Analysis of the Implementation of Network Slicing in 5G Radio Networks

Sérgio Ferrão Batarda Marinheiro

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. Luís Manuel de Jesus Sousa Correia

Members of Committee: Prof. António José Castelo Branco Rodrigues

Eng. Ricardo Dinis

January 2021

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

To my beloved ones.

Acknowledgements

I would like to express my appreciation to Professor Luís M. Correia for the opportunity to develop my master thesis in collaboration with one of the telecommunications operator in Portugal under his supervision and allowing me to be part of GROW. I am very thankful for all the availability of Professor Luís M. Correia, scheduling weekly meetings, for the last year, to guide me in my work. These meetings helped me academically and personally by making me want to always achieve perfection.

I thank all GROW members, in special to Frederico Maia e Moura, Nuno Silva and Behnam Rouzbehani for all the valuable advice and time spent together.

To Eng. Ricardo Dinis, from NOS, for the opportunity to work closely with the industry and providing me with valuable information, meetings and always being available to help me with technical and critical support in case of need.

To all the friends I made through my journey at Instituto Superior Técnico, that were there through good and bad times.

To my family, my mother, Ana Marinheiro, father, Marcos Marinheiro, and brother, Pedro Marinheiro, for all the love, support, understanding, through my academic and personal life, always supporting my choices that make me the person I am today.

To my girlfriend, Beatriz Marques, for all the emotional support, love, always believing in me and helping me to pursuit my dreams.

Abstract

The main goal of this thesis was to analyse one of the main technologies used in 5G radio networks, network slicing. To achieve this, several scenarios, resulting from making variations of a reference scenario, were deployed and further analysed. This work focuses on how the data rate should be allocated among different slices and users to maximise cell capacity. Several types of services, Service Level Agreements and priorities were taken into consideration. Alongside the parameters regarding the scenario with all the slices and their specifications, the model takes as input the necessary parameters to compute the maximum achievable cell data rate. The model considers six parameters to evaluate data rate allocation, including the percentage of total assigned data rate, the virtual capacity share, the total data rate of each service, the percentage of served users, the data rate of each user and users' satisfaction. The results show that the models' serving weights enable slice isolation and proper data rate allocation according to its services. Also, one can confirm that the total assigned data rate is always 100% independently of the scenario and number of users/services.

Keywords

5G, Network Slicing, VNO, SLA, eMBB, URLLC.

Resumo

O objetivo principal desta tese foi o de analisar uma das principais tecnologias utilizadas em redes rádio 5G, *network slicing*. Para alcançar este objetivo, vários cenários, resultantes de variações ao cenário de referência, foram desenvolvidos e analisados. Este trabalho foca-se em como a transmissão de dados deve ser distribuída pelas várias fatias e utilizadores para maximizar a capacidade da célula. Vários tipos de serviços, *Service Level Agreements* e prioridades são tomados em consideração ao estudar este problema. Para além dos parâmetros relativos ao cenário, que contêm todas as fatias e as suas devidas especificações, o modelo considera ainda como parâmetros de entrada, os parâmetros necessários para calcular a taxa de transmissão máxima de uma célula. O modelo considera seis parâmetros para avaliar a atribuição da taxa de transmissão incluindo a percentagem de taxa de transmissão atribuída, a divisão de capacidade virtual, a taxa de transmissão total de cada serviço, a percentagem de utilizadores servidos, a taxa de transmissão a cada utilizador e a satisfação dos utilizadores. Os resultados obtidos mostram que pesos de serviços permitem o isolamento entre fatias e uma atribuição da taxa de transmissão apropriada de acordo com os seus serviços. Também é possível confirmar que a taxa de transmissão atribuída total é sempre 100% independentemente do cenário e número de utilizadores/serviços.

Palavras-chave

5G, Slicing da Rede, VNO, SLA, eMBB, URLLC.

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5GC	5G Core
API	Application Programming Interface
AR	Augmented Reality
BE	Best Effort
BG	Best Effort with Minimum Guaranteed
BkG	Background
BPSK	Binary Phase-Shift Keying
BS	Base Station
CAPEX	Capital Expenditures
CN	Core Network
CNF	Core Network Functions
CP	Control Plane
CPr	Cyclic Prefix
DFT-s-OFDM	OFDM with Discrete Fourier Transform Precoding
DL	Downlink
E2E	End-to-End
eMBB	Enhanced Mobile Broadband
eNodeB	E-UTRAN Node B
EPC	Evolved Packet Core Network
EPS	Evolved Packet System
E-UTRAN	Evolved UTRAN
FDD	Frequency Division Duplex
FR1	Frequency Range 1
GB	Guaranteed Bit Rate
gNodeB	Fifth Generation NodeB
HSS	Home Subscription Server
IA	Interactive
IMS	IP Multimedia Subsystem
IMT-2020	International Mobile Telecommunications 2020
InP	Infrastructure Provider
IP	Internet Protocol

ITU-R	International Telecommunication Union Radiocommunication Sector
KPI	Key Performance Indicator
LTE	Long Term Evolution
MIMO	Multiple Input Multiple Output
MLPS	Multi-Protocol Label Switching
MM	Mobility Management
MME	Mobility Management Entity
mMIMO	Massive MIMO
mMTC	massive Machine-Type Communications
MOS	Mean Opinion Score
NFV	Network Function Virtualisation
ng-eNodeB	LTE Next Generation eNodeB
NR	New Radio
NSA	Non-Standalone
NSaaS	Network Slicing as a Service
NSSF	Network Slice Selection Function
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
ONF	Open Networking Foundation
OPEX	Operating Expenditures
PAPR	Peak-to-Average Power Ratio
PCRF	Policy and Charging Resource Function
PDB	Packet Delay Budget
PELR	Packet Error Loss Rate
P-GW	Packet Data Network Gateway
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifiers
QoE	Quality of Equipment
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RA	Radio Access
RAN	Radio Access Network
RB	Resource Block
RE	Resource Element
RNF	Radio Network Functions
RRM	Radio Resource Management
RS	Reference Signal
RT	Real time
SA	Standalone
SC-FDMA	Single Carrier Frequency Division Multiple Access

SCMA	Sparse Code Multiple Access
SCS	Subcarrier Spacing
SDN	Software-Defined Networks
S-GW	Serving Gateway
SLA	Service Level Agreement
SNR	Signal to Noise Ratio
SP	Service Provider
TDD	Time Division Duplex
TE	Terminal Equipment
UE	User Equipment
UICC	Universal Integrated Circuit Card
UL	Uplink
UMTS	Universal Mobile Telecommunication System
UP	User Plane
URLLC	Ultra-Reliable and Low Latency Communications
USIM	Universal Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VM	Virtual Machine
VNF	Virtual Network Function
VNO	Virtual Network Operator
VoIP	Voice over IP
VR	Virtual Reality
VRRM	Virtual Radio Resource Management

List of Symbols

γ_v	Priority defined by InP and assigned to VNO v
δ_s	Serving weight, assigned to service s
λ_{v_s}	Tuning weight associated with service s, provided by VNO v, to prioritise data rate assignment
μ_L	Lagrange multiplier corresponding to the inequality constraint
$f^{(j)}$	Scaling factor
J	Number of aggregated component carriers in a band or band combination
M_{UMIMO}	Multi-user MIMO factor
$N_{v_s}^{usr}$	Number of users performing service s, from VNO v
$N_{PRB}^{BW(j),\mu}$	Maximum number of RB allocation in bandwidth $BW^{(j)}$ with numerology μ
N^{srv}	Number of services
$O_h^{(j)}$	Overhead
$Q_m^{(j)}$	Maximum supported modulation order
$p_{v_s[\%]}^{usr_{net}}$	Percentage of served users
$p_{VRRM[\%]}^{tot}$	Percentage of total assigned data rate
$R_{v_s,i\ [\mathrm{Mbps}]}^{usr}$	Data rate of each user
$R_{v_s[\mathrm{Mbps}]}^{srv_{tot}}$	Total data rate of each service
$R_{VRRM[\%]}^{VNO_v}$	VRRM capacity share
$R_{cell\ [\mathrm{Mbps}]}$	Total available data rate of the cell
R_{max}	Maximum coding rate
$R_{vs\ [Mbps]}^{srv_{max}}$	Maximum assignable data rate to the user of service s, from VNO v
$R_{vs\ [Mbps]}^{srv_{min}}$	Minimum assignable data rate to the user of service s, from VNO v
$R_{vs\ [Mbps]}^{srv}$	Total served data rate of service v_s

$\boldsymbol{R^{srv}}$	Vector of serving data rates
$S_{v_s,i}^{usr}$	Users' satisfaction
T_s^{μ}	Average OFDM symbol duration in a subframe for numerology μ
$v_{Layers}^{(j)}$	Maximum number of supported MIMO layers
$w_{v_s,i}^{usr}$	Assigned weight to user i, performing service s, from VNO v
$\boldsymbol{w^{usr}}$	Vector of users' weights, to obtain the long-term average data rate of users

Chapter 1

Introduction

In this chapter, a brief overview of the 5G radio system, its requirements and the emergence of network slicing is presented. Next, one describes the motivation for the study of the subject relative to network slicing and the contents in this work.

1.1 Overview

There have been four generations of mobile networks and we are currently in the beginning of the fifth, 5G, which comes with a lot of expectations regarding its capabilities, like very high peak data rates, ultra-low latency and being able to support massive amounts of devices connected to the network at the same time. Several applications are expected to appear or be enhanced with 5G capabilities, such as autonomous driving, remote surgery, smart power grids, smart cities, industrial automation and many more. To support all these demands, new technologies need to be used.

Major growth in traffic demand, mainly due to the increased number of smartphones and high-quality video resolution, is being verified. In the third quarter of 2019, mobile data traffic grew 68% year-on-year [Eric19]. Figure 1.1 supports this statement by showing the total global monthly data and voice traffic from the first quarter of 2014 until the third quarter of 2019 and the year-on-year percentage change for mobile traffic.

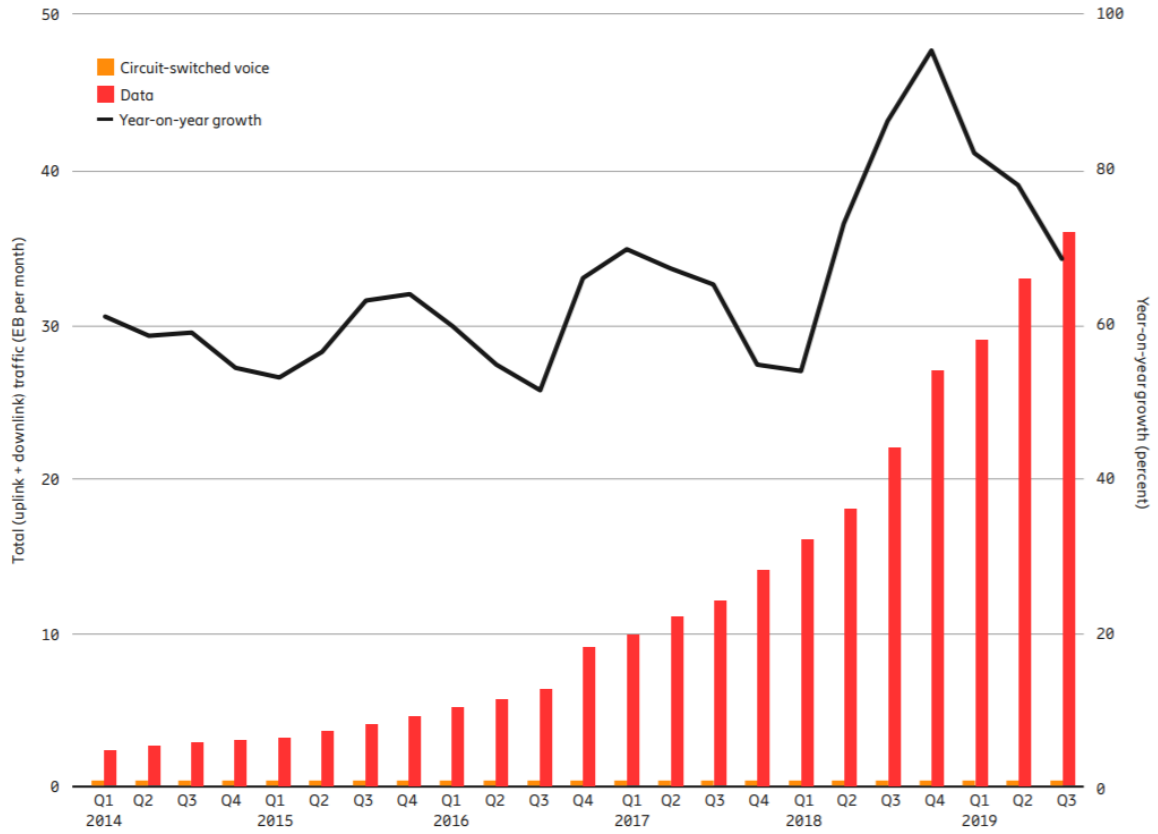


Figure 1.1 - Global mobile data traffic and year-on-year growth (extracted from [Eric19]).

The clear need for improvement demands a network with higher scalability flexibility and with a more efficient usage of its resources. Network slicing appears to tackle these problems and it is one of the most important and innovative technologies in 5G. This technology enables operators to deploy full networks in the form of logically separated and independent slices, each slice with its on configuration and customised according to one specific use case. This way, instead of trying to optimise one network

for all users and all services, operators can focus on its time optimising slices for specific services [HTSc17].

Services like online gaming and video streaming have become more popular in the last years. Consider the following scenario: a user is listening to music online while playing an online game and downloading a large file from a cloud storage. The three services are sharing the same radio access network, thus there is the possibility of quality degradation, such as intermittent music, delay in game interactions and low download speeds. Nevertheless, the lack of resources is not the main concern in this situation, because each of the described services have specific performance demands. Online music streaming services tolerate a high buffering times in the beginning, but while playing music there is a minimum threshold of data rate with high availability and high retainability that must be fulfilled. On the other hand, online games need low latency with high retainability and reliability to ensure a good gaming experience without lag spikes. Cloud storage synchronisation requires high channel capacity. The fact that the network is not optimised for each service may lead to scenarios where it is just not possible to support all new services, resulting in poor user experience. It is in scenarios like this that network slicing helps solve problems with more flexible and efficient resource allocation. Logical slices can be allocated physical network resources according to its services demands [HTSc17].

Current solutions exploited over 4G systems are not able to perform resource allocation in slicing environments, which means that mobile phones receive resources from the same traffic with equal priority levels. This problem arises from the way resource allocation is managed in 4G, which is by associating a priority to a service requested by the user. In contrast, in 5G multiple users can be associated to multiple slices, where each slice has a designated priority, resulting in a priority that considers not only the priority of the slice as well as the priority of the service the user is requiring. Hence, higher user experiences are achieved in 5G compared to 4G due to traffic being satisfied by different slices [JiCM16].

1.2 Motivation and Contents

In order to provide better services that meet the demands that come with the new generation of mobile communications, Software-Defined Networks (SDN), Network Function Virtualisation (NFV), network virtualisation and network slicing technologies have been proposed. By combining these technologies, more efficient networks are accomplished, leading to a better performance concerning data rate, latency, flexibility and connectivity capacity. By using network slicing it is possible to deploy multiple logical networks over the same infrastructure. The main advantage is that each slice is focused on one service (with the possibility of serving multiple services) and so provide a better service experience to the user. The resource management greatly benefits from network slicing as well due to the ability of only investing the necessary resources for each slice. The challenges that come with this technology are the proper resource allocation to each slice and the proper isolation among slices which is the main objective of this thesis.

This report is organised in five chapters.

Chapter 1, the current chapter, presents the introduction to this work. A brief overview of the subject and the respective motivation for the work are here described.

Chapter 2 provides an overview of 4G and 5G regarding its architecture and interface, and coverage and capacity are addressed. The various network architecture deployment options and the novelties of the 5G radio interface are described. Entering the network virtualisation domain, the concept is explained as are the technologies SFN, SFV and network slicing. A characterisation of the performance parameter is done and the degree of relevancy depending on the type of service is evaluated. The services and applications for 5G and of its QoS classes are presented. The chapter ends with the state of the art for the thesis subject.

Chapter 3 gives the model overview with the three main parts of the model. First, the model inputs composed of cell inputs used to calculate the maximum achievable cell data rate, and the network inputs that define a scenario with multiple slices services and users that need to be allocated data rate. The second part describes the model calculations and optimisations that are performed to solve the proposed problem. The third part contains the model outputs with all the metrics meant to study the model behaviour and with relevant values to study. Like the inputs, the model outputs are also divided into two categories: the network category with metrics relevant from the network viewpoint, and the user category with metrics relevant from the user viewpoint. Next follows the model development subsection with the purpose of describing all steps and calculations performed in the model. The model implementation is explained using flowcharts that describe the way the model behaves and reacts to different situations. Finally, the last subsection is dedicated to assessing the model.

Chapter 4 presents the several scenarios developed and studied. The scenarios suffer some variations that intend to comprehend/evaluate/analyse how the model reacts to real world scenarios and what can be expected from 5G use cases.

Chapter 5 concludes the thesis by summarising the main conclusions of the work and some final remarks regarding future work are discussed.

Chapter 2

Fundamental Concepts

This chapter provides an overview of 4G and 5G systems. Section 2.1 presents a study of the 4G system in terms of its architecture, interface, coverage and capacity. Section 2.2 starts by clarifying the architectural options for SA and NSA, and clarifies the main differences between the 5G radio interface and the 4G one. Section 2.3 defines SDN, network virtualisation, NFV, network slicing and how these technologies interact with each other. Section 2.4 clarifies what type of services and applications are to be expected for the 5G network. Section 2.5 addresses the performance parameters of the 5G network. Section 2.6 finalises the chapter with the state of the art.

2.1 LTE Concepts

2.1.1 Network Architecture

This subsection is based on [HoTo11]. 4G, Long Term Evolution (LTE), is an end-to-end (E2E) all IP network, that allows low latency and low-cost, the network architecture being described in Figure 2.1, and it is divided into four main high-level domains: User Equipment (UE), Evolved UTRAN (E-UTRAN), Evolved Packet Core Network (EPC) and the Services domain.

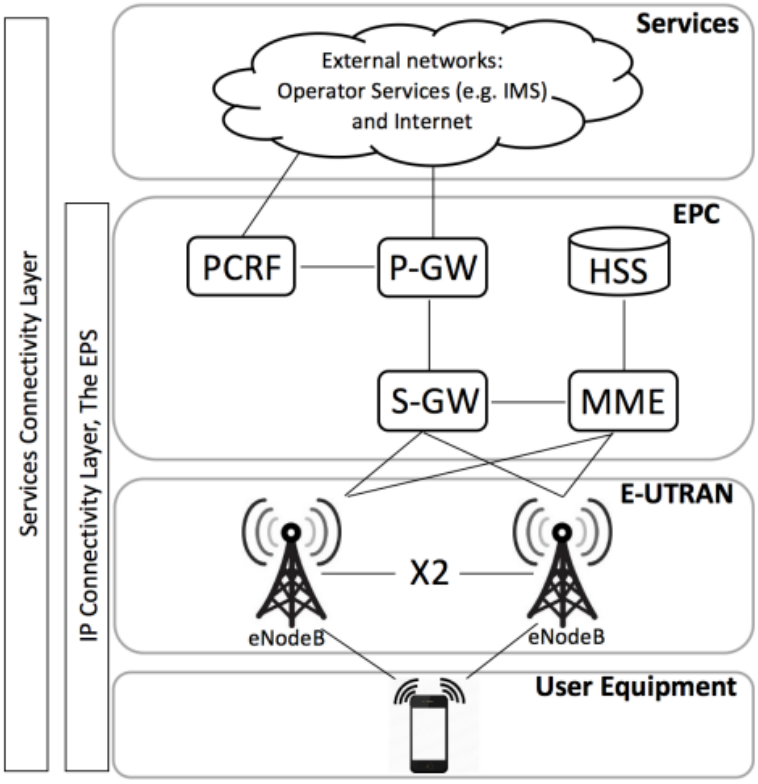


Figure 2.1 - System architecture for E-UTRAN (adapted from [HoTo11]).

Evolved Packet System (EPS) is the part of the network that includes the UE, the E-UTRAN, and the EPC. Together these domains represent the Internet Protocol (IP) Connectivity Layer, which the main purpose is to provide IP based connectivity. The functionalities of each node are explained in what follows.

The UE is the end terminal that the user uses for communication and includes both the Universal Subscriber Identity Module (USIM) and the Terminal Equipment (TE). USIM is used to identify and authenticate the user and to derive security keys for protecting the radio interface transmission. USIM

is an application placed into a removable smart card called the Universal Integrated Circuit Card (UICC). The UE allows the user to set up, maintain and remove the communication link with the network. All the mobility management functions, such as handovers and reporting the terminal location, are done according to network instructions.

The UE and the EPC are connected through the only node in the E-UTRAN: the E-UTRAN Node B (eNodeB). This is the termination point of all the radio protocols towards the UE and relays data between radio connection and the corresponding IP based connectivity towards the EPC. X2 is the interface among eNodeBs and can be divided into X2-CP, which consists of a signalling protocol, and the X2-UP, which is used to support loss-less mobility (packet forwarding). Several Control Plane (CP) functions, such as the Radio Resource Management (RRM) and the Mobility Management (MM), are controlled by the eNodeB. The RRM controls the usage of the radio interface, by constantly monitoring the resource usage situation, prioritising and scheduling traffic according to required Quality of Service (QoS), and allocating resources based on request. Regarding the MM, the eNodeB determines whether it should perform a handover by comparing his radio signal level measurements to the UEs. The eNodeB also establishes the routing from the UE and the Mobility Management Entity (MME) in case there is a new UE wanting to create a connection to the network, or the previous MME serving the UE becomes unavailable. It is also worth noting that the UE can only be connected to one eNodeB at the time, but the eNodeB can be serving multiple UE. The eNodeB can also be connected to multiple MMEs and Serving Gateways (S-GWs). Because each UE is only served by one MME and S-GW at a time, the eNodeB keeps track of this association.

The EPC domain includes several elements that are explained in what follows:

- MME: as the main control element in the EPC, the MME only operates in the CP and is not involved in the path of the User Plane (UP) data. It serves the purpose of operating some functions in the basic System Architecture Configuration, such as Authentication and Security, Mobility Management, and Managing Subscription Profile and Service Connectivity.
- S-GW: The S-GW manages his own resources and allocates them based on requests from MME, Packet Data Network Gateway (P-GW) or Policy and Charging Resource Function (PCRF). It establishes routes and forwards user data packets, while also acting as the mobility anchor for the user plane during inter-eNodeB handovers and between LTE and other 3GPP technologies.
- P-GW: The P-GW connects the EPC to the Services domain. It provides to the UE an IP, which it uses to communicate to other IP hosts in external networks. In addition to this, it also has the purpose of performing traffic gating and filtering functions.
- PCRF: based on QoS parameters, the PCRF decides how data flow is processed in the P-GW and ensures that it follows the users' subscription profile.
- Home Subscription Server (HSS): a database that stores users' data.

The last domain is the Services Domain. This domain includes various systems like the IP Multimedia Subsystem (IMS) based operator services, non-IMS based operator services and other services that are not provided by mobile network operators.

2.1.2 Radio Interface

This subsection is based on [HoTo11], [SeTB11], [Corr20], [Okub11] and [ANAC19a]. LTE uses two radio access methods: Orthogonal Frequency Division Multiple Access (OFDMA) for downlink (DL), and Single Carrier Frequency Division Multiple Access (SC-FDMA) for uplink (UL). By using OFDMA, it is possible to transmit very high data rate signals, that have very high-quality, in multipath mobile communication environments. SC-FDMA is used as the UL radio access method because it reduces the signal's Peak-to-Average Power Ratio (PAPR), resulting in lower costs regarding transmitter power amplifiers, and for power savings in the UE. ODFMA is a method that allows subcarriers to overlap with each other, in the frequency domain, by making so that the neighbour sub-carriers have no value at the centre frequency of the selected sub-carrier. This orthogonality uses a spacing of 15 kHz between each subcarrier, which allows a good tolerance for Doppler shift due to implementation imperfections and velocity.

LTE supports both Time Division Duplex (TDD) and Frequency Division Duplex (FDD). In Portugal, only FDD is used and the respective frequency bands are presented in Table 2.1.

Table 2.1 - Frequency bands use for LTE in Portugal (extracted from [ANAC19a]).

	Uplink [MHz]	**Downlink [MHz]**
LTE 800	[832, 862]	[791, 821]
LTE 1800	[1710, 1770]	[1805, 1865]
LTE 2600	[2510, 2570]	[2630, 2670]

Signals have three types of modulations for user data, these being Quadrature Phase Shift Keying (QPSK) and Quadrature Amplitude Modulation (QAM): 16QAM and 64QAM. 64QAM is the only modulation of the three that is only available depending on the UE capability. Both QPSK and 16QAM are available on all devices. In order to have good channel quality signalling and information and to avoid the resulting excessive overhead, OFDMA needs to use the same modulation for all the sub-carriers, even though it could use different modulations for each one. QPSK uses 2 bpsymbol, 16QAM uses 4 bpsymbol and 64QAM uses 6 bpsymbol.

Multiple Input Multiple Output (MIMO) operation including transmit diversity, pre-coding, and spatial multiplexing were some of the key technologies released with LTE. Spatial multiplexing consists of sending signals from two or more different antennas with different data streams. Using signal processing, in the receiver, it is possible to separate the data streams and thus increase the peak data rates by a factor of 2 or 4 depending on if the antenna configuration is 2×2 or 4×4, respectively. Pre-

coding, weights all the signals being transmitted from the antennas to ensure maximum Signal to Noise Ratio (SNR) at the receiver. Transmit diversity exploits the gains from independent fading between the antennas by sending multiple copies of the same signal, with some coding, from different antennas.

Resource Blocks (RB) correspond to the LTE physical channels that transport information in the radio interface. A unit of radio frame as 10 ms and is divided into 10 1 ms subframes, these being split into 2 RBs. As shown in Figure 2.2, each RB is the smallest unit that can be allocated to a user, as 0.5 ms, that corresponds to a time slot, and 12 sub-carriers, making a total of 180 kHz bandwidth (BW) required to be allocated. The slots are further divided into 7 symbols. There is a total of 84 Resource Elements (REs) in an RB each consisting of one sub-carrier with the duration of a symbol. While most of the REs are used for data transmission, some are reserved for synchronisation signals, control signalling and critical broadcast system information, and Reference Signals (RSs).

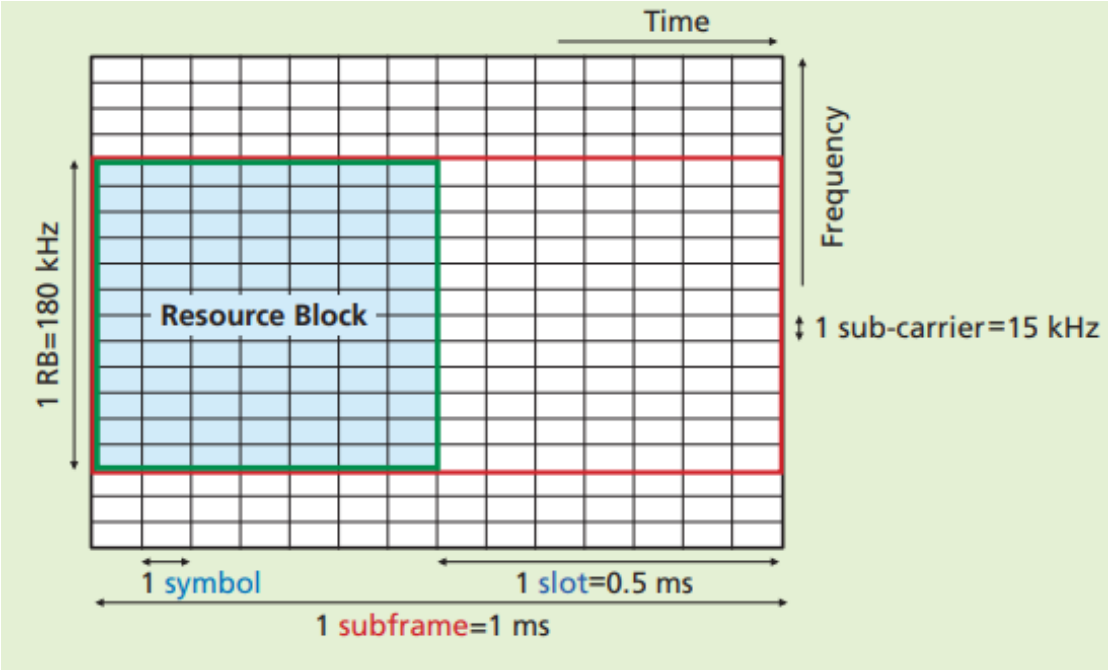


Figure 2.2 - Sub-frame configuration and RB placement (extracted from [Okub11]).

