TÉCNICO LISBOA

Implementation of VoLTE considering different QoS and QoE overall approaches

João Filipe Monteiro Rodrigues Cardoso

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. Henrique Manuel Oliveira Ribeiro

November 2017

To my family and friends

Acknowledgements

I would like to thank my thesis supervisor, Professor Luis M. Correia, for the opportunity to write this Master Thesis under his precious guidance. All the time spent and shared knowledge through weekly meetings certainly influenced my work culture and helped to shape my future professional life. I want also to thank him for allowing me to do this thesis in collaboration with one of the major network operators in Portugal and the chance to be part of the GROW research group. This allowed me to stay in touch with several cutting-edge subjects in the world of telecommunications.

I also thank the whole GROW team for the support, and address special acknowledgements to my Master Thesis colleagues Cristina Dias, Daniel Almeida, Diogo Martins, Hugo Martins, Miguel Ramos and Sónia Pedrinho for the shared experiences, which contributed to a much greater work experience and environment.

My acknowledgments extend also to Eng. Luis Santo, Eng. Henrique Ribeiro and Eng. João Pires, from *NOS*, who provided valuable insights and suggestions from a network operator's point of view, and contributed to a better understanding of several technical aspects that helped to improve the quality of this work.

To my long-time friends and colleagues, a special thanks for the support and encouragement throughout all these years at Instituto Superior Técnico. The constant companionship and good moments during this important period were crucial for the completion of this thesis.

Finally, but most importantly, I would like to thank to my family, my Father, Filipe Cardoso, for his wisdom and character, my mother, Ana Paula Cardoso, for her creativity and encouragement, my sister Ana Filipa Cardoso, for her confidence and loyalty, and my grandmother, Elvira Rodrigues, for all the love. I thank them for all the unconditional support, as without them any of this would not have been possible.

Abstract

The main goal of this thesis is to study the impact in other services of deploying VoLTE over an existing LTE network, while monitoring VoLTE call quality. The QoS performance of seven distinct services is analysed in terms of allocation delay, packet failure and throughput satisfaction. QoE for VoLTE users is assessed using the E-model. For that purpose, a single-cell model for the DL resource allocation in LTE was proposed and implemented on a time-based simulator. Several parameters were analysed, namely number of users, type of environment, bandwidth and other service-related parameters. As VoLTE is the highest priority service, call quality is barely affected for a realistic number of users. For the reference urban scenario being considered, one concludes that up to 60 and 36 simultaneous VoLTE users are supported for the AMR-WB and EVS codecs, respectively, without significant impact on other services. Regarding system bandwidth, 10 and 20 MHz were compared, being observed that the offered throughput scales linearly with system bandwidth. The cell throughput for the urban scenario is 38% above the rural one, and no significant difference is observed between urban and suburban ones. Finally, service penetration influence was analysed by defining a VoLTE centric scenario and a Video centric one. In the Video centric scenario, the percentage of satisfied users for the services with a priority lower than video streaming stay below 50%. For the VoLTE centric scenario, user satisfaction levels are above 90%.

Keywords

LTE, VoLTE, Resource Allocation, QoS, Delay, Satisfied Users.

Resumo

O principal objetivo desta tese consiste no estudo do impacto noutros serviços de implementar VoLTE numa rede LTE, monitorizando também a qualidade das chamadas VoLTE. O desempenho de sete serviços é analisado através de atraso, falhas de pacotes e satisfação do ritmo de transmissão. A QoE dos utilizadores de VoLTE é avaliada recorrendo ao *E-model*. Para tal, desenvolveu-se um modelo da atribuição de recursos em DL para uma célula LTE e implementou-se este num simulador temporal. Analisaram-se vários parâmetros, nomeadamente número de utilizadores, tipo de ambiente, largura de banda e outros parâmetros característicos dos serviços. Sendo o VoLTE o serviço de prioridade mais alta, a qualidade da chamada não é afetada para números realistas de utilizadores. Para o cenário urbano de referência considerado, conclui-se que é possível suportar até 60 e 36 utilizadores de VoLTE em simultâneo para os codecs AMR-WB e EVS, respetivamente, sem um impacto significativo nos restantes serviços. Comparando as larguras de banda de 10 e 20 MHz, verificou-se que o ritmo binário atingido varia linearmente com estas. O ritmo binário da célula para o cenário urbano é 38% superior ao registado num cenário rural e não existem diferenças significativas entre os cenários urbano e suburbano. Finalmente, analisou-se a penetração dos serviços, definindo-se dois cenários distintos. Num cenário centrado em Vídeo, a percentagem de utilizadores satisfeitos para os serviços de prioridade inferior à do *streaming* de vídeo permanece abaixo de 50%. Para um cenário centrado em VoLTE, os níveis de satisfação dos utilizadores permanecem acima de 90%.

Palavras-chave

LTE, VoLTE, Atribuição de Recursos, QoS, Atraso, Utilizadores Satisfeitos.

Table of Contents

List of Figures

List of Tables

List of Acronyms

1G	1st Generation of Mobile Communications Systems
2G	2nd Generation of Mobile Communications Systems
3G	3rd Generation of Mobile Communications Systems
3GPP	3rd Generation Partnership Project
4G	4th Generation of Mobile Communications Systems
AMBR	Aggregate Maximum Bit Rate
AMR	Adaptive Multi Rate
AMR-WB	AMR-wideband
AN	Access Network
ANACOM	*Autoridade Nacional de Comunicações*
APN	Access Point Names
ARP	Allocation and Retention Priority
AS	Application Servers
CP	Cyclic Prefix
CSCF	Call State Control Function
CSFB	Circuit-switched Fall Back
CQA	Channel and QoS Aware
DL	Downlink
DRX	Discontinuous Reception
DTX	Discontinuous Transmission
eNodeB	evolved Node B
EPC	Evolved Packet Core Network
EPS	Evolved Packet System
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
EVS	Enhanced Voice Services
FDD	Frequency Division Duplex
FIFO	First-In-First-Out
FTP	File Transfer Protocol
GBAR	Gamma Beta Auto-Regressive
GBR	Guaranteed Bit Rate
GSA	Global mobile Suppliers Association
GSM	Global System for Mobile Communications
GSMA	GSM Association
GSM EFR	GSM Enhanced Full Rate

