and queuing delays in the nodes, or to create a more direct connection between the RU and CU nodes.

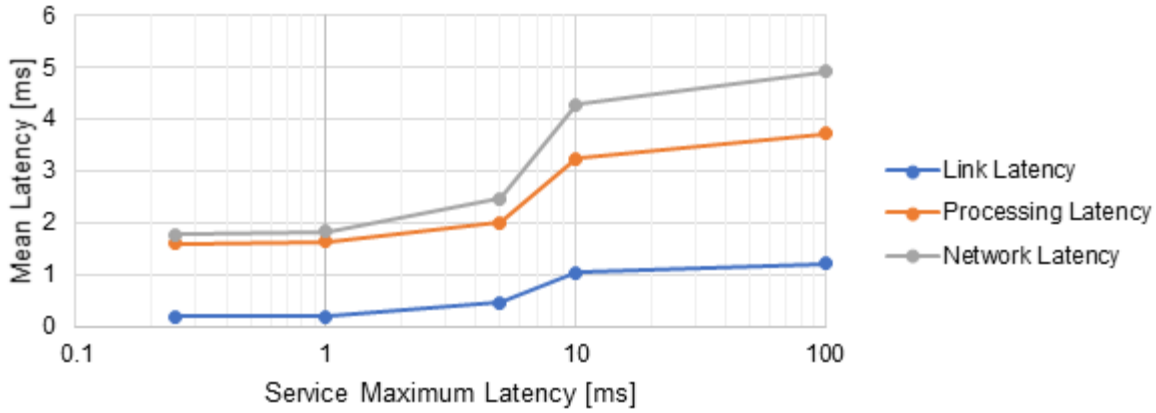


Figure 4.27. Network Latency variation with maximum network latency on the north of Portugal.

Regarding the north of Portugal, when the allowed maximum E2E latency on the network is reduced from 10 ms, the capacity for the network to support the requirements on all cell sites is reduced because of the impact of the processing delay on the nodes mainly due to queueing on the nodes. To support the 5 ms maximum latency, the network required 54.5% of CU nodes converted into MEC in the north of Portugal, only to support 88.3% of the network, mainly failing on northeast of Portugal due to the high network delay between the RU and the aggregation nodes. The 1 ms network latency required to support eMBB and URLLC services is only currently supported in 20% of the cell sites in the north of Portugal, which represent the sites near Porto, closer where the CN node is located. With the MEC nodes on the network, it is only achieved 57.9% services coverage, which means it is necessary to reduce node processing delay. To conclude the analysis of the north of Portugal, it is illustrated in **Erro! A origem da referência não foi encontrada.** the state of the network for the typically used E2E latency requirements on the main 5G use cases, representing the cell site support assuming the current state of the network, the location of the required MEC nodes on the network to support the latency requirement and, finally, the state of the RU support assuming the implementation of those MEC nodes.

In the analysis of the city of Lisbon, the current average E2E latency is 1.74 ms, so it is possible to support 10 ms of maximum latency without the implementation of any MEC node. Since the average E2E network distance is already small, around 50 km, the implementation of MEC nodes does not provide a substantial improvement of the network performance to support eMBB and URLLC services compared with the performance improvement achieved by reducing processing delay on the nodes. To support a maximum latency of 5 ms, it is required to convert 8 CU nodes into MECs, and reducing the allowed network E2E latency to 1 ms drastically increases the number of MECs in the network, converting 50 of the total 55 aggregation nodes in Lisbon just to achieve a 92% RU coverage. In the case of 0.25 ms maximum E2E latency allowed in the network, which represents the factory automation use cases, one concludes that it is required to implement a MEC node at the industrial area at a maximum distance of 10 km of the cell site that supports the factory automation use case.

Chapter 5

Conclusions

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

The main goal of this thesis was to analyse the performance of the deployment of 5G edge networking in order to understand the advantages and disadvantages of the different architecture scenarios for the multiple use cases accounted for in this work. The model was developed to have the LTE network characteristics as the input along with the use cases under study and provide results for the critical parameters that influence the performance of the network, as latency, capacity, and cost for the different network architectures of the 5G edge network.

In Chapter 1, one presents a brief overview of the current status of mobile communication systems, giving emphasis to the need of the transition from the 4G network to the 5G one. The main benefits of the 5G network capabilities are explained, as well as the motivation for the present work, considering that the study of the different C-RAN architectures is key to optimise the network for each service.

Chapter 2 gives an overview of the main background essential to better understand the thesis. The Non-Standalone architecture of 5G is presented since the first wave of 5G networks will be supported by existing 4G infrastructures. Next, one gives an overview of the new radio interface of the network, presenting a comparison between the technologies of 5G NR with the 4G radio interface. 5G is a service-oriented network, which means the resources provided by the network are adapted to the QoS and QoE demanded by each use case in real time. To understand the fundamental technologies in the implementation of a flexible network some technologies are presented, including cloud networking, SDN, virtualisation and NFV, which are essential to create a Network Slicing environment specific for the required demand. The structure of C-RAN and the advantages of this approach are described, and a subsection of the chapter is reserved to explain the basic aspects of Edge Networking on the 5G network and compare them with the existing 4G C-RAN, providing the main architecture option available on the new 5G C-RAN that is analysed in this thesis, giving emphasis to the MEC technology implemented in 5G networks. This chapter also includes a list of base station UL and DL functions, that are divided between the RU, DU and CU nodes of the 5G C-RAN. A description of the 5G services and applications is presented, dividing those into three services types: eMBB, requiring extremely high data rates, low-latency and reliable broadband access over a high user density; mMTC, requiring wireless connectivity for millions of devices worldwide, with scalable connectivity and high energy efficiency devices; URLLC, using ultra-reliable low-latency and resilient communication links primary between machines. The performance parameters demanded by the network for each specific service type are presented, which are essential to provide users with the high expected QoE and reliability required in sensitive 5G new use cases. The state of the art is present at the end, summarising the most relevant work related to this thesis.

Chapter 3 describes the model developed in the thesis. First, it offers an overview of the model introducing the input and output parameters. Secondly, the characteristics of the different 5G C-RAN architecture scenarios are illustrated, depending on the different BS functions splitting between RU, DU and CU nodes. Since there is no information on the traffic profile of 5G, a model was created to simulate 5G traffic based on the existing 4G traffic profiles and characteristics of new network services, with a detailed explanation of the computation of the input traffic used. Each output parameter of the model is presented with the relevant expressions underlying the model. One gives a detail explanation of the

implementation of the model, using illustrating flowcharts to better understand the procedure of the main tasks. At the end, the model assessment is presented, illustrating empirical test for each output parameter of the simulator.

The model has been divided into seven steps. The model starts with the computation of the traffic and throughput arriving at the RU depending on the number of users on the cell site and the use cases specifications, the traffic computation being based on 4G traffic profiles along with 5G use cases specification. Next, it considers the architecture scenario characteristics under study in order to compute the node capacity to support the different architecture scenarios, the splitting options of the model being divided into two separate considerations: one is the physical splitting characteristics among RU, DU, CU, CN and MEC nodes that identify the characteristics of the FH, MH and BH distances; another is the different FH splitting options between the physical layer of BS functions. In the case of the implementation of DU nodes, which is used to offload the CU nodes primary in dense traffic areas, or the MEC nodes that are essential to reduce network distances to support the ultra-low E2E latency, the model follows an additional step based on a machine learning algorithm called K-means used to optimise the DU or MEC implementation location on the network since those nodes do not have an already existed fix position. The K-mean algorithm identifies the required MEC on the network for a specific maximum latency allowed by the service, providing the MEC locations with the smaller combined distance between the node and the connected CU nodes assign for each MEC cluster. Depending on the architecture requirements, namely distance restriction, the model executes the aggregation process of the node. Finally, the status of the network nodes is updated, computing the load of the nodes and calculated depending on the link delay and process delay of the E2E latency.

The model considers five output parameters to analyse network performance. Latency parameters are defined in the link delay of the network, the processing delay because of the BS functions processing requirements, and the queuing delay on the nodes from the input data. The latency is the main constraint for the implementation of URLLC services with 1 ms of E2E latency, and it is the constraint that is responsible for limiting the network distance that demands the implementation of MEC nodes. Next, it introduces the processing capacity in the nodes, which depends on the BS functions assigned to each C-RAN node, this parameter being measured in GOPS. The throughput from the different use cases is defined on the input and output of the nodes for the different splitting options, this parameter being strongly related to the signal compression of the nodes, which depends on the BS functions previously executed on the signal. The centralisation gain is defined, dividing it into two main measurements: the aggregation gain compares the impact of the peak processing capacity on the nodes with the peak processing capacity of the aggregation node of the different splitting options; the process gain measures the level of functions centralisation on the network. Finally, to evaluate the investment that needs to be made when developing a C-RAN architecture, one created a model to compute the CAPEX and OPEX of the network.

Chapter 4 presents the analysis of the model output parameters. Firstly, it defines the scenarios under analysis, taking into consideration RU and CU locations and the traffic information from NOS: the scenarios of Minho, Lisbon, North of Portugal and the whole of Portugal. In this study, a reference

situation was created using a RU-DU+CU architecture since it is the one used in 4G C-RAN, with a 7.1 FH splitting option in order to achieve lower FH throughput while still benefiting from centralisation. In the reference scenario it was assumed that the network nodes achieve an 80% maximum load, so the reference values of the processing capacity on the nodes are established to support this initial assumption. Regarding the aggregation process of the nodes of the network it was considered that for all architecture scenarios, the first connection made by the model would be between the nodes closest to the user (e.g. RU to CU), and a balancing algorithm is used to balance the number of nodes connected to each aggregation node in order to balance the network traffic and load of the network, and in the second or third connection (e.g. CU to CN), an algorithm is used to minimise the network delay.

With the reference scenario already defined, one evaluates the response of the output parameters with a variation of the input parameters one at a time. Regarding the network architecture, one considered the different splitting options among nodes. Regarding the implementation of new nodes, network performance was analysed with the implementation of a different number of DUs and MECs in the network. The variation of the processing power in the nodes is also considered, which is directly related to the GOPS processing delay component of the total network latency. Finally, it was analysed the impact of the number of users of each network service on the overall network performance.

First, all the output parameters for the Minho scenario were analysed, since the average running time of the program when assuming the Minho scenario is only one minute, while the running time on the Portugal scenario is around 30 minutes, making it possible to elaborate a deeper analysis with the output parameters of the Minho scenario which can then be used to evaluate the Portugal scenario since the parameters follow similar behaviours.

Regarding the centralisation gain, one concludes that a higher splitting option achieves higher aggregation gains since the signal is more compressed in the RU nodes and the data rate on the CU nodes are reduced but, on the other hand, the process gain is reduced since fewer functions are assigned to the central node, achieving less benefits from the centralisation of resources. Splitting the physical layer will achieve great traffic and throughput aggregation gains, substantially reducing the requirements of the FH link capacity compared with splitting option 8, which leads that to support the high traffic demands of the 5G network it is necessary to assign more processing power on RU nodes, even though the gains of resource centralisation are reduced.

Concerning the processing capacity of the nodes, two parameters were analysed. The first one considered the throughput on the nodes measure in Gbps, and one concludes that changing the splitting option from option 8 to a higher one greatly reduces the input throughput on the central node, since in option 8 the signal is not compressed so the throughput that arrives at the CU node is extremely high, and increasing from option 8 to option 7.1 shows an approximately 94% reduction on the throughput of the CU node. Another option to offload the central node data rate is to use DU nodes. In this case, the throughput on the DU is 25% lower and the throughput on the CU is 50% lower than in the reference scenario. In the UL scenario, the splitting option 7.1 cannot achieve the same data compression as in DL, and in this case, the best splitting option is 7.3, which achieves a 75% reduction compared with option 8. The second parameter was the required processing power on the nodes measured in GOPS.

Higher splitting options increase the processing capacity requirements on the RU node, which reduces the benefits of centralisation in the node, increasing RU complexity. Since using a splitting option 7.1 is recommended to reduce throughput in the CU node, one verifies that this option requires a 48% higher processing capacity in the RU. The processing power in the RU also substantially increases from option 7.2 to 7.3 since MIMO encoding and baseband modulation is performed in the RU (29%).