2.1.3 Coverage and Capacity

This subsection is based on [Shar18] and [Rouz19]. When designing a cellular network there are two concepts often associated one to another, in the sense that there is a trade-off to be made between them: coverage and capacity. Coverage is the geographical area where the base station (BS) can communicate and broadcast information to users, while capacity is the highest peak data rate, served by a cell at a given time, for that given area, and it is limited by bandwidth. With higher frequencies, it is possible to achieve higher levels of throughput but covering less area. There are other several aspects that influence a radio cell coverage distance, such as bandwidth and physical obstructions. In order to accommodate these restrictions, there are four types of cell sizes, represented in Table 2.2: macro-,

micro-, pico- and femto-cells.

Table 2.2 - Comparison of the different cell types (adapted from [Corr20]).

Cell		R [km]	P_{EIRP} [dBm]
Macro	Large	> 3	[50, 60]
	Small	1 - 3	[47, 50]
Micro		0.1 - 1	[30, 47]
Pico		< 0.1	[22, 33]
Femto		< 0.05	[10, 25]

Macro-cells are used as outdoor cells and offer a large coverage area that may go up to a radius of the order of tens of kilometres. The antennas are typically divided into 3 sectors, each covering 120º. Micro-cells cover smaller areas, such as malls or special events, with a radius of up to 1 km. Pico-cells have omnidirectional antennas, cover a maximum of 100 m, making them ideal for offices or railway stations, and support up to 100 users. Finally, femto-cells have omnidirectional antennas, that transmit low power, around 100 mW, and cover an area with a radius of less than 30 m.

In LTE, capacity can be described as the total among of aggregated data. In order to accommodate more users, it is necessary to have more RBs and consequently a higher bandwidth. Table 2.3 shows the number of RBs for each bandwidth.

Table 2.3 - Number of RB for each bandwidth (extracted from [Corr20]).

Bandwidth [MHz]	1.3	3	5	10	15	20
Number of RBs	6	15	25	50	75	100

Using the maximum bandwidth of 20 MHz, 64 QAM and MIMO, it is possible to achieve high peak data rates. Table 2.4 and Table 2.5 present the peak data rates for DL and UP, respectively. It is important to notice that signalling and control are not being considered. If they were the values would be around 20% lower.

Table 2.4 - DL peak data rates, in Mbps (adapted from [HoTo11]).

Resource blocks			Bandwidth [MHz]					
Modulation and coding	**Bpsymbol**	**MIMO usage**	**1.4**	**3**	**5**	**10**	**15**	**20**
QPSK ½	1.0	-	1.0	2.5	4.2	8.4	12.6	16.8
16QAM ½	2.0	-	2.0	5.0	8.4	16.8	25.2	33.6
16QAM ¾	3.0	-	3.0	7.6	12.6	25.2	37.8	50.4
64QAM ¾	4.5	-	4.5	11.3	18.9	37.8	56.7	75.6
64QAM 1/1	6.0	-	6.0	15.1	25.2	50.4	75.6	100.8
64QAM ¾	9.0	2 x 2	9.1	22.7	37.8	75.6	113.4	151.2
64QAM 1/1	12.0	2 x 2	12.1	30.2	50.4	100.8	151.2	201.6
64QAM 1/1	24.0	4 x 4	24.2	60.5	100.8	201.6	302.4	403.2

Table 2.5 - UL peak data rates, in Mbps (adapted from [HoTo11]).

Resource blocks		Bandwidth [MHz]					
Modulation and coding	**Bpsymbol**	**1.4**	**3**	**5**	**10**	**15**	**20**
QPSK ½	1.0	1.0	2.5	4.2	8.4	12.6	16.8
16QAM ½	2.0	2.0	5.0	8.4	16.8	25.2	33.6
16QAM ¾	3.0	3.0	7.6	12.6	25.2	37.8	50.4
16QAM 1/1	4.0	4.0	10.1	16.8	33.6	50.4	67.2
64QAM ¾	4.5	4.5	11.3	18.9	37.8	56.7	75.6
64QAM 1/1	6.0	6.0	15.1	25.2	50.4	75.6	100.8

2.2 5G Aspects

2.2.1 Standalone vs. Non-Standalone

This subsection is based on [GSMA18]. With 5G, a new radio interface is introduced, the New Radio (NR). With this new technology, unlike other generations where it is mandatory that both access and core network be part of the same generation, operators can use one of the following architecture deployment options: a 5G non-standalone (NSA) network or a full 5G standalone (SA) one. There are a total of 6 deployment options for SA and NSA, which are represented in Figure 2.3.

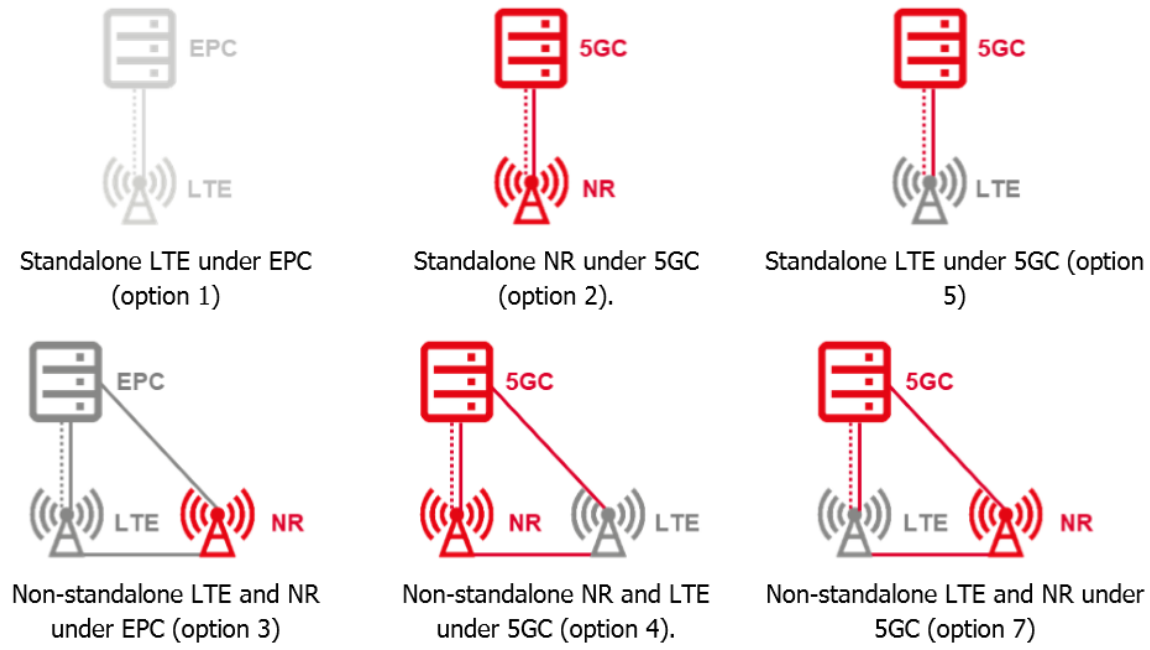


Figure 2.3 - SA and NSA deployment options (extracted from [GSMA18]).

As the name suggests, in an NSA architecture the NR and LTE radio cells are used together in order to provide Radio Access (RA), which is called dual connectivity. By combining the 5G Radio Access Network (RAN) and its NR with the existing LTE and EPC infrastructure Core Network (CN), 4G LTE services are supported with the benefit of using the 5G NR capacities, like lower latency, without replacing the network. This is one of the three options defined in 3GPP and is known as “Architecture Option 3”, where one has a dedicated LTE eNodeB acting as a master and an NR en-gNodeB acting as secondary. The other options are:

- Architecture Option 4: uses 5G Core (5GC) and a 5G NR base station (gNodeB) acting as a master and an LTE Next Generation eNodeB (ng-eNodeB) acting as a secondary.
- Architecture Option 7: uses 5GC and an LTE ng-eNodeB acting as a master and an NR gNodeB

acting as a secondary.

The NSA architecture is especially useful for transitioning to a full 5G deployment. The user experience is affected depending on which technology is being used.

On the other hand, an SA scenario implies a solo radio technology system where radio cells are both for user and control plane. This allows operators to deploy a network that uses inter-generation handover between 4G and 5G. Like in the NSA, there are three variations defined in 3GPP:

- Architecture Option 1: EPC and LTE eNodeB access.
- Architecture Option 2: 5GC and NR gNB access.
- Architecture Option 5: 5GC and LTE ng-eNodeB access.

2.2.2 Radio Interface

This subsection is based on [Eric18], [ANAC19b], [3GPP19] [ETSI18], and [BNKZ18]. 5G NR interface is already fully standardised since release 15 of 3GPP. It has a lot of similarities to LTE's radio interface, the differences being addressed in this section.

As a radio access method, 5G NR uses OFDMA with Cyclic Prefix (CPr) for DL. Unlike LTE, OFDMA is also possible to use for UL. OFDMA with Discrete Fourier Transform precoding (DFT-s-OFDM) with lower PAPR, can be used for the UL as a complement waveform in order to improve coverage, even though it is restricted to single-layer transmission only. To make sure it is possible to provide a variety of services on a wide range of frequencies, a scalable OFDMA is used in the sense that each subcarrier is separated 15×2^n kHz, with n ranging from 0 to 4, opposed to LTE's fixed 15 kHz.

NR supports two ranges of frequency bands, the low and mid bands ranging from 410 MHz to 7.125 GHz, and the high-frequency bands ranging from 24.250 GHz to 52.600 GHz. The low-frequency bands, that include the 600 MHz, 700 MHz, 800 MHz and 900 MHz, will mainly be used as a coverage layer, essential to support most NR usage in a wide area and deep indoor environments. The mid-frequency bands are the 1.5 GHz, 1.7 GHz, 1.8 GHz, 1.9 GHz, 2.0 GHz, 2.1 GHz, 2.3 GHz, 2.5 GHz 2.6 GHz, 3.5 GHz and 4.7 GHz, constitute the coverage and capacity layer, providing the best compromise between capacity and coverage. The high-frequency bands are the 26 GHz, 28 GHz and 39 GHz, being the super data layer, meaning they will provide extremely high data rates. The high-frequency band usage is limited due to the restrictions that millimetre waves impose. These waves only propagate in a line-of-sight path and suffer a lot of attenuation when contacting any kind of obstacle, like buildings, trees, and even heavy rain. For Portugal, the frequencies that are going to be allocated are the 700 MHz, 900 MHz, 1800 MHz, 2.1 GHz, 2.6 GHz and 3.5 GHz

5G NR supports all modulation schemes supported by LTE. In addition, it also supports 256 QAM, in DL, and Binary phase-shift keying (BPSK) π/2-BPSK, in UL, in order to improve power efficiency at lower data rates and reduce PAPR.

Massive MIMO (mMIMO) is supported in 5G NR, as an extension of MIMO technology. This technology uses a higher number of antennas at the BS to fulfil two main objectives. The first is to solve the problem

of coverage. Once can expect much higher signal attenuation for the higher frequency bands of NRB. Comparing to LTE's 2 GHz to 3 GHz frequency range, transmissions used for NR, over 28 GHz, are going to suffer a signal attenuation 100 times higher. The second big objective is to have a spectral efficiency 3 times higher than the current one for LTE. This is especially important for the sub-6 GHz spectrum since NR is competing with LTE in this spectrum.

2.3 Virtualisation

2.3.1 Network Virtualisation

This subsection is based on [WeTL13], [Rouz19], and [YLJZ14]. Virtualisation technologies have moved from server virtualisation to network virtualisation. Server virtualisation, or computer virtualisation, is a technology where multiple users are able to be connected and use the same server through different Virtual Machines (VMs). This is achieved by decoupling and abstracting the computing functionalities from the hardware. Server virtualisation allows on-demand and flexible management of computer resources, but does not include the virtualisation of the network fabric, e.g., switches and routers. Network virtualisation is a technology where multiple virtual networks share the same physical network infrastructure. This is achieved by decoupling the network infrastructure from the services that it provides, which allows maximum network reusability and different types of services to share the same infrastructure as presented in Figure 2.4.

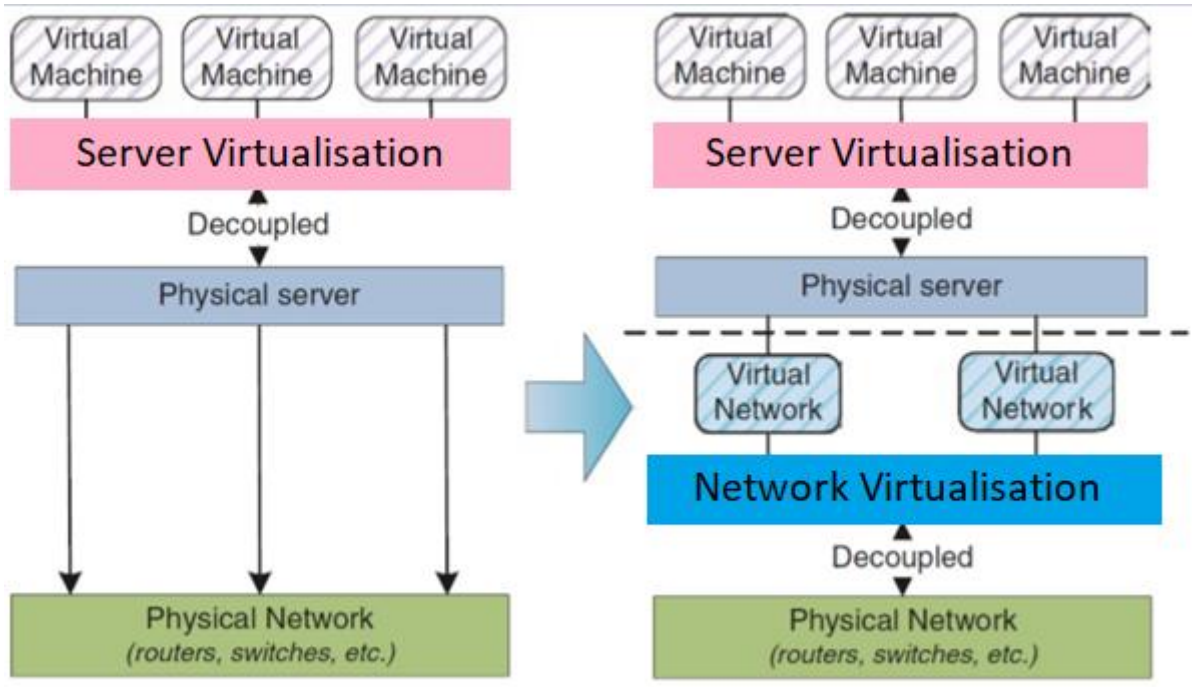


Figure 2.4 - Server virtualisation versus network virtualisation (adapted from [WeTL13]).

Network virtualisation and SDN technologies can significantly complement each other and efficiently address the challenges that come with wireless networks, such as Network Function Virtualisation (NFV). SDN separate the data from the control planes, with a centralised control plane, enhances programmability and customisation of network virtualisation. Network Virtualisation augments flexibility and scalability, that directly positively impact the resource utilisation in SDN. These two technologies are mutually beneficial however using both is not mandatory, meaning they are independent of each other.

Network virtualisation has also become involved in cloud computing applications, so that it is possible to have a more flexible management of the network between physical servers in a large datacentre.

2.3.2 Software-Defined Networks

This subsection is based on [JSSA14], [JZHT14] and [KFJK17]. Modern days Internet applications demand networks to carry high amounts of traffic, to be fast and to deploy several dynamic applications and services. Because the decision-making capability or network intelligence is spread among various hardware components, introducing a new network device involves a reconfiguration of all network nodes, which is something complicated. Legacy networks are difficult to automate due to their rigid structure that does not always allow the necessary programmability to satisfy the client requirements resulting in fragile and complex deployments that often need manually changes to the network components. There is a clear need for a network that is flexible, agile, efficient and scalable. To help mitigate these problems, Software-Defined Networks (SDN) are introduced. SDNs main purpose is to separate the forwarding/data plane from the control plane while ensuring programmability on the control plane. The idea is to have a centralised control plane intelligence with a separate data plane, which allows the administrator to configure the network hardware from the controller. This provides the ability to quickly respond to changing network conditions, business, market and end-users needs.

SDN bring several advantages, from which four are here highlighted. First, a granular network traffic control with high performance across devices from different network vendors. With the control of the network, network administrators can set different Quality of Equipment (QoE)/QoS policies at the application levels and network devices. Second, network applications can take advantage of centralised network intelligence and ultimately provide a better user experience. They can also adapt, according to users' needs, to network conditions. Third, because there is not a need to configure individual devices for new network services and capabilities, the network flexibility increases. Finally, the enhanced reliability of network security done by autonomic management of the network devices.

Open Networking Foundation (ONF), as presented in Figure 2.5, separates the SDN architecture in three layers: application, control and infrastructure. The application layer comprises the end-user business applications that consume the SDN network services and communications capabilities. The control layer comprises the software based SDN controllers that receive the application's requirements and transmits them to the networking components; it also extracts and flows information regarding statistics from the hardware devices to the SDN applications. The infrastructure layer includes network elements, such as physical and virtual switches.

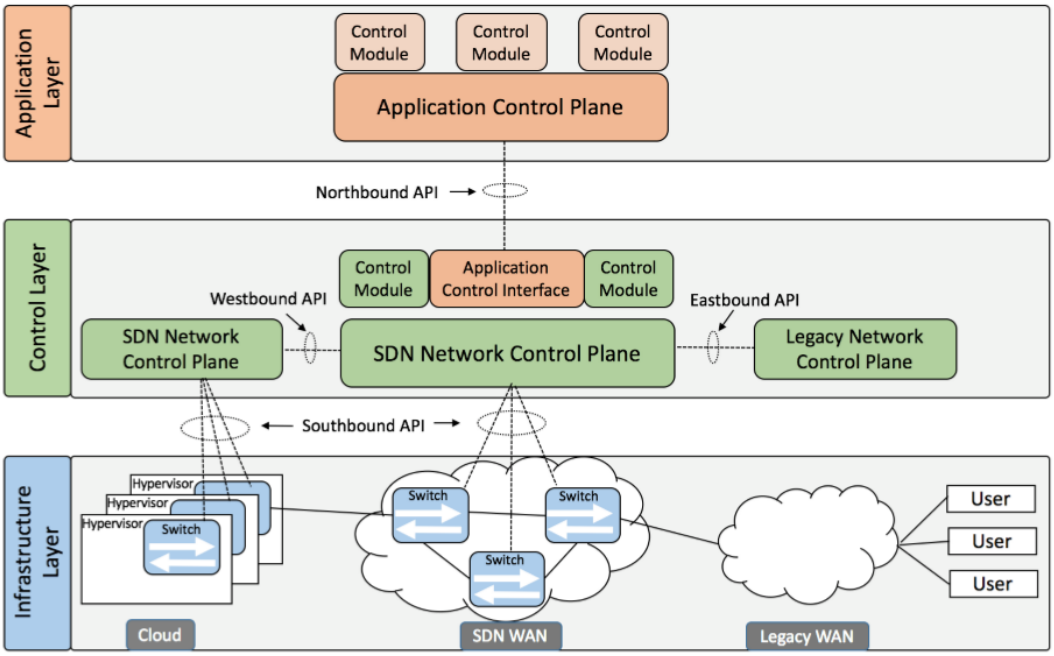


Figure 2.5 - Interfaces and layers of SDN (adapted from [JZHT14]).

The four key Application Programming Interfaces (APIs) are represented in Figure 2.5.

- Southbound-API is the interface between control and data planes. It enables the externalisation of the control plane. In order to control and manage the interface between various pieces of network equipment, it uses a protocol, being the most common the OpenFlow protocol.
- Northbound-API exchange information between the SDN controller and the application control planes, with the last running on top of the network. There is not a universal standardised Northbound API because the type of information exchanged, it is form and frequency depend on the application.
- Westbound-API routes the information between the SDN control plane of different network domains. This exchange of information serves the purpose of influencing routing decisions of each controller while enabling the seamless setup of network flow across multiple domains.
- Eastbound-API is the interface between the SDN control plane and legacy network control planes, for example, a Multi-Protocol Label Switching (MLPS) control plane. The technology used in the non-SDN domain will condition the implementation for the interface.

2.3.3 Network Functions Virtualisation

This subsection is based on [ETSI14], [HGLL15], [SaIF19], [JSSA14] and [HSMA14]. A network function is a network infrastructure component, usually physical with the need for manual installation into a computer network, that fulfils a specific role such as firewalls, routing and others. Combining with the virtualisation technology that was developed to complement the physical infrastructure appears the concept of NFV. NFV separates network functions from their physical hardware converting them into

software-based applications. At the core of NFV are virtual network functions (VNFs), that run on virtual machines and handle the specific network functions such as firewalls or routing. Several VNFs connected together make a virtualised environment. This results in a minimisation of the costs associated with hardware equipment, power, and cooling making the overall capital expenditures (CAPEX) and operating expenditures (OPEX) lower. NFV also provide scalability, flexibility, operating performance improvement, and shorter development cycles.

The framework of a high-level architectural NFV can be classified with 4 different blocks, presented in Figure 2.6:

- The orchestrator manages all software resources and the virtualised hardware infrastructure to provide networking services.
- The VNF manager controls all the events during the life cycle of a VNF like the instantiation, scaling, termination, and update.
- The virtualisation layer abstracts the physical resources and anchors the VNFs to the virtualised infrastructure. Offering standardised interfaces makes it so the VNF life cycle is independent of the underlying hardware platforms.
- The virtualised infrastructure manager is responsible for the virtualisation and management of the configurable computer, network, and storage resources, and control their interaction with VNFs. It also analyses the root cause of performance issues and gathers information about infrastructure fault for capacity planning and optimisation.

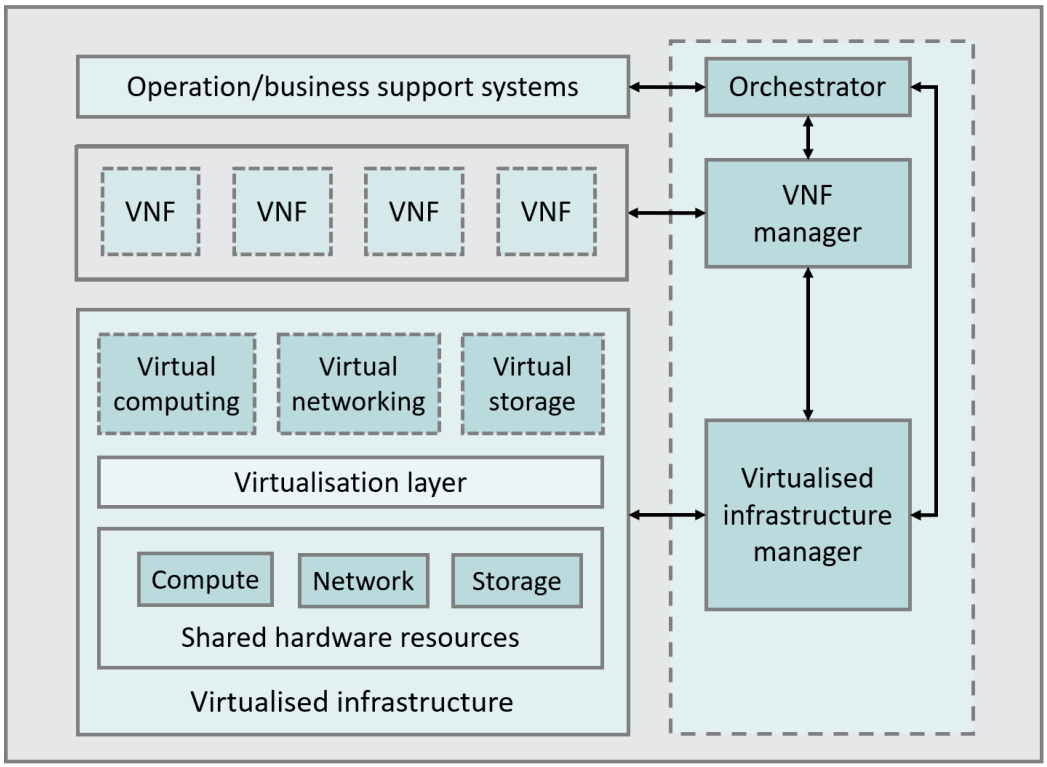


Figure 2.6 - NFV architectural framework (adapted from [HGLL15]).

Regarding NFV relationship with SDN, they are independent technologies that mutually benefit from each other, in the sense that combined they can help to mitigate the challenges of legacy networks. The OpenFlow switches can be controlled using NFV software, meaning that the SDN controller would be deployed as virtual functions. On the other hand, SDN can create a set of tunnels and virtual switches that do not allow any interaction between different virtual network functions software network making it so that NFV can support the use of a software overlay network. Combining these two technologies it is possible to replace expensive hardware equipment with software.

2.3.4 Network Slicing

This subsection is based on [ATSK18], [3GPP17], and [MBQB18]. 5G aims to support a lot of services for the vertical industries: transportations, factories and health, among others. Each of these sectors has a different set of needs with a different set of requirements. Although each sector may vary in the types of requirements it demands, these may be categorised as several key performance indicators (KPIs) such as user experienced data rate, E2E latency, reliability, communication efficiency, availability, and energy consumption. These KPIs must be fulfilled by different parameters, specific to the environment, like mobility, expected data traffic, density and types of network nodes, position accuracy, and others. Network slicing appears as a way to tackle these challenges and is one of the main enablers to support the required level of flexibility that comes with the requirements and desired KPIs in 5G networks. According to 3GPP, a network slice is “a logical network that provides specific network capabilities and network characteristics”, meaning that each slice is a logical network and all of the slices are being deployed over the same physical infrastructure and can be used simultaneously. Figure 2.7 presents a comparison between 4G and 5G systems.

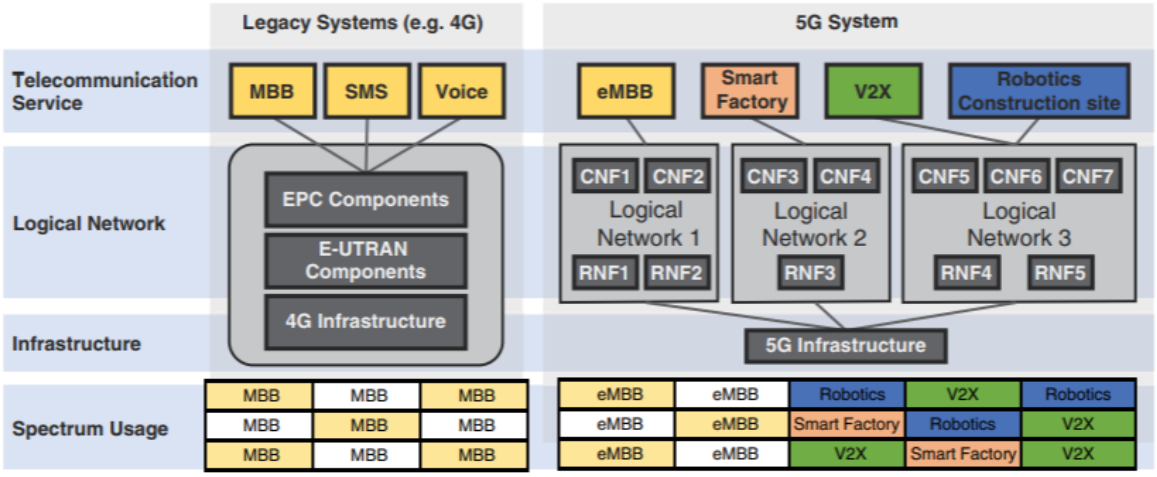


Figure 2.7 - Comparison between 4G and 5G networks (extracted from [MBQB18]).

As Figure 2.7 illustrates, the 4G system uses the same logical network independent of which service it

is providing. This type of solution is not possible for 5G services because there is no viable way to deploy KPIs like ultra-high reliability and ultra-low delay for a vertical industry. Contrasting with 4G, and other legacy systems, in order to support all the requirements from the vertical industries, 5G has multiple dedicated logical networks, the so-called slices, on top of the infrastructure. Running on the same physical network components, each logical network has core network functions (CNF) and radio network functions (RNF). The dedicated spectrum is allocated for each slice or several slices and manages to meet the service-level agreements (SLAs) with the vertical industries. NFV and SDN are crucial for the dynamic deployment of slices. NFV with its capabilities of virtualising multiple NF and further organise them into virtual blocks that create communication services. SDN capable of separating the control and data planes allowing full programmability of the network.

3GPP defined three types of predefined slices: type 1, type 2 and type 3. Network slice type 1 is dedicated to the support of enhanced mobile broadband (eMBB), type 2 for the support of ultra-reliable and low latency communications (URLLC) and type 3 is for the support of massive machine-type communications (mMTC). A new NF dedicated for network slices is also introduced: the "Network Slice Selection Function" (NSSF), with the purpose of selecting the most fitting slice. Network slicing supports roaming scenarios and the UEs may use multiple slices at the same time.