HD	High Definition
HOL	Head of Line
HSS	Home Subscription Server
HTTP	Hyper Text Transfer Protocol
HTTPS	Hyper Text Transfer Protocol Secure
ICS	IMS Centralised Service
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPv4	Internet Protocol version 4
ISI	Inter-Symbol Interference
ITU-T	International Telecommunication Union-Telecommunication
KPI	Key Performance Indicator
LTE	Long Term Evolution
MBR	Maximum Bit Rate
MIMO	Multiple-Input Multiple-Output
MM	Mobility Management
MME	Mobility Management Entity
MOS	Mean Opinion Score
MRF	Multimedia Resource Function
MSC	Mobile Switching Centre
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PAPR	Peak-to-Average Power Ratio
PBCH	Physical Broadcast Channel
PCEF	Policy Control Enforcement Function
PCRF	Policy and Charging Resource Function
PDCCH	Physical Downlink Control Channel
PDN	Packet Data Network
PDSCH	Physical Downlink Shared Channel
PESQ	Perceptual Evaluation of Speech Quality
P-GW	Packet Data Network Gateway
POLQA	Perceptual Objective Listening Quality Analysis
PRACH	Physical Random Access Channel
PSS	Priority Set Scheduler
PSTN	Public Switched Telephone Network
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QoE	Quality of Experience

QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAM	Random Access Memory
RB	Resource Block
RE	Resource Element
RLC	Radio Link Control
RoHC	Robust Header Compression
RRM	Radio Resource Management
RTP	Real-time Transport Protocol
SC-FDMA	Single Carrier Frequency Division Multiple Access
S-GW	Serving Gateway
SID	Silence Insertion Descriptor
SIP	Session Initiation Protocol
SLF	Subscription Locator Function
SNR	Signal-to-Noise Ratio
SR-VCC	Single Radio Voice Call Continuity
TAS	Telephony Application Server
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
V2X	Vehicle-to-everything
VAD	Voice Activity Detection
VAF	Voice Activity Factor
VBR	Variable Bit Rate
ViLTE	Video over LTE
VoIP	Voice over IP
VoLGA	Voice over LTE via Generic Access
VoLTE	Voice over LTE
VoWi-Fi	Voice over Wi-Fi

List of Symbols

α	Probability of transition from the silent state to the talking state
β	Probability of transition from the talking state to the silent state
$\Gamma_{failure}$	Packet failure ratio
Δ	Speed deviation
λ	Propagation wavelength
μ	Average value
ρ_N	SNR
σ	Standard deviation
τ_{dec}	Voice decoding delay
τ_{enc}	Voice encoding delay
τ_{HOL}	HOL delay
$\overline{\tau_{HOL,k}}$	Average HOL delay of the packets received by user *k*
$\overline{\overline{\tau_{HOL,s}}}$	Average HOL delay of users performing service *s*
τ_k	End-to-end delay experienced by user *k*
τ_{proc}	Processing delay in the UE and in eNodeB
τ_{trans}	Transport delay
τ_{UL}	Radio interface delay in the UL
A_n	Auto-correlation function
B_n	Marginal distribution
B_{pl}	Robustness against packet loss
D	Reading time
D_d	Parsing time between two consecutive packets
D_{pc}	Reading time between two consecutive packet calls
$I_{d,wb}$	Delay impairment factor
$I_{e,eff,wb}$	Equipment impairment factor
$I_{e,wb}$	Equipment impairment factor without any packet loss
L_{cell}	Cell load
$L_{p\ ind}$	Indoor attenuation
N_D	Number of embedded objects
N_{pc}	Number of packet calls

$\widetilde{N}_{RB}$	Estimate of the number of RBs needed to accommodate traffic in queue
$N_{RB,k}$	Number of RBs allocated to user *k*
$N_{RB,max}$	Maximum number of RBs per TTI
$N_{RB,reduce}$	Sum of all the RBs that must be reduced
$N_{RB,usersub}$	Number of RBs that must be reduced for each user
$N_{received}$	Number of packets received during the time interval under analysis
N_{sent}	Number of packets sent during the time interval under analysis
N_u	Total number of active users in the cell
$N_{u,s}$	Number of users performing service *s*
N_Z	Number of samples
p_{gen}	Signal that a packet was generated during the current frame
P_{max}	Higher Priority level
P_S	Packet size
Q	Total number of bits in queue
R	Transmission rating factor
$R_{b,eNodeB}$	Total cell throughput
$\overline{R_{b,k}}$	Average throughput of user *k*
$R_{b,RB}$	Achievable RB throughput
R_x	Corrected transmission rating factor
S_E	Embedded object size
S_M	Main object size
$t_{activity}$	Instant where the current ON or OFF state started
$t_{arrival}$	Instant of arrival of a user to the network
T_c	Coherence time
T_{call}	Call/session duration
t_{frame}	Frame duration
t_{last}	Last instant where a packet was generated
$t_{last,SNR}$	Last instant where the SNR for a given user changed
T_{ON}	Active state duration
T_{OFF}	Silent state duration
T_p	Parsing time
t_{silent}	Instant where the silent state started
T_{sim}	Simulation time
t_{state}	Duration of the ongoing state
T_{TTI}	TTI in LTE
v	User speed

V_{av}	Average user speed
V_D	Object data volume
V_{max}	Maximum user speed
V_{min}	Minimum user speed
X_n	Gamma distributed stationary stochastic variable
z_i	Value of sample i

List of Software

Mathworks MATLAB R2013a	Numerical computing software
Microsoft Excel 2013	Calculation and graphical chart tool
Microsoft Word 2013	Text editor software

Chapter 1

Introduction

This chapter describes the context for the thesis inside the scope of the last generation of mobile communication systems, giving a high-level description of the evolution of voice services. The motivation and the thesis structure are also presented.

1.1 Overview

Mobile communications systems were designed from the very beginning, since the analogue First Generation (1G), with the goal of providing voice services. With the arrival of the Second Generation (2G) using digital transmission, a wider range of voice services became available to a large number of users, with Global System for Mobile Communications (GSM) being the most popular system in Europe, at the time, based on circuit-switched technology. At a later stage, data services were added, but voice remained as the main traffic source. The Third Generation (3G) emerged with the intent of providing high-speed packet-switched data transfer following the abrupt growth of popularity of the Internet in fixed networks, bringing Universal Mobile Telecommunications System (UMTS) as the evolution of GSM according to the specifications of the Third Generation Partnership Project (3GPP).