Afterwards, an analysis of the latency on the network is presented. Considering the network distance, one verifies that assuming a RU+DU-CU architecture leads to a 48% increase in network distance since the maximum MH distance is 40 km instead of 10 km on the maximum FH distance. Looking at the other network architectures, network distances are approximately similar even though the RU+DU+CU architecture only has a BH link that can bring latency benefits to the network, since the RU is directly connected to the CN. Regarding the capacity of the network to support the different 5G use cases, it is important to notice that, when using the references scenario architecture, it is not possible to achieve full coverage on the ultra-low 2E2 latency use cases, so it is necessary to introduce MEC nodes in the network. It should be noticed that URLLC services are more dependent on the network distance, so it is best to reduce network distance with the introduction of MEC nodes, since the processing delay of these services is already very low due to the high priority level. In contrast, eMBB services are more sensitive to the processing capacity in the node, primarily in the Gaming use case, since this requires a lot of network resources. The performance of eMBB services in the network benefits from increasing node capacity. One concludes that increasing the connected devices on the network due to the mMTC services does not decrease neither the latency network performance nor the required node capacity due to the low data rate requirements and low priority level of the service. Increasing the number of URLLC devices in the network affects the overall performance of the network, for all network services since these use cases have high priority levels.

In order to achieve full network coverage on 5G use cases, the implementation of MEC nodes in the network is required; in this case, the information of 5G use cases does not go to the CN nodes, being processed in the MEC node, thus reducing network latency. Considering the Minho scenario, one concludes that it is required to have at least 5 MEC nodes in order to achieve a maximum network latency below 1 ms. In the special case of the factory automation use case, due to its 0.25 ms latency requirement, it is necessary to use an RU+DU+CU architecture, creating a direct connection between RU and the MEC node with a maximum 11.5 km of fibre link.

Concerning the network cost, one considered relative values on the analysis due to the lack of detailed information on the hardware cost of the nodes since the 5G network is still in its first stages. As expected, increasing the BS functions on the RU nodes increases the overall cost of the network, since the majority of the cost of the nodes on the network comes from the RUs. From splitting option 8 to option 7.1, the initial investment in the network is 21% higher just due to the increased cost of the RU nodes. Assuming a higher splitting option will result in a substantial initial node investment increment, so these architectures should only be considered if a direct connection between the RU and the CN or MEC node is required to, for example, support the factory automation use case. The OPEX of the network follows a similar behaviour as CAPEX regarding the different network architectures, and one concludes that the

majority of the OPEX in the network is due to node maintenance, mainly the RU maintenance in high splitting architectures.

Finally, the model simulates the results for the Portugal scenario, in which case 53% of RUs are in rural areas, which, compared with only 42% in the Minho scenario, contributes to an increase of the average network latency, aggravated by the fact that the CN nodes are only in Porto and Lisbon. Because of the lower node density in Portugal, the centralisation gain on the network is reduced and the network is less balanced overall than in the Minho scenario, since a maximum processing capacity equal for all RU nodes is assumed, while on average the load of the nodes is lower.

The results in the Portugal scenario show that when considering the reference scenario without using MEC nodes in the network the average network latency is 1.33 ms, which is not acceptable to support the 5G use cases besides the Road safety ITS, that due to the 10 ms latency requirement can achieve full network coverage. However, it is impossible not to notice that this study does not take into account the high reliability required in the ITS use case.

Finally, regarding the simulation using as input for the model the network latency information provided by NOS, it is important to highlight that the latency values on the network are higher than the reference scenario ones considered in the previous simulations, mainly because the network connections are already implemented, reducing network distance optimisation, and the processing delay, since the connection between RU and CU node is not a direct link, with multiple wireless and fibre links between the nodes. The processing delay represents 75% of the initial total network delay in the north of Portugal. Furthermore, to achieve lower maximum values of latency in north of Portugal and Lisbon, although it will be required to convert the CU nodes further away from the core into MEC nodes, it is essential to reduce processing delays on the network.

Regarding future work, the present simulations will be able to achieve better accuracy on the output parameters of the model when the future 5G use cases start to roll out and have better traffic profile predictions, depending on the time of the day and location. Updated traffic profiles will optimise the load distribution and level traffic output throughout the day.

Some approximations were made on the cost model for the implementation of nodes, due to lack of information on the cost parameters for the new 5G nodes and MEC DC, so it would be interesting to consider a more in-depth cost analysis when the 5G new technology starts to be commercialised.

The present model only allows the implementation of one network architecture on the scenario per simulation. However, the model should consider a hybrid architecture, capable of changing the network and considering an adaptable solution depending on the real time information on the performance of the sites. This solution would be more accurate in representing a virtual network infrastructure based on network slicing technology.

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 a 2018a version with the most common Matlab's toolboxes.

To run the model simulator, first it is necessary to configure the Input_File.xlsx which is divided into four main reconfigurable sheets:

- Parameters – It is used to define the network parameters.
- Flags – It is used to define the different running option of the simulator.
- Use Cases – It is used to define the use cases parameters.
- Cost – It is used to define the cost parameters.

After configuring the Input_File_xlsx one should be able to run the simulator by running the script "main.m" in the Matlab_Files folder.

When the simulator finishes, the output files will be in the Output_Files folder with the output performance parameters.

A.2 Simulator Configuration

The parameters sheet allows the user to change the path of the files, required to run the simulator in a specific work station, and the parameters of the network. Table A.1 explain the input parameters of the excel sheet.

Table A.1. Network configuration parameters.

Path_Input	Specification of the Input_Files folder location.
Path_Ouput	Specification of the Output_Files folder location, where the outpufiles will be printed.
Path_Network_Info	Specification of the Network_Info location where the network scenario information is located.
Path_Matlab_Files	Specification of the input scenario file.
Configuration_File	Propagation speed on the fibre.
Propagation_Fiber	Propagation speed on the microwave link.
Propagation_MicroWave	Radius use for the computation of the node density
Percentage_of_DUs	Percentage of RU nodes that will be converted into DU nodes.
Percentage_of_MECs	Percentage of CU nodes that will be converted into MEC nodes. Used to test the network for a specific number of MEC nodes.

Maximum_Link_Radius	Specification of the maximum link distance, considering the FH, MH and BH links.
Maximum_Node_Capacity	Specification of the maximum capacity allowed in the pools, measure in GB or GOPS.
Flatness_Timestamps	Number of hours to be considered in the simulator.
Efficiency_Factor	Efficiency conversion from GOPS to Watt
Years_for_Opex	Number of years considered in the OPEX calculation.
Splitting_Option	FH splitting option used in the network architecture.
Penetration_ratio	Percentage of the penetration ratio.
Usage_ratio	Percentage of the usage ratio.
Max_Latency	Maximum latency allowed in the network for all use cases.
Throughput_multiplier	Throughput capacity of the node. This parameter can be used to adjust the queuing delay of the node.
Processing_Power_multiplier	BS processing power of the node. This parameter can be used to adjust the GOPS delay of the node.
Service_Users_multiplier	Parameter used to adjust the number of users in a specific 5G service.

The Flags sheet allows the user to change network architecture and running option for the simulator test different case studies. Table A.2 explain the input flags of the excel sheet

Table A.2. Flags configuration parameters.

Flag_CoLocatedRUDU	Boolean value to indicate if the RU and DU node are in the same location.
Flag_CoLocatedDUCU	Boolean value to indicate if the DU and CU node are in the same location.
Flag_MEC	Boolean value to indicate if it is allowed MEC nodes in the network.
Flag_UL(0)/DL(1)	Boolean value to specify the DL or UL traffic considered.
Flag_Test_Max_Latency	Boolean value equal to 1 if one wants to test a specific maximum latency of the network. The maximum latency value is specified in the parameters sheet.
Flag_Test_Distance_FH	Boolean value equal to 1 if one wants to test all the possible DU location on the FH at the same time.

Flag_DU_Reference_Scenario	Boolean value equal to 1 if one wants to assume the reference configuration of DU nodes.
Flag_Test_Services	Boolean value equal to 1 to test the network for a specific service. Only on specific use case can be tested at the same time.
Flag_Virtual_reality	Boolean value equal to 1 to test the Virtual reality use case
Flag_Gaming	Boolean value equal to 1 to test of the Gaming use case
Flag_Factory_automation	Boolean value equal to 1 to test of the Factory automation use case
Flag_ITS	Boolean value equal to 1 to test of the Road safety ITS use case
Flag_Remote_surgery	Boolean value equal to 1 to test of the Remote surgery use case

The Use cases sheet presents the values of the use cases parameters computed in the model. The first column presents the use cases names consider in the study, and the first line of the table present the different parameters consider for the specification of the use cases. The use case input table is a combination of Table 3.1 and Table 4.4.

Next, the input file presents the three cost sheets, in the first one, it is possible to configure the initial investment on the network, considering the on the initial cost of the microwave and fibre link, independent of the link distance, and the initial investment on the nodes. The RU initial cost depends on the used bandwidth of the node, but in this initial investment, it is not considered the different costs depending on the processing capacity of the nodes. The second sheet represents the variable cost of the network, this includes the cost of fibre per kilometre, the node cost depending on the processing capacity and the number of interfaces used in the nodes. The third sheet is used to specify the OPEX cost of the network, considering a percentage of the CAPEX cost of the equipment.

The following sheets are used to specify network parameters on the simulator, the values presented on the last Input_Files sheets are illustrated in Annex C and Annex D of the thesis.

A.3 Output Files

After finishing running the simulator it will be created output files in the Ouput_Files folder presenting the network output performance parameters required to analysed the simulation.

The Output folder is composed of:

- Output folders for each pool presenting an xlsx file for each node pool with the information of the index and location of each node aggregated to the aggregator node.
- Node_clusters_output.txt – Cluster configuration and traffic profile with the nodes assigned to each aggregator node.
- Network_Link_output.txt – General information about the network link and the presentation of multiple output results of the simulation run such as central gain, distance, latency, and node capacity measure in GB, Mbps and GOPS. There are on txt file for the FH, MH and BH link.
- Network_Total_output.txt – General information about the overall network specification used in the simulation, and the presentation of the output results such as total network distance, latency, and use cases performance.
- Cost_Total_output.txt – Presentation of the output costs of the network.
- Stand_Alone_output.txt – Output information about the nodes that could not be connected to an aggregator node.

A.4 Simulator of the real case scenario

The simulator was developed in a Matlab environment, it was used a 2018a version with the most common Matlab's toolboxes.

To run the simulator for the real case scenario, first it is necessary to configure the Input_File_Real.xlsx to choose the use case under analysis. It is only possible to select one use case at a time.

Secondly, it is necessary to configure the initial path and the input network path to run the simulator. One should open the "Real_Latency.m" in the Matlab_Files folder and input the right path on the path_initial variable and the name of the network file from the Portugal regions that one wants to analyse. It is possible to consider five different input network files, including the archipelago of Açores, the archipelago of Madeira, Lisbon, north of Portugal and the scenario considering all the Portugal regions.

After configuring the Input_File_Real.xlsx and the correct initial path and input network file one should be able to run the simulator by running the script "Real_Latency.m" in the Matlab_Files folder.

When the simulator finishes, the output file is the "network_Real_output.txt" that is located in the Output_Files folder with the output performance parameters.

Annex B

LTE and 5G QoS Characteristics

This annex summarises the standardise QCI and QoS classes characteristics for an LTE and 5G network.

Table B.1 Standardised QCI characteristics for LTE and 5G (based on [3GPP18]).

QCI	Resource Type	Priority	PDB [ms]	PELR	Example Services
1	GBR	2	100	10^{-2}	Conversation Voice (Live Streaming)
2		4	150	10^{-3}	Conversation Video (Buffered Streaming)
3		3	50	10^{-3}	Real-time Gaming
4		5	300	10^{-6}	Non-Conversational Video
65		0.7	75	10^{-2}	Mission Critical user plane Push To Talk voice (e.g., MCPTT)
67		1.5	100	10^{-3}	Mission Critical Video user plane
75		2.5	50	10^{-2}	V2X messages
82		1.9	10	10^{-4}	Discrete Automation
83		2.2	10	10^{-4}	Discrete Automation
84		2.4	30	10^{-5}	Intelligent Transport Systems
85		2.1	5	10^{-5}	Electricity Distribution-high voltage
5	Non-GBR	1	100	10^{-6}	IMS Signalling
6		6	300	10^{-6}	Video (Buffered Streaming), TPC-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
7		7	100	10^{-3}	Voice, Video (Live Streaming), Interactive Gaming
8		8	300	10^{-6}	Video (Buffered Streaming), TPC-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
9		9			
69		0.5	60	10^{-6}	Mission Critical delay sensitive signalling (e.g., MC-PTT signalling, MC Video signalling)
70		5.5	200	10^{-6}	Mission Critical Data (e.g. example services are the same as QCI 6
79		6.5	50	10^{-2}	Video (Buffered Streaming), TPC-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
80		6.8	10	10^{-6}	Low latency eMBB application (TCP/UDP-based), Augmented Reality

Table B.2 Characteristics of QoS classes (extracted from [Corr18]).