There are seven fundamental principles that define network slicing and its related operations:

- Automation: third parties are able to make a request for a slice creation that meets their needs besides the conventional SLA, meaning that not only the slice would reflect the desired features such as capacity, latency, jitter, among others, but also information regarding the duration of the network slice. Through signalling-based mechanisms, it is possible to have an on-demand configuration of network slicing without the need for fixed contractual agreements and manual intervention.
- Isolation: it is a property meant to guarantee performance and security requirements for each tenant. The notion of isolation involves both the data and the control planes, and the different degrees of isolations require different degrees of resource separation.
- Customisation: it guarantees that the resources allocated to a particular tenant are being used in an efficient way, that meets the respective service requirements. Network slices may be customised in a network-wide level, on the data plane or on the control plane. The different customisations vary according to network slices characteristics.
- Elasticity: it relates to the allocated resources of a particular network slice, in order to assure the desired SLA under varying conditions. Resource elasticity can be accomplished with a redistribution of the allocated resources or by altering the amount of initially allocated resources by modifying physical and VNF. Elasticity may require an interslice negotiation since changes to one slice may affect the performance of other slices that share the same resources.
- Programmability: it enables the control of slice resources, meaning networking and cloud resources, through open APIs.
- End-to-end: it is one of network slicing major characteristics, that facilitates the delivery from the provider to the end-user. A slice can use resources from different infrastructure providers in

order to deliver the service, and it includes various network layers and technologies, ranging from the core network to RAN.

- Hierarchical abstraction: it is a property of network slicing that has its roots on recursive virtualisation, wherein the resource abstraction procedure is repeated on a hierarchical pattern with each successively higher level, offering a greater abstraction with a broader scope. In other words, a lower level tenant can only partially control the resources of the network slice, and as the hierarchy level increases so does the control over the network slice resources.

With 5G network slicing the complexity regarding delivery and maintenance will increase, forcing the management of the network as a whole, in an integrated and coordinated unit, opposing to treating it as individual boxes and layers. This concept is known as network-wide orchestration and it is applied to 5G, contrasting with the node-by-node fashion services were managed in legacy systems. To realise this, it is necessary to configure every component of the network service at once with a higher-lever of abstraction and automated procedures

Network-wide orchestration brings the advantage of giving a single point of integration, providing a centralised representation of the distributed network. Having overall control of the network, independent of the number of resources involved or their physical locations, offers many benefits. This functionality makes the deployment of complex services across the network much easier due to the facilitated access to automation and KPI measurements. All configuration and monitoring tasks inherent to each service can be performed from the orchestration module. To fully benefit from network-wide orchestration, 5G networks should consider two levels:

- Inter-slice orchestration, which deals with the orchestration of resources across different network slices in the network.
- Intra-slice orchestration, referring to the orchestration of resources within each network slices.

These levels could be further divided by the operator responsible for the slices, thus creating intra-operator slice orchestration and inter-operator slice orchestration.

2.4 Services and Applications

This subsection is based on [ITUR15], [BPNA16], [Conne18] and [Mark99]. With the new radio system, 5G enables very high data rate speeds that benefit a lot of services and applications. There are a lot of real use cases already being thought and developed base on the new system, such as autonomous driving and remote surgery, but there are much more that have not even been considered yet. ITU-R defined three types of services, as illustrated in Figure 2.8: eMBB, mMTC and URLLC.

eMBB aims to target the traffic growth demand in multi-media content, services and data. This category covers usage scenarios such as Virtual Reality (VR), Augmented Reality (AR), wide-area coverage and hotspot. Each scenario demands different requirements, for example, regarding the wide-area coverage case, here seamless coverage with medium to high mobility is crucial. Although it is important, it does

not require as much data rate as a hotspot, which has a very high user density, a big necessity for high traffic capacity, but a much lower mobility requirement. There are three main aspects to considering in order to satisfy the data rate demand: densifying networks with small cell deployments, delivering high spectral efficiency and gaining access to more spectrum.

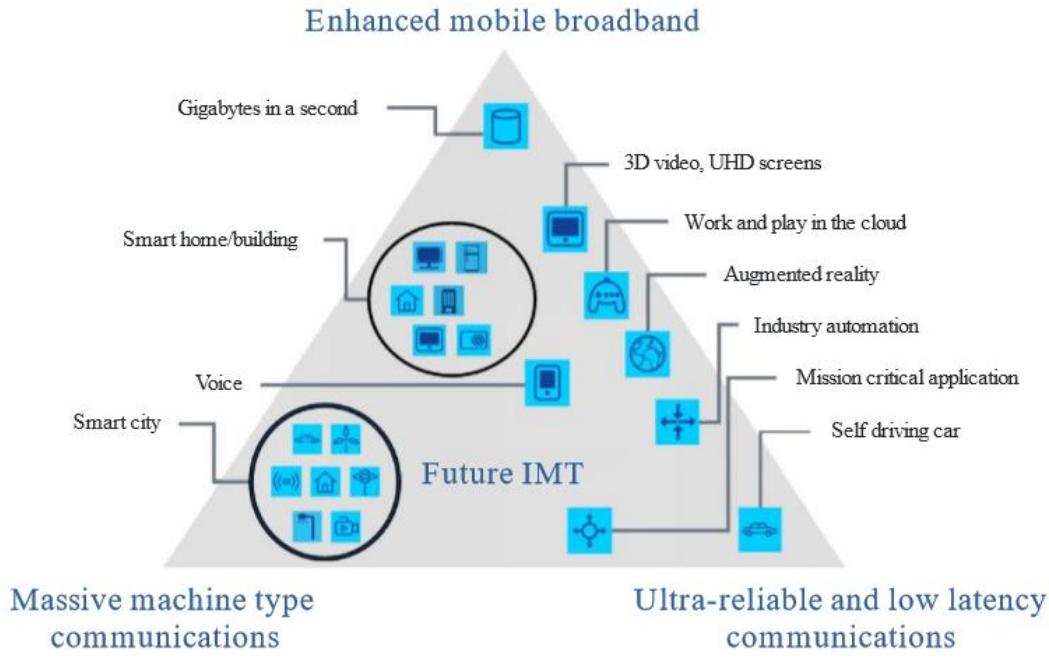


Figure 2.8 - Usage scenarios of IMT for 2020 and beyond (extracted from [ITUR15]).

mMTC is reserved for a very high number of low power devices that transmit low volumes of non-delay sensitive data. The most common use case is sensors that communicate among each other in other to achieve the desired goal. The typical requirements for mMTC are low user data rates (around 10 kbps), a large number of users (up to 300 000 per cell), mostly UL transmissions, low power consumption devices, and scattered user activity. Because eMBB and mMTC have such different targets, with the first heavily relying on DL communication and the second relying more on UL with low data rates, the technologies to support them are also different. To address the need for massive amounts of connected devices Sparse Code Multiple Access (SCMA) and Compressed Sensing based Multi-User Detection are possible solutions. URLLC targets all the applications that require very high levels of throughput, availability, reliability and ultra-low latency. The applications that fall under this category are autonomous vehicles, public and mass transit systems, drones, remote healthcare and smart grid monitoring.

5G uses the same QoS used for legacy systems, and these can be categorised as four different classes.

- Conversational: this class is meant to evaluate the QoS like telephony speech, voice over IP (VoIP) and video conferencing tools. Due to the nature of the services being real-time conversation schemes, the required characteristics are given by human perception. In order to make the conversation seems as natural as possible, the latency must be low. Failing to provide a low enough delay results in an undesired low-quality service, therefore the maximum transfer

delay, that is given by the human perception of video and audio conversation, must be very strict. The real-time conversation fundamental characteristics are to preserve time relation between information entities of the stream and to have a low delay conversational pattern.

- Streaming: this class is not heavily affected by some transfer delay, but the delay variation of the end-to-end flow must be limited to ensure the time relation between video and audio. There are time-alignment applications that correct small variations at the receiving end of the stream, so the maximum delay variation is defined by what type of application is controlling it. The real-time streams fundamental characteristic is the preserve time relation between information entities of the stream.
- Interactive: this class is applied when there is an interaction between the user (either human or machine) and remote equipment. This can involve activities like web browsing, server access, and polling for measurement records. Interactive traffic is characterised by the request-response pattern of the end-user. The interactive traffic fundamental characteristics are the request-response pattern and the preserve payload content.
- Background: this class is applied when, for example, a computer sends and receives data-files in the background. This is a classical data communication scheme and it is characterised by the destination end-user not expecting data at a given time. Because of this, traffic is not delivery time-sensitive and must be transparently transferred. The background fundamental characteristics are the destination not expecting data within a certain time and preserve payload content.

Regarding virtual network operators (VNO) contract with the infrastructure provider, there are three levels defined. The first and highest priority level, is the guaranteed bitrate (GB), where the data rate allocated to users performing services from that VNO are within a determined fixed range of values. This means that there is a minimum and maximum value for the data rate allocated to that user, guarantying the users is always served. VNOs with this SLA can expect their services to have better QoE. Next, the best effort with minimum guaranteed (BG), where only the minimum value of allocated data rate to a user is guaranteed. For this SLA, the user is allocated data rate according to the priority associated with him, and the available capacity of the cell, never going bellow an established minimum. The best effort (BE) is the third and lowest priority level defined, where services have no minimum or maximum data rate values associated with them. The data rate allocated to users from VNOs with this type of SLA follow a best effort fashion. These users are the first to get penalised with high traffic loads. [Rouz19].

QoS Class Identifiers (QCI) are used to measure the performance of each service. QCI are divided in two resource types, the Guaranteed Bit Rate (GB) ideal for real time services, and the Non-GB used for non-real-time services. Each service is associated with a order of priority, a Packet Delay Budget (PDB) and the maximum Packet Error Loss Rate (PELR). Table 2.6 provides the standardise QCI characteristics for 4G and 5G networks.

Table 2.6 - Reference services characteristics (extracted from [Corr20]]).

QCI	Resource Type	Priority	PDB [ms]	PELR	Example Services
1	GB	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)
66		2	100	10^{-2}	Non-Mission Critical user plane Push To Talk voice
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 (small packets)
83		2.2	10	10^{-4}	Discrete Automation (big packets)
84		2.4	30	10^{-5}	Intelligent Transport Systems
85		2.1	5	10^{-5}	Electricity Distribution-high voltage
5	Non-GB	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		86.8	10	10^{-6}	Low latency eMBB application, Augmented Reality

2.5 Performance Parameters

The International Telecommunication Union Radiocommunication Sector (ITU-R) International Mobile Telecommunications 2020 (IMT-2020) provides far more enhanced capabilities than the ones described in Recommendation ITU-R M.1645 [ITUR15] and has key performance parameters that must be achieved. Table 2.7 summarises these parameters.

Table 2.7. 5G Performance parameters identified by ITU-R IMT 2020 (extracted from [Domi18]).

Parameters	Values
Peak data rate	20 Gbps
User experienced data rate	100 Mbps
Latency	1 ms

There are also other parameters important to consider for this work:

- **Peak data rate**, that refers to the maximum data rate per device under ideal conditions (Gbps).
- **Latency**, the time that takes a packet to be transmitted from the source and received at the destination (ms).
- **Allocated slice capacity,** that is the capacity allocated from the virtual radio resource management (VRRM) to a slice (Mbps or Gbps).
- **Efficiency of slice sharing**, that is the efficient allocation of resources to slices, by monitoring the status of the network slices with respect to their SLAs.
- **Percentage of total assigned data rate**, one of the most important metrics, showing the total network throughput in terms of data rate allocation, the values closer to 100% obviously leading to a better VRRM performance.
- **VRRM capacity share**, the percentage of capacity allocated to each VNO, out of the total available VRRM one, being a key performance metric from VNOs and VRRM perspective.
- **Total data rate of each service**, it shows the total data rate assigned to each service slice of a VNO, being important from both VRRM and VNOs' viewpoints (Mbps or Gbps).
- **Percentage of served users**, the percentage of served users performing a specific service to the total number of offered ones, which is an essential metric for VNOs.
- **Users' satisfaction,** which is important from a VNO perspective because it reflects the user satisfaction. This metric is measured differently according to the service class.
- **Data rate of each user**, the data rate allocated to a user is an important QoS metric from both users' and VNOs' viewpoints, having a direct impact on the satisfaction level of the served users (Mbps or Gbps).

Although all these parameters are important, their degree of relevancy can vary depending on which scenario is being taken into consideration. Each parameter is classified, regarding one scenario, according to one of three importance levels: low, medium and high. For eMBB it is very important to

take the peak data rate, user experienced data rate and mobility. It is important to notice that although mobility and user experienced data rate are both classified in an equal manner, one can be higher or lower than the other considering the specific usage. For example, in a hotspot scenario, it would be required more user experienced data rate than mobility. In mMTC, the massive number of connected devices in the network justify the high need of connection density and it is also essential that the devices have a low energy consumption. URLLC needs very high latency and mobility capabilities because of the nature of its applications. Autonomous driving is an example of an application where not only the device is constantly moving but need to make decisions immediately.

2.6 State of the Art

The authors of [JiCM16] aim to tackle future conflicting demands that will appear in slice allocation. Network slicing will have conflicts regarding traffic prioritisation in the sense that it demands simultaneous management for priority among different slices and for priority among the users in the same slice. To solve this problem and with maximum user satisfaction as a goal, the authors propose a novel heuristic-based admission control mechanism that is capable of dynamically allocate network resources to different slices, while guaranteeing each slice meets its requirement. The model created has four main elements: the service slice layer, the virtual network layer, the physical resources, and the admission control manager. The service slice layer contains different services that need resources. The virtual network layer gives an abstraction of the physical network resources. The physical resources dictate what radio resources are available for the virtual network. Finally, the admission control manager is in charge of new incoming slices and optimising resource allocation. The authors demonstrate that their model achieves higher user experience in individual slices, optimises network resource usage, and provides higher scalability when the number of users per slice increase.

Similar to the work described above, a solution for optimising resource allocation in network slices is implemented in [SSCB17]. This greatly benefits infrastructure providers (InPs) due to simplifying the work of implementing new admission control policies each time a request is made by a network slice, depending on their SLA. InPs can use traffic forecasting techniques to make the proper resource allocation to each slice and still meet the slice's SLAs. Three main blocks are described in [SSCB17], these being: traffic analysis and prediction per network slice, admission control decisions for network slice requests, and adaptive correction of the forecasted load based on measured deviations.

Network slices can be quickly deployed by recurring to SDN and NFV technologies. This will result in simplified management and optimal resource allocation, however knowing how to efficiently allocate, manage and control the network slice resources will prove to be a challenging task. The algorithmic side of these challenges is focused on [Vass17].

The isolation challenges of network slicing are addressed in [KaSa18]. One of the main advantages of network virtualisation and network slicing is the capability of deploying multiple logical networks using

the same physical infrastructure. Guaranteeing the proper resource usage of each slice and ensuring that each slice is isolated from the other is not an easy task that, if not done properly, may lead to congestion between slices. [KaSa18] discuss these challenges in the context of a wireless system having a time-varying number of users that need reliable low latency and self-managed network slices. A novel control framework for stochastic optimisation based on the Lyapunov drift-plus-penalty method is proposed in [KaSa18]. The main advantages of this framework are minimising the power consumption of the system, slice isolation and providing low latency end-to-end communication for RLL slices. The InP provides to the service provider (SP) one network slice with E2E requirement in terms of rate, maximum tolerable delay and reliability number of the users. It is in the SPs best interest and responsibility to control the number of users per slice because exciding a certain threshold would result in a QoS drop. The model described in the paper takes into consideration the number of physical RBs and the number of users per slice during a certain time interval.

Two challenges regarding network slicing are emphasised in [SPFA17], these being sharing network resources and the influence that the transmitter may have on the receiver. With this in mind, the authors identify the problem in a multi-cell network concerning radio resource management (RRM) functionalities capabilities to support splitting the radio resources among the network slices. Four types of network slicing are presented: Spectrum Planning, Inter-Cell Interference Coordination, Packet Scheduling, and Admission Control. Furthermore, the different types of network slicing are compared regarding the granularity in frequency/time/special domain, its degree of customisation, radio-electrical isolation and traffic isolation.

Chapter 3

Model and Simulator Description

This chapter focuses on the developed model and its main stages. Firstly, one gives an overview containing the model inputs, the model calculations and the model outputs. Secondly, the model development subsection follows with all the calculations and used values discretised. Thirdly, the model flowcharts are presented, and at last the assessment of the model is made, considering a simple test scenario.

3.1 Model Overview

The purpose of this thesis is to develop a model capable of analysing the implementation of network slicing in 5G radio networks with service differentiation, by defining the resources allocated to each slice and the corresponding SLA in order to achieve near-optimum performance. Table 3.1 presents an overview of the model developed and both the inputs and outputs.

Table 3.1. Model overview.

Input	Model	Output
• **Cell** ➢ MIMO layers ➢ Numerology ➢ Bandwidth ➢ Multi-user MIMO • **Network** ➢ VNO ➢ Service type ➢ Service class ➢ Priorities ➢ Service data rates ➢ Contracted SLA ➢ Service mix ➢ Number of users	• Maximum achievable cell data rate calculation • Admission control and delay process • Maximise the usage of the available capacity, using VRRM optimisation • Output parameters calculation and analysis	• **Network** ➢ Percentage of total assigned data rate ➢ VRRM capacity share ➢ Total data rate of each service ➢ Percentage of served users • **Users** ➢ Data rate of each user ➢ Users' satisfaction

The model inputs consist of two classes: cell and network. Each of these classes have different parameters that are worth looking at.

Regarding the first class, Cell, it considers the following parameters to calculate the maximum data rate available at the cell: MIMO layers, Numerology, Bandwidth and Multi-user MIMO. These inputs are used in the equation for throughput calculation for NR provided by 3GPP. The remaining equation parameters are not listed as inputs because they are calculated with the ones that are listed as inputs. For example, to know how many RBs are available, one needs to know which numerology is being used and which is the available bandwidth. All these calculations are described in Section 3.2.1.

The second class, Network, is composed of several parameters that define one slice and are used as inputs for the VRRM model, these being: VNO, Service Type, Service Class, Priorities, Service data rates, Contracted SLA, Service mix and Number of users. VNO is the identification tag of VNO. Service Type refers to which type of service is being provided (e.g., voice). Service class is the class in which

the service is inserted (e.g., conversational). Priorities refers to the priority level given to the service. This parameter is composed of two variables, one regarding the priority given by the VNO and the other given by the InP. The Service data rates, which is the acceptable range of data rate for a given service. The Contracted SLA is the type of SLA defined between VNO and InP (GB, BG or BE). The Service mix, which is the percentage of users that are assigned to each service. Finally, the Number of users that is the total number of users allocated to the slice.

The model consists of four main stages. First the maximum achievable cell data rate calculation. This stage uses the cell input parameters detailed above, to make this calculation. The output serves as an input of the third stage, the VRRM optimisation. The second stage is the admission control and delay process, which is essential to guarantee that the VRRM problem is possible to solve. When the network is congested, and the total minimum thresholds of guaranteed demands get higher than the available VRRM capacity, it is necessary to delay the BE users because they do not have a guaranteed bit rate associated with their SLA. Then, low priority users is delayed one by one until the capacity is enough. This stage uses the capacity calculation of the first stage and network input parameters, specifically it makes the summation of the minimum demanded capacities of all services, in order to compare the two and make the admission. The third stage is solving the VRRM problem, which aims to maximise the usage of available capacity in order to satisfy the contracted SLAs to the highest possible level, considering services' priority, while distributing capacity in a fair manner, subject to some constraints, including maximum achievable capacity, predefined SLA thresholds [Rouz19]. This stage uses the calculation of the maximum achievable cell capacity provided by the first stage and the network input parameters.

The model outputs also divide into two different classes: network and users. The first one has parameters that are important from the network viewpoint, being: Percentage of total assigned data rate, VRRM capacity share, Total data rate of each service, Percentage of served users and Percentage of served users in each service. Percentage of total assigned data rate is the total network throughput in terms of data rate, meaning that if it is close to 100% it reflects an optimal VRRM performance. VRRM capacity share is the total capacity assigned to each slice out of the total capacity available to VRRM. Total data rate of each service is like the name suggests, the total data rate assigned to a given service. Percentage of served users is the percentage of users that are allocated to a given service out of the total number of users of that service. The second class contains parameters that are important from the users' viewpoint: Data rate of each user and Users' satisfaction. Data rate of each user is an important parameter because it has a direct impact on the satisfaction level of the served users. Users' satisfaction is a classification method that intends to determine if the data rate allocated to a user will have a good or a bad experience.

3.2 Model Development

3.2.1 Maximum Achievable Cell Data Rate

In order to calculate how the data rate distribution among the services is done, first one needs to know the maximum achievable data rate. For that purpose, one uses an adapted expression provided by 3GPP that takes into consideration several parameters, such as number of MIMO layers and modulation, making the calculation for the DL or UP data rate [3GPP20].

$$R_{cell\,[\mathrm{Mbps}]} = \sum_{j=1}^{J} \left(v_{Layers}^{(j)} \; Q_m^{(j)} \; f^{(j)} \; R_{max} \; \frac{12 \, N_{PRB}^{BW(j),\mu}}{T_{s\,[\mathrm{s}]}^{\mu}} \left(1 - O_h^{(j)}\right) M_{UMIMO} \right) \tag{3.1}$$

where:

- R_{cell}: Achievable data rate in one cell.
- J: Number of aggregated component carriers in a band or band combination.
- R_{max}: Maximum coding rate.
- For the j-th component carrier:
    - $v_{Layers}^{(j)}$: Maximum number of supported MIMO layers.
    - $Q_m^{(j)}$: Maximum supported modulation order.
    - $f^{(j)}$: Scaling factor.
    - $N_{RB}^{BW(j),\mu}$: Maximum number of RB allocation in bandwidth $BW^{(j)}$ with numerology μ.
    - T_s^{μ}: Average OFDM symbol duration in a subframe for numerology μ.
    - $O_h^{(j)}$: Overhead.
    - M_{UMIMO}: Multi-user MIMO factor.

The added parameter, M_{UMIMO}, takes into consideration Multi-user MIMO, which was lacking from the original expression, which allows the base station to serve multiple users, that are close to it, with the same RB making a much more efficient usage of resources.

In this work, one does not consider carrier aggregation, thus J is equal to 1, although it can go up to 16 [3GPP20]. MIMO layers, $v_{Layers}^{(j)}$, have a maximum value of 8 in DL and 4 in UL, but in this work one only considers 2x2 and 4x4, meaning 2 and 4 layers. The scaling factor, $f^{(j)}$, is signalled per band and per band combination, being set to 1, but it can take the values 0.4, 0.75, 0.8 and 1. The modulation order, $Q_m^{(j)}$, of a modulation scheme is determined by the number of different symbols that the modulation scheme can transmit, therefore, the modulation orders of QPSK, 16QAM, 64QAM, and 256QAM are respectively 2, 4, 6, and 8. One must keep in mind that for single carrier NR SA operation, the UE needs to support a data rate for the carrier with the product of $v_{Layers}^{(j)} \; Q_m^{(j)} \; f^{(j)}$ greater of equal to 4. This means that in order to use QPSK modulation, that has $Q_m^{(j)} = 2$, the minimum number of MIMO layers must be also 2, $v_{Layers}^{(j)} = 2$ [3GPP20].

In Table 3.2, one shows the coding rates of each modulation. These values are SINR dependent and, because the goal is to use the value that achieves the maximum possible data rate, the ones that are considered are the best coding rate values for the respective modulation. This means 449/1024 for QPSK, 616/1024 for 16QAM, 873/1024 for 64QAM, and 948/1024 for 256QAM.

Table 3.2 – Modulation schemes and code rate (adapted from [3GPP20]).

Modulation scheme	Code rate R_{max}
QPSK	78/1024
QPSK	193/1024
QPSK	449/1024
16QAM	378/1024
16QAM	490/1024
16QAM	616/1024
64QAM	466/1024
64QAM	567/1024
64QAM	666/1024
64QAM	772/1024
64QAM	873/1024
256QAM	711/1024
256QAM	797/1024
256QAM	885/1024
256QAM	948/1024

The overhead, $O_h^{(j)}$, assumes different values according to the frequency range and link. Because only frequency range 1 (FR1), that ranges from 410 MHz to 7125 MHz, is considered in this work the values taken by $O_h^{(j)}$ are:

- $O_h^{(j)} = 0.14$ in DL.
- $O_h^{(j)} = 0.08$ in UL.

5G uses a scalable OFDM in the sense that each subcarrier is separated 15×2^n kHz ensuring it is possible to provide a variety of services on a wide range of frequencies. As such, the available numerologies for FR1 are 0, 1, and 2 that correspond to a subcarrier spacing (SCS) of 15 kHz, 30 kHz

and 60 kHz respectively. With μ one can calculate T_s^μ as shown in

$$T_{s\,[s]}^{\mu} = \frac{10^{-3}}{14 \times 2^{\mu}} \tag{3.2}$$

To obtain the number of RB it is necessary to know both the BW as well as the numerology being used. Table 3.3 presents, according to 3GPP, the number of RB available for a given bandwidth and numerology. For Portugal, the frequencies that are allocated for LTE and NR and their respective available bandwidth per operator are listed below.

- For LTE [ANAC15]:
  - Frequency: 800 MHz; Bandwidth: 10 MHz.
  - Frequency: 1.8GHz; Bandwidth: 20 MHz.
  - Frequency: 2.6 GHz; Bandwidth: 20 MHz.
- For NR [ANAC19a]:
  - Frequency: 700 MHz; Bandwidth: 10 MHz.
  - Frequency: 3.5GHz; Bandwidth: 100 MHz.

Table 3.3 – RB per Numerology per Bandwidth for Frequency Range 1 [3GPP17b].

Bandwidth [MHz]	**SCS [kHz]**		
	15 (μ=0)	**30 (μ=1)**	**60 (μ=2)**
5	25	11	N/A
10	52	24	11
15	79	38	18
20	106	51	24
25	133	65	31
30	160	78	38
40	216	106	51
50	270	133	65
60	N/A	162	79
70	N/A	189	93
80	N/A	217	107
90	N/A	245	121
100	N/A	273	135

For the frequency of 3.5 GHz, TDD is used, so it is necessary to multiply the final value by the respective defined slot format. Usually, it is used a format that favours DL over UL like format 31 with 87.5% of the slots being dedicated to DL and the remaining 12.5% dedicated to UL. This is the ratio that is used in this work.

3.2.2 VRRM Optimisation

VRRM aims to maximise the usage of aggregated capacity and then allocate the resources to the different services of the VNOs by considering the set of available radio resources. This model was developed by [Rouz19] and its explanation is given bellow.

The model is formulated as a constrained concave optimisation problem. The objective function balances the efficiency and fairness when allocating resources to a multi service network. This function is a measure of efficiency and quantifies the expected users' satisfaction. The chosen utility function, in order to formulate the problem as a convex optimisation, is the logarithmic one. This function copes with the criterion of proportional fairness, by being an effective function to maximise the average long-term users' data rate while sharing the capacity among users in accordance with predefined weights [Rouz19]. The logarithmic utility function makes the balance between two competing interests, these being providing the maximum capacity to a user and at the same time providing all users with the minimal level of service. Therefore, the problem of the objective function of VRRM, $f_{VRRM}(\boldsymbol{R^{srv}})$, is formulated as the logarithm of the normalised weighted sum of the total data rate for different services:

$$f_{VRRM}(\boldsymbol{R^{srv}}) = \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log \frac{R^{srv}_{v_s\,[\mathrm{Mbps}]}}{R_{cell\,[\mathrm{Mbps}]}} \tag{3.3}$$

where:

- $\boldsymbol{R^{srv}}$: Vector of serving data rates, which can be written as $\boldsymbol{R^{srv}} = \left[R^{srv}_1, \dots, R^{srv}_{N_{srv}}\right]^T$.
- N^{srv}: Number of services.
- λ_{v_s}: Tuning weight associated with service s, provided by VNO v, to prioritise data rate assignment.
- $R^{srv}_{vs\,[Mbps]}$: Total served data rate of service v_s.
- $R_{cell\,[\mathrm{Mbps}]}$: Total available capacity at the cell.