Long Term Evolution (LTE) appears as 3GPP's Fourth Generation (4G) solution, based on an entirely packet-switched oriented architecture, aimed at fulfilling the exponential demand for Internet Protocol (IP) data services. It was initially designed to provide a peak data rate of 100 Mbps in the downlink (DL) and 50 Mbps in the uplink (UL), while ensuring low latency to allow real-time applications, like voice and interactive gaming [HoTo11]. High spectral efficiency as well as high levels of mobility, security and terminal power efficiency constituted also the main targets for LTE. In this context, the need to move voice services from circuit-switched to an entirely packet-switched architecture raised.

In 2000, 3GPP standardised the IP Multimedia Subsystem (IMS), a framework aimed at providing multimedia services like voice, over 3GPP systems. However, implementation aspects like session setup, authentication or bearer setup were left for service providers and vendors to decide upon. In November 2009, the One Voice initiative was established by 12 telecommunications companies, from operators to manufacturers, proposing a standard solution for an IMS-based Voice over LTE (VoLTE) service [YiCh12]. In March 2010, the GSM Association (GSMA) supported by more than 40 companies all around the globe published an improvement of the One Voice profile, specifying a set of requirements to provide Voice over IP (VoIP) calls over LTE in their official document *IR.92 –IMS Profile for Voice and SMS* [GSMA16a].

The inexistence of a clear standardised solution for VoLTE at the time of the initial deployments of LTE networks in 2009 led to the development of temporary alternatives to provide voice services, namely Circuit-switched Fall Back (CSFB) and Voice over LTE via Generic Access (VoLGA) as illustrated in Figure 1.1. CSFB appears as the most popular transitional solution, allowing operators to use their GSM/UMTS network by moving voice calls originated on a LTE terminal to a GSM or an UMTS cell where the need for changes to support this solution on the existing networks is minimal. VoLGA, which in this case is not standardised by 3GPP, is yet another solution that allows a terminal to reach the GSM/UMTS core network through a generic access network, like for example a wireless local area network. This solution was not significantly supported by the mobile communications sector.

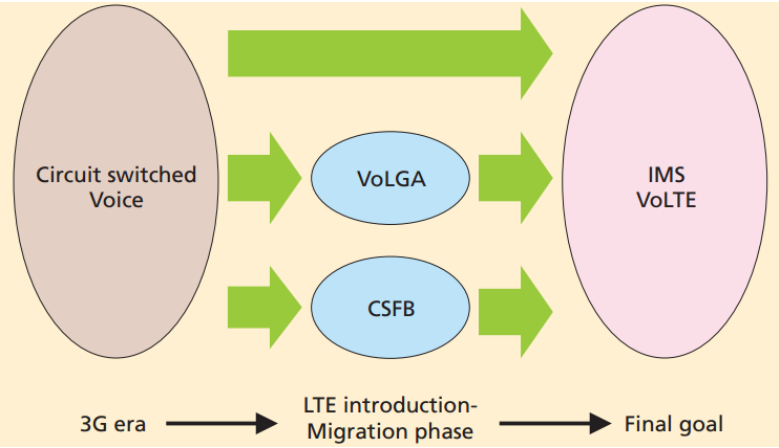


Figure 1.1. Voice service migration (extracted from [TaKo11]).

According to [Noki17], operators have three main strategies to deploy voice services for LTE users:

- If LTE coverage is good, VoLTE and multimedia can be implemented using IMS with the goal of having an all-IP architecture. 3G coverage and capacity are insufficient to provide the required QoS.
- CSFB is implemented for voice, while the remaining multimedia services are offered through IMS. This approach requires good 3G coverage to support voice subscribers and advanced services simultaneously, while using LTE to provide extra data capacity when needed.
- CSFB is implemented in a long-term perspective and the evolution to VoLTE is dictated by market demand. LTE is destined only for data, and coverage is usually limited.

VoLTE constitutes an operator and user-friendly solution compared to third-party VoIP, which poses several disadvantages. Traffic generated by applications like, for example, Skype does not differ from any other IP-based application, turning service performance dependent of what the Internet can provide. This of course has an impact in the Quality of Experience (QoE) perceived by the end user, as the network is not able to ensure a minimum guaranteed bit rate, as well as a maximum value of end-to-end delay. On the other hand, VoLTE using the IMS allows Quality of Service (QoS) control as the User Equipment (UE) is able to specify multiple performance requirements. With this functionality, the network can prioritise voice packets over data ones, which are not time critical.

The first successful deployments of VoLTE occurred in 2012 with *SK Telecom* and *LG U+* in South Korea and *MetroPCS* in the USA. According to a Global mobile Suppliers Association (GSA) report dated from August 2017 [GSA17], 179 operators are currently investing in VoLTE in 81 countries around the world, from which 121 operators have commercially launched VoLTE in 60 countries. In Portugal, *Vodafone* introduced VoLTE for the first time in 2015 and it is still up to this date the only operator providing the service. According the same report from GSA, there are currently 1 313 devices that support VoLTE, of which 92.3% are mobile phones. Over 30% of these devices are manufactured by Samsung and LG, as part of the 88 different vendors offering VoLTE devices all around the world. As depicted in Figure 1.2, it is expected that VoLTE has a significant market take-off in the upcoming years in terms of subscriptions all around the globe.

It is important to mention that several complementary technologies are being developed along with VoLTE. Voice over Wi-Fi (VoWi-Fi) aims also at providing VoIP calls using IMS technology, but over a Wi-Fi access network, allowing a seamless handover of calls between LTE and Wi-Fi. This service has a strong potential from the network operator's point of view, as this system provides a solution for indoor coverage without the need of installing additional base stations and taking advantage of the high amount of Wi-Fi routers currently installed. GSMA specified a set of requirements in their official document *IR.51 – IMS Profile for Voice, Video and SMS over Wi-Fi* [GSMA16b].