	Conversational	**Streaming**	**Interactive**	**Background**
Real-time	Yes	Yes	No	No
Symmetric	Yes	No	No	No
Guaranteed Rate	Yes	Yes	No	No
Delay	Minimum Fixed	Minimum Variable	Moderate Variable	High Variable
Buffer	No	Yes	Yes	Yes
Bursty	No	No	Yes	Yes
Example	Voice	video–clip	www	email

Annex C

Processing Power Reference Scenario

This annex presents auxiliary values for the calculation of the Processing Power in the nodes.

The reference numbers of the parameters along with the scaling exponents for the components used are in Table C.1 and Table C.2.

Reference scenario:

- Bandwidth: 20 MHZ.
- Single antenna and single stream.
- Modulation: 64-QAM.
- No channel coding.
- Load of 100%.
- Quantisation of 24 bits.

Table C.1. Reference processing power for the components (adapted from [DDLo15]).

Operation name	DL [GOPS]	UL [GOPS]
RF front-end	7	7
OFDM Modulation	5	5
Antenna and resource mapping/demapping	2	2
MIMO encoding/decoding	1.3	3.3
Baseband Modulation	1.3	2.7
Channel Coding	1.3	8
MAC	2.7	3
RLC	2.7	1
PDCP	2	2
Core Network interface	8	5.3

Table C.2. Scaling exponents for the processing capacity (extracted from [DDLo15]).

Operation name	e1	e2	e3	e4	e5	e6
RF front-end	1	0	1	0	0	1.2
OFDM Modulation	1.2	0	1	0	0	1.2
Antenna and resource mapping/demapping	1	1.5	0	1	1	1.2
MIMO encoding/decoding	1	0	2	1	1	1.2
Baseband Modulation	1	0	1	0.5	0	1.2
Channel Coding	1	1	0	1	1	1.2
MAC	0	0	0.5	0	1	1
RLC	0	0	0.5	0	0.2	0.2
PDCP	1	0	0	0	0.2	0.2
Core Network interface	1	1	0	1	0	0

Annex D

Link Capacity Reference scenario

This annex presents auxiliary values for the calculation of the Link Capacity for the different splitting options.

Table D.1. All options link capacity (extracted from [3GPP16]).

Protocol Split option	FH BW	Reference values	Calculation
Option 1 (RRC-PDCP)	**DL: 4000 Mbps**	R_p: 150 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 2 M: 256 M_c:64	$R_{1[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right)$
	UL: 3000 Mbps	R_p: 50 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 1 M: 64 M_c:16	$R_{1[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right)$
Option 2 (PDCP-RLC)	**DL: 4016 Mbps**	R_p: 150 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 2 M: 256 M_c:64 signalling: 16	$R_{2[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right) + signaling_{[Mbps]}$
	UL: 3024 Mbps	R_p: 50 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 1 M: 64 M_c:16 signalling: 24	$R_{2[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right) + signaling_{[Mbps]}$
Option 3 (Intra-RLC)	**lower than option 2 for UL/DL**	See option 2	$R_{3[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right) + signaling_{[Mbps]}$

Option 4 (RLC-MAC)	**DL: 5226.7 Mbps**	R_p: 196 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 2 M: 256 M_c:64	$R_{4[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right)$
	UL: 4500 Mbps	R_p: 75 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 1 M: 64 M_c:16	$R_{4[Mbps]} = R_{p[Mbps]}\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right)$
Option 5 (Intra-MAC)	**DL: 5626.7 Mbps** **UL: 7140 Mbps**	See option 6	$R_{5[Mbps]} = R_{6[Mbps]}$
Option 6 (MAC-PHY)	**DL: 5626.7 Mbps**	R_p: 196 R_c: 5 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 2 M: 256 M_c:64	$R_{6[Mbps]} = (R_{p[Mbps]} + R_{c[Mbps]})\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right)$
	UL: 7140 Mbps	R_p: 75 R_c: 44 B: 100 B_c: 20 N_L: 8 $N_{L,c}$: 1 M: 64 M_c:16	$R_{6[Mbps]} = (R_{p[Mbps]} + R_{c[Mbps]})\left(\frac{B_{[MHz]}}{B_{c[MHz]}}\right)\left(\frac{N_L}{N_{L,c}}\right)\left(\frac{\log_2 M}{\log_2 M_c}\right)$
Option 7 (Intra-PHY)	**3** **DL: 9.8 Gbps**	N_{SC}:1200*5 N_{SY}: 14 N_Q: 8	$R_{7-2[Mbps]} = (N_{SC}N_{SY}\,N_{Q[bits]}N_L 2 * 1000) + MAC_{info[Mbps]}$

			N_L: 7 MAC_{info}: 713.9	
		UL: **15.2 Gbps**	N_{SC}:1200*5 N_{SY}: 14 N_Q: 8 N_L: 10 MAC_{info}: 120	$R_{7-2[Mbps]} =$ $(N_{SC}N_{SY}\ N_{Q[bits]}N_L 2 * 1000)$ $+ MAC_{info[Mbps]}$
	2	**DL:** **9.2 Gbps**	N_{SC}:1200*5 N_{SY}: 14 N_Q: 8 N_L: 7 MAC_{info}: 121	$R_{7-2[Mbps]} =$ $(N_{SC}N_{SY}\ N_{Q[bits]}N_L 2 * 1000)$ $+ MAC_{info[Mbps]}$
		UL: **60.4 Gbps**	N_{SC}:1200*5 N_{SY}: 14 N_Q: 32 N_A: 10 MAC_{info}: 80	$R_{7-2[Mbps]} =$ $(N_{SC}N_{SY}\ N_{Q[bits]}N_A 2 * 1000)$ $+ MAC_{info[Mbps]}$
	1	**DL:** **9.8 Gbps**	Same as 7-3	$R_{7-1[Mbps]} = R_{7-3[Mbps]}$
		UL: **60.4 Gbps**	Same as 7-2	$R_{7-1[Mbps]} = R_{7-2[Mbps]}$
8) PHY-RF	**DL:** **157.3 Gbps**		S_r: 30.72 N_Q: 32 N_A: 32	$R_{8[Mbps]} = S_{r[sample/s]}\ N_{Q[bits]}N_A 5$
	UL: **157.3 Gbps**		S_r: 30.72 N_Q: 32 N_A: 32	$R_{8[Mbps]} = S_{r[sample/s]}\ N_{Q[bits]}N_A 5$

Annex E

Model Main Tasks Flowcharts

This annex illustrates the flowcharts with the procedure to the main task of the model.

In order to connect the network nodes and distribute the BS functions between the nodes, one uses the procedure in Figure E.1.

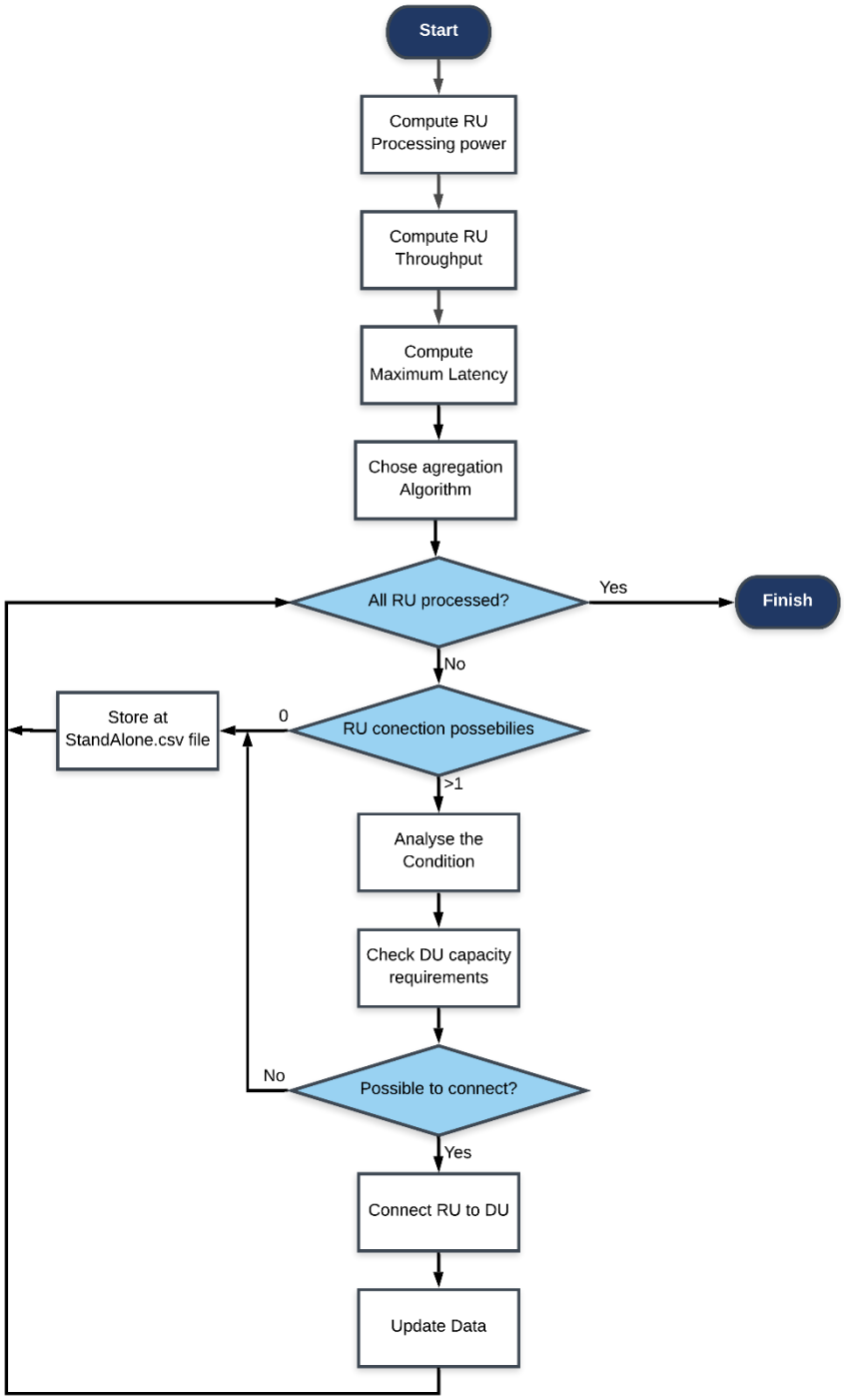


Figure E.1 Aggregation Flowchart.

Figure E.2 describes the process used by the model to create new nodes on the network. The Flowchart on Figure E.2 describes an example of the implementation of a DU node, but the same procedure is used to implement the new MEC nodes.

The K-means algorithm used to optimise the location of the new nodes on the network is described in Figure E.3

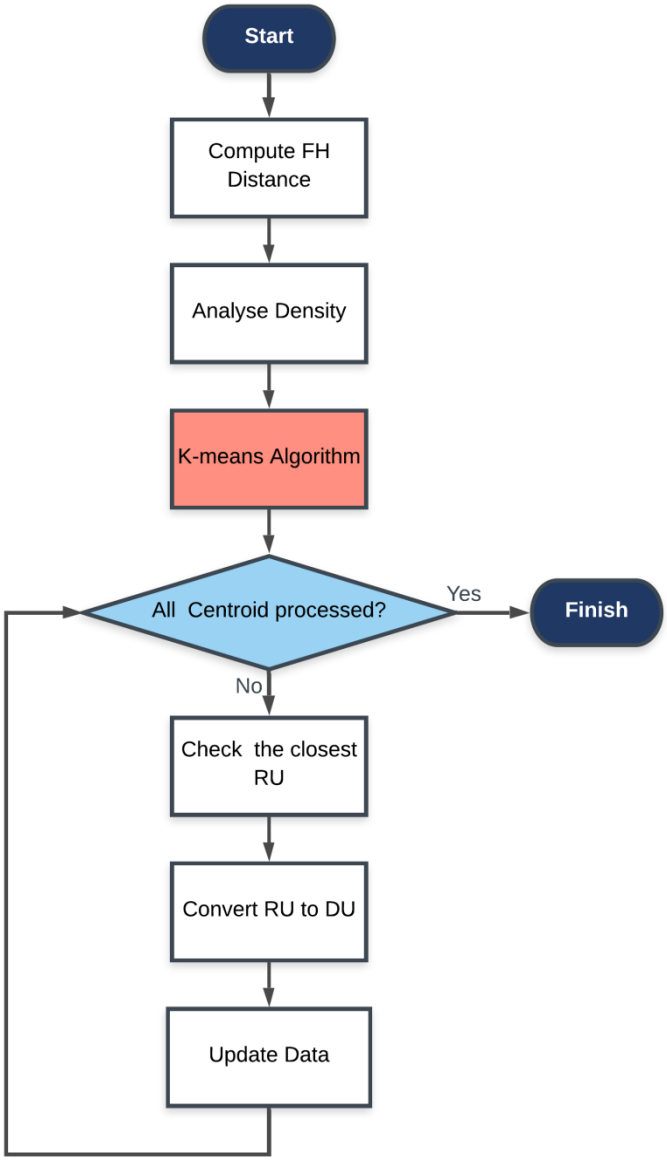


Figure E.2. New DU implementation process Flowchart.