In order to solve this problem, a standard technique based on Lagrange multipliers is used. This way one can compute the maximum value of the function. Since both objective and constraint have continuous first partial derivatives, a new variable, the Lagrange multiplier, is introduced to form the $L(\boldsymbol{R^{srv}}, \mu_L)$ Lagrangian:

$$L(\boldsymbol{R^{srv}}, \mu_L) = \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log \frac{R^{srv}_{vs\,[\mathrm{Mbps}]}}{R_{cell\,[\mathrm{Mbps}]}} + \mu_L \left(1 - \sum_{v_s=1}^{N^{srv}} \frac{R^{srv}_{v_s\,[\mathrm{Mbps}]}}{R_{cell\,[\mathrm{Mbps}]}}\right) \tag{3.4}$$

where:

- μ: Lagrange multiplier corresponding to the inequality constraint.

by taking the derivatives with respect to the variables, one obtains:

$$\begin{cases} \dfrac{\partial L(\boldsymbol{R^{srv}}, \mu_L)}{\partial R_{vs}^{srv}} = 0 \\ \dfrac{\partial L(\boldsymbol{R^{srv}}, \mu_L)}{\partial \mu_L} = 0 \end{cases} \rightarrow \begin{cases} \dfrac{\lambda_{v_s}}{R_{vs\,[Mbps]}^{srv}} = \mu_L \qquad v_s \in \{1, \dots, N_{srv}\} \\ R_{[Mbps]}^{cell} - \sum_{v_s=1}^{N^{srv}} R_{v_s\,[Mbps]}^{srv} = 0 \end{cases} \tag{3.5}$$

Finally, solving the two equations one obtains the allocated data rate of each service proportional to its serving weight:

$$R_{vs\,[\text{Mbps}]}^{srv} = \frac{\lambda_{v_s}}{\sum_{v_s=1}^{N^{srv}} \lambda_{v_s}} R_{cell\,[\text{Mbps}]} \tag{3.6}$$

By maximising the concave utility function in the form of a weighted logarithmic function, one can cope to the criterion of proportional fairness and this way maximise the usage of aggregated capacity, while maintaining a level of fairness in the process of resource allocation among users according to their priority of requested services. Expression (3.3) is then rewritten as (3.7) in order to further differentiate users' weights in each slice. Then the goal becomes to find the vector $\boldsymbol{w^{usr}}$ that maximises f_{VRRM}:

$$Max\ f_{VRRM}(\boldsymbol{w^{usr}}) = Max \sum_{v_s=1}^{N^{srv}} \lambda_{v_s} \log\left(\sum_{v_s=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} \frac{R_{v_s\,[\text{Mbps}]}^{srv_{max}}}{R_{cell\,[\text{Mbps}]}} \right) \tag{3.7}$$

where:

- $\boldsymbol{w^{usr}}$: Vector of users' weights, to obtain the long-term average data rate of users, which can be written as $\boldsymbol{w^{usr}} = \left[w_{1,1}^{usr}, \dots, w_{N_1^{usr}}^{usr}, \dots, w_{N^{srv},1}^{usr}, \dots, w_{N^{srv},N_{N^{srv}}^{usr}}^{usr} \right]$
- $N_{v_s}^{usr}$: Number of users performing service s, from VNO v.
- $R_{vs\,[Mbps]}^{srv_{max}}$: Maximum assignable data rate to the user of service s, from VNO v.
- $w_{v_s,i}^{usr}$: Assigned weight to user i, performing service s, from VNO v, ranging in [0,1].

Regarding the tuning weight, this serves the purpose of prioritising data rate assignment to each service. This parameter also isolates InP policies from VNOs decisions for capacity sharing. It is the combination of two independent positive integer numbers: γ is defined by InP and assigned to VNO v according to the type of its SLA agreements to VNOs' priorities in capacity sharing; δ is a serving weight, assigned to service s, performed by VNO v, to project the internal policy of each VNO in distributing capacity among the services provided by that VNO [Rouz19]. This way, the higher the serving weight number, the higher his priority is regarding capacity allocation and the lowest value provided by each VNO is always 1,

$$\lambda_{v_s} = \gamma_v \delta_s \tag{3.8}$$

There are two constraints associated with the problem of VRRM and the objective function has to be solved respecting these constraints. The first constrain considers that the average long-term data rate

assigned to each user has to fall within this acceptable data rate interval due to VNO policies:

$$R_{v_s\,[\text{Mbps}]}^{srv_{min}} \leq w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{v_s\,[\text{Mbps}]}^{srv_{max}} \tag{3.9}$$

where:

- $R_{vs\,[Mbps]}^{srv_{min}}$: Minimum assignable data rate to the user of service s, from VNO v.

The second constrain is a logical constraint, which indicates that the whole bandwidth allocated to all users cannot exceed the total aggregated cell capacity. Therefore, the entire VRRM bandwidth assigned to all users are subject to an upper bound defined by the InP:

$$\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \leq R_{cell\,[\text{Mbps}]} \tag{3.10}$$

There are two ways of capacity allocation between InP and VNO: static and dynamic. This thesis considers static capacity allocation, as illustrated in the left part of Figure 3.1. In this approach each VNO, represented as “slice” in Figure 3.1, pays the InP for a certain fixed capacity. This capacity is different for each VNO, depending on the type of contract made with the InP, but when the InP fails to deliver the contracted capacity, which is called a violation and then slack variables need to be consider for resource allocation. This scenario, while worth mentioning, is not part of this thesis study. On the right side of Figure 3.1, one represents the allocation of resources to each service made by the VNO.

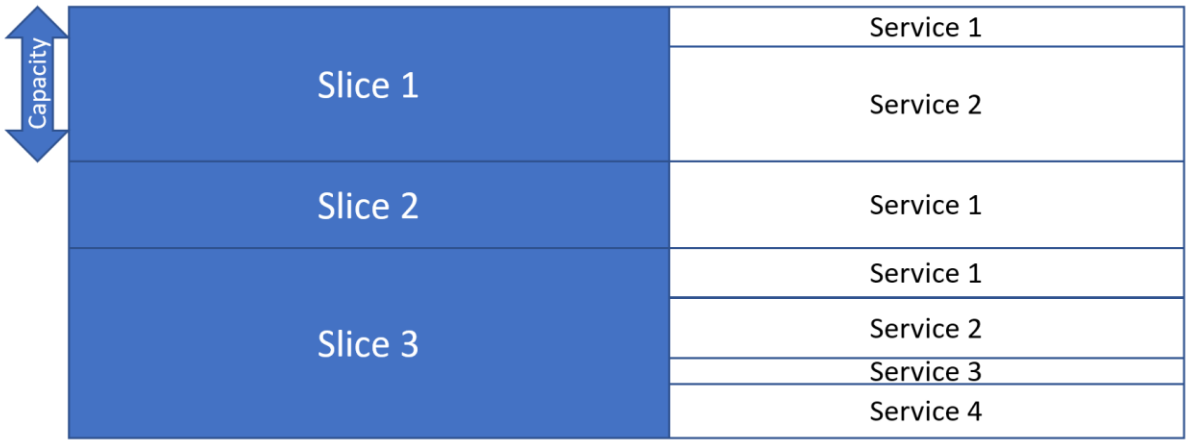


Figure 3.1 – Slice and service static resource allocation example.

3.2.3 Output Parameters

It is important to know if the overall user experience and network performance are working as intended. Several evaluation metrics are defined with the goal of measuring these performance requirements. The metrics used for network performance evaluation are as follows:

- Percentage of total assigned data rate – one of the most important metrics, showing the total network throughput in terms of data rate allocation, the values closer to 100% obviously leading to a better VRRM performance:

$$p_{VRRM[\%]}^{tot} = 100 \ \frac{\sum_{v_s=1}^{N^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}}}{R_{cell\,[\text{Mbps}]}} \tag{3.11}$$

- VRRM capacity share – the percentage of capacity allocated to each VNO, out of the total available VRRM one, being a key performance metric from VNOs and VRRM perspective:

$$R_{VRRM[\%]}^{VNO_v} = 100 \ \frac{\sum_{v_s=1}^{N_{VNO_v}^{srv}} \sum_{i=1}^{N_{v_s}^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}}}{R_{cell\,[\text{Mbps}]}} \tag{3.12}$$

- Total data rate of each service – it shows the total data rate assigned to each service slice of a VNO, being important from both VRRM and VNOs' viewpoints:

$$R_{v_s[\text{Mbps}]}^{srv_{tot}} = \sum_{i=1}^{N_k^{usr}} w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \tag{3.13}$$

- Percentage of served users – the percentage of served users performing a specific service out of the total number of users from that service, which is an essential metric for VNOs:

$$p_{v_s[\%]}^{usr_{net}} = 100 \ \frac{N_{v_s}^{usr}}{N^{usr_{tot}}} \tag{3.14}$$

The metrics that are important from a users' viewpoint are as follows:

- Data rate of each user – the data rate allocated to a user is an important QoS metric from both users' and VNOs' viewpoints, having a direct impact on the satisfaction level of the served users:

$$R_{v_s,i\,[\text{Mbps}]}^{usr} = w_{v_s,i}^{usr} R_{v_s\,[\text{Mbps}]}^{srv_{max}} \tag{3.15}$$

- Users' satisfaction, $S_{v_s,i}^{usr}$ – it is important from a VNO perspective because it reflects the user satisfaction. This metric is measured differently according to the service class and/or type. For voice, one uses AMR-WB codecs with the respective voice quality mean opinion score (MOS) values provided by NOS, presented in Table 3.4. For the remaining services, a similar way of classification is used. Depending on the service type, it is defined five levels of user satisfaction, where just like MOS one represents bad quality and five represents excellent quality. For Background there is no expression defined due to the service nature.

Table 3.4 – Voice codecs and respective MOS [Dini20].

Codec Name	Nominal MOS
AMR WB Mode 0 (6.6k)	3.39
AMR WB Mode 1 (8.85)	3.81
AMR WB Mode 2 (12.65)	4.04
AMR WB Mode 3 (14.25)	4.09
AMR WB Mode 4 (15.85)	4.11
AMR WB Mode 5 (18.25)	4.14
AMR WB Mode 6 (19.85)	4.18
AMR WB Mode 7 (23.05)	4.18
AMR WB Mode 8 (23.85)	4.18

3.3 Model Implementation

In this section, the explanation of the overall model flowchart is given as well as the several algorithms used. The model was implemented in MATLAB and CVX, which is used for the VRRM optimisation.

Figure 3.2 illustrates the flowchart of the model. The input parameters are loaded from an excel data sheet and the maximum achievable cell capacity is calculated with the cell input parameters. This is done using (3.1) and following all the necessary steps detailed in the Section 3.2.1. Then the programme proceeds to check if there is enough capacity to serve all users and if not, the delay process algorithm starts. Once it is verified that all users can be served with at least the minimum demanded capacity for their service, the VRRM optimisation algorithm is initiated. After running the VRRM optimisation, the programme computes the several output parameters that reflect the overall network performance and user satisfaction: percentage of total assigned data rate, VRRM capacity share, total data rate of each service, percentage of served users, data rate of each user and the users' satisfaction.

After calculating the maximum achievable cell capacity, the admission control and delay process algorithm take place, as illustrated in Figure 3.3. There are two possible scenarios at this point.

- First scenario, the previously calculated cell capacity is enough to serve all users, with at least the minimum service requirements. In this scenario the algorithm admission control and delay process algorithm are not executed.
- Second scenario, the cell capacity is not enough to serve all users. This last scenario is the reason why it is necessary to implement a delay process algorithm and ensure that higher priority services are served with the minimum demanded capacity.

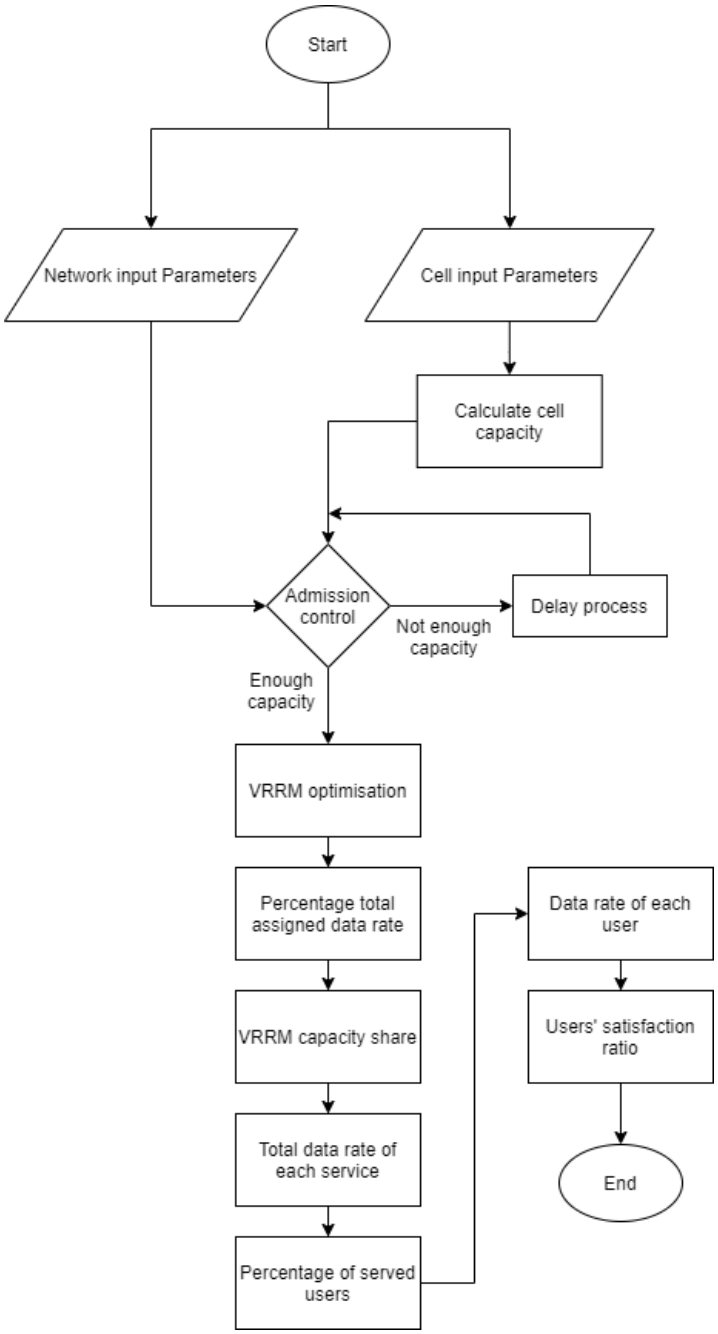


Figure 3.2 – Model flowchart.

This process aims to delay low priority users in order to achieve a state where the high priority users can be served. Naturally, the first type of users to be delayed are BE ones, as they do not have any minimum contracted level of data rate. Next, both VNO slices and services are ordered from lowest to highest priority. Starting from the lowest priority slice and service, 1 user is delayed. After every delayed user it is checked if it is possible to serve the remaining users with minimum demands. In case it is not possible the programme continues to delay users until all BG users are delayed. Next starting again from lowest to highest priority, the programme delays GB users until there is enough capacity. Note that if the scenario being studied is the third one, then the VRRM optimisation is not executed as it would not make sense to optimise something that is already using the minimum demanded requirements.

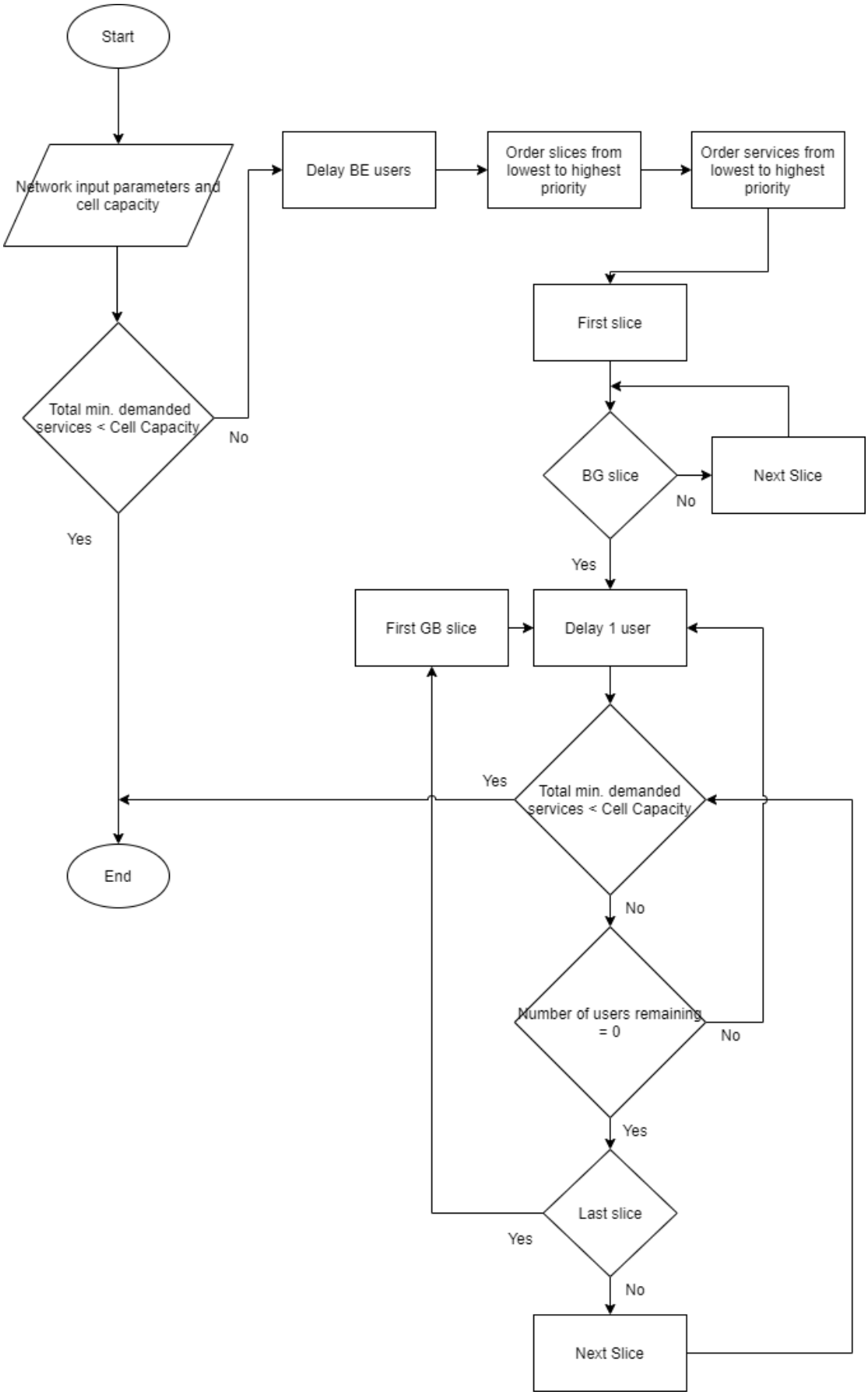


Figure 3.3 – Admission control and delay process algorithm.

Once it is verified that the scenario being studied is the third one, the VRRM optimisation step is started, which optimises (3.1). This is a convex optimisation problem where finding the vector of users' weights, $\boldsymbol{w^{usr}}$, that maximise the function is the goal. The solution is obtained using CVX in MATLAB. CVX is a tool implemented in MATLAB, that turns MATLAB into an optimisation modelling language. The specifications of the model are formulated using MATLAB syntax, which allows for more flexibility while implementing the problem. The constraints the problem was subject to were that $\boldsymbol{w^{usr}}$ should be between 0 and 1, 0 meaning that the user is not allocated any data rate and 1 meaning the user is being allocated the maximum amount of data rate demanded by the performing service. The second constraint states that the product between the service maximum allocated data rate and $\boldsymbol{w^{usr}}$ should not be lower than the minimum amount needed to perform the given service. This prevents the solver to choose extremely low values for $w_{v_s,i}^{usr}$ of low priority services just because it maximises the overall function. The last constraint states that the total amount of allocated data rate, meaning the product between $\boldsymbol{w^{usr}}$ and each user maximum amount of demanded data rate, should not be higher than the available cell capacity.

3.4 Model Assessment

The purpose of the model assessment is to validate the implemented model, described in the previous Section 3.2. This is done by defining a set of tests in which the output was previously calculated. Valid outputs will confirm the model validation.

In order to test the model and check if the VRRM optimisation is being done correctly, a set of values was given to multiple parameters to solve a more simplified version of the problem. This is necessary because it makes the problem solvable by hand. The chosen problem for this test considers the variables presented in Table 3.5.

Table 3.5 – Values of the test to assess the model.

Parameters	Service A	Service B
SLA	GB	BG
$\boldsymbol{R_{min}}$ **[Mbps]**	10	10
$\boldsymbol{R_{max}}$ **[Mbps]**	30	30
$\boldsymbol{\lambda}$	1	1
Number of users	1	1
$\boldsymbol{R_{cell}}$ **[Mbps]**	40	

Table 3.6 shows the set of chosen tests.

Table 3.6 - Module Assessment Tests.

Number	Description.
1	Validation of the input file read, by verifying if the type of variable is correct.
2	Validation of the input variables, by verifying if the parameters are correctly stored in memory.
3	Validation of the computation of the cell capacity: • Check the computation of T_s^{μ}. • Check if the overhead, the number of resource blocks, and the modulation order are correct. • Check if the computation R_{cell} is correct.
4	Validation of the admission control and delay process: • Check if the available capacity is not enough to serve all users with minimum demanded data rate. • Check if all BE users are delayed. • Check if the rest of delayed users were delayed based on their slice and service priorities. Check if the new minimum demanded capacity is equal or approximate to the previously calculated cell capacity.
5	Validation of the VRRM optimisation: • Check if the size and values of the created array with the minimum demanded capacity for all users are correct. • Check if the CVX status is solved. • Check if the computed values of the users' weights give the optimal solution to the problem. • Check if the imposed restrictions are being complied with.
6	Validation of the output files, by checking if they are correctly printed and plotting the output results.

In this problem there are two users, one performing service A and other performing service B. The cell capacity is known, the SLA, the minimum and maximum demanded data rate, and the priority of each service. Note that the maximum value for service B is set to $R_{cell\ [\mathrm{Mbps}]}$, because service B is BG meaning it only guarantees a minimum level of data rate. The equation that represents the problem is as follows:

$$f_{VRRM}(w_1, w_2) = 1 \times \log\frac{w_1 \times 30}{40} + 1 \times \log\frac{w_2 \times 30}{40} \tag{3.16}$$

The constraints of the problem are:

- 1 - $0 < w < 1$.
- 2 - $w_1 \times 30 + w_2 \times 30 = 40$.
- 3 - $w_1 \times 30 \geq 10$.
- 4 - $w_2 \times 30 \geq 10$.

The first constraint ensures that the allocated data rate to each service falls within the interval of 0 and the maximum demanded service data rate, $R_{\max\ [\mathrm{Mbps}]}$. The second constraint guarantees that the sum of allocated data rate of both services does not exceed the cell capacity limit. The third and fourth constraints limit the lower boundary of service A and B respectively.

The solution that maximises the objective function is $w_1 = 0.67$ and $w_2 = 0.67$. After executing the MATLAB programme, the status presented by CVX is solved and the solution obtained is the same, as expected. The several output parameters obtained as well as the $\boldsymbol{w^{usr}}$ values are presented in Table 3.7.

Table 3.7 - Impact of λ in the output parameters and $\boldsymbol{w^{usr}}$.

	$\boldsymbol{\lambda_1 = 1; \lambda_2 = 1}$	$\boldsymbol{\lambda_1 = 2; \lambda_2 = 1}$	$\boldsymbol{\lambda_1 = 4; \lambda_2 = 1}$
$\boldsymbol{(w_1, w_2)}$	$[0.67, 0.67]$	$[0.89, 0.44]$	$[1, 0.33]$
$\boldsymbol{p^{tot}_{VRRM[\%]}}$	100	100	100
$\boldsymbol{R^{VNO_v}_{VRRM[\%]}}$	$[50, 50]$	$[66.67, 33.3]$	$[75, 25]$
$\boldsymbol{R^{srv_{tot}}_{v_s[\mathrm{Mbps}]}}$	$[20, 20]$	$[26.67, 13.3]$	$[30, 10]$
$\boldsymbol{p^{usr}_{v_s[\%]}}$	$[50, 50]$	$[50, 50]$	$[50, 50]$
$\boldsymbol{R^{usr}_{v_s,i\ [\mathrm{Mbps}]}}$	$[20, 20]$	$[26.67, 13.3]$	$[30, 10]$
$\boldsymbol{S^{usr}_{v_s,i}}$	$[4, 4]$	$[5, 2]$	$[5, 1]$

By analysing the results, one can check if they are as expected. $p^{tot}_{VRRM[\%]}$ shows that 100% of the maximum capacity limit is being used. $R^{srv_{tot}}_{v_s[Mbps]}$ represents the capacity allocated to each service, these values are easily validated, multiplying w_1 and w_2 by their respect $R_{\max\ [\mathrm{Mbps}]}$. $R^{VNO_v}_{VRRM[\%]}$ is the percentage of allocated data rate of each VNO. $p^{usr}_{v_s[\%]}$ is showing that 100% of the users were served, 50% in each service. $R^{usr}_{v_s,i\ [Mbps]}$ is the capacity allocated to each user. $S^{usr}_{v_s,i}$ and being the users' satisfaction which in this example the boundaries of the five levels of user satisfaction are divided linearly, meaning if $R^{usr}_{v_s,i\ [Mbps]}$ is between 10 Mbps and 15 Mbps the respective $S^{usr}_{v_s,i}$ is 2 (poor), between 15 Mbps and 20 Mbps it is 3 (fair), between 20 Mbps and 25 Mbps it is 4 (good), and between 25 Mbps and 30 Mbps it is 5 (excellent); they comply with the expected results. One can observe that, with two services with the exact same requirements, they were allocated the same capacity and in a way that maximised the threshold defined as maximum cell capacity.

Now, using the same example but changing λ_1 to 2, it is noticeable that there is a slight change with (w_1, w_2) and as a result all the other output parameters. Because service A now has more priority than service B it is only natural it gets more data rate allocated. This increase is done until λ_1 achieves a

value of 4. For this example, a ratio of 4:1 in the service priorities, means that the most priority service, service A, now is receiving the maximum demanded data rate it requires. So, increasing λ_1 more does not have an impact in the final solution.

A good way to test the model regarding the delay process is by varying the number of users. For this, three cases were studied with a differing number of users. Using the case presented in Table 3.6, the number of users allocated to service A is 1 for the first test, 4 for the second, and 5 for the third. The results are presented in Table 3.8.

Table 3.8 - Impact of the number of users in the output parameters and $\boldsymbol{w^{usr}}$.

	A	B	A	B
Number of users	1	1	5	1
$(\boldsymbol{w_1}, \boldsymbol{w_2})$	$[0.67, 0.67]$		$[0.33, 0.33, 0.33, 0.33]$	
$\boldsymbol{p^{tot}_{VRRM}}[\%]$	100		100	
$\boldsymbol{R^{VNO_v}_{VRRM}}[\%]$	$[50, 50]$		$[100, 0]$	
$\boldsymbol{R^{srv_{tot}}_{v_s}}[\mathbf{Mbps}]$	$[20, 20]$		$[50, 0]$	
$\boldsymbol{p^{usr}_{v_s}}[\%]$	$[50, 50]$		$[100, 0]$	
$\boldsymbol{R^{usr}_{v_s,i}}\ [\mathbf{Mbps}]$	$[20, 20]$		$[10, 10, 10, 10, 10]$	
$\boldsymbol{S^{usr}_{v_s,i}}$	$[4, 4]$		$[2, 2, 2, 2, 2]$	

The values chosen for Table 3.8 perfectly represent the two scenarios addressed in Section 3.3. In the first scenario, the cell has capacity to serve all the users with at least the minimum demanded capacity, so it runs the CVX optimisation algorithm and allocates capacity accordingly. In the second scenario, there are more users than it is possible to serve. So, the 1 user from service B is delayed, because it has a BG SLA.

Chapter 4

Result Analysis

This chapter provides the description of the reference scenario and all the variations made to it. Then it provides the results obtained by the developed model and a study of all these variations.

4.1 Reference Scenario

The scenario, chosen by NOS that is studied in this chapter, represents a case study with multiple VNOs sharing the resources of a cell. Each VNO is characterised with an SLA and an assigned priority level, γ_s, provided by the InP. Within the VNO there are multiple services that the VNO intends to provide, each characterised by the class of service it belongs to, the minimum and maximum required data rate to provide the service, and the priority assigned to the service, δ_s, by the VNO. The last parameter, Mix, shows how the users are distributed among the several services of all VNOs. Table 4.1 shows the class and data rate necessary to provide a given service.

Table 4.1 – Services classes and demanded data rates.

Service	Service type	Class	Data rate [Mbps]
Email	4G	Background	[0, R_{cell}]
IoT		Background	[0, R_{cell}]
Voice		Conversational	[0.0066,0.024]
Music		Streaming	[0.015,0.32]
Web browsing (Web)		Interactive	[0.5, R_{cell}]
File Sharing (FS)		Interactive	[1, R_{cell}]
Social networking (SN)		Interactive	[2, R_{cell}]
Video		Streaming	[2,13]
Real-time gaming (RTG)	eMBB	Streaming	[20, R_{cell}]
VR/AR	eMBB	Streaming	[50, R_{cell}]
Factory automation (FA)	URLLC	Interactive	[5, 50]
Road safety ITS (RSI)	URLLC	Interactive	[10, 50]
Remote surgery (RS)	URLLC	Interactive	[10, 100]
Smart meters (SM)	mMTC	Background	[0.0066,0.020]

The cell input parameters are listed in Table 4.2. These inputs suffer variations that are described later in the subsection.

The chosen scenario intents to study a real case in the 5th generation of mobile communications. This scenario covers eMBB services and it is shown in Table 4.3 marked as green. The other set of colours represent variations to this scenario which is covered next. These parameters are the cell network input parameters. An acronym was created to each VNO to facilitate the writing and reading when referring to a particular one.

Table 4.2 – Cell input parameters.

MIMO layers, $v_{Layers}^{(j)}$	4
Numerology, μ	1
Bandwidth [MHz]	100
Multi-user MIMO, M_{UMIMO}	2

Table 4.3 – Reference, URLLC and mMTC scenarios.