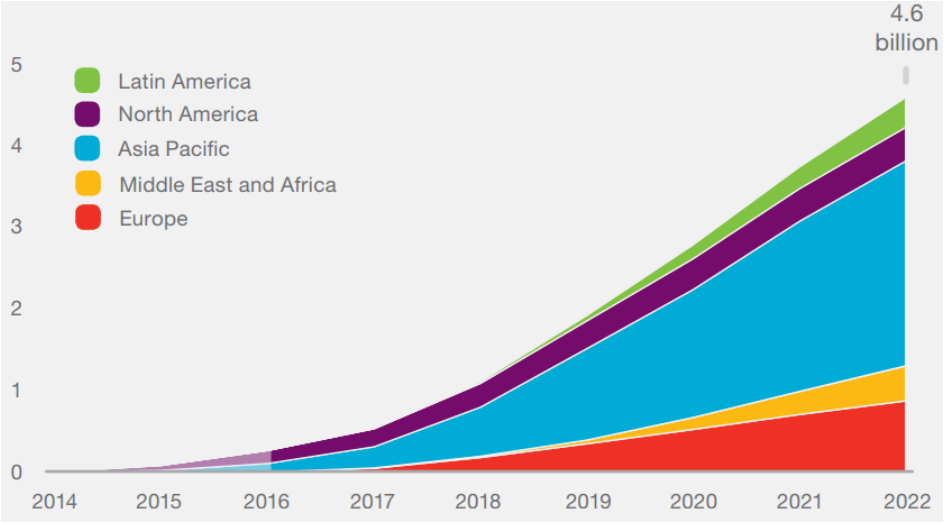


Figure 1.2. Prediction of VoLTE subscriptions by region, in billions (extracted from [Eric17]).

To understand the current trends for these technologies, Figure 1.3 shows the expected volume of minutes of use for the multiple alternatives of delivery of VoIP calls in the upcoming years. It is expected that in the short-term both VoLTE and VoWi-Fi surpass third-party VoIP as the main solutions to provide packet-switched voice calls. Moreover, it is also expected that in this period VoWi-Fi will become the main service, mostly due to the fact that some operators are making it available before VoLTE as a proof of concept for the use of an IMS-based solution to deliver calls. However, there is a general trend of growth for all these technologies, which are slowly replacing the older circuit-switched technologies.

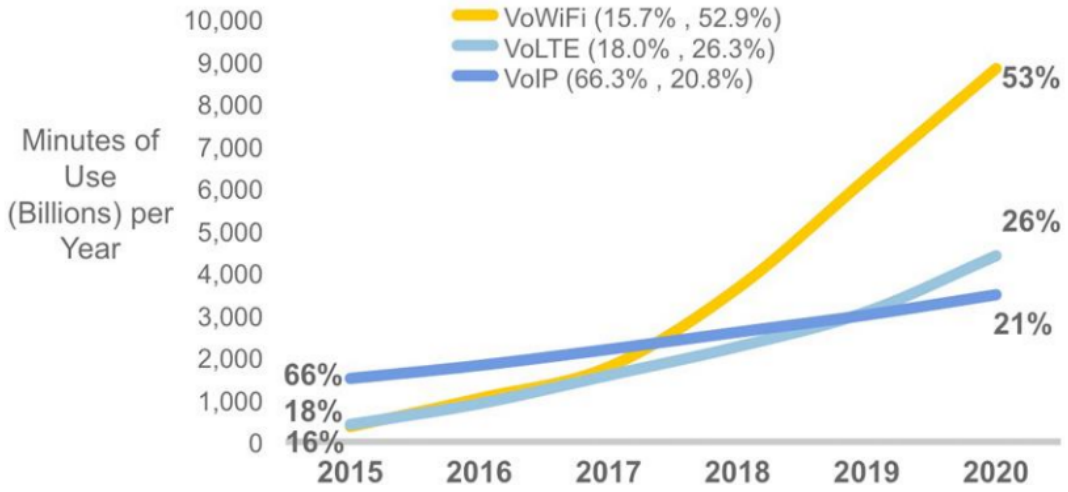


Figure 1.3. Mobile voice minutes of use – VoWiFi, VoLTE and VoIP (extracted from [Cisc16]).

VoLTE should be considered not only as just a migration from the traditional voice services to an all-IP technology but as way to integrate several services taking advantage of the IMS capabilities. These services can be: High Definition (HD) voice, video communications, IP messaging, content sharing between calls, among others [Eric17]. With an increasing demand for video services, Video over LTE (ViLTE) is an example of an extension of VoLTE designed to enhance the voice service. It is aimed at providing a high-quality video channel and registered a growing interest from the industry in recent years with 10 operators around the world that have already launched or are deploying the service according to data from August 2017 [GSA17].

1.2 Motivation and Contents

Operators face several challenges in the time of replacing the current voice service delivered through the circuit-switched technologies of systems prior to LTE as they have proven to be successful in the past. However, this replacement is somehow inevitable as the demand for data services is growing at a high pace and, from a long-term perspective, merging all services to packet-switched technologies will result in a more cost-effective solution. At least, the implementation of VoLTE is supposed to deliver the same levels of performance compared to the existing technologies. The same is not guaranteed in third-party VoIP for example, where the lack of priority for voice packets in the network makes it only a Best Effort service.

This thesis focuses on evaluating the impact in other services of deploying VoLTE over an existing LTE network, and monitoring the performance of VoLTE calls. One of the most relevant aspects that determines the performance of voice calls is the end-to-end delay, which refers to the difference between the instant of speech and the instant of listening between two users. The estimation of end-to-end delay, and also the packet failure rate, allows the evaluation of the speech quality through the mean opinion score (MOS), which is an indicator of the call quality perceived by the end user. The method to evaluate these parameters is through a model of DL resource allocation in LTE considering a set of seven different services including VoLTE.

The developed model aims at the estimation of performance metrics that deal with the overall capacity of the network and the degree of satisfaction of the served users performing multiple and distinct services. This model takes several input variables into account that are related to cell environment, cell bandwidths and parameters related to users' behaviour, like users' arrival rate to the cell and the usage pattern of the considered set of services. The main output of the thesis consists of the software implementation of these models over MATLAB numerical computing software. While monitoring the performance of VoLTE is not a new topic, there is interest especially from network operators in having a systematic approach that enables them to parametrise networks in an efficient way and to understand the impact of deploying the service over the existing LTE networks under different conditions.