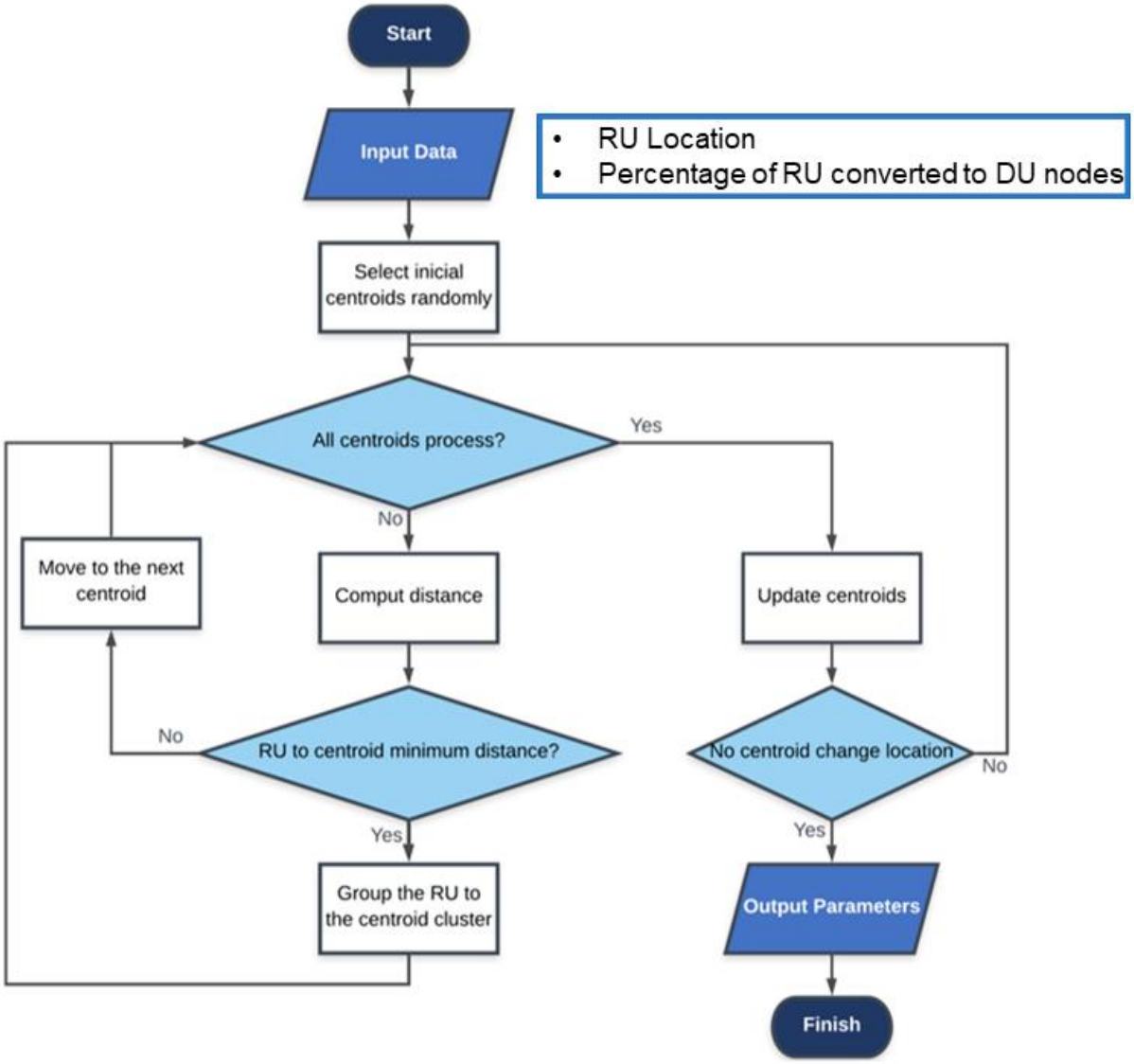


Figure E.3. K-means algorithm Flowchart.

.

Annex F

Model Assessment Tests Results

This annex illustrates the figures with the empirical test results of the model assessment.

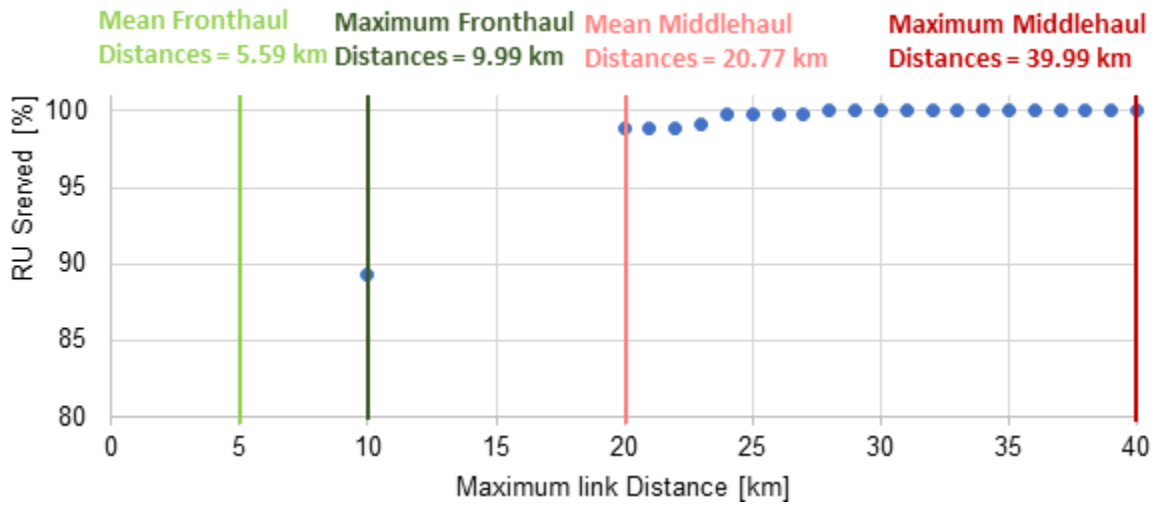


Figure F.1. Served function of RUs with the maximum link distance.

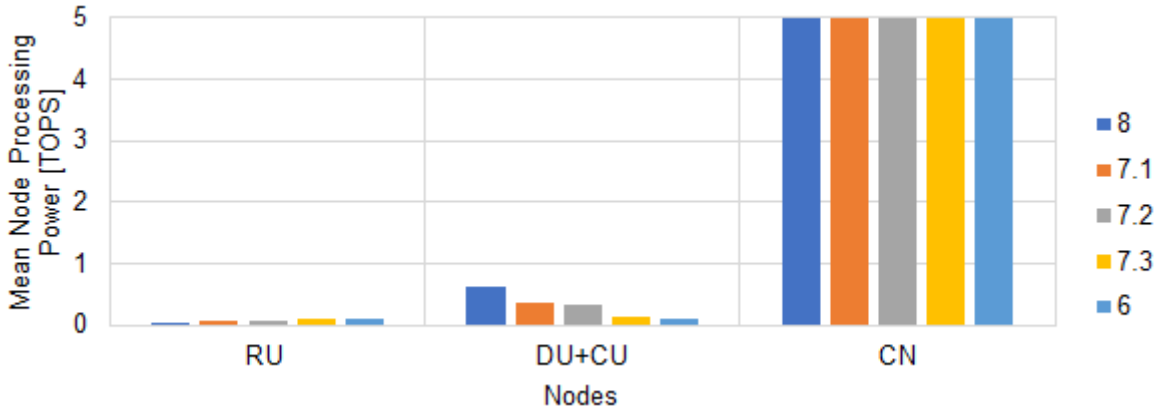


Figure F.2. Nodes processing power evolution with the splitting options.

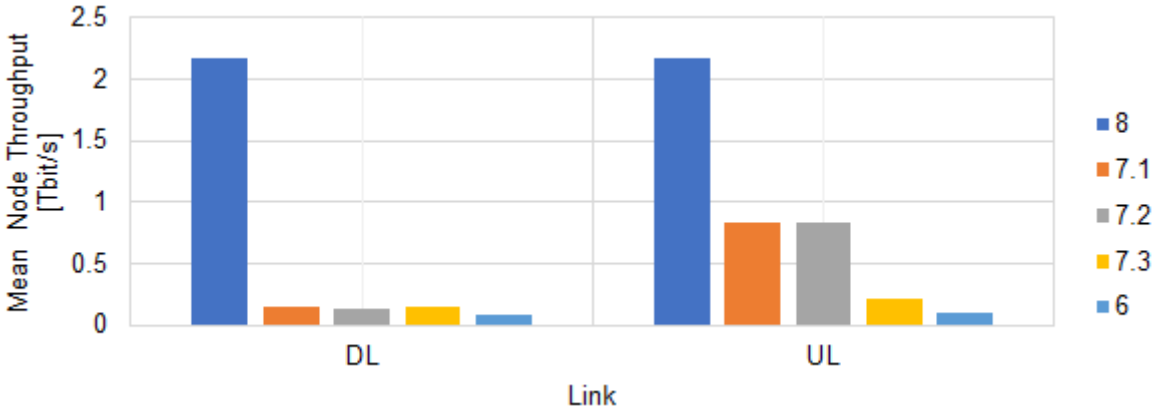


Figure F.3. Evolution of the throughput on the FH with the splitting options on the DL and UL.

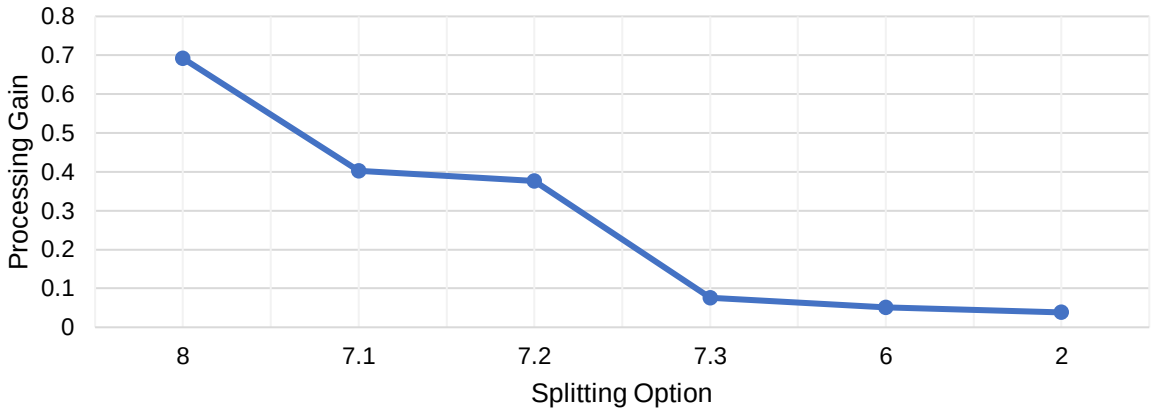


Figure F.4. Process gain between the RU and CU node for different splitting options

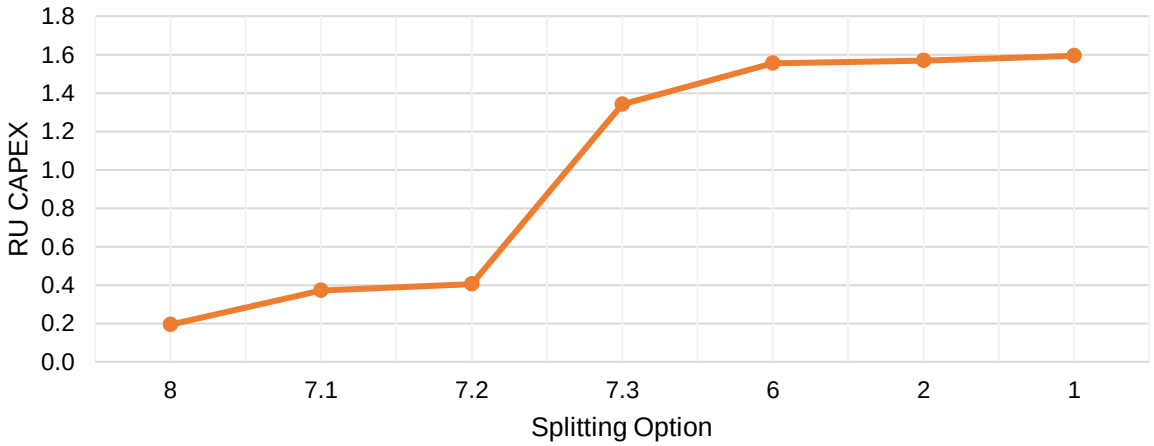


Figure F.5. RU CAPEX for different splitting options

Annex G

Reference Scenario Configuration

This annex presents the input parameters used in the reference scenario.