VNO	Service	γ_s	δ_s	SLA	Mix RS [%]	Mix URLLC [%]	Mix mMTC [%]		Users
							Relative	Global	
GB Real-time (RT)	Voice	30	10	GB	20	18	20	0.83	100
	Video		8		35	30	35	1.45	
	Music		9		5	5	5	0.21	
BG Interactive (IA)	Social networking	20	5	BG	10	8	10	0.41	
	Web		6		10	8	10	0.41	
	File Sharing		4		3	3	3	0.12	
BG RT	VR/AR	40	7	BG	2	3	2	0.083	
	Real-time gaming		6		2	5	2	0.083	
BE Background (BkG)	Email	10	3	BE	3	3	3	0.12	
	IoT		2		10	8	10	0.41	
GB URLLC (L)	Factory automation	40	10	GB	-	2	-	-	
	Road safety ITS		11		-	5	-	-	
	Remote surgery		12		-	2	-	-	
BE uMMTC (MM)	Smart meters	1	1	BE	-	-	100	95.85	2311

For the mMTC scenario, there are some differences regarding the number of users. The number of smart meters is much greater than the number of other services users', so the concept of mix needs to be approached in a different manner. To solve this problem, the mMTC scenario is the same as the reference scenario, including number of users and mix plus VNO GB MM with 2311 users. This number is explained in the respective section. The relative and global value of this scenario is presented, with

the global value being the actual percentages of how the 2 411 users are distributed among the services and the relative value being an easier way to understand how the 100 users are distributed among the services from first five VNOs and the remaining 2 311 from VNO GB MM.

With the intent of better analysing and representing other 5G cases, some variations are made to the reference scenario. Besides the cell input parameters, each modification has a dedicated section intended for its study. The cell input parameters changes are studied in the subsection reserved for the study of the reference scenario. These inputs are the BW and the Multi-user MIMO. BW assumes three different values that correspond to the allocated BW for the frequencies used in 5G. Multi-user MIMO also assumes three values but only when the BW is 100 MHz BW as it is the only case where this factor is relevant. The variation on the Number of MIMO layers and numerology intend to increase the scientific coherence of this study (a comparison between the results is not relevant). MIMO layers changes according to the BW. For 10 MHz and 20 MHz, MIMO layers takes a value of 2, whilst for 100 MHz, it takes a value of 4. The numerology is maintained at 1, except in the URRLC scenario where it changes to 2. Table 4.4 shows the summary of the variations performed to the input parameters.

Table 4.4 – Reference scenario variations.

MIMO layers, $v_{Layers}^{(j)}$	{2, 4}
Numerology, μ	{1, 2}
Bandwidth [MHz]	{10, 20, 100}
Multi-user MIMO, M_{UMIMO}	{1, 2, 3}
Number of VNOs	[4, 8]
Priority, γ_s	[10, 100]
Service mix	[0, 100]
Number of Users, $N_{v_s}^{usr}$	[50, 1100]

The reference scenario intends to study eMBB type situations, so it contains the third VNO with heavy data rate demanding applications. This BG VNO has the highest assigned priority by the InP, γ_s, to ensure its demands are met. Concerning the variations to the network input parameters, the first is to the VNOs. VNO GB L is added with URLLC services with the goal of studying how the model behaves in a URLLC scenario. The priorities of each services change, δ_s, but the priority assigned by the InP and the SLA remains the same. The mix is changed as well, however, since it affects all services and not only the new ones, a description is provided in the respective subsection. The second variation to the VNOs aims to study the impact of mMTC services in the model. To achieve this goal, a 6th VNO is added to the reference scenario with smart meters as the only service. These variations are shown in Table 4.3 and are the Mix URLLC and Mix mMTC.

The next set of variations is to the mix and number of users. First it is necessary to, maintaining the mix, increase the number of users from 50 to 100, 150, 200, and so on until the programme reaches its limit. It is beneficial to understand not only how the model reacts to an increasing number of users allocated to a specific cell but also determine the maximum number of users one cell can serve simultaneously, for a given mix. From this point on, all scenarios are subjected to this variation in the number of users. As for the mix variations, two new scenarios were conceived and are presented in Table 4.5. These new scenarios are direct variations of the reference and URLLC scenarios: the first representing a football game scenario and the second representing a hospital scenario. This way one can study how the programme responds to people's needs in several different locations.

Table 4.5 – Service mix variation scenarios.

VNO	Service	γ_s	δ_s	SLA	Mix Football [%]	Mix Hospital [%]	Users
GB RT	Voice	30	10	GB	5	30	100
	Video		8		1	5	
	Music		9		1	5	
BG IA	Social networking	20	5	BG	25	3	
	Web		6		21	10	
	File Sharing		4		1	8	
BG RT	VR/AR	40	7	BG	30	8	
	Real-time gaming		6		5	3	
BE BkG	Email	10	3	BE	1	10	
	IoT		2		10	8	
GB L	Factory automation	40	10	GB	-	0	
	Road safety ITS		11		-	0	
	Remote surgery		12		-	10	

The final variation is related to the priorities assigned from the InP to the VNO, γ_s. This subsection has two goals, the first being the study of two similar VNOs where one has a much higher priority than the other, creating a "premium" and "low-cost" slices type scenario, this scenario being presented in Table 4.6, and the second goal is to increase the priority assigned to one VNO, specifically VNO GB L, to study the behaviour of the model and determine the ideal value for this priority.

Table 4.6 – Premium and low-cost slices scenario.

VNO	Service	γ_s	δ_s	SLA	Mix [%]	Users
GB RT	Voice	30	10	GB	10	100
	Video		8		17	
	Music		9		3	
GB RT 2	Voice	15	10	GB	10	
	Video		8		17	
	Music		9		3	
BG IA	Social networking	20	5	BG	5	
	Web		6		5	
	File Sharing		4		2	
BG IA 2	Social networking	10	5	BG	5	
	Web		6		5	
	File Sharing		4		2	
BG RT	VR/AR	40	7	BG	1	
	Real-time gaming		6		1	
BG RT 2	VR/AR	20	7	BG	1	
	Real-time gaming		6		1	
BE BkG	Email	10	3	BE	1	
	IoT		2		5	
BE BkG 2	Email	5	3	BE	1	
	IoT		2		5	

4.2 Reference Scenario Study

This subsection intends to study the reference scenario and to make some variations to the cell input parameters and evaluate their effects on the results. More specifically, a study on the impact that different values of BW have in the reference scenario and the usage of Multi-user MIMO when BW = 100 MHz. The cell input parameters changes were already addressed but there are other necessary parameters for the calculation of the maximum achievable cell data rate, shown in Table 4.7.

Table 4.7 – Maximum achievable cell data rate parameters.

J	1
$f^{(j)}$	1
$O_h^{(j)}$	0.14
$Q_m^{(j)}$	{2, 4, 6, 8}
R_{max}	{449/1024, 616/1024, 873/1024, 948/1024}
$N_{PRB}^{BW(j),\mu}$	{24, 51, 273}

J is set to 1 because carrier aggregation is not being used. The scaling factor, $f^{(j)}$, is set to 1 for maximum data rate and the overhead, $O_h^{(j)}$, is 0.14 which is the defined value for DL FR1. For the sake of keeping the values as real as possible, the calculation of the maximum achievable cell data rate is the result of averaging the values of the maximum achievable cell data rate of all four considered modulations (QPSK,16-,64-,256QAM). This means that $Q_m^{(j)}$ assumes the values of 2, 4, 6 and 8, and R_{max} will assume the respective values of 449/1024, 616/1024, 873/1024 and 948/1024. The number of RB depends on the used BW.

The cell input parameter μ is 1 because the services in study are eMBB ones, therefore latency is not the main priority, so a subcarrier spacing of 30 kHz is the minimum value for 100 MHz and simultaneously provides more RB, thus resulting in more data rate. With this value one can calculate T_s^{μ} and obtain $3.57 \times 10^{-5}\ s$.

As previously mentioned in Section 3.2.1, NOS is currently using three different BWs for their frequencies: 10 MHz, 20 MHz and 100 MHz. For 5G networks, only the 700 MHz and the 3.5 GHz frequencies are used, with a BW of 10 MHz and 100 MHz, respectively. Nevertheless, since there is a possibility that, in the future, frequencies currently allocated to LTE are reused for NR, the 20 MHz BW is also studied.

Regarding BW = 100 MHz, one cannot forget that this BW is allocated to a frequency that is working in TDD, meaning that not all slots are dedicated to DL. In this work, one considers that 87.5% of the slots are dedicated to DL and 12.5% to UL. So, when making the final calculation for this BW it is necessary to multiply the final value by 0.875. Table 4.8 presents the calculated value of the cell data rate.

Table 4.8 – Calculation results of R_{cell}.

	BW				
	10 MHz	20 MHz	100 MHz		
			$M_{UMIMO} = 1$	$M_{UMIMO} = 2$	$M_{UMIMO} = 3$
R_{cell} [Mbps]	45.58	96.86	888.67	1777.30	2666.00

For the reference scenario, the minimum demanded data rate by all users is 238.21 Mbps, which means that the cell can serve all users. The results of the VRRM optimisation and output parameter calculations are listed in Table 4.9.

Table 4.9 – Output Parameters and weights for BW = 100 MHz with $M_{UMIMO} = 2$.

VNO	Service	w^{usr}	p^{tot}_{VRRM}[%]	$R^{VNO_v}_{VRRM}$[%]	$R^{srv_{tot}}_{v_s}$[Mbps]	$p^{usr}_{v_s}$[%]	$R^{usr}_{v_s,i}$ [Mbps]	$S^{usr}_{v_s,i}$
1	Voice	1	100	25.71	47.00	100	0.024	4.18
	Video	1			455.00	100	13	5
	Music	1			1.60	100	0.32	5
2	Social networking	0.020		48.08	350.20	100	35.02	5
	Web	0.024			420.24	100	42.024	5
	File Sharing	0.160			84.05	100	28.02	5
3	Virtual Reality	0.055		20.49	196.11	100	98.06	3
	Real-time gaming	0.047			168.10	100	84.23	4
4	Email	0.006		5.71	31.51	100	10.5	-
	IoT	0.004			7.00	100	7	-

Because there is enough data rate to serve all users, all the VNOs were allocated the maximum data rate demanded by each service to its users, therefore having the best user satisfaction possible, except for VNO BG RT with only scores fair and good for VR and RTG. VNOs GB RT and BG IA are being allocated the most data rate because of the combination of two factors. First, the higher number of users and second because the fact that VNO BG IA is a BG VNO, meaning it has no maximum limit for allocated data rate, because the GB VNO is already being served with maximum demanded data rate.

In this scenario, using $M_{UMIMO} = 3$ would be relevant because it enables VNO BG RT to provide its services with better user satisfaction, with all services having the best score in user satisfaction except VR with a score of good at 164.06 Mbps. Thus, using $M_{UMIMO} = 3$ would be more useful when the number of users and the data rate demands increase. On the other hand, using $M_{UMIMO} = 1$ is not enough to achieve good user satisfaction among all services. VNO BG RT can only achieve a fair and good user satisfaction for virtual reality and real time gaming, respectively. Figure 4.1 depicts the results obtained by varying M_{UMIMO}, with BW fixed to 100 MHz.

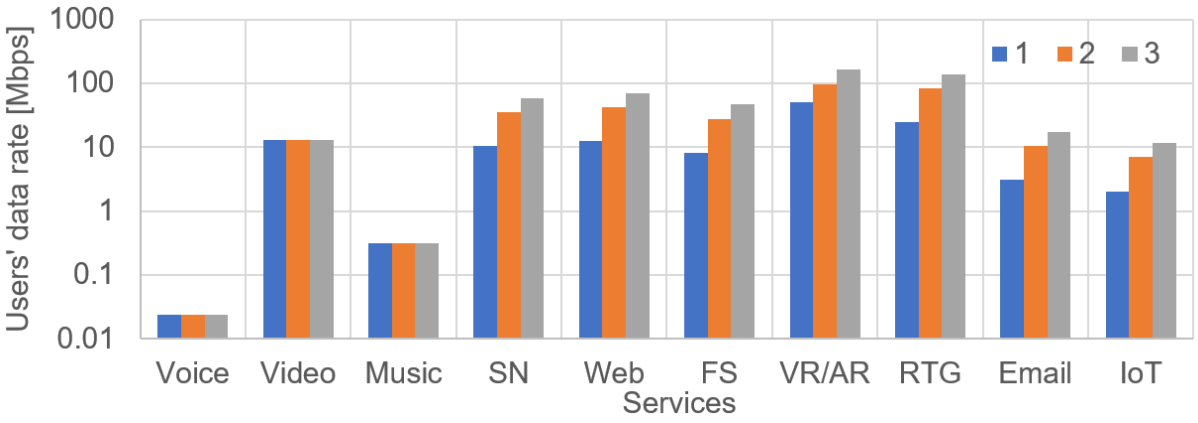


Figure 4.1 - Comparison between users' data rates for different M_{UMIMO} values.

When BW = 10 MHz, it is not possible to serve all users due to the lack of resources, Figure 4.2. As such, the delay process algorithm is executed delaying all users from VNO BE BKG, which has a BE SLA, and then starts delaying users one by one based on priorities and SLAs. Because the algorithm was executed, the VRRM optimisation did not happen, therefore $\boldsymbol{w}^{usr}$ has no assigned values. VNO BG IA is the first to be delayed, followed by VNO BG RT, seeing as both have BG SLA. The last VNO, VNO GB RT, with a GB SLA has all its users served with minimum demanded data rate for each service. One can infer that the lack of resources is due to VNO BG RT, which includes services with high demands of data rate. Due to the high priority of γ_s set between the VNO and the InP, VNO BG IA is affected with all its users not being served, which is why only 59.87% of the cell resources are being used. Because all services are being served with minimum demands, the user satisfaction is bad except for Voice being classified with a fair value according to MOS. The 700 MHz frequency is going to be used mainly in rural areas where this problem will not have impact.

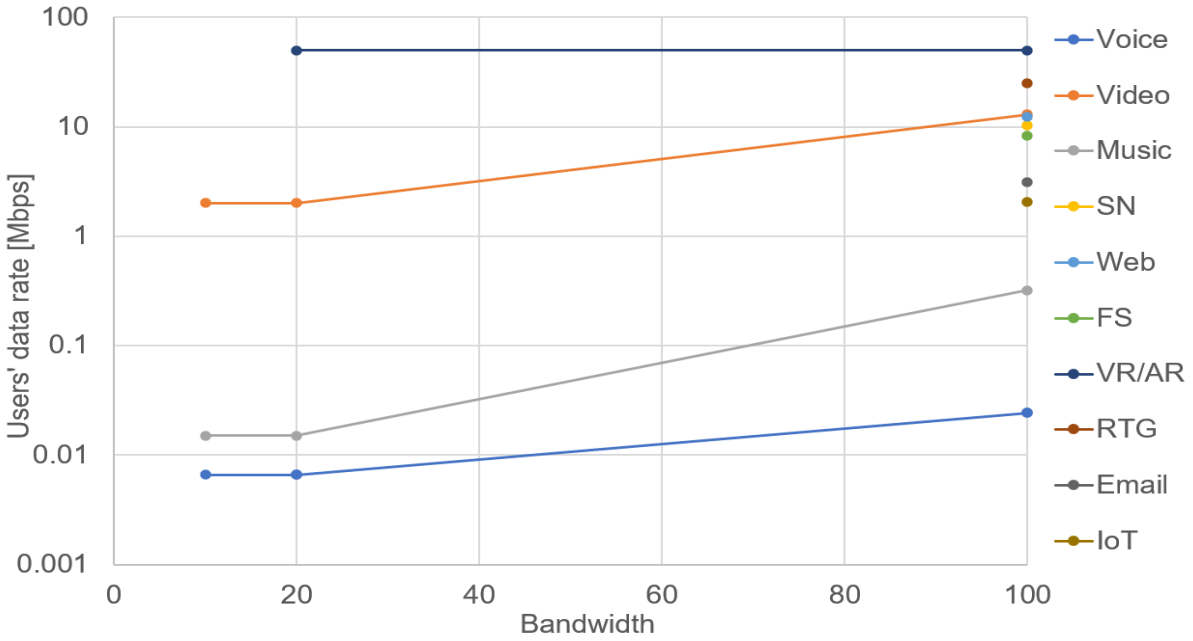


Figure 4.2 - Comparison between users' data rates for different BWs.

For BW = 20 MHz, the scenario is similar to the previous one although, in this case, it is possible to serve VNO BG RT VR service at minimum demanded data rate. Figure 4.2 shows the comparison between the users' data rate for these three cases. For comparison purposes, note that when the BW is 100 MHz $M_{UMIMO} = 1$. As expected, with the increase of BW, it is possible to allocate more data rate to users. Several services do not have allocated data rate, when considering 10 MHz and 20 MHz BWs, because of the third VNO being too demanding. VNO GB RT is the only VNO that can guarantee its services despise the BW in use due to the low data rate requirements and SLA type.

4.3 Influence of Incrementing the Number of Users

In this subsection, the number of users is linearly incremented until the point of saturation and its effects on the reference scenario are studied. Figure 4.3 illustrates the impact of increasing the number of users in the reference scenario. The model behaves as expected and the decision making for the allocated data rate for each service follows some principles. Firstly, the data rate range, which is particular to each service, defined by the VNO. This data range value, together with the priority, λ_{v_s}, work together to decide how much data rate is allocated to a specific user. Secondly, the SLA type that each VNO has, which is a key factor in deciding what services get delayed first if there are not enough resources to serve all users. This figure uses (3.15) for the calculation of each point.

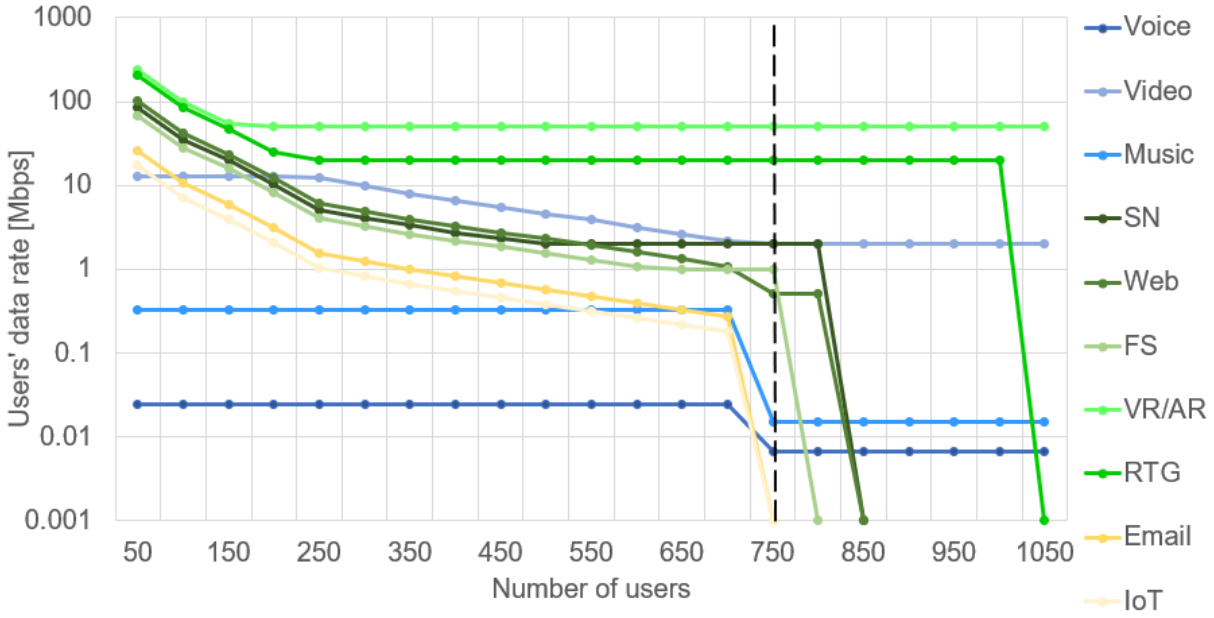


Figure 4.3 – Impact of the increment of number of users in the reference scenario.

In this scenario, there are three services in the VNO with the highest-level SLA, GB, five with the intermediate SLA, BG, and another two with the lowest level SLA, BE. Note that for the last two, BG and

BE, the users' data rate is set to the maximum value the cell can provide while keeping all GB VNO services with maximum allocated data rate. With the increased number of users, the allocated data rate of these services starts to decrease while the GB ones remain stable. VR/AR and RTG are the first services to reach the minimum threshold of allocated data rate, due to their very high demanded values, with VR/AR starting at 200 users and RTG at 250. The other BG services will not reach their respective minimums until 500, when SN does. The only GB service that is also affected by these numbers of users is video, due to its data rate requirement. Marked by the dotted line, in Figure 4.3, is the beginning of the delay process algorithm. At this point, all services are being served with the minimum defined data rate and the cell does not have enough capacity to serve more users. Both email and IoT are delayed first since they do not meet the minimum demands associated with their service. Next, by order of the lowest priority service from the lowest priority VNO, users are delayed one by one until the cell has enough capacity to serve the remaining users.

Table 4.10 shows the summary of the thresholds for each service. Note that, for BG VNO services, the maximum is defined as the maximum number of users that are being served with maximum satisfaction, because these types of service do not have a maximum associated with them. For example, VR/AR has a threshold of maximum satisfaction above 200 Mbps. The maximum number of users that is possible to serve with this restriction is 50 users, where each user is being served with 227.98 Mbps. For BE VNO services the maximum is not defined as this type of background services has no associated user satisfaction.

Table 4.10 – Thresholds of users to achieve maximum and minimum data rate as well as delay.

	Maximum (≤)	Minimum (≥)	Delay (≥)
Voice	700	750	1100
Video	200	750	1100
Music	700	750	1100
SN	200	500	850
Web	300	750	850
FS	200	650	800
VR/AR	50	200	1100
RTG	50	250	1100
Email	-	-	750
IoT	-	-	750

Figure 4.3 alone is not enough to study the scenario because it only shows the data rate of the served users, meaning it does not encompass the full spectrum of users that are getting delayed and not having their needs fulfilled, so Figure 4.4 and Figure 4.5 are presented as well.

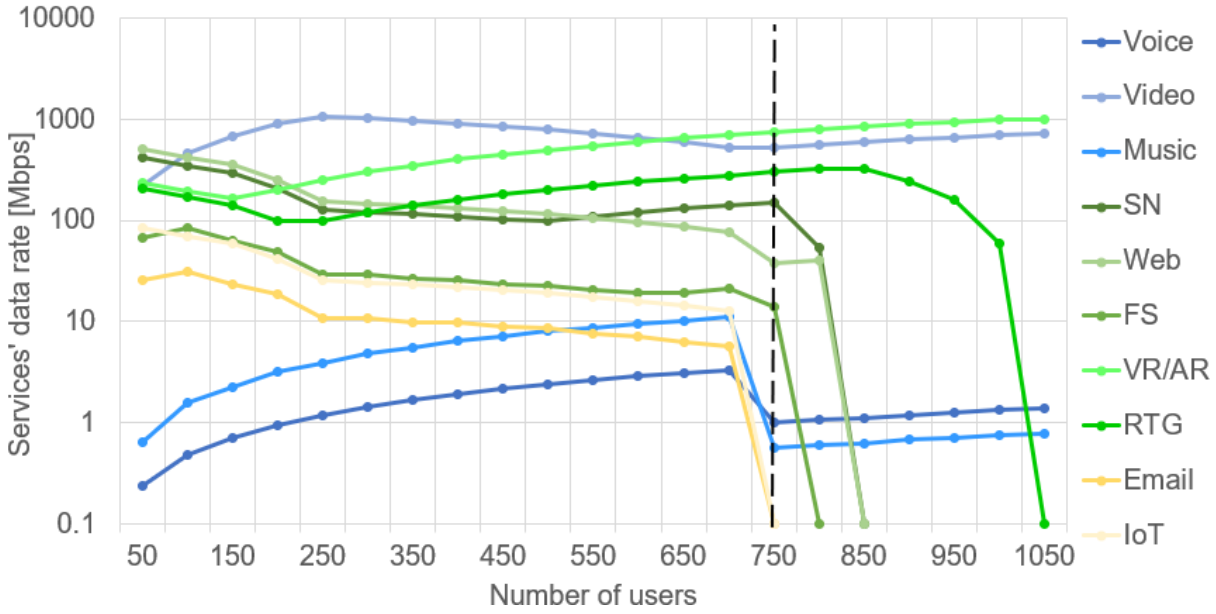


Figure 4.4 – Services' data rate evolution with the increment of number of users.

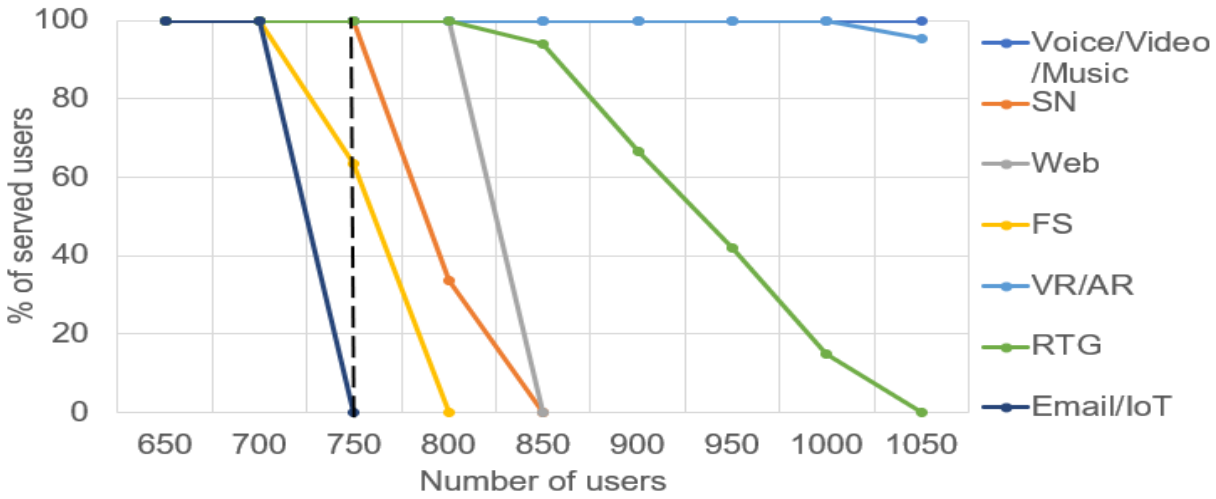


Figure 4.5 – Percentage of served users with the number of users increasing.

Figure 4.4 illustrates the data rate allocated to each service and Figure 4.5 the percentage of served users. Until the point of delay the curvatures behave as expected. Voice, music, VR/AR and RTG have an overall growth to the data rate each service is allocated (note that for VR/AR and RTG this is only true when the number of users is greater than 150 and 250 respectively). The common factor among all these services is the fact that they are either being served with maximum data rate (voice and music) or minimum data rate (VR/AR RTG). The other services even serving more users, also see a decrease in the data rate each user is being allocated. From 750 users onward, all users are being served with minimum demands. The result is all services that do not need to delay users see an increase to the total data rate allocated to them. This includes voice, music, video and VR/AR. The other services percentage of served users start tending to 0 as the users from its services start to get delayed. RTG is notably

slower than the others due to its high priority value. In this situation one can also see the isolation factor, in the sense that VNO GB RT is not affected even though other VNOs are seeing decreases in the amount of served users. This figure uses (3.13) for the calculation of each point.

Figure 4.5 details the exact percentage of the served users from each service. The first "wave" of delays is for email, IoT and FS, where the first two are completely delayed and the last needs to delay 40% of its users. The second "wave" of delays, affects all FS users and around 75% of web users. This process repeats itself like the figure is showing. Voice, video and music are the only services that are never delayed due to the combination of the three principles mentioned at the beginning of the subsection. Both are within a GB VNO, have the highest priorities among the services in that VNO, and have low data rate demands. This figure uses (3.14) for the calculation of each point.

For a better understanding on how the cell resources are being distributed among the VNOs, Figure 4.6 demonstrates the evolution of the VNOs' capacity share with the number of users increasing. This figure uses (3.12) for the calculation of each point. VNO GB RT, BG RT, BG IA and BE BkG start with a share of 12.48%, 56.43%, 24.87% and 6.22%, respectively. These values result from the combination of multiple factors, such as the SLA each VNO has with the InP, the number of served users and the priorities associated with them and their services. From this point, all BG and BE VNOs (2, 3, and 4) start decreasing their share, while the GB VNO increases.