This study is the result of a collaboration with the Portuguese operator *NOS* motivated by the industry

demand to deploy this technology over existing LTE networks. Assistance was provided through industry guidelines for the implementation of VoLTE, as well as useful feedback on several technical details that heavily determined the studied scenarios in this thesis.

This thesis is composed of 5 chapters and 3 annexes. The current chapter introduces the context and motivations to study the implementation of VoLTE over existing LTE networks, presenting an overview of the current maturity of the technology around the world and specifically in Portugal. The structure of the thesis is also described.

Chapter 2 starts by introducing the main aspects that characterise LTE, namely its network architecture and the correspondent radio interface. Then, an overview of the functionalities of VoLTE is provided, specifying its advantages in comparison with a regular VoIP service, and how that poses several additional challenges in terms of network implementation and performance measurement. The chapter finishes with the state of the art, where the work developed by other authors related to this subject is presented.

An overview and a detailed description of the developed single-cell model for resource allocation in LTE are detailed in Chapter 3. One presents also the mathematical formulation for the metrics considered to evaluate the performance of several services, including VoLTE, over an LTE network. The implementation on a simulator is described in terms of its modular structure and the generated output information. Finally, the assessment and statistical validation of the simulator is presented.

Chapter 4 starts by presenting the description of the reference scenario, the configuration of all the parameters of the simulator and the simulation strategy being considered in this thesis. The remaining sections refer to the analysis of results, which includes the assessment of the VoLTE call quality and a performance analysis of the remaining services.

Chapter 5 provides the main thesis conclusions, by highlighting the most relevant results and providing an overview of the main aspects that were addressed. This chapter closes by pointing out guidelines for further improvement of this work and additional approaches for research regarding this subject.

The annexes provide additional information on the models used in this work. Annex A deals with the mathematical formulation used to relate the experienced SNR and the achievable throughput in LTE. Annex B presents the traffic source models used to generate traffic for the seven services under study. In the end, Annex C describes the mobility model which consists of a random generator for user speeds.

Chapter 2

Fundamental aspects

This chapter provides an overview of the fundamental aspects of LTE, namely its network architecture and radio interface. A description of the implementation of VoLTE, by considering the specifications of a generic VoIP network service is made. A characterisation of the typical services and applications is provided, and models to evaluate system performance are addressed. Finally, the State of the Art is presented, concerning the work carried through by several authors on this subject.

2.1 LTE

This section includes a description of the network architecture and the radio interface that characterise the LTE system based on [HoTo11], [SeTB11], [PoHo12] and [Cox14].

2.1.1 Network Architecture

In contrast with previous systems that rely on a circuit-switched mode of operation, LTE is optimised for packet-switched services which, associated with a need of improving performance and reducing latencies, led to the development of its flat architecture. This designation resembles the fact that the Access Network (AN) of LTE, named Evolved UMTS Terrestrial Radio Access Network (E-UTRAN), is entirely located at the base station level without the need for a centralised controller.

Figure 2.1 shows the most basic architecture, which only includes E-UTRAN and that supports other architectures with additional functions. This architecture integrates four main high-level domains: UE, E-UTRAN, Evolved Packet Core Network (EPC), and Services. It also shows the main nodes that compose each of these domains and the interfaces responsible for the inter-connection between them.

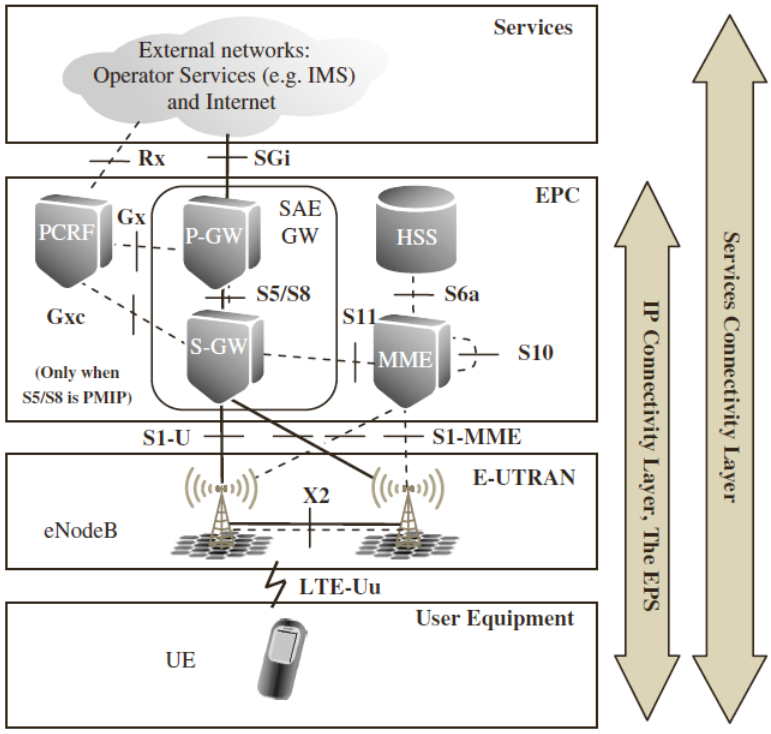


Figure 2.1. E-UTRAN system architecture (extracted from [HoTo11]).