Table G.1. Service Mix reference values

Service name	Service Mix [%]	MEC Supported
Voice	10	No
Video conference	5	No
Video streaming	25	No
Music streaming	7	No
Web browsing	10	No
Social networking	11	No
File sharing	5	No
Email	3	No
Virtual reality	1	Yes
Realtime gaming	10	Yes
Smart Meters	7	No
Factory automation	1	Yes
Road safety ITS	4.9	Yes
Remote surgery	0.1	Yes

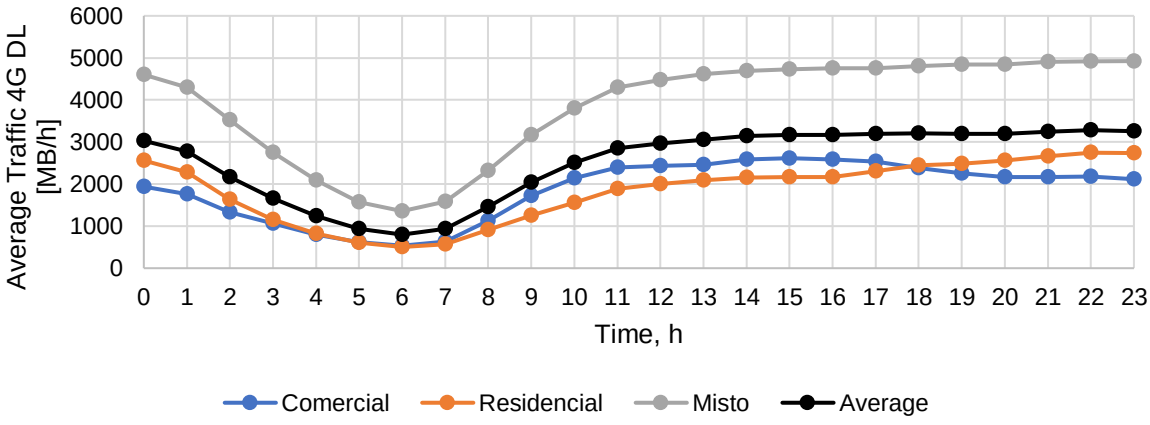


Figure G.1. Average DL 4G traffic on the RUs on Minho scenario (extracted from [Silva16]).

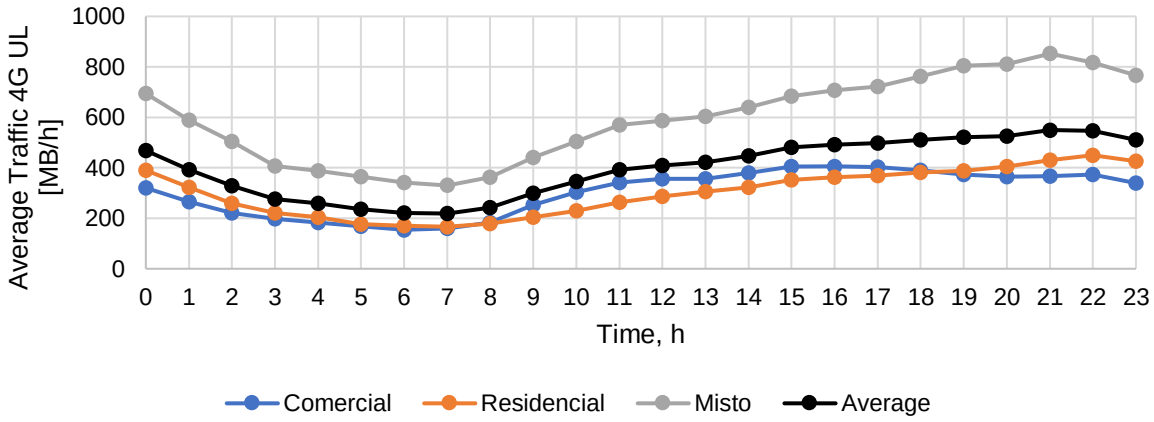


Figure G.2. Average UL 4G traffic on the RUs on Minho scenario (extracted from [Silva16]).

The information from Table G.1 and the traffic profiles from the Figures above represent the input parameters of the model for the calculation of the average number of users or connected devices in the network per hour.

Table G.2. Average number of users on the reference scenario.

Average Total Users	21 250
Average eMBB Users	2 340
Average mMTC Users	1 490
Average URLLC Users	1 275

Table G.3. Reference values of Bandwidth for the different RU density type.

RU Density type	**Bandwidth [MHz]**
Dense Urban	100
Urban	50
Rural	20

It is assumed different bandwidth values on the RU depending on the node density, which has a direct impact on the processing power of the nodes.

Table G.4 summarise the input parameters of the reference scenario used by the simulator.

Table G.4. Network reference values

Scenario	Minho
Architecture	RU-DU+CU
Splitting Option	7.1
Number of DUs	50
Number of MECs	0
Max FH link [km]	10
Max MH link [km]	40
Max BH link	N.D.
RU Processing Capacity [TOPS]	174
RU throughput Capacity [Tbps]	20
DU Processing Capacity [TOPS]	12
DU throughput Capacity [Tbps]	62
CU Processing Capacity [TOPS]	13
CU throughput Capacity [Tbps]	41

The cost model used in the thesis divided the CAPEX cost between constant cost, which does not depend on the output parameters of the model, and the variable cost, which depends on the output parameters. This structure was used to better simulate the impact of the characteristics of the network nodes between the different network architectures.

Table G.5. Assumption values in reference scenarios for CAPEX.

Constant Costs	$C_{fibre,i[€/link]}$	5000
	$C_{microwave,i[€/link]}$	5000
	$C_{MEC[€]}$	2000
	$C_{RU100[€]}$	1000
	$C_{RU50[€]}$	750
	$C_{RU20[€]}$	500
	$C_{DU[€]}$	500
	$C_{CU[€]}$	500
Variable Costs	$C_{fibre[€/km]}$	5000
	$C_{Inter[€/Interface]}$	500
	$C_{GOPS\_RU[€/GOPS/s]}$	0.1
	$C_{GOPS\_DU[€/GOPS/s]}$	0.1
	$C_{GOPS\_CU[€/GOPS/s]}$	0.1
	$C_{GOPS\_MEC[€/GOPS/s]}$	0.1

Table G.6. Assumption values in reference scenarios for OPEX.

Renting	$C_{A[€/m^2/month]}$	10
	$A_{MEC[m^2]}$	6
	$A_{RUDUCU[m^2]}$	4
	$A_{DUCU[m^2]}$	4
	$A_{RUDU[m^2]}$	4
	$A_{CU[m^2]}$	3
	$A_{DU[m^2]}$	3
	$A_{RU[m^2]}$	3
	$C_{GOPS[GOPS/W]}$	2.5
Energy	$C_{E[€/kW/h]}$	0.16
Maintenance	$m_{node[\%]}$	0.02
	$m_{fibre[\%]}$	0.001

Annex H

Centralisation Gain Results

This annex illustrates the centralisation gain results for the different network architectures of the model for the Minho scenario.

Table H.1. DL Centralisation Gain for RU-DU-CU architecture.

RU-DU-CU		Traffic Aggregation Gain	Throughput Aggregation Gain	GOPS Aggregation Gain	Process Gain
FH	8	0.887	0.887	0.446	0.692
	7.1	14.29	14.29	1.488	0.402
	7.2	15.77	15.77	1.658	0.376
	7.3	14.29	14.29	12.24	0.076
	6	25.34	25.34	18.46	0.051
MH	8	2.940	2.940	5.641	0.151
	7.1	1.361	1.361	3.554	0.219
	7.2	1.301	1.301	3.489	0.223
	7.3	1.365	1.365	1.300	0.434
	6	0.950	0.950	0.951	0.513
BH	8	1.025	1.025	0.884	0.531
	7.1	1.043	1.043	0.821	0.549
	7.2	1.046	1.046	0.783	0.561
	7.3	1.041	1.041	0.437	0.696
	6	1.049	1.049	0.412	0.708

Table H.2. DL Centralisation Gain for RU+DU-CU architecture.

RU+DU-CU	Traffic Aggregation Gain	Throughput Aggregation Gain	GOPS Aggregation Gain	Process Gain
MH	43.33	43.33	25.10	0.038
BH	1.050	1.050	0.265	0.791

Table H.3. DL Centralisation Gain for RU+DU+CU architecture.

RU+DU+CU	Traffic Aggregation Gain	Throughput Aggregation Gain	GOPS Aggregation Gain	Process Gain
BH	45.51	45.51	6.914	0.126

Table H.4. UL Centralisation Gain for RU-DU+CU architecture.

RU-DU+CU		Traffic Aggregation Gain	Throughput Aggregation Gain	GOPS Aggregation Gain	Process Gain
FH	8	1.223	1.223	0.376	0.726
	7.1	3.185	3.185	1.104	0.475
	7.2	3.185	3.185	1.159	0.463
	7.3	12.66	12.66	6.672	0.130
	6	26.94	26.94	13.169	0.070
BH	8	1.103	1.103	19.47	0.049
	7.1			12.73	0.073
	7.2			12.45	0.074
	7.3			3.790	0.209
	6			2.086	0.324

Annex I

Node Throughput Results

This annex illustrates the node throughput results for the different network architectures of the model for the Minho scenario.

This annex presents the input throughput on the nodes of the network for the different network architectures. Increasing the FH splitting option leads to a reduction on network links, and consequently the reduction of the input throughput on the aggregation node. Another option to reduce the input throughput on the aggregation node is the implementation of a RU-DU-CU architecture, where the DU nodes are used to offload computation resources from the central node.

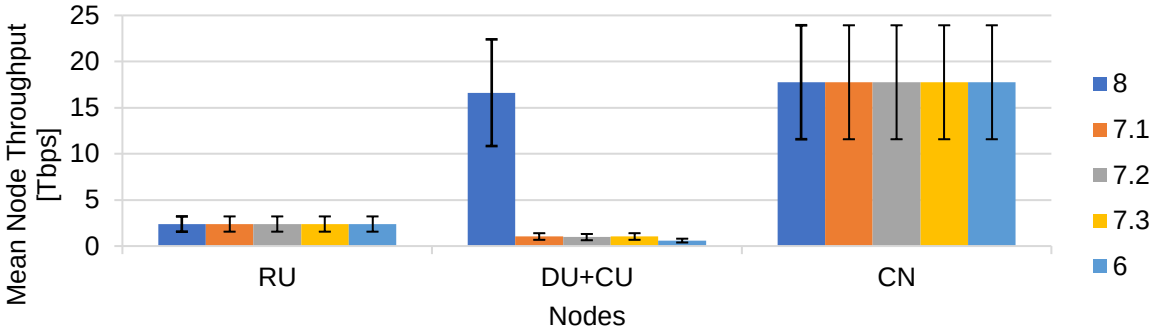


Figure I.1. Mean input throughput on the network nodes in different splitting options for RU-DU+CU architecture on DL.

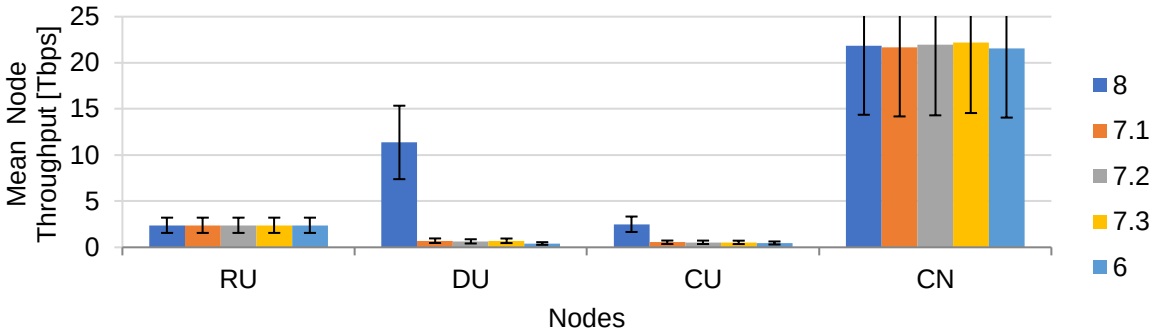


Figure I.2. Mean input throughput on the network nodes in different splitting options RU-DU-CU architecture on DL.