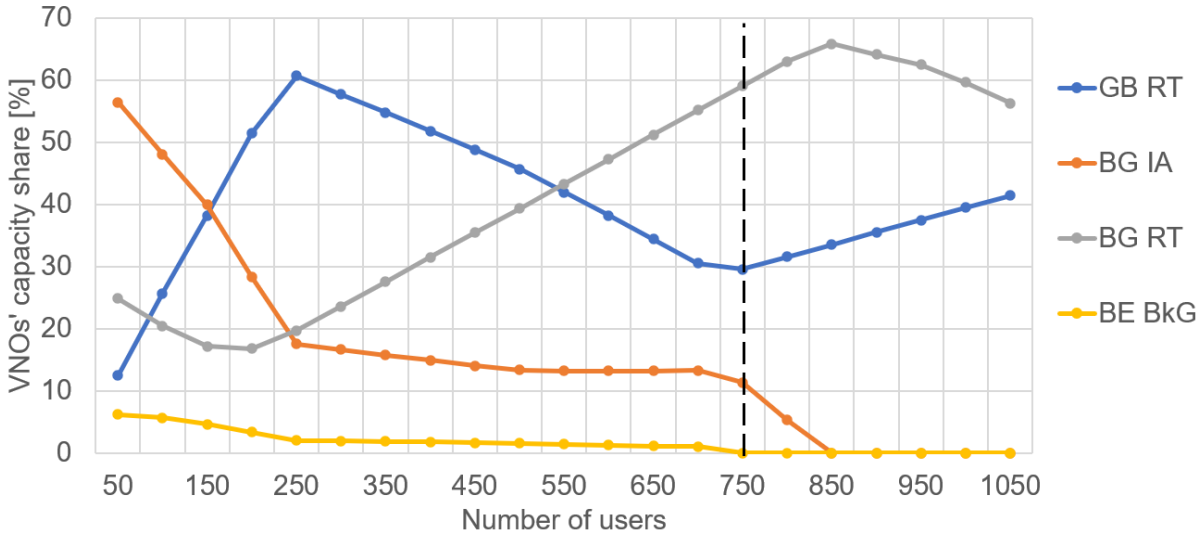


Figure 4.6 – VNOs' capacity share with the number of users increasing.

With the increase in the number of users, the demanded data rate also increases. With more users to serve and with the same amount of allocated data rate per user, the GB VNO sees an increase in its capacity share until the point it starts reducing the allocated data rate for its users. More specifically, when the number of users surpasses 250, video users no longer are allocated 13 Mbps but rather a lower value that continues decreasing with the increment of the number of users. The capacity share of VNO GB RT continues to decrease until the number of users is 750, point where CVX optimisation stops

and the delay algorithm starts delaying users. BE VNO users are all delayed, and BG VNO users start being delayed one by one. The natural consequence is to see an increase of VNO GB RT that maintains all its users with minimum demanded data rate. VNO BG RT appears to behave differently, which is true to some extent, but most of its behaviour is similar to VNO GB RT with a shift in the curve. The start is different because VNO BG RT has a BG SLA that does not limit the amount of data rate allocated to its service. This means that with the increase of the number of users, the allocated data rate per user starts decreasing until it reaches the minimum demanded by the service. For the specific case of VNO BG RT, the minimum occurs when the number of users is 200 and all VR/AR users are being allocated 50 Mbps. From here, it assumes the same behaviour that VNO GB RT did for 750 users, with all users at a fixed minimum capacity and with the number of users increasing its natural that the capacity share also increases. The second point worth mentioning is at 850 users when all VNO BG IA users have been delayed and the first RTG users start to be delayed, resulting in the decrease of VNO BG RT capacity share.

The last output of the model to be analysed is the users' satisfaction. Figure 4.7 summarises the evolution of the satisfaction of each served user. This satisfaction, defined in the Subsection 3.2.3, directly results from the data rate each user has been allocated. From this figure one can conclude that the ideal number of users that guaranties all users have at least a rating of 3 (fair) user satisfaction is at 350 users. Although VR/AR and RTG are the first to drop, they stay constant until users start to get delayed because the minimum data rated for these services was defined such that the served users would always get at least a fair satisfaction.

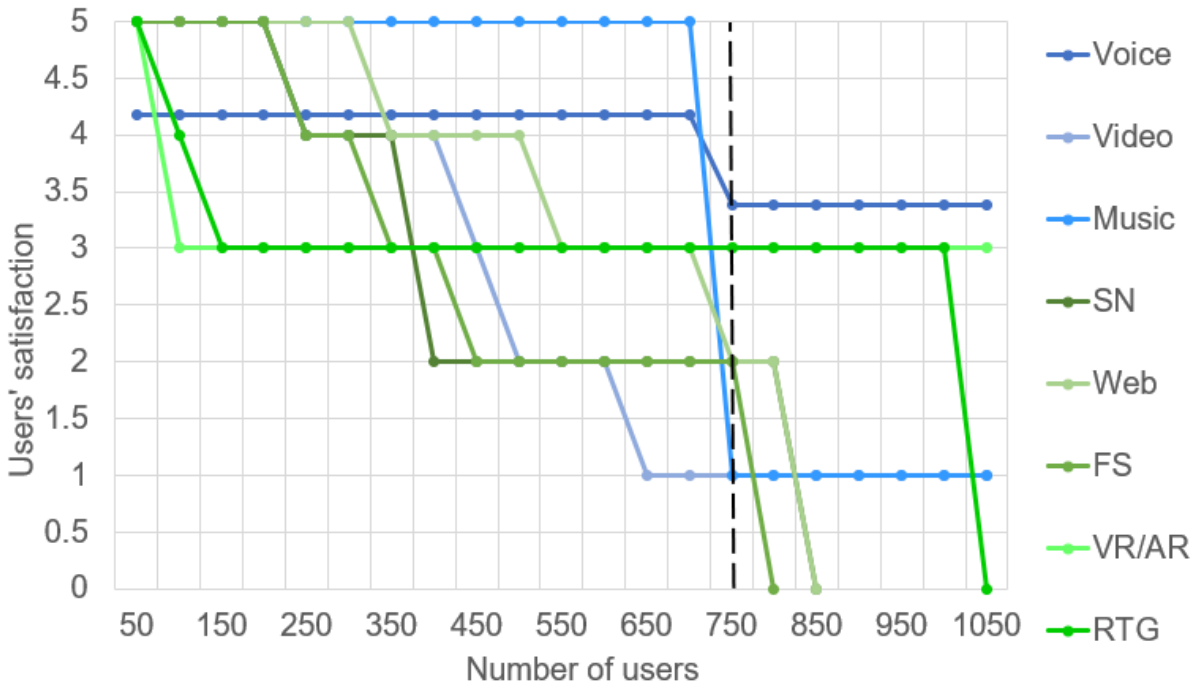


Figure 4.7 – Users' satisfaction with the increment of the number of users.

4.4 Impact of the URLLC VNO

This subsection is meant to study the impact URLLC applications have in the network. Because these types of services have extremely low latency requirements μ is set to 2, the maximum value for frequencies within frequency range 1. A subcarrier spacing of 60 kHz and a symbol length of only 0.25 ms is perfect to serve these types of services. With all the other parameters remaining the same, one needs only to calculate T_s^μ and then $R_{cell\,[\mathrm{Mbps}]}$. With $\mu = 2$, $T_s^\mu = 1.79 \times 10^{-5}\, s$, and $R_{cell} = 1.758$ Gbps. Note that the data rate values are slightly lower than the previous calculated ones, because when $\mu = 2$ the number of RB is also lower (135). Table 4.3 shows the scenario represented by Mix URLLC, that encompasses not only the previous services but also VNO GB L with URLLC type services. In this scenario the minimum demanded data rate for all services is 413.19 Mbps, whose increase from the reference scenario is due to the new added VNO, and the new mix that has more users in VNO BG RT.

In this scenario, the worst performing VNO is VNO BG RT, whose services are scoring only a fair user satisfaction. VNO GB L manages to have excellent user satisfaction for RSI and FA and a good for RS due to its SLA type and high priority. This is extremely important because it ensures the possibility to serve URLLC services with the desired latency and good values of data rate while serving more common services like voice and video that will most likely be used in the same scenarios URLLC services will. Not only that but it is also possible to serve more data rate demanding services like real-time gaming, for the considered mix and number of users. The VNO that got more data rate allocated was VNO GB L which was allocated 30.48% of the cell data rate. VNOs GB RT, BG IA and BG RT all share about the same data rate allocation percentage, around 22%, and VNO BR BkG with background services was only allocated 2.75%. Were it the case where one would use higher levels of M_{UMIMO}, the result would be an increase of the percentage allocated to the BG and BE VNOs, as well as decrease in GB VNOs. This feature becomes more relevant when the number of users, or demand of services increases.

Figure 4.8 compares the results with the previous obtained for the reference scenario. The change in the mix and the new added VNO affected the reference scenario mostly for VNO BG RT. While VNO GB RT maintains all levels of data rate and VNO BG IA despise the drop, the user satisfaction remains unchanged, VNO BG RT has a 43.89 Mbps decrease in VR and a 37.8 Mbps decrease in RTG.

Like in the previous subsection, Figure 4.9, portrays the evolution of the users' data rate with the increasing number of users. Comparing it to the reference scenario, one can observe a similar behaviour, but more crowded due to the added VNO that brings new services. As a result, the delay process starts at 450 users, as opposed to 750 users. This figure, however, does not represent the percentage of allocated users to each service. For example, between 450 and 550 RTG users' data rate maintains constant, but the percentage of served users is dropping, which results in an overall less data rate allocated to RTG.

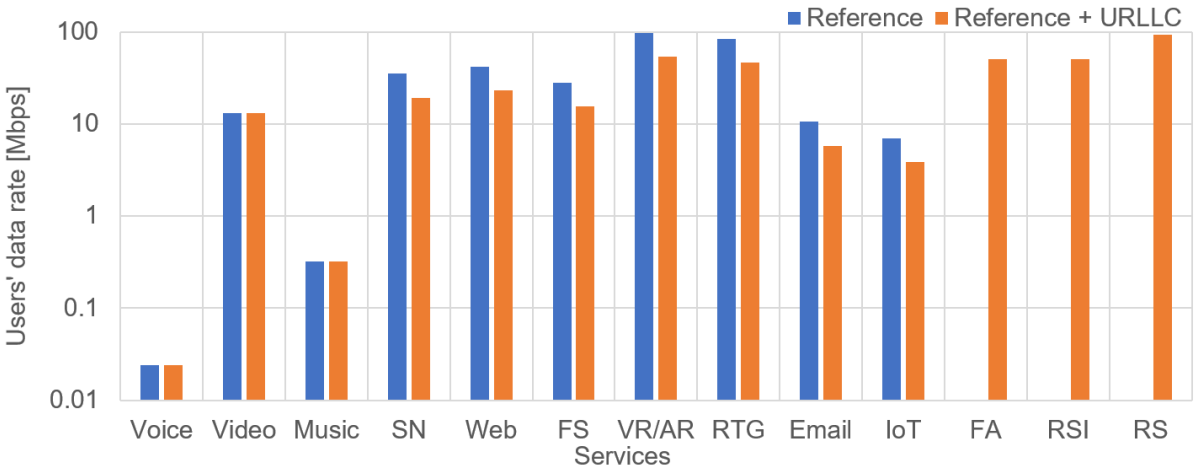


Figure 4.8 - Comparison between users' data rates for the reference scenario and the URLLC scenario.

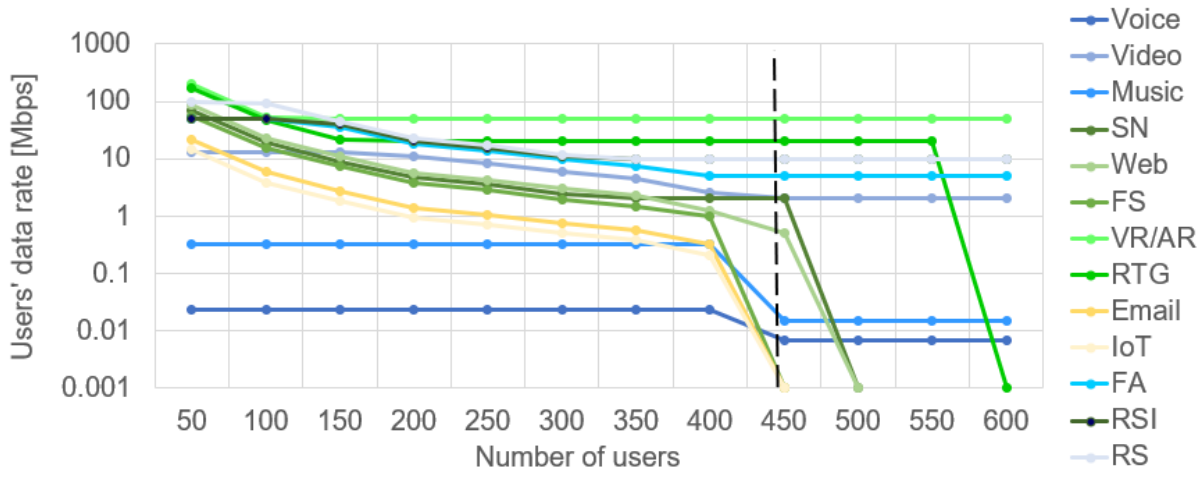


Figure 4.9 - Impact of the increment of number of users in the URLLC scenario.

The summary of the thresholds of each service, regarding maximum and minimum allocated capacity as well as delay, are presented in Table 4.11. For the BG VNOs the same logic as the previous subsection one was applied.

An important output to analyse, from the VNO perspective, is how the capacity share is distributed among the VNOs. Figure 4.10 illustrates how the VNOs compete for the capacity with the number of users increasing. The two GB VNOs behave in a similar fashion, starting with a small percentage of the cell capacity and increasing up until the point where one of its services starts getting allocated less data rate than its maximum. For GB RT peaks at 200 users while GB L peaks at 100 users. Comparing with the precious scenario, GB RT has roughly less 20% of the capacity share because of the new added VNO also using the cell resources. Both VNOs gradually decrease their capacity share until their services start serving with minimum demands of data rate and start increasing again from this point onwards. BG RT manages to peak at 62%, being the VNO with more capacity allocated at a given point,

the reasoning being the high demand services it has, even when being served minimum demands. From this point, there is a drop on its share resulting from the delay of RTG users. BG IA and BE BkG see a gradual decrease in the data rate allocated to them, due to the nature of their SLA and priorities. When the delay process starts, at 450 users, BG IA has only 2.28% of the cell capacity.

Table 4.11 – Thresholds of users to achieve maximum and minimum data rate as well as delay for the URLLC scenario.

	Maximum (≤)	Minimum (≥)	Delay (≥)
Voice	400	450	1100
Video	150	750	1100
Music	400	450	1100
SN	100	350	500
Web	250	450	500
FS	150	400	450
VR/AR	50	150	1100
RTG	50	200	600
Email	-	-	450
IoT	-	-	450
FA	100	400	1100
RSI	100	350	1100
RS	50	350	1100

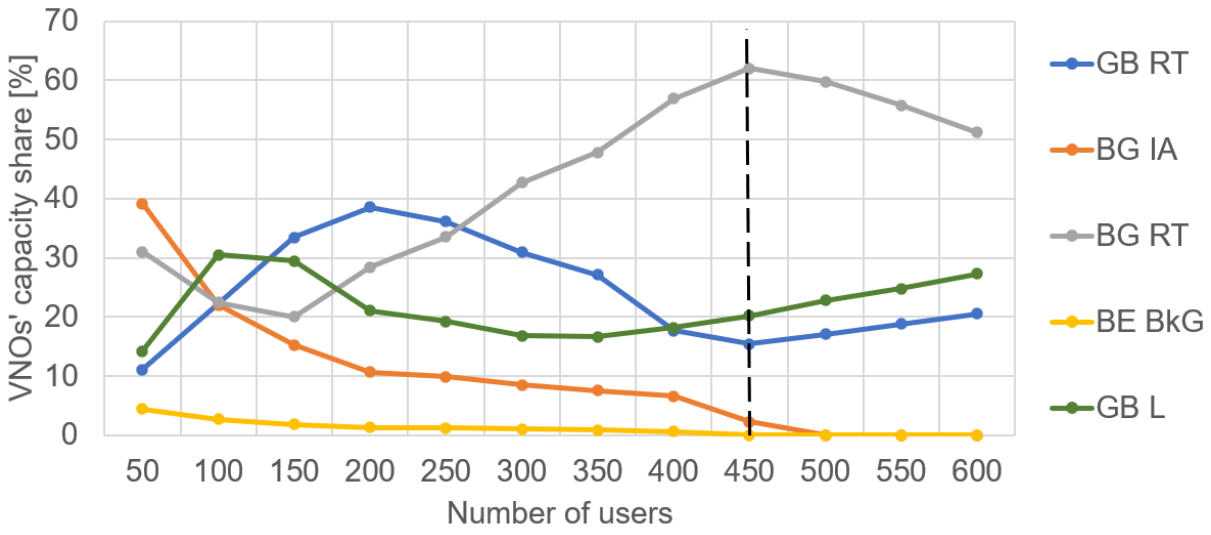


Figure 4.10 - VNOs' capacity share with the number of users increasing.

Besides the delay threshold starting earlier, all other thresholds see a shift as well. The most important conclusion to take from the observation of Table 4.11 is that URLLC type services can be served even with high number of users, but one should not expect great performances as they are being served with minimum demands. The nature of these services makes it so that their priority levels are very high, so this is the reason why they never get delayed. Nevertheless, the ideal number of users for these services is around 50 and 100, to ensure maximum user satisfaction and the best performance out of these services. Figure 4.11 shows how users' satisfaction changes with the increase of served users.

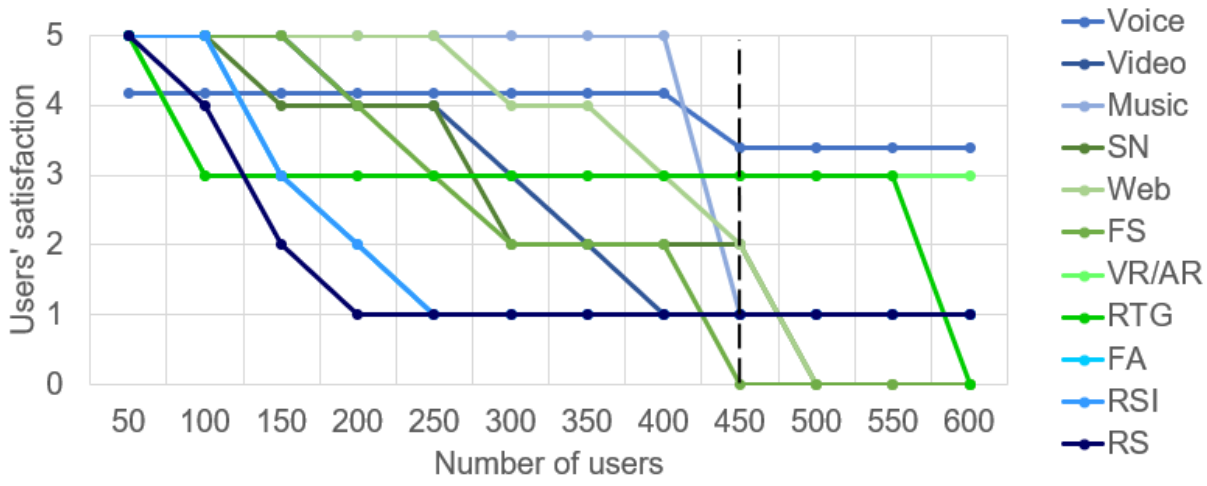


Figure 4.11 - Users' satisfaction with the increment of the number of users.

4.5 Impact of the mMTC VNO

This subsection is reserved for the study of the impact that many smart meters have in the reference scenario. In this scenario the number of users distributed by the several services is different from the previous considered scenarios. The number of smart meters was obtained based on statistical data of the number of people that live in Benfica, one the Lisbon's districts. Benfica has a population of 36 985 [WIKI20] and an area of 8.03 km^2, which translates into approximately a population of 4 623 per km^2. A household has 2.5 people, on average, [PORD19] so the number of people per house is rounded down to 2. This means that in Benfica there are 2 311 houses per km^2. Assuming each house has 3 smart meters (water, electricity and gas), one concludes that there are 6 933 smart meters per km^2. In this work, 3 cell sites are being considered, thus dividing 6 933 by 3 one gets the final value of 2 311 smart meters per cell. Due to the number of smart meters being much greater than the number of users, the mix is divided into two categories for better understanding. A global category that includes the 2 411 users considered in this scenario, and a relative category where there are two mixes: the mix of the GB MM VNO and the mix of the 100 users from remaining VNOs. Comparing the two categories, regarding smart meter percentages, for the relative category smart meters have 100% of the 2 311 while in the

global category they represent 95.85% of the 2 411 total users. Table 4.3 shows the scenario represented by Mix mMTC. The calculation of the maximum achievable data rate is the same as in the reference scenario, thus $R_{cell} = 1.777$ Gbps. The minimum demanded data rate is 253.45 Mbps.

In practice, this scenario leads to results similar to the reference scenario. Even with thousands of devices connected to the cell, their demands are so low that the outputs of the model are almost the same. In the end, all the data rate that gets allocated to this VNO is 46.22 Mbps which is roughly equivalent to one VR/AR user. Even with such low data rate requirements, the devices are not always served with maximum data rate, because the priority associated to this service is 1, i.e., 300 times lower than the priority of voice for example. Therefore, the impact these devices have in the maximisation of the objective function is much lower. Resulting from the explanation above, rather than showing the same figures as shown in the previous subsections, the difference between the VNOs' capacity share from the reference scenario and this one is displayed in Figure 4.12, which uses

$$R_{VRRM[\%]}^{VNO_v} = R_{VRRM[\%]}^{VNO_{v\_Ref}} - R_{VRRM[\%]}^{VNO_{v\_mMTC}} \tag{4.1}$$

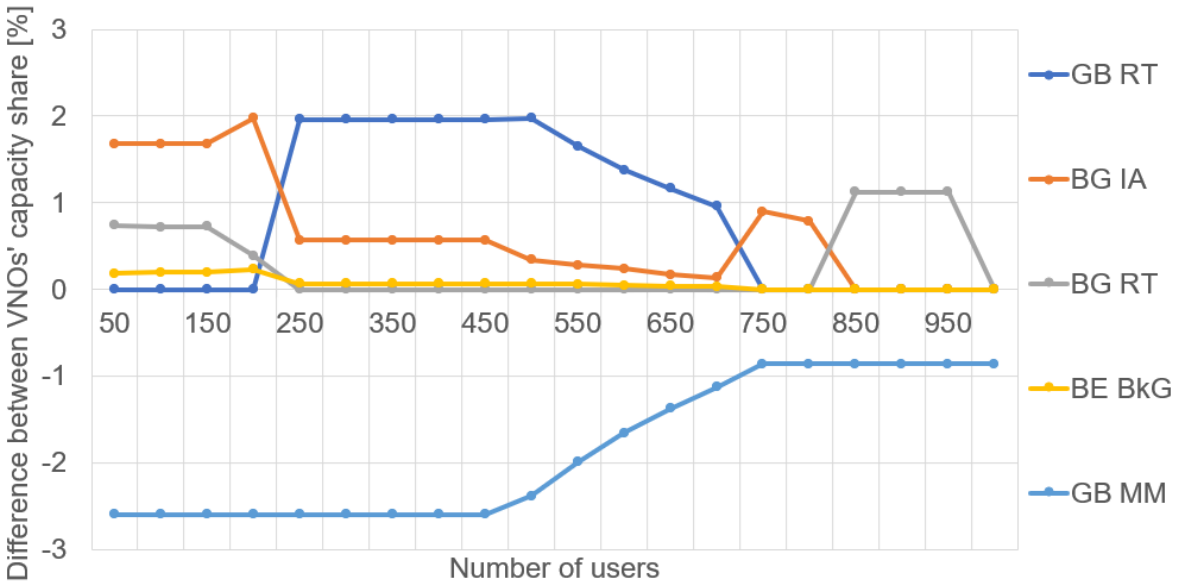


Figure 4.12 – Evolution over the number of users of the difference between the reference scenario and mMTC scenario VNOs' capacity share.

By direct analysis of the figure, the first thing that stands out is the negative value line. This results from the fact that the reference scenario did not have VNO GB MM. Also, regarding this VNO capacity share, one can see that not only it never occupies a big part in the division of data rate among VNOs (highest percentage being 2.6%), but it also starts decreasing when the number of users is greater than 450 (lowest percentage 0.8%). That is due to the priority set to this VNO, as explained above. Further important conclusions, concerning the remaining VNOs' capacity share, are the following. VNO GB RT is only different when the number of users ranges from 250 to 750. This is because the allocation of data rate for video is being optimised, by the CVX solver, in this range of values. The result is not the same because of the existence of VNO GB MM in one of the scenarios. Outside this range of values,

the difference is zero because all services are being served with either maximum or minimum data rate. For the other VNOs the same logic applies. When all the services belonging to a VNO do not suffer changes to their allocated data rate, the value maintains constant, whilst when there is optimisation or delaying being done the difference between VNOs capacity share changes. All these differences are almost overlooked with the biggest difference, outside the new added VNO, being 2%.

4.6 Influence of Service Mix

This subsection addresses the study of service mix variations. The reference scenario and the URLLC scenario will suffer variations to their mixes to better represent other real-world scenarios. This way VNOs can understand how flexible they can be when serving different sets of mixes. The behaviour of the model with the increment of the number of users is also a subject of study. Two scenarios were considered, represented in Table 4.5, the first one being a football scenario and the second being a hospital scenario.

The football scenario is a direct variation from the reference scenario, where the process on acquiring the mix values was as follows: VNO GB RT saw a significant reduction in all its services because it is not expected for people to be listening to music, watching videos or making phone calls during a football match but, because it is not impossible, these number are not zero; on the contrary, VNO BG IA saw an increase on almost all of its services mixes, which is expected because there is always a percentage of people streaming and sharing the event on social media and web browsing, while file sharing was reduced to one because it is not expected to be a service people use in this scenario. VNO BG RT is also very different mainly because of the increase from 2% to 30% VR/AR service. The thought on this service is that in a futuristic scenario it is likely people is using VR/AR, mainly AR, to augment their experience of the match. Finally, VNO BE BkG is the more constant out of the four VNOs, with IoT remaining the same and Email dropping from 3% to 1%.

The hospital scenario is an indirect variation from the reference scenario, meaning it is a variation from the URLLC scenario which was a variation from the reference scenario. Unlike the football scenario, where people have the same purpose, to watch the match, in the hospital scenario it is easier to identify the groups of people that are going to be using different services. Firstly, and more importantly, to emphasise the doctors performing RS, this service was increased from 2% to 10%. VR/AR is also increased from 3% to 8% because it is expected that doctors use these tools to better aid them in their needs. Secondly, the hospital staff, such as assistants and secretaries for example, or people in the waiting room, are expected to be using a lot of voice, but not nearly as much video as before. The remaining services are similar but with few variations to encompass what one could expect from a scenario like this. Note that FA and RSI are set to zero since they do not fall under this scenario's expected services.

Starting with the football scenario, Figure 4.13, illustrates the comparison between the users' data rate

for each service from the football and the reference scenarios.

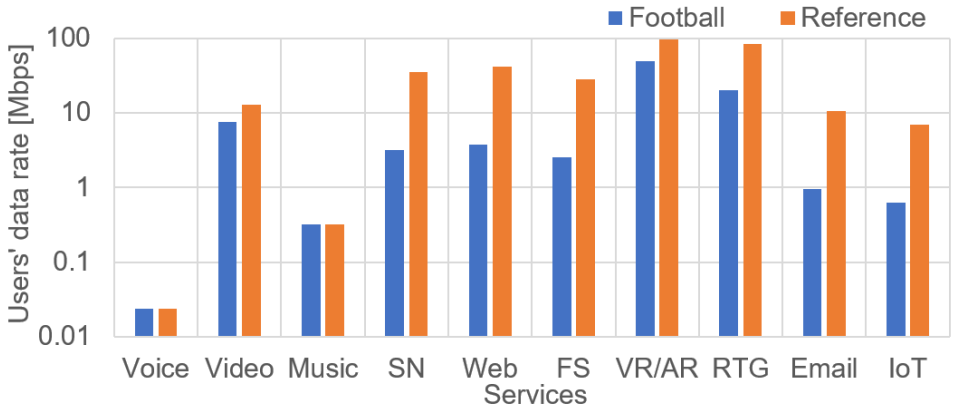


Figure 4.13 - Comparison between users' data rates for the football and reference scenarios.

The obtained results show that this is a very demanding scenario where most services are being served with low data rate, which results in a worse user satisfaction. The main service for this scenario VR/AR is being served with minimum demanded data rate, 50 Mbps, because there are a lot of users allocated to this service. Also, when studying the impact of adding more users, the problem becomes automatically impossible to solve due to the lack of capacity in the cell. The delay process starts in the first iteration, where the number of users is 150. VNOs should take into consideration that increasing the number of people using this type of service requires more capacity than usual due to its high demands. Using higher levels of M_{UMIMO} and carrier aggregation are some possibilities that will help improve scenarios like this one.

Regarding the hospital scenario, Figures 4.14 and 4.15, illustrate the users' allocated data rate and the VNO capacity share.