The UE represents the device used by the end user to perform communications. Each UE connects with E-UTRAN through the radio interface, described in detail in Section 2.1.2. E-UTRAN is composed of an inter-connected network of nodes through the X2 interface, each of them named evolved Node B

(eNodeB). An eNodeB is basically a radio base stations that is responsible for controlling all the radio related functions of the network, including Radio Resource Management (RRM) ones. They are also responsible for Mobility Management (MM) to take decisions for handover UEs between cells and monitor resource usage, as well as Control Plane functions. The eNodeB works as a layer bridge between the UE and the EPC, by allowing data transmission between the radio connection from the UE and the corresponding IP based connectivity from the EPC. As shown in Figure 2.1, the EPC is composed of five nodes with the following elements:

- Serving Gateway (S-GW). It serves as a mobility anchor for handover situations as it receives all user IP packets. S-GW also performs administrative functions, like charging the volume of data exchanged and lawful interception.
- Packet Data Network Gateway (P-GW). It is responsible for IP address allocation for the UE, as well as for traffic gating and filtering functions according to the QoS requirements of each service. This function allows interconnection between the EPC and external IP networks, like for example IMS. Each of these networks is commonly called a Packet Data Network (PDN) and is identified by an Access point name (APN).
- Mobility Management Entity (MME). It operates in the Control Plane, constituting the main control element. It interacts with most of the nodes of the architecture, namely with a direct control connection to the UE, and it includes the functions of authentication and security, mobility management and subscription management.
- Home Subscription Server (HSS). It includes all the permanent System Architecture Evolution (SAE) subscription data of the users.
- Policy and Charging Resource Function (PCRF). It is responsible for policy control and for controlling how data flow is regulated by the Policy Control Enforcement Function (PCEF), located in the P-GW, to be in accordance with the profile of each user.

Together, UE, E-UTRAN and EPC form the Evolved Packet System (EPS), which designates the layer that provides IP connectivity, allowing services to be entirely offered through IP. EPS uses bearers, which are basically IP packet flows with a determined QoS, to route IP traffic between the P-GW and the UE. Bearers are set up and released by E-UTRAN and EPC, depending on each application's requirements. The MME automatically defines a default bearer to provide IP connectivity to the UE, while dedicated bearers can be set up to provide multiple simultaneous services with specific QoS requirements. This subject and its implications for VoLTE are discussed in more detail in Section 2.2.2.

On top of EPS there is the Services layer, which constitutes the top level domain of the network being divided in three different subsystems related to IMS: IMS based operator services, Non-IMS based operator services, and other services not provided by the operator, such as those using the Internet, for example.

The aggregation of these four architectural layers forms the Services Connectivity Layer. IMS works as an example of a system that works on top of it, being fundamental to provide a service like VoIP through IP connectivity. As implementing VoLTE requires the deployment of IMS, Figure 2.2 shows the high-level IMS architecture and its interfaces, as well as how it connects with the LTE network architecture.

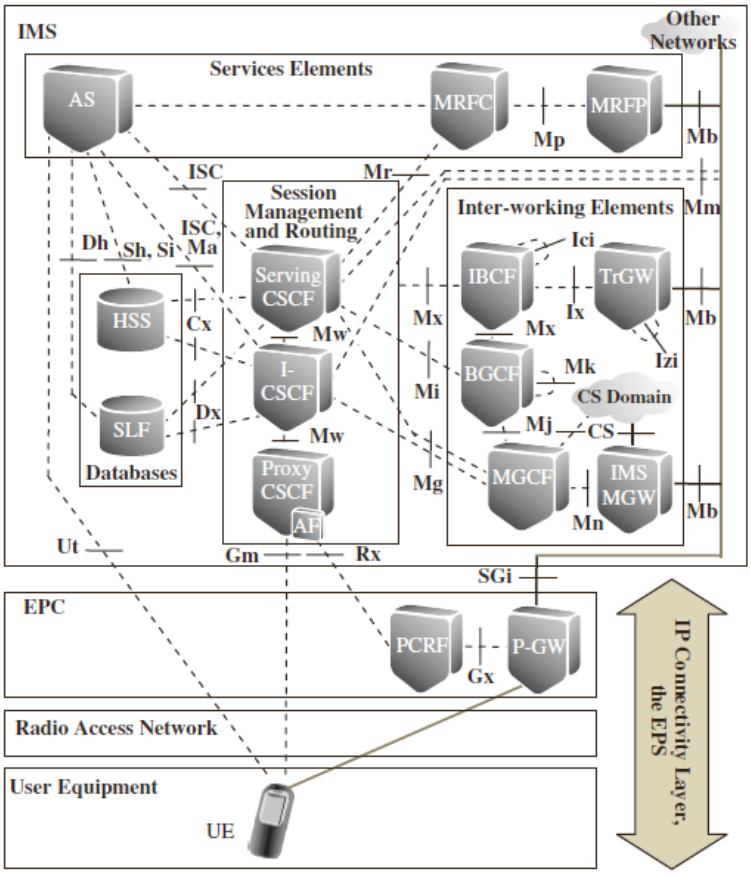


Figure 2.2. IMS architecture (extracted from [HoTo11]).

The IMS architecture is composed of four main entities:

- Session Management and Routing. It is based on the Call State Control Function (CSCF), which deals with the UEs registration to the IMS and the management of the service session, using Session Initiation Protocol (SIP) signalling.
- Inter-working Elements. These are responsible for the interoperation with other IMS or Circuit-switch systems.
- Services Elements. They contain the Application Servers (AS) which contain the service logic to provide different services. The Telephony Application Server (TAS) is an example of a standardised AS used to provide VoIP services. Media is handled by the Multimedia Resource Function (MRF).
- Databases. IMS uses HSS as the main database, while the Subscription Locator Function (SLF) can be used when there is more than one addressable HSS.

2.1.2 Radio Interface

Communication between users and eNodeBs is carried through radio transmission and reception using a specific radio interface. LTE supports both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) for any of its multiple access techniques. LTE uses similar radio frame structures for both FDD and TDD, the main difference being the fact that in TDD there is a need for additional fields and a guard

period to control the switching between UL and DL. 3GPP specifies the radio frame structures Type 1 and Type 2, applicable to FDD and TDD, respectively [3GPP16b]. This thesis focuses on the Type 1 FDD frame structure, as it is the most common duplex system currently used by network operators.

According to the current 3GPP specifications, there are 32 frequency bands available for FDD and 14 for TDD [3GPP16a]. Portugal follows the European spectrum regulation, where the main bands used for LTE are 800, 900, 1 800 and 2 100 MHz. The Portuguese communications regulator, *Autoridade Nacional de Comunicações* (ANACOM), auctioned the rights of use of the 800, 1 800 and 2 600 MHz bands in the end of 2011 to the three Portuguese operators *NOS*, *MEO* and *Vodafone*. Table 2.1 shows the auction results, including the current LTE FDD bands used for UL and DL with the total price paid for each one and where the 800 MHz band sorts out as the most appealing for operators.

Table 2.1. Frequency bands for LTE in Portugal (extracted from [ANAC12a] and [ANAC12b]).