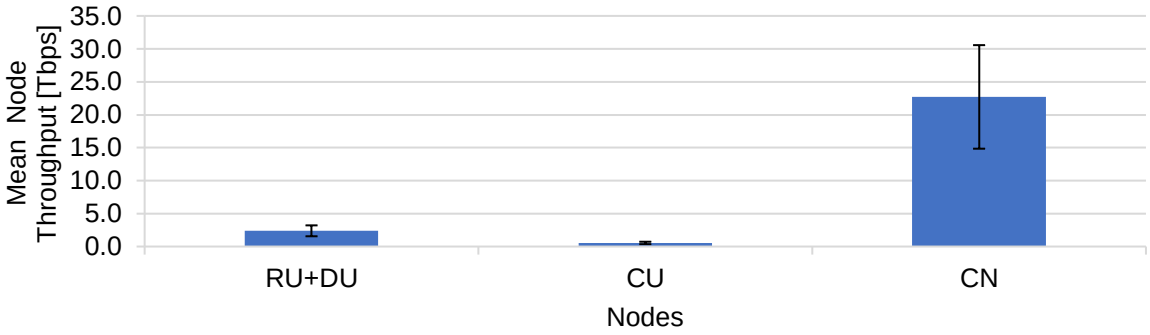


Figure I.3. Mean input throughput on the network nodes in different splitting options RU+DU-CU architecture on DL.

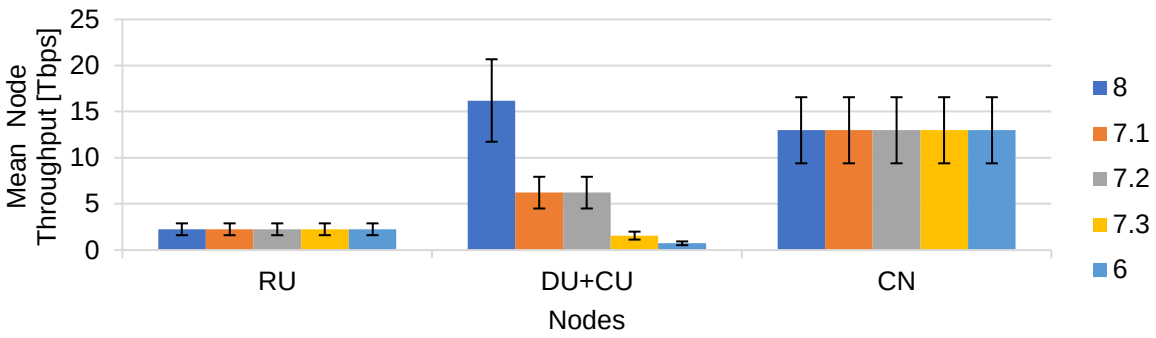


Figure I.4. Mean input throughput on the network nodes in different splitting options RU-DU+CU architecture on UL.

Annex J

Node Processing Power Results

This annex illustrates the node processing power results for the different network architectures of the model for the Minho scenario.

The following Figure illustrates the processing power on the nodes of the network for the different network architectures on the DL and UL. Increasing the splitting option of the network leads to an increase of BS function on the RU node, providing more processing capacity on the edge of the network in order to process the signal and reduce link throughput on the network.

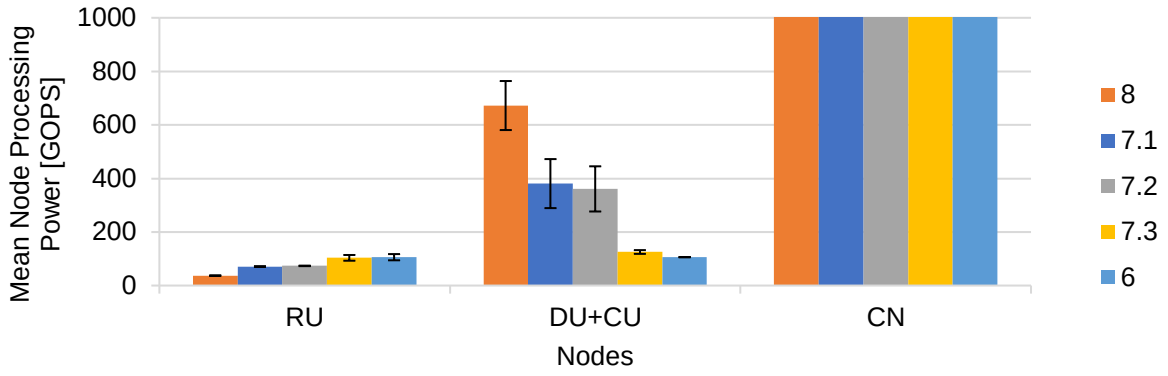


Figure J.1. Mean processing power on the network nodes in different splitting options for RU-DU+CU architecture on DL.

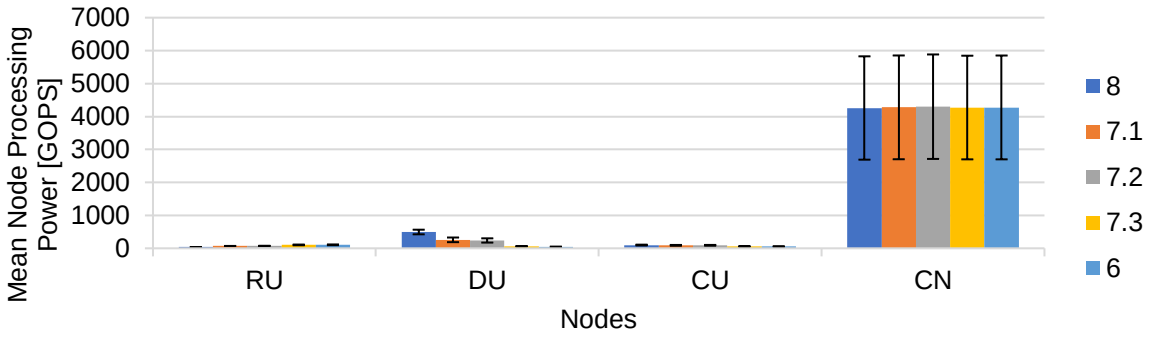


Figure J.2. Mean processing power on the network nodes in different splitting options RU-DU-CU architecture on DL.

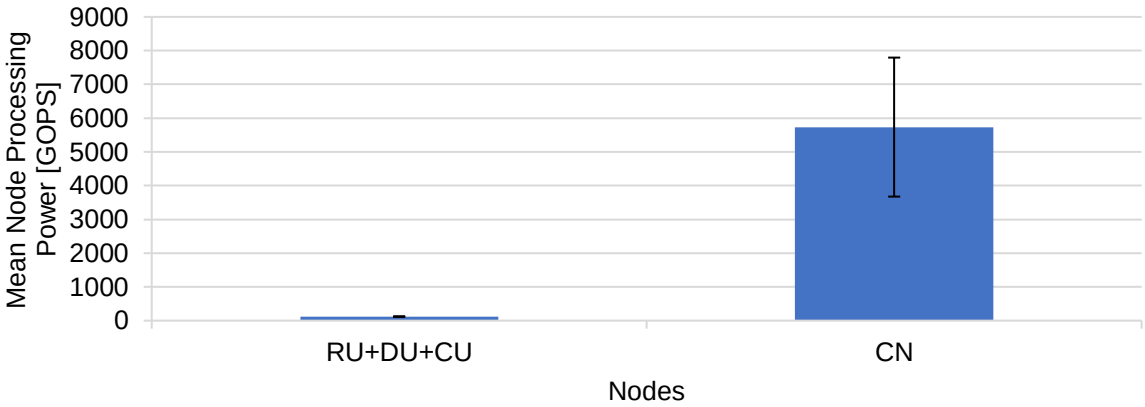


Figure J.3. Mean processing power on the network nodes for RU+DU-CU architecture on DL.

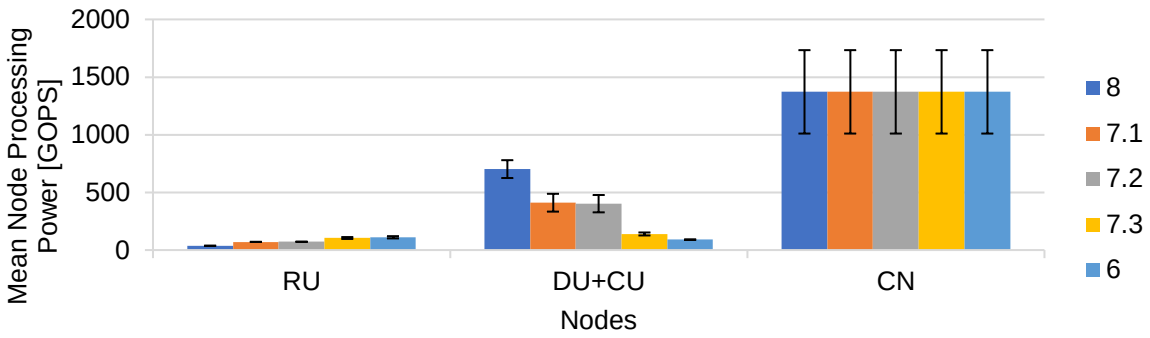


Figure J.4. Mean processing power on the network nodes in different splitting options RU-DU+CU architecture on UL.

Annex K

Latency Impact of the Users

This annex illustrates the results of the impact of the number of users connected to the network and the impact of different user profiles used on the model for the Minho scenario.

In the next analyse it is studied the impact of the number of eMBB users on the E2E network latency, the following analysis considers a constant number of connected devices on all the other use cases which means that increasing the number of eMBB devices will consequently increase the number of total connected devices on the network. Figure K.1 presents the results where one notice that the variation of eMBB users on the site has a strong impact on the performance of the network, increasing the number of users to 2 times the reference values will increase 17.4% of the total latency. The eMBB users have the strongest impact on the network since the service demands are very high (i.e. 1 Gbps), so in order to support the future increasing on the 5G eMBB use cases the network needs to support more throughput on the nodes, this can be achieved assigning more processing capacity on the edge of the network, or using an all collocated nodes architecture RU+DU+CU.

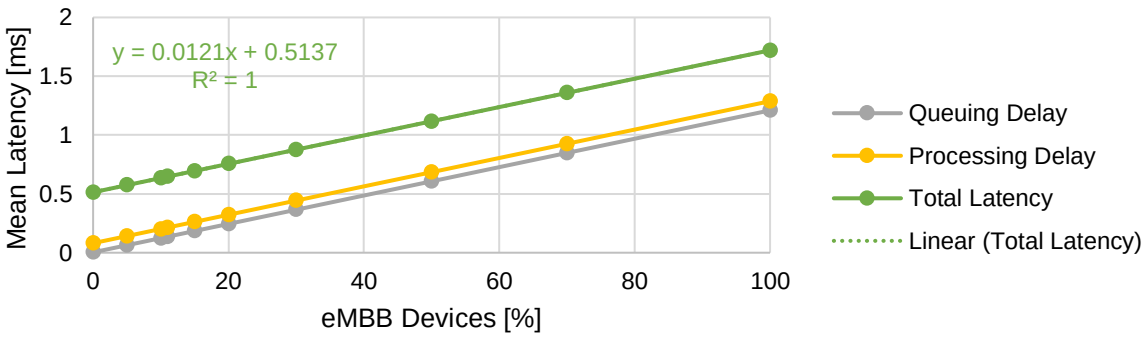


Figure K.1. Mean network latency for RU-DU+CU architecture with variable eMBB devices for a fix total number of devices.

Regarding the mMTC and URLLC service types, the variation on the user profile for these services has a similar effect on the overall performance on the network. Since the mMTC and URLLC services use fewer network resources than eMBB, increasing the number of users of these services will benefit on the performance of the network since the number of users of eMBB services and Video streaming, that are the more demanding use cases, are being reduced. Since the mMTC use cases use the fewer resources on the network, increasing the number of mMTC devices, while maintaining the same total number of devices on the network, will achieve lower network latency, as illustrated in Figure K.2.