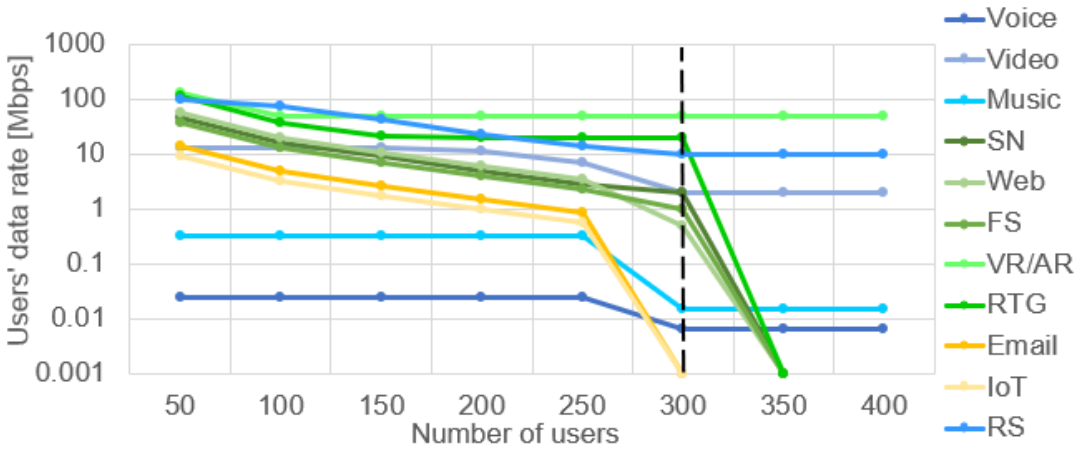


Figure 4.14 – Impact of the increment of number of users in the reference scenario.

This scenario is better than the previous one in the sense that the optimisation problem can run during more iterations of the number of users. In practice this means that the cell is not as overloaded as the previous scenario. From the analysis of the obtained results, one can observe that, for the reference number of users (100), the allocated data rate each service receives is satisfactory. The most important service in this scenario, RS, gets each of the 10 users allocated 76.29 Mbps, which corresponds to a "fair" users' satisfaction just shy of "good" at 80 Mbps. Also, it is worth mentioning that the model is capable of always allocating data rate to this service even when the delay process starts. The remaining services are also allocated data rates such that its users fall between the classification of "fair" and "excellent", which is also very positive.

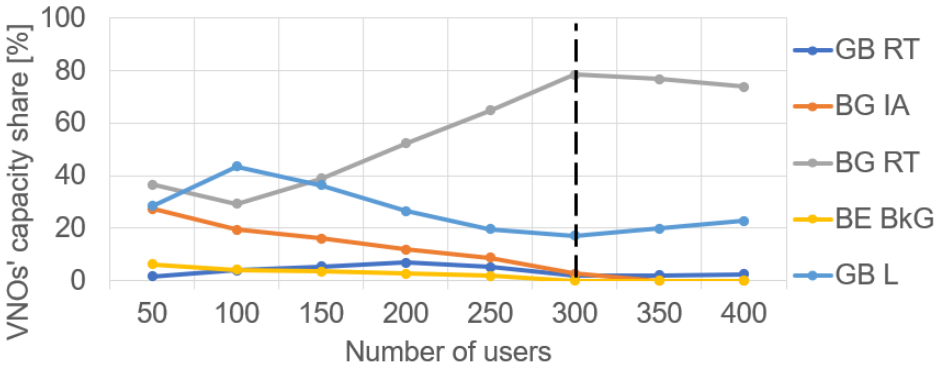


Figure 4.15 – VNOs' capacity share with the number of users increasing for the hospital scenario.

In Figure 4.15, the VNOs' capacity share evolution is presented, with the increase of served users. For the reference number of users, the most important VNO of this scenario, GB L, is being allocated the biggest percentage out of all the VNOs. At the point of delay, BG RT has almost 80% of the cell capacity, because the minimum demands of data rate of its services and number of users are much greater than the other services. This result is inevitable, because it does not depend on the priority associated with the service. Nevertheless, it is possible for VNO GB L to have a bigger share of the cell capacity prior to this threshold by increasing the priority γ, meaning paying more to the respective InP.

4.7 Influence of the Priorities

This subsection is reserved for the last change made to the reference scenario, regarding services and VNO priorities. As described before, the priority associated with a service, λ_{v_s}, that ultimatly defines the weigh that the user will have in the solution of convex optimisation problem, is the result of the product of $\gamma_s$with δ_s. γ_s being the priority that results from the contract between the VNO and InP and δ_s the priority assigned to a service by the VNO. The objective is to change these values and study how these variations impact the previously studied scenarios and developed model. As such, two changes were

made. The first one is described in Table 4.5.

In this first modification of the reference scenario, each VNO was divided into two identical VNOs and the mix was also divided between them. For example, in this new scenario there is the VNO GB RT and the VNO GB RT 2 each with the same mix among their users. The same applies to the other VNOs. All the duplicate VNOs have their γ_s value halved, allowing a study where two VNOs providing the same service compete for the cell capacity, one being the "low cost" version of the other. This study is helpful to understand if a VNO should or should not invest in priority for its slice depending on the services that it is providing.

After running the programme and getting the results, Figure 4.16 displays the evolution of the users' data rate through the increase in the number of users.

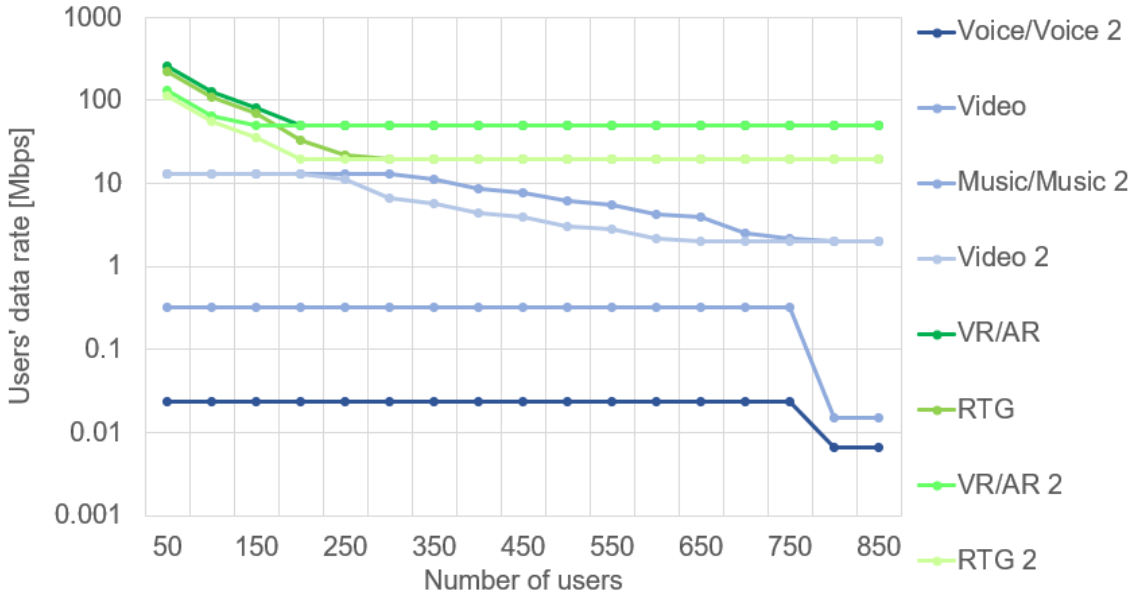
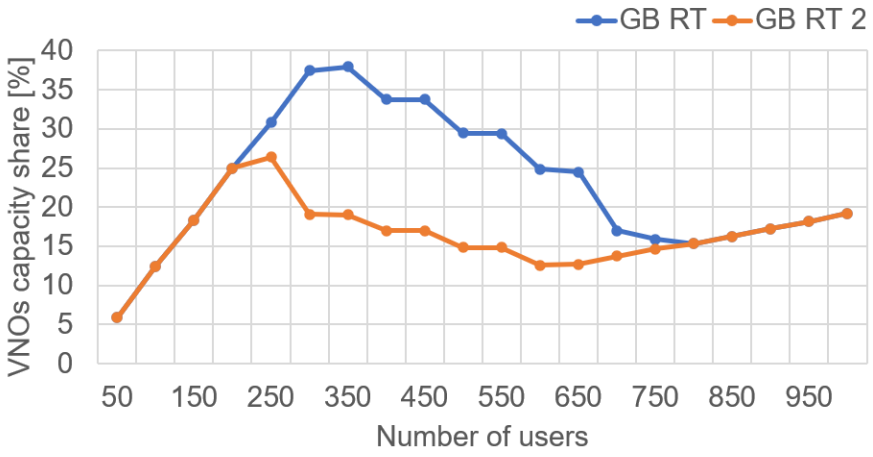


Figure 4.16 – Impact of the increment of number of users in the premium and low-cost slices scenario.

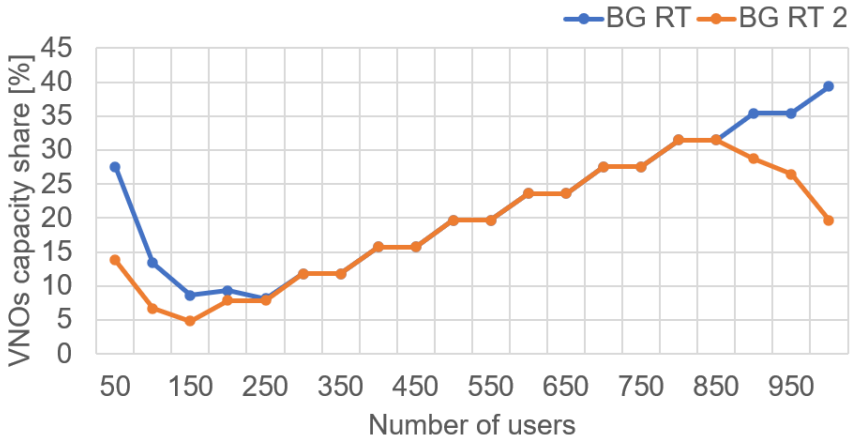
Four out of the eight VNOs were chosen to be represented: GB RT, BG RT and their respective duplicates. VNO GB RT, that comprises voice, video and music, sees that both voice and music do not have any difference regarding the data rate serving its users. Both have the same minimum and maximum demand and one has half the priority of the other. This happens because these services are extremely low data rate demanding so the cell has no problems serving both the same way. On the other hand, the same does not apply to video. The higher priority service has a significant amount of more data rate allocated to him, between 200 and 800 served users. In this zone, the cell needs to manage its resources and prioritises the service of the VNO that paid more. Outside this range both services have the same allocated data rate because either there is enough capacity to serve both at maximum demanded data rate (<200 users) or there are so many users that both services need to operate at the minimum demanded data rate (>800 users). As for the BG RT VNO and its duplicate

lower quality VNO, the main difference is verified prior to 200 users, for VR/AR, and 300 users, for RTG. These are high demanding services, so it is natural that they start being served with minimum demanded data rate sooner than other services. Regarding the remaining non displayed services, the evolution of these is the same across all of them: the duplicate VNO services are allocated half the data rate of the main VNO until they converge to the minimum demand or zero.

Another interesting result to analyse is the VNOs' capacity share. The results are illustrated across Figure 4.17 and Figure 4.18, for a better visualisation and comparison of the outputs between the main VNO and its duplicate. Note that, although only two VNOs are represented per graphic, the summation of all eight VNOs from all the graphics is what gives the 100% capacity of the cell. Figure 4.17 shows the evolution of the VNOs GB RT and BG RT.



(a) GB RT VNO



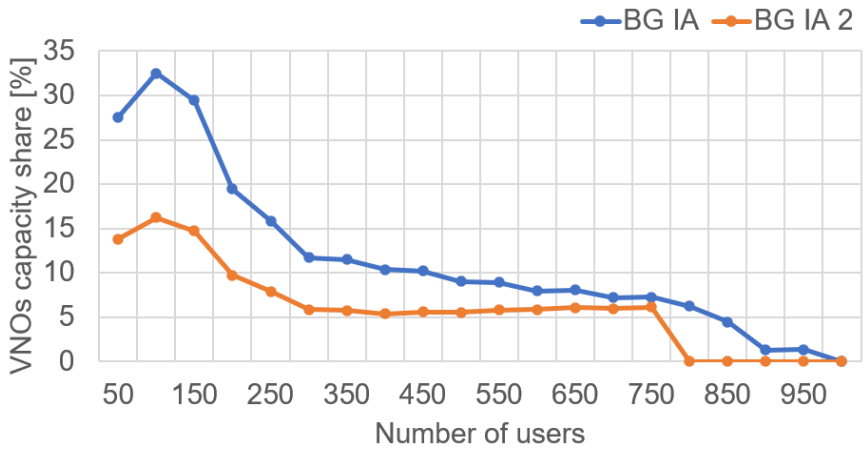
(b) BG RT VNO

Figure 4.17 - VNOs' capacity share with the number of users increasing of VNOs GB RT and BG RT.

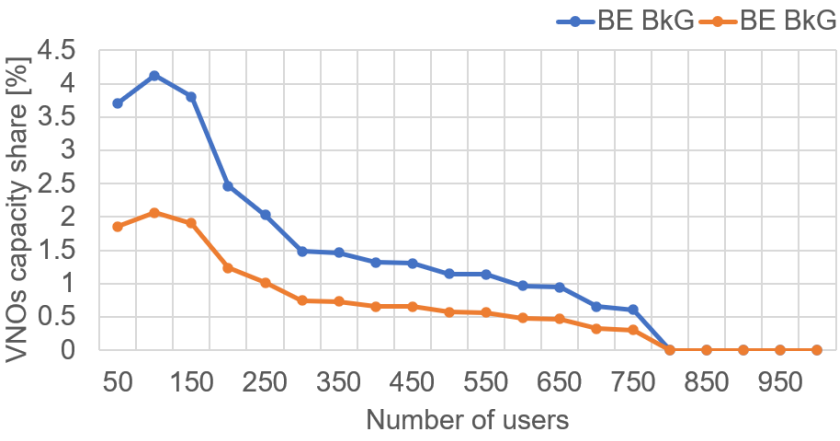
With the analysis of these results the VNO can determine whether it should or should not invest in a certain slice depending on the target audience. For example, if the average number of users of a certain

cell is 350, VNOs that want to provide video should invest in better quality slices because the difference in the capacity share is around 18%, the biggest gap verified. On the other hand, if the intended service is VR/AR or RTG, this is irrelevant as all users will get the same data rate, thus resulting in the same capacity share. For these types of services, the main difference is with lower users, where it impacts the most in terms of user satisfaction, and in saturation when users get delayed first if the VNO has a worse contract with the InP.

Figure 4.18 displays the evolution of VNOs BG IA and BE BkG.



(a) BG IA VNO.



(b) BE BkG VNO.

Figure 4.18 – VNOs' capacity share with the number of users increasing of BG IA and BE BkG VNOs.

These two VNOs were grouped together since their behaviour is very similar which leads to the same conclusions. VNOs should opt for the quality that they intend to give to their users, which is true for all VNOs regardless of the service they are providing. But, for these types of services, VNOs should not consider the number of users a relevant factor. In other words, the number of users matters but the QoS being provided is always proportional to the priority of that VNO. So, no matter how many users are being served, if a VNO has paid more to have better quality services the result will show regardless of

the users, unlike the previous services where for a certain number of users the quality and respective capacity share was the same. The only exception to this is for BG IA VNOs when the number of users is very high, the lower priority VNO users are delayed first while the others keep getting served with minimum demands resulting in this higher quality VNO still having some capacity share while the other is at zero. Note that this study assumes fixed values for minimum and maximum data rate ranges for all services from GB and BG VNOs. If a VNO that paid more not only got associated a higher priority but also an increased data ranged, this study would no longer apply. In that case what would happen, for voice for example, is that the higher priority VNO would have its users at his max, which can be called $voice_{vno1\_max}$, and the lower priority VNO would also be serving its voice users at his predefined maximum, $voice_{vno2\_max}$.

The study of how the priorities affect the reference scenario and the developed model is directed to the URLLC scenario. Now the goal is to change γ_s, assigned to VNO GB L, from values ranging 10 and 100 with increments of 10, while maintaining constant the number of users. This way, one can understand what the best value for this VNO is. This specific VNO was chosen for this study due to being crucial to secure the maximum performance of services like RS and RSI. Figure 4.19 shows the evolution of the users' allocated data rate with changing γ_s and a constant number of 100 users. FA and RSI converge to their respective maximums when $\gamma_s = 30$, but only when $\gamma_s = 50$, the same is true for RS, because of the higher maximum demanded data rate of the service. The other services also see some changes resulting from more data rate being allocated to VNO GB L, with the ones being more affected being services from BG RT.

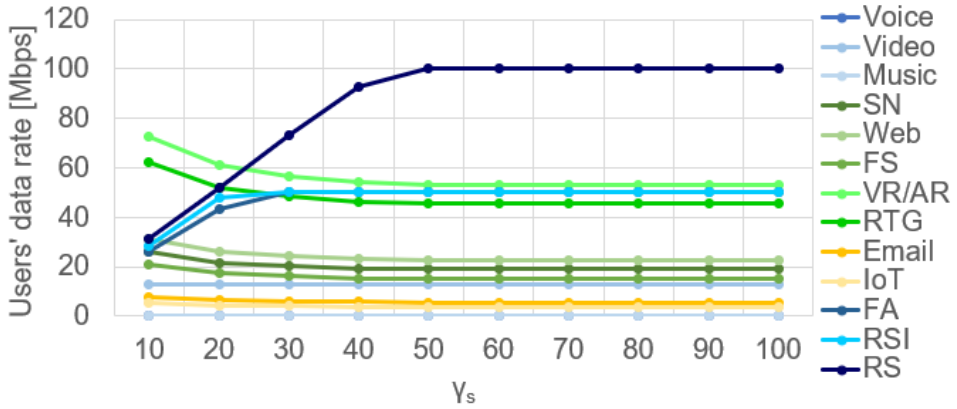


Figure 4.19 – Users' data rate with γ_s increasing.

Figure 4.20 presents the VNOs' capacity share and its impact with the changing γ_s. It is easy to conclude that the ideal value for this VNO to have its priority set is 50 as it allows its services to get the maximum data rate each demands. Any value below that will result in worse performances, and any value above is pointless.

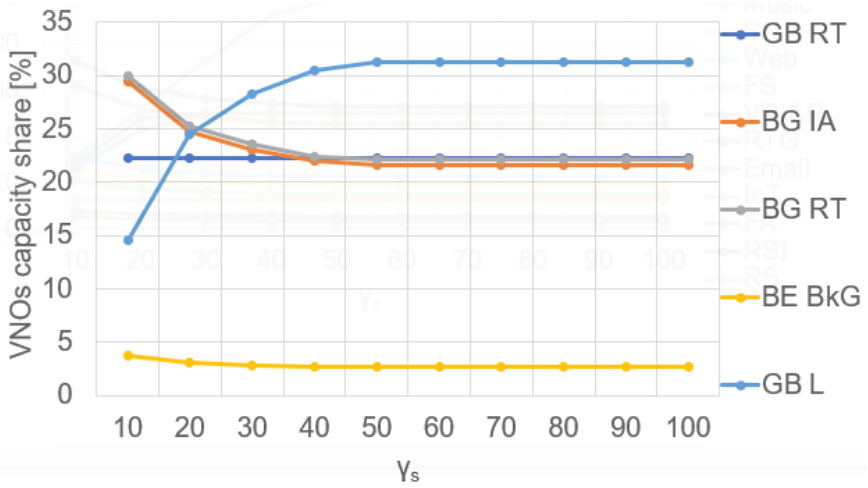


Figure 4.20 – VNOs' capacity share with γ_s increasing.

Chapter 5

Conclusions

This chapter finalises this this thesis, presents the main conclusions and future work in the area.

The main goal of this work was to analyse the implementation of network slicing in 5G radio networks, more specifically, to understand how the radio resources should be allocated among slices and their isolation. The model was developed in a such way that the inputs consist of cell parameters that are used to calculate the amount of data rate available in one cell, and the scenario with all slices, services and users that are the subject of study. The results provide information regarding the network status such as the VNOs' capacity share and services' allocated data rate, and more useful parameters from the user's viewpoint such as the users' allocated data rate.

Chapter 1 gives an overview to this work, presenting the current situation of mobile communication systems and the evolution in data consumption. The new requirements services have been addressed as well as the role of network slicing in 5G radio systems. The motivation and content structure are also presented in this chapter.

In Chapter 2, the fundamental concepts needed to understand the work of this thesis are given. The contents within this chapter were also structured in a progressive way in the sense that first LTE concepts are presented, encompassing topics like the network architecture, radio interface, and coverage and capacity. This is necessary because, even though this work focuses on network slicing, one is in a period of transition between 4G and 5G. Following this line of thought, the same concepts are presented but regarding 5G systems and further compared to the differences in 4G ones. SA and NSA architectures are introduced in this chapter, because the first wave of 5G networks is deployed over existing 4G network infrastructures. Having to conceptual contents covered, the next topic that follows is virtualisation, being very important for this work. One provides a description of network virtualisation as well as SDN and NFV, with the role and impact each technology will have in network slicing. The section ends with the most important topic of this work, network slicing, presenting its definition and fundamental principles. 5G services and applications follow, first addressing the three main categories of services: eMBB representing very high values of data rate, URLLC with ultra-low latency applications and mMTC for massive numbers of devices connected to the network. A description of the legacy system classes is also presented alongside VNOs SLA types: GB for high priority services with guaranteed data rate, BG for services where only the minimum data rate threshold is assured and BE without any compromise in data rate allocation where services are allocated the available data rate in a best effort fashion. The chapter concludes with the essential performance parameters to consider in 5G and network slicing, and presenting the state of the art summarising relevant work being developed in the topic.

Chapter 3 covers the description of the model developed. The chapter starts by making an overview of the model, including input and output parameters. Next the calculations of the cell maximum achievable data rate are presented with all the parameters explained. The VRRM optimisation mathematical expressions are described and the output parameter equations and their relevance are also given. The explanation of the model implementation is done with help from flowcharts that can provide an easy visualisation of how the model works. The final section of Chapter 3 assesses the model with a small test scenario.

The model consists of four steps. The first step is to compute the maximum achievable cell data rate.

To achieve this, an adapted version of an expression from 3GPP was used. This expression takes into consideration several parameters, like carrier aggregation, modulation, MIMO layers and so on. There is one added parameter to this expression, which is the Multi-user MIMO factor that allows the base station to serve multiple users, that are close to it, with the same RB making a much more efficient usage of resources. After this stage, the output is the cell data rate and the model will check if this value is enough to serve all users from the scenario in study. This stage is called the admission control. If the result is negative, meaning there is not enough data rate to serve all users, the next step is the delay process algorithm. This algorithm delays all BE users, because they have no minimum contracted data rate values, and after that starts delaying BG users one by one until there is enough data rate. On the other hand, if the result from the admission control is positive, then the model has gathered sufficient conditions to proceed with the VRRM optimisation. Using CVX to solve a convex optimisation problem, the model allocates data rate to each user according to is priority. From this optimisation results a vector of values that is then used to compute the output parameters.

There are six output parameters, divided between two classes. The first class, named Network, contains parameters that are relevant from the network viewpoint. Starting with these, the first one is the percentage of total assigned data rate, which is the total network throughput in terms of data rate, meaning that if it is close to 100% it reflects an optimal VRRM performance. Next, the VRRM capacity share is the total capacity assigned to each slice out of the total capacity available to VRRM. The total data rate of each service is the total data rate assigned to a given service. Finally, the percentage of served users is the percentage of served users out of the total number of users of a specific service. The second class, named Users, contains two parameters that are important from the users' viewpoint, the first being the data rate of each user and the second the users' satisfaction, which is obtained based on bit rates from codec AMR-WB for voice, and the remaining services are classified in a similar fashion where each service has 5 levels from "Bad" being the worst to "Excellent" being the best.

Chapter 4 contains the result analysis with all defined scenarios, variations and results from the developed model. Starting with the study of the reference scenario, the approach was to make variations to the BW, studying the values used in Portugal by operators, to understand if it was possible to support the given scenario. The results show that for 10 MHz and 20 MHz of BW it is not possible to serve all users from the reference scenario. For 10 MHz BW only services from the GB VNO are served, and when using 20 MHz BW it starts being possible to serve VR/AR. When BW is 100 MHz all users are served, but users from BG RT VNO are almost at the minimum threshold of data rate, due to their high demands. Multi-user MIMO, being one of the new features that improves the cell capacity, is supported for 100 MHz BW, so the values 1, 2 and 3 were tested. In this case, using a value of 3 is indeed beneficial especially for the heavy data rate demanding VNO (BG RT), as it has boosted the data rates to almost the maximum threshold of user satisfaction.

Regarding the increment of the number of users to the reference scenario the number of users ranges from 50 to 1100. This variation allows for some very interesting results and to display the data obtained in a graph. The results from this test confirm what was to be expected regarding the behaviour of user's allocated data rate. GB VNOs have their users served with maximum data rate, for the respective

service, and BG and BE VNO have the maximum possible data rate according to their priorities. When the number of users increases first one can see a decrease in the data rate of BG and BE VNOs and only then the GB VNO. After 750 users the cell can no longer support the increasing rate of users, so it starts delaying them. The VNO GB RT never has any user delayed because of its SLA and priority. In this situation one can also see the isolation factor between slices that protects VNO GB RT from the other users in the sense that when users from other VNOs are delayed this does not affect the users from the first VNO.

The capacity share among VNOs is also another very important output parameter to analyse as it describes how VNOs compete for the available data rate. Here one can conclude that a VNO capacity share depends on the combination of multiple factors. First the SLA contract which determines the quality level of the contract between VNO and InP and if the VNO users is delayed first. Second the priority of each VNO and each service that determines the amount of data rate allocated to one user. And third, the number of users being considered. VNO GB RT, BG RT, BG IA and BE BkG start with a share of 12.48%, 56.43%, 24.87% and 6.22%, respectively. With the increase of the number of users the percentage of capacity share among VNOs changes accordingly. VNO GB RT has the most allocated data rate when the number of users is between 150 and 550 due the culmination of factors just described. After this point, the number of users is so much and the minimum data rate of VNO BG RT is so high that ends up surpassing all VNOs reaching 65% of the cell capacity.

The users' satisfaction is directly associated with the data rate that gets allocated to one user. Voice and music users are fully satisfied until the delay point where their data rate drops to the minimum demanded to maintain the service. Users from VR/AR and RTG services start by having excellent satisfaction values, but they rapidly decrease due to the nature of its services. Nevertheless, the minimum data rate allocated to these services is enough to guarantee a fair score. As for the other services, they see a piecewise decrease as the number of users increases.

Concerning the URRLC VNO, this added VNO brings users with considerable amounts of data rate. The results show the impact this VNO has in the reference scenario, BG and BE VNO services are allocated much less data rate. Considering 100 users, VR/AR has a 43.89 Mbps decrease and RTG 37.8 Mbps decrease. The delay point now is at 450 users as opposed to 750 users. Something to be noted is that the allocated data rate to the URLLC VNO, VNO GB L, is always above the minimum threshold due to its high priority level, only reaching it at 400 users.

Regarding the capacity share comparing with the precocious scenario, GB RT has roughly 20% less of the capacity share because both VNOs gradually decrease their capacity share until their services start serving with minimum demands of data rate and start increasing again from this point onwards. BG RT manages to peak at 62%, being the VNO with more capacity allocated at a given point, the reasoning being the high demand services it has, even when being served minimum demands.

Afterwards the mMTC VNO which adds to the reference scenario a VNO providing only one service, this one being smart meters. The number of devices connected to this VNO is 2 311 and they add on the 100 from the reference scenario making a total of 2 411 users. When increasing the number of users, the number of smart meters remains constant only changing the other services' users. The results

may be surprising as not much has changed from the reference scenario. There is a slight variation to the data rate allocated to each user, slightly less, but not enough to have an impact. This happens because even though thousands of devices are connected, their data rate requirements are so low, they end up not having an impact on the network traffic. One can conclude that if the network is prepared to connect thousands of devices, as a 5G network should be, they will offer no difficulties or problems to other existent VNOs.

In order to study the service mix influence, two new scenarios were created: a football scenario and a hospital scenario. In the football scenario the cell has barely enough resources to serve all users because the incredibly amount of VR/AR users are consuming almost all the resources. It is not even possible to make a study incrementing the number of users because the model automatically enters the delay process algorithm which means the cell is overloaded. This scenario proves that in order to serve high demanding data rate services such as VR/AR, not only the 100 MHz BW needs to be used, but also other techniques that increase the data rate such as carrier aggregation, multi-user MIMO and so on.

The hospital scenario is a variation of the URLLC scenario, where RS and voice have seen a big increase in its mix and the other services from VNO GB L were removed. The users from RS, the most important service in this scenario, are allocated 76.29 Mbps which corresponds to a “fair” users’ satisfaction just shy of “good” at 80 Mbps. The service is always being allocated data rate even when the delay process starts. The remaining services are also allocated data rates such that its users fall between the classification of “fair” and “excellent”, which is also very positive. Regarding the capacity share, for 100 users VNO GB L, is being allocated the biggest percentage out of all the VNOs. At the point of delay, BG RT has almost 80% of the cell capacity. Nevertheless, it is possible for VNO GB L to have a bigger share of the cell capacity prior to this threshold by increasing the priority γ, meaning paying more to the respective InP.