Operator	**Frequency band [MHz]**	**Total Bandwidth [MHz]**	**DL band [MHz]**	**UL band [MHz]**	**Final price [M€]**
NOS	800	2 × 10	811 – 821	852 – 862	90
	1 800	2 × 14	1 825 – 1 845	1 730 – 1 750	11
	2 600	2 × 20	2 530 – 2 550	2 650 – 2 670	12
MEO	800	2 × 10	791 – 801	832 – 842	90
	1 800	2 × 14	1 845 – 1 865	1 750 – 1 770	11
	2 600	2 × 20	2 550 – 2 570	2 670 – 2 690	12
Vodafone	800	2 × 10	801 – 811	842 – 852	90
	1 800	2 × 14	1 805 – 1 825	1 710 – 1 730	11
	2 600	2 × 20	2 510 – 2 530	2 630 – 2 650	12

Regarding multiple access techniques, LTE uses Orthogonal Frequency Division Multiple Access (OFDMA) for DL and Single Carrier Frequency Division Multiple Access (SC-FDMA) for UL. OFDMA is based on Orthogonal Frequency Division Multiplexing (OFDM), which ensures multicarrier transmission where the band is partitioned into a set of sub-carriers that are overlapped but are orthogonal with each other, in the frequency domain. This means that each of the centre frequencies for the sub-carriers is selected such that the neighbouring sub-carriers have a zero value at the sampling instant of the determined sub-carrier. For LTE, these sub-carriers have a constant frequency difference of 15 kHz between each other, eliminating the need of having guard-bands to separate carriers and making OFDM a highly spectrally efficient solution suited for high-rate mobile data transmission. OFDMA is the extension of this method by assigning subcarriers to different users at the same time such that they receive data simultaneously. OFDMA is only used for DL, because of its low power conversion efficiency, associated with the high Peak-to-Average Power Ratio (PAPR) of the OFDM signal.

SC-FDMA is used for UL to improve the battery duration of the mobile terminals, as it is more power efficient. The main difference comparing to OFDMA is that in SC-FDMA, for the same time instant, each sub-carrier is modulated with a signal that is a linear combination of all the symbols transmitted at that instant, providing its single-carrier characteristics. It uses the same structure that OFDMA, with the same

15 kHz sub-carrier spacing. Figure 2.3 compares both schemes for the transmission of four data symbols, with a Quadrature Phase Shift Keying (QPSK) modulation.

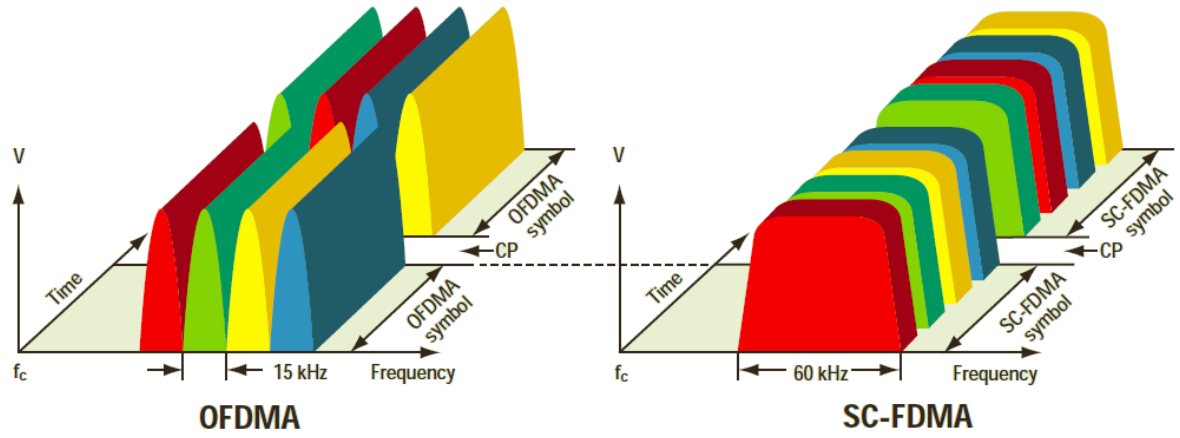


Figure 2.3. Symbol transmission in OFDMA and SC-FDMA (adapted from [MRum08]).

The resource allocation of LTE is based on the concept of Resource Blocks (RBs), which consist of a representation of information as a function of frequency and time simultaneously, defining a resource grid. This allows a signalling resolution, which if it was based on an individual sub-carrier, could not support the allocation of different sub-carriers to multiple users. As shown in Figure 2.4, each RB is composed of 12 sub-carriers corresponding to a bandwidth of 180 kHz. In the time domain, each slot has a 0.5 ms duration being composed of 7 symbols and using a normal Cyclic Prefix (CP) to avoid inter-symbol interference.

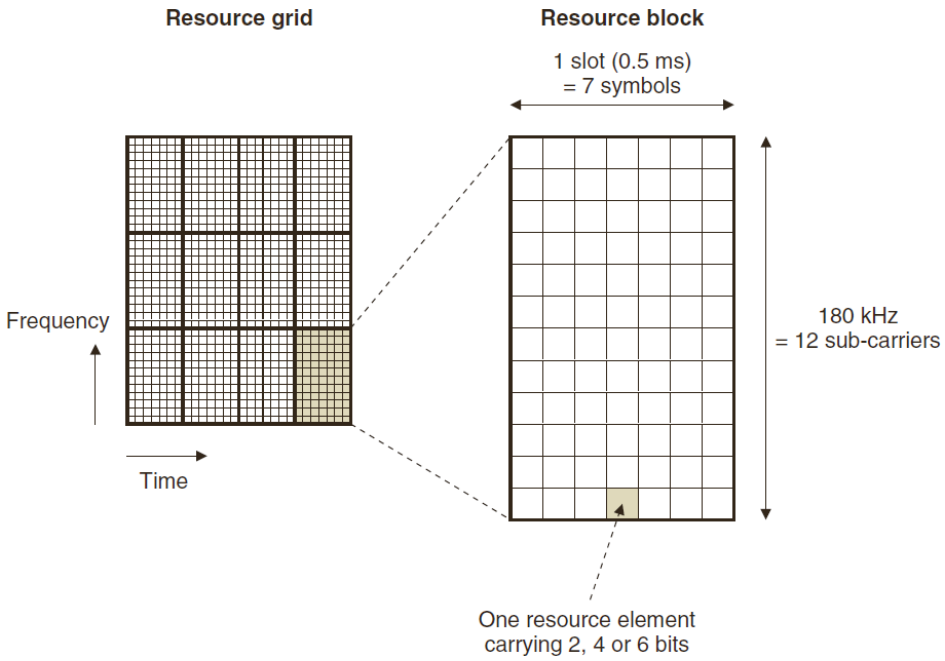


Figure 2.4. LTE resource grid in time and frequency domains (extracted from [Cox14]).