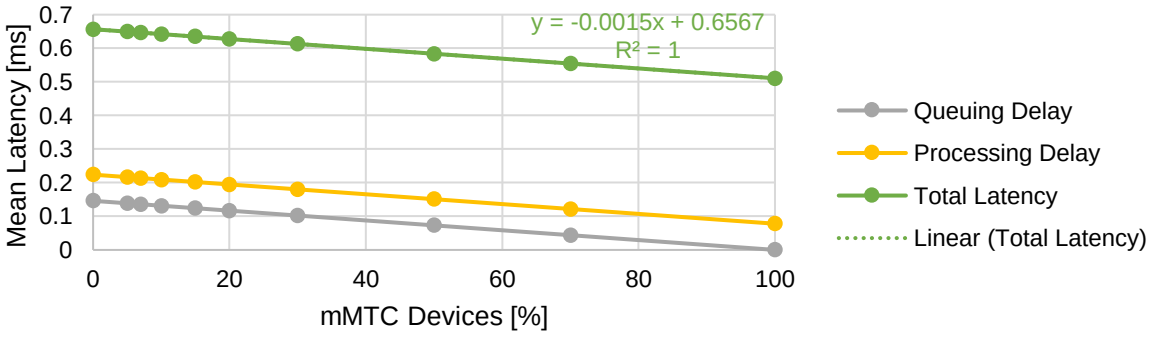


Figure K.2. Mean network latency for RU-DU+CU architecture with variable mMTC devices for a fix total number of devices.

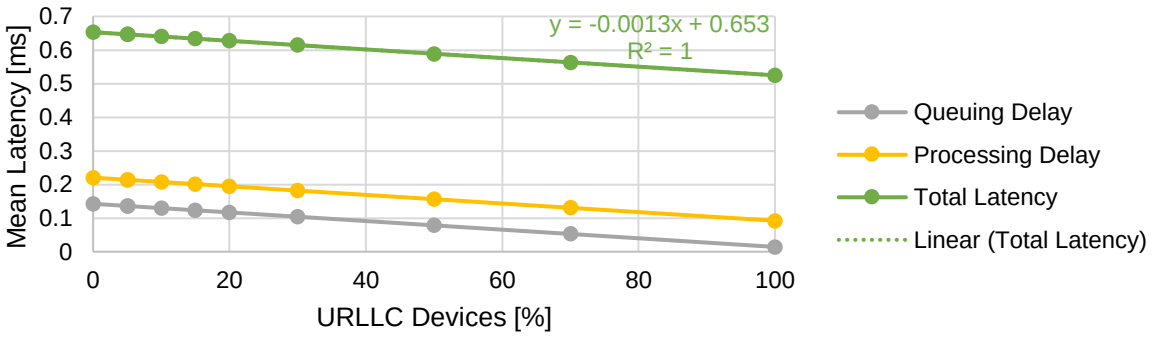


Figure K.3. Mean network latency for RU-DU+CU architecture with variable URLLC devices for a fix total number of devices.

The URLLC use cases generate on average around 85 times more traffic than the mMTC services so URLLC services load the network more than the mMTC and the latency results are slightly higher but, the most important fact to consider is that the mMTC latency requirement is 100 ms and the URLLC can be lower than 1 ms so in other to achieve an confidence interval of 68.3% all below 1 ms the URLLC devices should be 50% of the total network devices or more, and, to achieve the 99.999% reliability of the URLLC services presented on [EFSZ16], it is possible to conclude that the network latency needs to be reduced using higher processing capacity on the nodes or increasing the number of MEC nodes on the network.

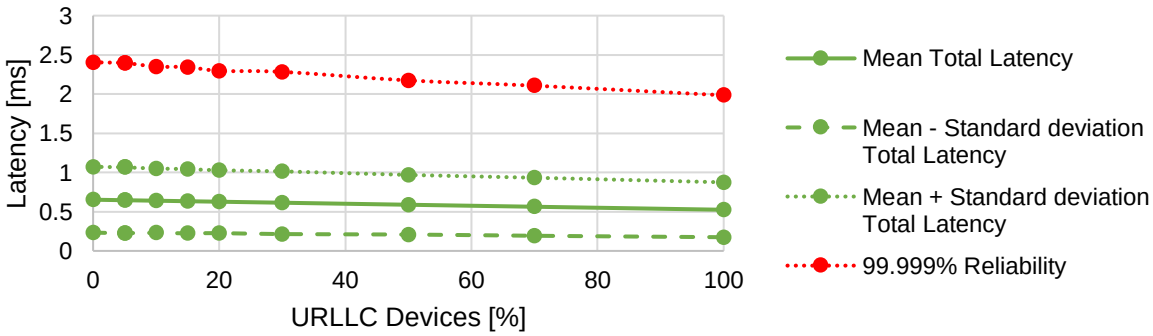


Figure K.4. Total network latency for RU-DU+CU architecture with variable URLLC devices for a fix total number of devices.

Annex L

Analysis of the Implementation of DU Nodes

This annex presents auxiliary analysis of the implementation of DU on the RU-DU-CU architecture to support the assumption of the DU reference scenario.

The main purpose of the implementation of DU nodes is to balance the processing capacity and network load between the CU nodes, which, in lower splitting option, can reach very high traffic demands on the network. It is considered that the DU can only be implemented in the location that the RU node is already located, converting the node into a RU+DU node.

It is considered a reference scenario in Minho with 50 initially implemented DUs on the network, distributed throughout dense urban and urban areas, it was considered this assumption in the reference scenario since the majority of the RU nodes, and consequently the traffic on the network is on the dense urban areas, and, since in dense urban area the maximum FH distance is not a limiting factor on the implementation of DU nodes, despite the necessity of these nodes to balance the traffic on dense population areas. This section aims to analyse the impact of the DU nodes without considering any density restriction, which is used in the reference scenario, and compared the results with the reference scenario.

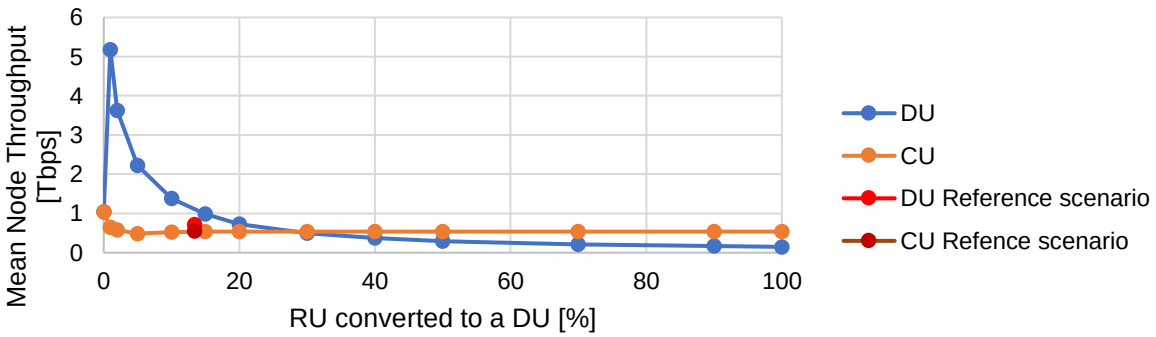


Figure L.1. Mean input throughput on the DU and CU with a different number of DU nodes on the network on DL.

Whit respect to the processing capacity on the nodes, Figure L.1 illustrates the throughput on DU and CU nodes as a function of the number of DUs on the network. With low DU nodes on the network, the DU nodes are very heavy so there are no benefits of implement less than 10% of DU implemented on the network. When comparing with the reference scenario, one verifies that with 50 DU on the network the average DU throughput is 22.7% higher than the average throughput on the CU so the nodes are already balanced.

Figure L.2 illustrates the results for the processing power on the nodes, the results are similar to the throughput on the nodes. On the reference scenario, the DU and CU variation is 60% which is much lower even compared with a scenario with 20% of DU on the network that has a 77.3% variation.

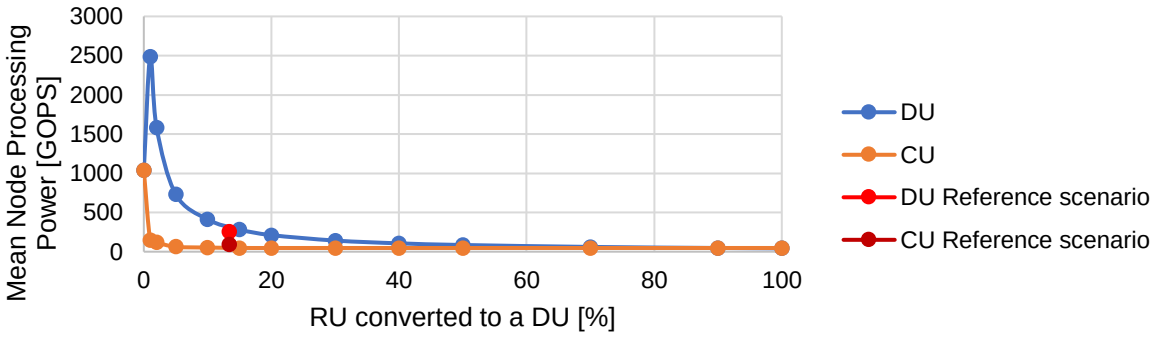


Figure L.2. Mean processing power on the DU and CU with a different number of DU nodes on the network on DL.

Regarding the network distance of the network, when the number of DU increases the average network distance increases since the all independent node architecture has FH and MH and the RU-DU+CU architecture only has FH. It is worth notice that even though the total network distance increases the FH distance decreases with the number of DU which can have benefits since with low splitting option the FH link is extremely heavy. Analysing the Figure L.3 one verifies that the reference scenario achieves around 26.2% reduction compared with the average network distance, this happens since in rural areas the RU node are very scattered so the benefits of the DU nodes are less noticeable. Figure L.3 illustrates the total network latency results for different number of DUs implemented on the network:

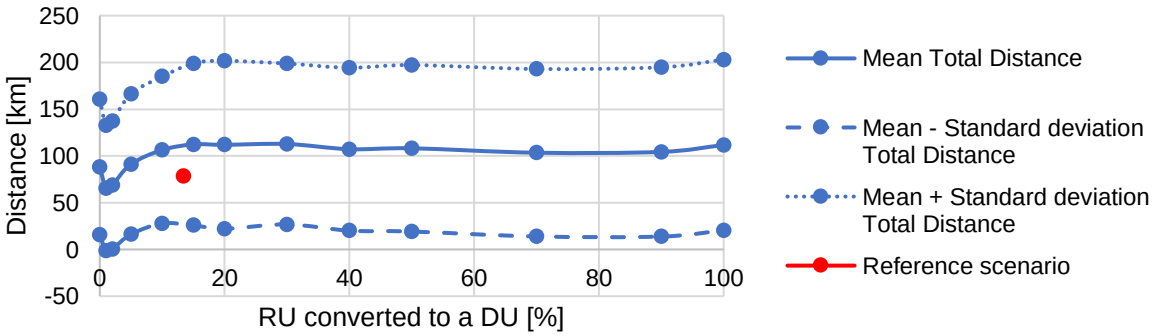


Figure L.3. Total network distance for different number of DU nodes.

Annex M

Confidential Information

This annex presents some confidential information related with this thesis, namely the reference location of sites and aggregation nodes in the North of Portugal.

Annex N

Confidential Information

This annex presents some confidential information related with this thesis, namely the reference location of sites and aggregation nodes in Lisbon.

References

[3GPP02] 3GPP, *Technical Specification Universal Mobile Telecommunications System (UMTS). Quality of Service (QoS) concept and architecture (Release 5),* Report TS 23.107 V5.6.0, Sep. 2002.

[3GPP16] 3GPP, *CU-DU split: Refinement for Annex A (Transport network and RAN internal functional split)*, R3-162102, Sophia Antipolis, France, Oct. 2016.

[3GPP17] 3GPP, *Study on new radio access technology: Radio access architecture and interface (Release 14)*, Report TR 38.801, V14.0.0, Mar. 2017.

[3GPP18] 3GPP, *Technical Specification Group Services and System Aspects,* Policy and charging control architecture, Report TS 38.300, V15.3.0, Jun. 2018.

[5GAm16] 5G Americas, *Network Slicing of 5G and Beyond*, White Paper, Bellevue, WA, USA, Nov. 2016 Available: http://www.5gamericas.org/files/3214/7975/0104/5G_Americas_Network_Slicing_11.21_Final.pdf.

[5GAm17] 5G Americas, *5G services & use cases*, White Paper, Bellevue, WA, USA, Dec. 2017. Available: https://blogs.intel.com/technology/2017/12/5g-americas-white-paper-explores-5g-services-use-cases-and-market-implications/.