Finally, the last variation was made to the priorities of the VNOs and two changes were made. The results obtained are very interesting as they provide VNOs the necessary information whether they should invest or not in certain slices. The evolution of the capacity share of the VNOs show that for GB RT the main difference between higher and lower priority VNOs is when the number of users ranges from 200 and 800 users due to the service video allocated data rate being optimised. This is the opposite of what happens for the VNO BG RT where between 250 users and 850 capacity share is the same due to its services being served with minimum demands of data rate. For the remaining VNOs, the capacity share follows a more linear pattern where the higher priority VNO has double the capacity share of his duplicate. With this information, VNOs can choose what contract to do with the InP, based on the average number of users and the type of traffic they intend to serve.

The second change was changing the priority value of a target VNO from 10 to 100 and with this conclude what would be the best value for that VNO for that given scenario. The target VNO was VNO GB L. One can see an increase in allocated data rate with the increase of the priority level, as to be expected, and a saturation of the allocated data rate when the VNO priority is 50. Thus, all users from VNO GB L experiencing maximum data rate and are fully satisfied.

For future work, it would be interesting to implement two new ideas that were not possible to implement due to the time constraints of a work of this nature. Instead of optimising the data rate according to VNO and services priorities, it would be interesting to make an optimisation of the priorities themselves and this way understand what division of data rate could benefit the most users. Some new constraints need to be developed to ensure this idea works as well as the objective function. Another interesting topic to explore would be how to improve the existing model to incorporate the latency factor that was not considered in the developed model.

Annex A

Additional results

This Annex presents additional results that were not shown in Chapter 4.

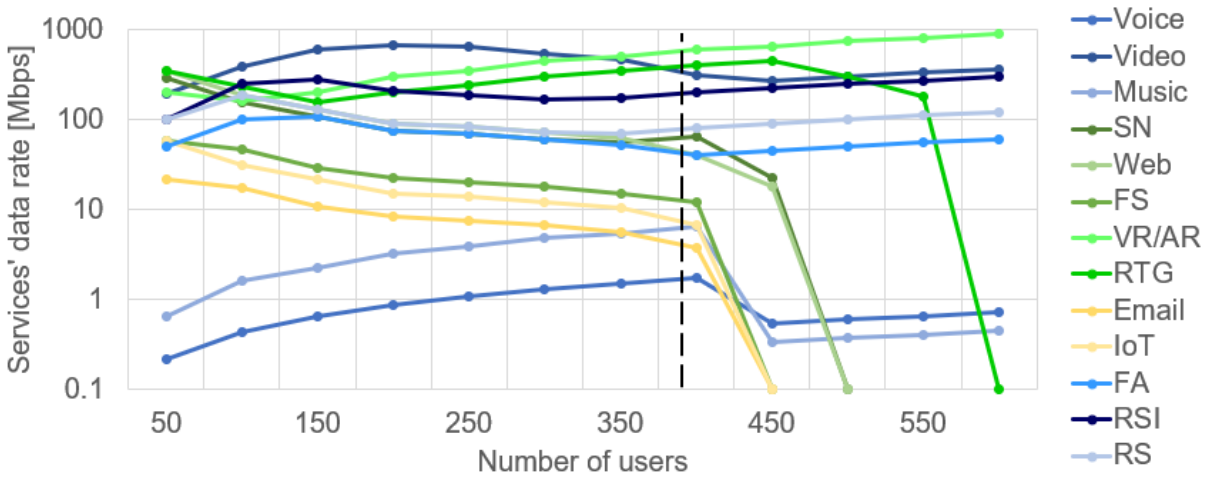


Figure A.1 – Services' data rate with the number of users increasing for the URLLC scenario.

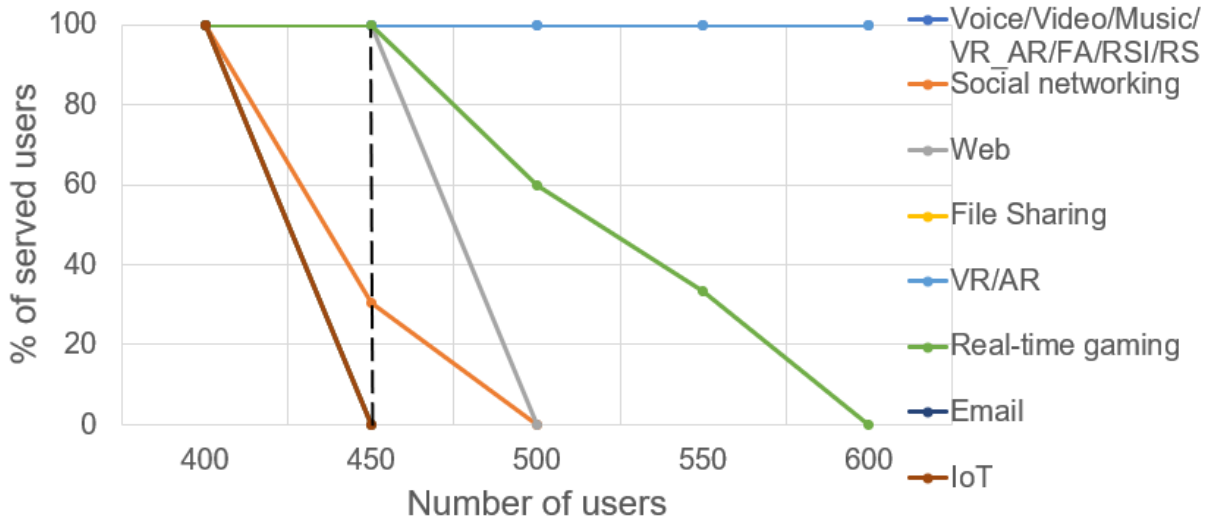


Figure A.2 – Percentage of served users with the number of users increasing for the URLLC scenario.

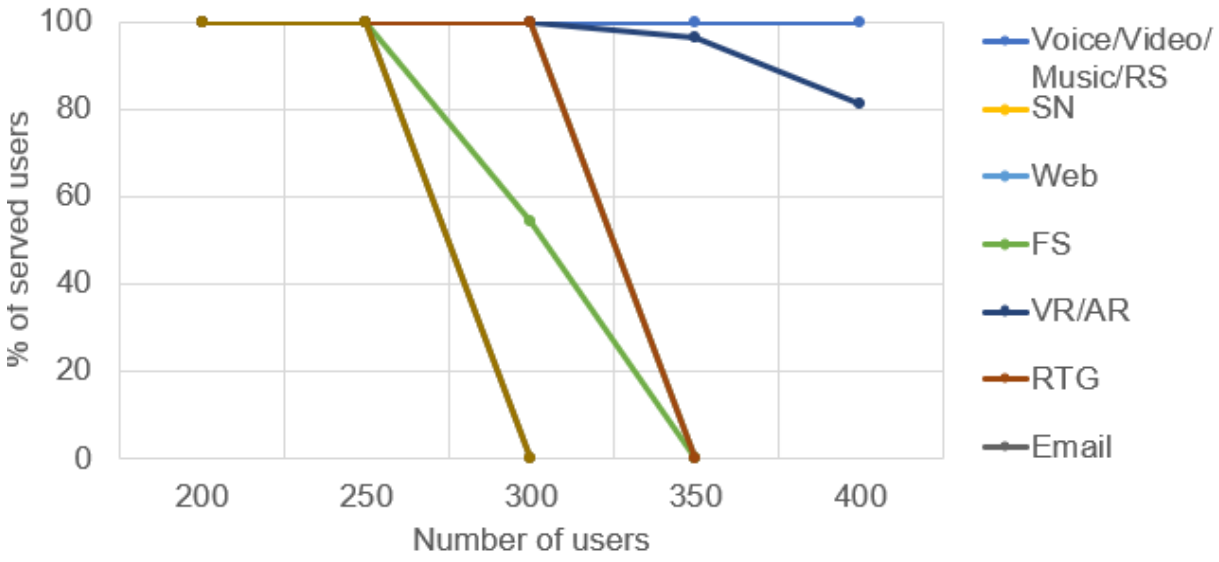


Figure A.3 – Percentage of served users with the number of users increasing for the Hospital scenario.

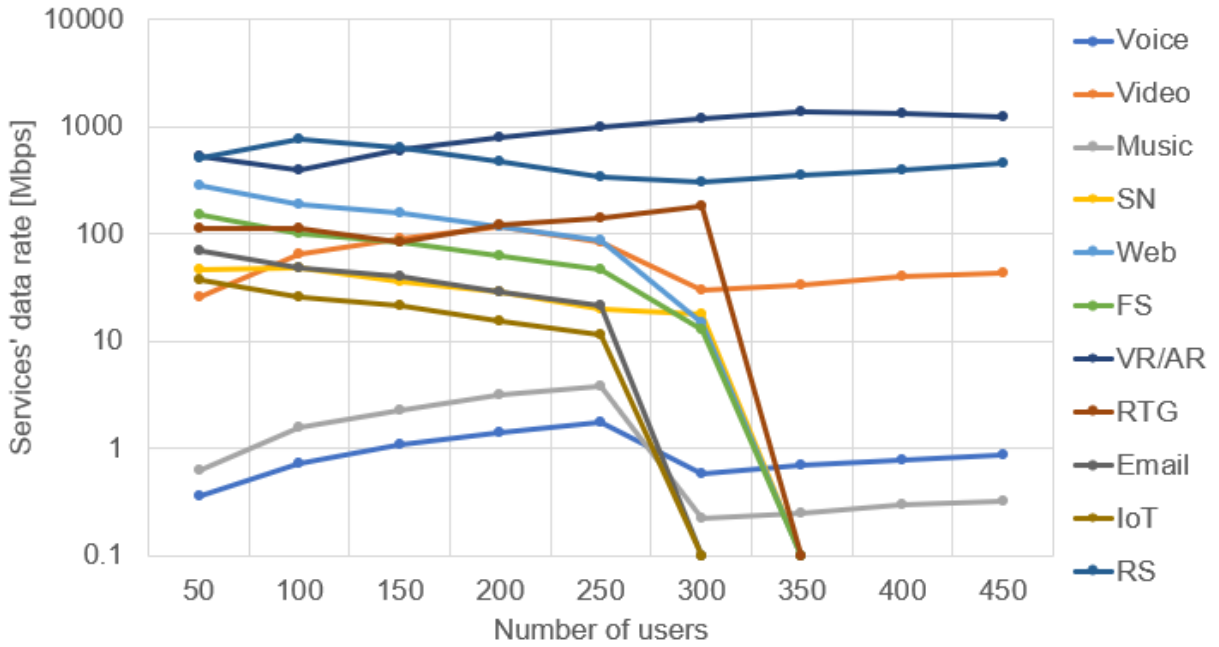


Figure A.4 – Services' data rate with the number of users increasing for the Hospital scenario.

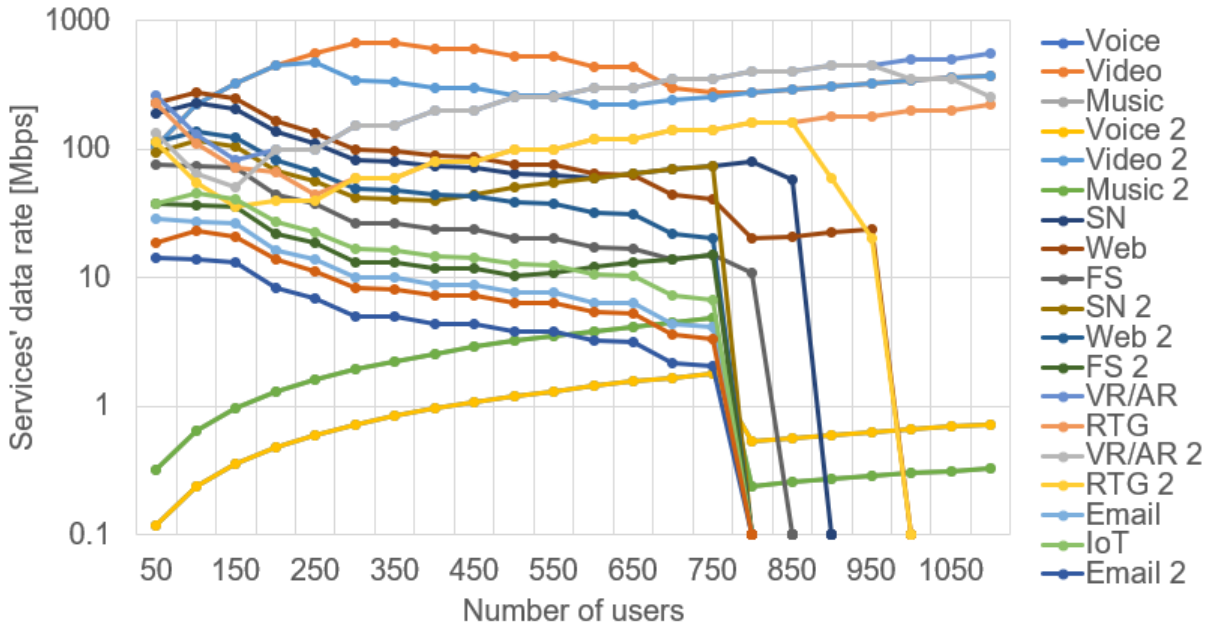


Figure A.5 – Services' data rate with the increase of the number of users for the scenario where services are duplicated.

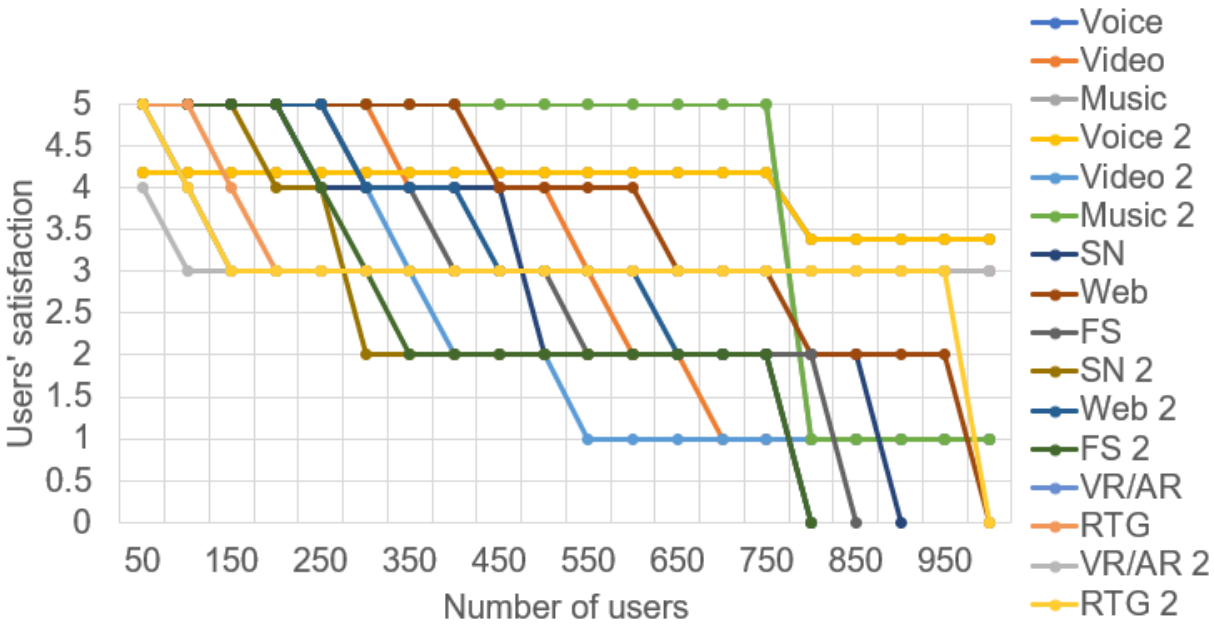


Figure A.6 – Users' satisfaction with the increase of the number of users for the scenario where services are duplicated.

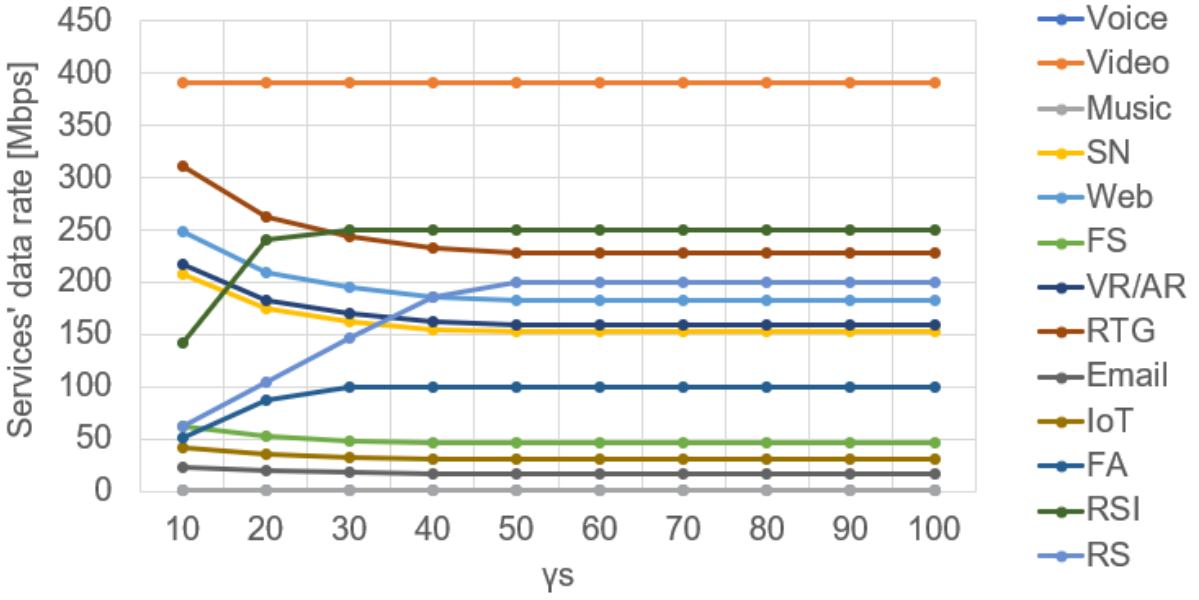


Figure A.7 – Services' data rate with the increase of γ_s.

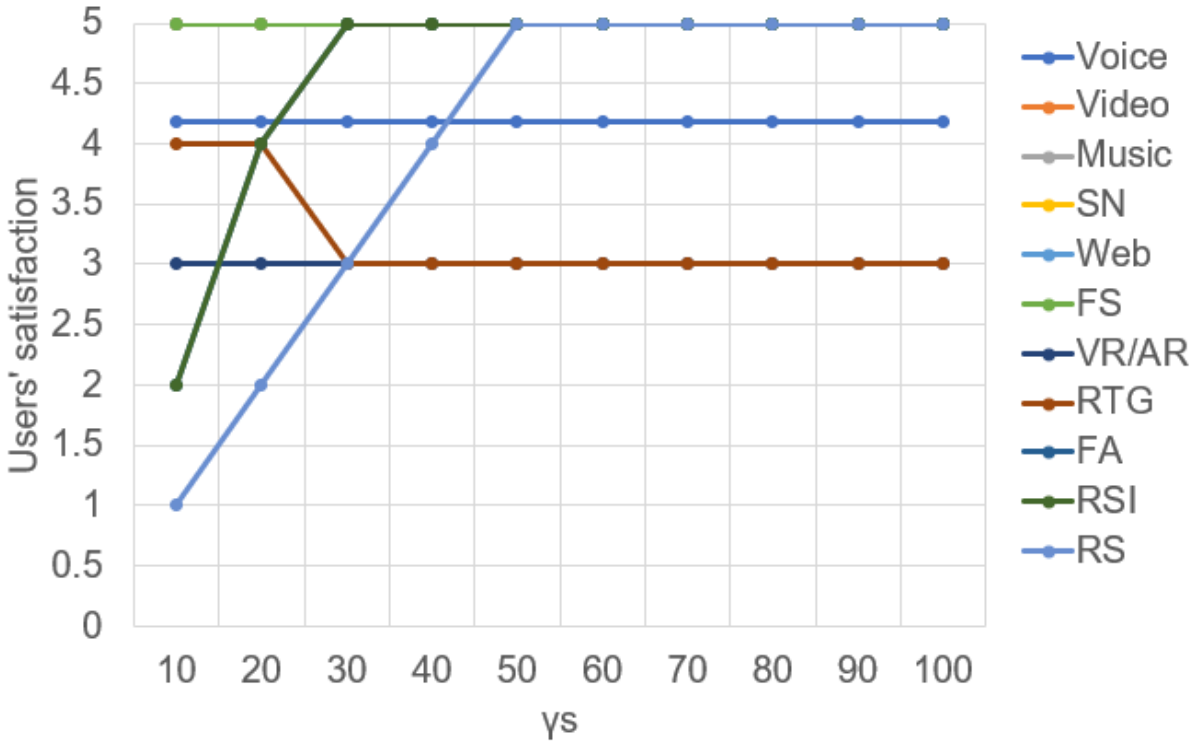


Figure A.8 – Users' satisfaction with the increase of γ_s.

References

[3GPP15] 3GPP, *Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Quality of Service (QoS) concept and architecture* (Release 13), ETSI TS, No. 23.107, Ver. 13.0.0, Dec. 2015

[3GPP17a] 3GPP, *System Architecture for the 5G System*, Release-15, v. 2.0.1, Dec. 2017.

[3GPP17b] 3GPP, *Technical Specification Group – User Equipment (UE) radio transmission and reception*; TR38.101-1 V1.0.0 Release 15, Dec. 2017 (https://goo.gl/hLmhBN).

[3GPP19] 3GPP, *Technical Specification Group Services and System Aspects; Release 15 Description,* Release 15, TR 21.915 V15.0.0, Sep. 2019

[3GPP20] 3GPP, *User Equipment (UE) radio access capabilities*, TS 38.3063, Mar. 2020.

[ANAC15] Mobile Communications System Utilized Frequencies, https://www.anacom.pt/render.jsp?categoryId=382989, Aug 2015, Accessed in May 2020

[ANAC19a] ANACOM, *Decision about the designation of the 700 MHz band for electronic terrestrial communications services* (in Portuguese), Dec 2019 (https://www.anacom.pt/streaming/dec23122019Atribuiao700_outrasfaixas.pdf?contentId=1498324&field=ATTACHED_FILE), Dec 2019

[ANAC19b] ANACOM, *ANACOM creates conditions for consistent and competitive development of 5G in Portugal*, Available: https://www.anacom.pt/render.jsp?contentId=1493002, Dec. 2019

[ATSK18] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini and H. Flinck, "Network Slicing and Softwarisation: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429-2453, thirdquarter 2018. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8320765&isnumber=8443345.

[BNKZ18] B. Bertenyi, S. Nagata, H. Kooropaty, X. Zhou, W. Chen, K. Younsun, X. Dai and X. Xu, *5G NR Radio Interface*, Journal of ICT Standardisation, Vol. 6, pp. 31-58, 2018.

[BPNA16] C. Bockelmann et al., "Massive machine-type communications in 5g: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59-65, Sep. 2016. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7565189&isnumber=7565175

[Conne18] T. Connectivity, "*Mass connectivity in the 5G era, 2018*," 2018.

[Corr20] L.M. Correia, *Notes from Mobile Communication course*, Instituto Superior Técnico,

University of Lisbon, Lisbon, Portugal, 2019.

[Dini20] R. Dinis, NOS Private Communication, Jun. 2020

[Domi18] S. Domingues, *Analysis of the Performance of Multi-Access Edge Computing Network Slicing in 5G*, M.Sc. Thesis, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2018.

[Eric18] Ericsson, *Mobility Report*, Jun. 2018, Available: https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf.

[Eric19] Ericsson, *Mobility Report*, Nov. 2019 Available: https://www.ericsson.com/4acd7e/assets/local/mobility-report/documents/2019/emr-november-2019.pdf.

[ETSI14] ETSI, *5G NR Network Functions Virtualisation (NFV); Architectural Framework*, GS NFV 002 V1.2.1, Dec. 2014, Available: https://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_NFV002v010201p.pdf

[ETSI18] ETSI, *5G NR Physical layer; General description*, TS 138 201 V15.0.0, Sep. 2018, Available: https://www.etsi.org/deliver/etsi_ts/138200_138299/138201/15.00.00_60/ts_138201v150000p.pdf

[GSMA18] GSMA, Road to 5G: Introduction and Migration, Apr. 2018.

[HGLL15] B. Han, V. Gopalakrishnan, L. Ji and S. Lee, "Network function virtualisation: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90-97, Feb. 2015. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7045396&isnumber=7045380

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

[HSMA14] H. Hawilo, A. Shami, M. Mirahmadi and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp. 18-26, Nov.-Dec. 2014. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6963800&isnumber=6963793

[HTSc17] B. Han, S. Tayade and H. D. Schotten, "Modeling profit of sliced 5G networks for advanced network resource management and slice implementation," in *Proc. of 2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, 2017, pp. 576-581

[ITUR15] ITU-R, *IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond*, Recommendation ITU-R M.2083-0, 2015.

[JiCM16] M. Jiang, M. Condoluci, and T. Mahmoodi, “Network Slicing Management & Prioritisation in 5G Mobile Systems,” Euro. Wireless 2016, Oulu, Finland, 2016, pp. 1–6. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7499297

[JSSA14] M. Jammal, T. Singh, A. Shami, R. Asal and Y. Li. “Software defined networking: State of the art and research challenges.” *Computer Networks* 72 (2014): 74-98.

[JZHT14] M. Jarschel, T. Zinner, T. Hossfeld, P. Tran-Gia and W. Kellerer, "Interfaces, attributes, and use cases: A compass for SDN," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 210-217, June 2014, Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6829966&isnumber=6829933

[KaSa18] A.T.Z. Kasgari and W. Saad, "Stochastic optimisation and control framework for 5G network slicing with effective isolation," in *Prof. of 52nd Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2018, pp. 1-6. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8362271&isnumber=8362188

[KFJK17] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir and Y. Wang, "Wireless Network Virtualisation with SDN and C-RAN for 5G Networks: Requirements, Opportunities, and Challenges," *IEEE Access*, vol. 5, pp. 19099-19115, 2017. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8025644&isnumber=7859429

[Mark99] R. B. Marks, "The IEEE 802.16 Working Group on broadband wireless," *IEEE Network*, vol. 13, no. 2, pp. 4-5, March-April 1999, Available: http://www.ieee802.org/16/sysreq/contributions/80216sc-99_28.pdf

[MBQB18] P. Marsch, O. Bulakci, O. Queseth and M. Boldi, *5G System Design: Architectural and Functional Considerations and Long Term Research*, John Wiley & Sons, Hoboken, NJ, USA, 2018.

[Okub11] N. Okubo, et al., *Overview of LTE Radio Interface and Radio Network Architecture for High Speed, High Capacity and Low Latency, Special Articles on "Xi" (Crossy) LTE Services-Toward Smart Innovation*, NTT DOCOMO Technical Journal vol. 13 No. 1, Jun. 2011. Available: https://pdfs.semanticscholar.org/008c/9ac4e0f6419abef4a13c4a3218a9fbd38839.pdf

[PORT20] Pordata base de Dados Portugal Contemporâneo, https://www.pordata.pt/Portugal/Dimens%C3%A3o+m%C3%A9dia+dos+agregados+dom%C3%A9sticos+privados+-511, Jun. 2020

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

[SaIF19] B.J. Santoso, R.M. Ijtihadie and M. al Fatih Abil Fida, "Docker-based Network Functions Virtualisation as a Learning Tool in Computer Network Course," in *Prof. of 12th International Conference on Information & Communication Technology and System (ICTS*), Surabaya, Indonesia, 2019, pp. 338-342. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8850969&isnumber=8850928.

[SETB11] S. Sesia, I. Toufik, and I. Baker, *LTE. The UMTS Long Term Evolution: From Theory to Practice* (2nd edition), John Wiley & Sons, Chichester, United Kingdom, 2011.

[Shar18] K. Sharma*, Comparison of energy efficiency between macro and micro cells using energy*

saving schemes, M.Sc. Thesis, University of Lund, Lund, Sweden, 2018.

[SPFA17] O. Sallent, J. Perez-Romero, R. Ferrus and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166-174, October 2017. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7891795&isnumber=8088405

[SSCB17] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia and A. Banchs, "Mobile traffic forecasting for maximising 5G network slicing resource utilisation," in *Proc. of IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, Atlanta, GA, 2017, pp. 1-9. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8057230&isnumber=8056940

[Vass17] S. Vassilaras et al., "The Algorithmic Aspects of Network Slicing," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 112-119, Aug. 2017. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8004165&isnumber=8004134

[WeTL13] H. Wen, P. Tiwary and T. Le-Ngoc, *Wireless Virtualisation*, Springer International Publishing, Heidelberg, Germany, 2013

[WIKI20] Benfica (Lisboa) in Wikipedia, https://pt.wikipedia.org/wiki/Benfica_(Lisboa), Jun, 2020

[YLJZ14] M. Yang, Y. Li, D. Jin, L. Zeng, X. Wu and A. Vasilakos, "*Software-Defined and Virtualised Future Mobile and Wireless Networks: A Survey*", Mobile Networks and Applications, Vol. 19, No. 1, Sep. 2014, pp. 1-15.