[AHGZ16] S. Abdelwahab, B. Hamdaoui, M. Guizani and T. Znati, “Network function virtualisation in 5G”, *IEEE Communications Magazine*, Vol. 54, No. 4, Apr. 2016, pp. 84-91. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7452271.

[Alca13] Alcatel-Lucent, *The LTE Network Architecture*, White Paper, Boulogne-Billancourt, France, 2013 Available: http://www.cse.unt.edu/~rdantu/FALL_2013_WIRELESS_NETWORKS/LTE_Alcatel_White_Paper.pdf.

[Alme13] D. Almeida, *Inter–Cell Interference Impact on LTE Performance in Urban Scenarios*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, Oct. 2013.

[Amaz17] Amazon Web Services, *Overview of Amazon Web Services,* Public Consultation, Seattle, Washington, EUA, Apr. 2017. Available: https://docs.aws.amazon.com/aws-technical-content/latest/aws-overview/aws-overview.pdf?icmpid=link_from_whitepapers_page.

[ATNO18] O. Arouk, T. Turletti, N. Nikaein, K. Obraczka, “Cost Optimisation of Cloud-RAN Planning and Provisioning for 5G Networks”, in *Proc. of ICC18-2018 IEEE International Conference on Communications*, Kansas City, MO, USA, May 2018. Available: https://ieeexplore.ieee.org/document/8422744.

[Bara17] M. Barahman, *Efficient Software-based Base Station Design for C-RAN*, Ongoing Ph.D. Thesis, IST, University of Lisbon, Lisbon, Portugal, Dec. 2018.

[BVCG17] J.Bartelt, N. Vucic, D. Camps-Mur, E. Garcia-Villegas, I. Demirkol, A. Fehske, M. Grieger, A. Tzanakaki, J. Gutiérrez, E. Grass, G. Lyberopoulos and G. Fettweis, "5G transport network requirements for the next generation fronthaul interface", *EURASIP Journal on Wireless Communications and Networking*, May 2017. Available: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-017-0874-7.

[CCYS14] A.Checko, H.L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M.S. Berger and L. Dittmann, "Cloud RAN for Mobile Networks – A Technology Overview", *IEEE Communications Surveys & Tutorials*, Vol. 17, No. 1, Sep. 2014, pp. 405-426. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6897914.

[Corr18] L.M. Correia, *Notes from Mobile Communication course*, Instituto Superior Técnico, Lisbon, Portugal, 2018.

[DDLo15] B. Debaillie, C. Desset, and F. Louagie, "Flexible and Future-Proof Power Model for Cellular Base Stations", in *Proc. of VTC Spring 2015 – 81st IEEE Vehicular Technology Conference,* Glasgow, United Kingdom, May 2015.

[EFSZ16] S. E. Elayoubi, M. Fallgren, P. Spapis, G. Zimmermann, D. Matín-Sacristán, C. Yang, S. Jeux, P. Agyapong, L. Campoy, Y. Qi, S. Singh and S. Singh, "5G service requirements and operation use cases: Analysis and METIS II vision", in *Proc. of EuCNC16 – IEEE European Conference on Networks and Communications*, Athens, Greece, Jun. 2016. Available: https://ieeexplore.ieee.org/document/7561024.

[EmFS18] M. Emara, M. Filippou and D. Sabella, "MEC-Assisted End-to-End Latency Evaluation for C-V2X Communications", in *Proc. of EuCNC18 – 2018 European Conference on Networks and Communications*, Ljubljana, Slovenia, Jun. 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8442825.

[Eric18A] Ericsson, *Ericsson Mobility Report*, Public Consultation, Jun. 2018 [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf.

[Eric18B] Ericsson, *Network Slicing*, Public Consultation. Available: https://www.ericsson.com/en/digital-services/trending/network-slicing.

[ErLG15] Ericsson and LG Electronics, "Service Provider SDN with Cloud Transformation", in *Proc of OVN15 – Open & Virtual Networking Conference 2015*, Seoul, South Korea, Feb. 2015.

[GAMZ10] A. Ghosh, J. G. Andrews, R. Muhamed and J. Zhang, "Overview and Channel Structure of LTE", *Fundamentals of LTE*, Prentice Hall, Upper Saddle River, Nova Jersey, EUA, 2010.

[HKPS18] S. Husain, A. Kunz, A. Prasad, K. Samdanis and J. Song, "Mobile edge computing with network resource slicing for Internet-of-Things", in *Proc of WF-IoT18 - 2018 IEEE 4th World Forum on Internet of Things*, Singapore, Singapore, Feb. 2018. Available:

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8355232.

[Hodg18] J. Hodges, *Transforming the Edge: The Rise of MEC*, Intel, Heavy Reading, New York, USA, Mar. 2018. Available: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/the-rise-of-multi-access-edge-computing-paper.pdf.

[HoTo11] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE Advanced (2nd edition)*, John Wiley & Sons, Chichester, United Kingdom, 2011.

[IEEE17] IEEE, *IEEE 5G and Beyond Technology Roadmap White Paper*, White Paper, IEEE Advancing Technology of Humanity, New Jersey, USA, Oct. 2017. Available: https://futurenetworks.ieee.org/images/files/pdf/ieee-5g-roadmap-white-paper.pdf.

[ITUT18] ITU-T, *Transport network support of IMT-2020/5G*, Technical Report, Stephen Shew, Ciena, Canada, Feb. 2018.

[Jone17] J. Jones, *Edge Computing: The Cloud, the Fog & the Edge*, Public Consultation, Apr. 2017. Available: https://www.solid-run.com/edge-computing-cloud-fog-edge, [Accessed: Oct. 2018].

[JPYK18] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, "Ultra Reliable and Low Latency Communications in 5G Downlink: Physical Layer Aspects", *IEEE Wireless Communication*, Vol. 25, No. 3, Jun. 2018, pp. 124-300. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8403963.

[Khat16] S. Khatibi, *Radio Resource Management Strategies in Virtual Networks*, Ph.D. Thesis, IST, University of Lisbon, Lisbon, Portugal, 2016.

[LAYG15] M. Liyanage, Ijaz Ahmad, M. Ylianttila, A. Gurtov, A. Abro and E. Oca, "Leveraging LTE Security with SDN and NFV", in *Proc. of ICIIS 2015 - 10th IEEE International Conference on Industrial & Information Systems*, Peradeniya, Sri Lanka, Dec. 2015. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7399014.

[LCCh18] L. Larsen, A. Checko, H. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Network", *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 1, Oct. 2018, pp. 146-172. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8479363.

[LHWW18] F. Lin, C. Hsiao, Y. Wen and Y. Wu, "Optimisation-Based Resource Management Strategies for 5G C-RAN Slicing Capabilities", in *Proc. of ICUFN18 – 10th International Conference on Ubiquitous and Future Networks*, Prague, Czech Republic, Jul. 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8436837.

[LiHW18] D. Lin, Y. Hsu and H. Wei, "A Novel Forwarding Policy under Cloud Radio Access Network with Mobile Edge Computing Architecture", in *Proc. of ICFEC18 –IEEE 2nd International Conference on Fog and Edge Computing*, Washington, DC, USA, May 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8358722.

[LLHC18] Y. Lin, Y. Lai, J. Huang and H. Chien, “Three-Tier Capacity and Traffic Allocation for Core, Edges, and Devices for Mobile Edge Computing”, *IEEE Transactions on Network and Service Management*, Vol. 15, No. 3, Sep. 2018, pp. 923-933, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8402110.

[LMWX17] T. Li, C. Magurawalage, K. Wang, K. Wang, K. Xu, K. Yang and H. Wang, “On Efficient Offloading Control in Cloud Radio Access Network with Mobile Edge Computing”, in *Proc. of ICDCS17 – 2017 IEEE 37th International Conference on Distributed Computing Systems*, Atlanta, GA, USA, Jun. 2017. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7980179.

[Mart17] H. Martins, *Analysis of CoMP for the Management of Interference in LTE*, M.Sc. Thesis, IST, University of Lisbon, Lisbon, Portugal, Nov. 2017.

[MFAM16] M. Mushtaq, Scott Fowler, “Brice Augustin and Abdelhamid Mellouk, QoE in 5G Cloud Networks using Multimedia Services”, in *Proc. of WCNC16 – 2016 IEEE Wireless Communications and Networking Conference*, Doha, Qatar, Apr. 2016. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7565173.

[MMBT15] J. F. Monserrat, G. Mange, V. Braun, H. Tullberg, G. Zimmermann and O. Bulakci, *METIS research advances towards the 5G mobile and wireless system definition*, 5G Wireless Mobile Technologies, Art. 53, Mar. 2015. Available: https://pdfs.semanticscholar.org/1e9d/c888b5b759646eb95df5dbb0a18d27909448.pdf.

[Mont16] T. Monteiro, *Implementation Analysis of Cloud Radio Access Network Architectures in Small Cells*, M.SC. Thesis, IST, Technical University of Lisbon, Lisbon, Portugal, Nov. 2016.

[MSGB15] R. Mijumbi, J. Serrat, J.Gorricho, N. Bouten, F. Turck and R. Boutaba, “Network Function Virtualisation: State-of-the-Art and Research Challenges”, *IEEE Communications Surveys & Tutorials*, Vol. 18, No. 1, Sep. 2015, pp. 236-262. Available: https://ieeexplore.ieee.org/document/7243304.

[NHHS18] M. Nasimi, M. Habibi, B. Han, H. Schotten, “Edge-Assisted Congestion Control Mechanism for 5G Network Using Software Defined Networking”, in *Proc. of ISWCS18 - 15th International Symposium on Wireless Communication Systems*, Lisbon, Portugal, Aug. 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8491233.

[OMGK18] OMGKRK, Krakow Startup Community, Available: http://www.omgkrk.com/apply-for-hubraum-low-latency-prototyping-program-edge-computing-and-5g-technology.

[PRRG18] I. Parvez, A. Rahmati, I. Guvenc, A. Sarwat, H. 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/stamp/stamp.jsp?tp=&arnumber=8367785.

[Qual16] Qualcomm Technologies, *Making 5G NR a reality*, Public Consultation, Sep. 2016.

Available: https://www.qualcomm.com/media/documents/files/making-5g-nr-a-reality.pdf.

[RoSh17] S. Routray and K. Sharmila, “Software defined networking for 5G”, in *Proc. of ICACCS17 – 2017 4th International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, Jan. 2017. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8014576.

[Rouz19] B. Rouzbehani, *On-demand RAN Slicing Techniques for SLA Assurance in Virtual Wireless Networks*, Ph.D. Thesis, IST, University of Lisbon, Lisbon, Portugal, 2019.

[RWNL17] C. Ranaweera, E. Wong, A. Nirmalathas, C. Jayasundara and C. Lim, “5G C-RAN architecture: A comparison of multiple optical fronthaul networks”, in *Proc. of ICONDM17 – IEEE International Conference on Optical Network Design and Modelling*, Budapest, Hungary, May 2017. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7958544.

[Sara18] R. Saracco, *Data at the edges*, IEEE Future Directions, Oct. 2018. Available: http://sites.ieee.org/futuredirections/2018/10/17/data-at-the-edges/.

[Saty17] M. Satyanarayanan, “Edge Computing”, *Computer*, Vol. 50, No. 10, Oct. 2017, pp. 36-38. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8057308.

[Silv16] H. Silva, *Design of C-RAN Fronthaul for Existing LTE Networks*, M.Sc. Thesis, IST, University of Lisbon, Lisbon, Portugal, 2016.

[SKRA18] I. Sarrigiannis, E. Kartsakli, K. Ramantas, A. Antonopoulos and C. Verikoukis, “Application and Network VNF migration in a MEC-enable 5G Architecture”, in *Proc. of CAMAD18 - 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, Barcelona, Spain, Sep. 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8514943.

[TSMF17] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta and D. Sabella, “On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration”, *IEEE Communications Surveys & Tutorials*, Vol. 19, No. 3, May 2017, pp. 1657-1681. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7931566.

[WZHW15] J. Wu, Z. Zhang, Y. Hong and Y. Wen, “Cloud radio access network (C-RAN): a primer”, *IEEE Network*, Vol. 29, No. 1, Jan. 2015, pp. 35-41.

[ZBAF17] A. A. Zaidi, R. Baldemair, M. Andersson, S. Faxér, V. Molés-Cases and Z. Wang, *Designing for the future: the 5G NR physical layer*, Ericsson Technology Review, Jun. 2017. Available: https://www.ericsson.com/assets/local/publications/ericsson-technology-review/docs/2017/designing-for-the-future---the-5g-nr-physical-layer.pdf.