In this resource grid, the elementary unit is the Resource Element (RE), which associates one symbol with one sub-carrier, Figure 2.4. LTE supports QPSK and Quadrature Amplitude Modulation (QAM) as modulation schemes. The adopted scheme defines if each RE carries 2, 4 or 6 bits for QPSK, 16-QAM

or 64-QAM respectively.

Figure 2.5 illustrates DL resource allocation in LTE. In the time domain, the basic unit for user scheduling is the sub-frame, with a 1 ms duration, which corresponds to two slots being frequently named the Transmission Time Interval (TTI). Ten of these sub-frames define a complete frame with a duration of 10 ms. With this kind of allocation, operators can configure cells to operate with different radio channel bandwidths, providing different numbers of RBs and sub-carriers, Table 2.2. This configuration is made in a way such that there is a remaining bandwidth to use as two guard bands, commonly with the same width.

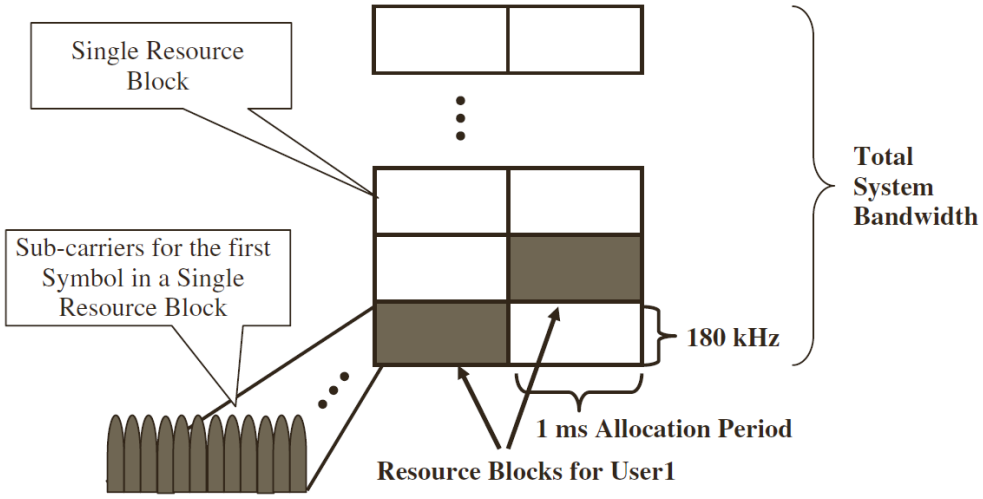


Figure 2.5. OFDMA resource allocation in LTE (extracted from [HoTo11]).

Table 2.2. Cell bandwidths supported by LTE (extracted from [Cox14]).

Total bandwidth [MHz]	Number of RBs	Number of sub-carriers	Occupied bandwidth [MHz]	Usual guard bands [MHz]
1.4	6	72	1.08	2 × 0.16
3.0	15	180	2.70	2 × 0.15
5.0	25	300	4.50	2 × 0.25
10.0	50	600	9.00	2 × 0.50
15.0	75	900	13.50	2 × 0.75
20.0	100	1 200	18.00	2 × 1.00

LTE uses several physical channels to transport system information from the radio channels. It is at this layer that dynamic resource allocation is made, by controlling the data rates and the resource division for the multiple users. For UL, Physical Uplink Shared Channel (PUSCH) carries user data, Physical Random Access Channel (PRACH) enables transmission to the UE, and Physical Uplink Control Channel (PUCCH) is used for control purposes. For DL, the main channels are: Physical Downlink

Shared Channel (PDSCH), used to enable transmission to the UE and carry user data; Physical Broadcast Channel (PBCH), which carries system parameters and Physical Downlink Control Channel (PDCCH), among others, used for control.

LTE can additionally use Multiple-Input Multiple-Output (MIMO) to increase the peak data rates by using spatial multiplexing along with pre-coding and transmit diversity. This technique consists of sending different signals from two or more antennas that can be separated in the receiver using appropriate signal processing. This allows increase factors of 2 or 4 on the peak data rate, depending if a 2-by-2 or 4-by-4 antenna configuration is used, respectively, among any other possibilities.

2.2 Voice service on LTE

In this section, a description of the basic aspects of a VoIP network service is presented, namely its main implementation constraints and protocols used for communication. A description of the main aspects of the implementation of VoLTE are provided, as it poses several additional constraints in comparison with regular VoIP. The main methods used to measure its performance are also addressed as well as considerations about the implementation of VoLTE.

2.2.1 VoIP

This section describes the basic aspects of a VoIP network service, based on [AhIl08] and [DPBK06]. The main challenge for the implementation of VoIP arises from the fact that it is a real-time application, distinct from the nature of most of the networks, which are best-effort ones. This means that data packets have to be delivered on time, with special constraints for the case of mobile communications systems that are based on wireless channels. Transmission through wireless channels is susceptible to a frequent occurrence of bit errors, resulting in a need to retransmit packets. Additionally, voice continuity has to be ensured when users move along different cells, as a transmission gap is unacceptable for voice.

Thus, delay is a critical aspect, being one of the main concerns in providing a real-time application. It is originated by multiple factors:

- Propagation. It is caused by the physical propagation of signals, which is not instantaneous.
- Processing delay. It is associated with the delay introduced by encoding, compression and packetisation.
- Serialisation delay. It is related to the time spent at the receiver to actually playback the received data and it is usually not taken into account in comparison with other sources of delay.

A maximum end-to-end delay of 200 ms is considered to be acceptable for a voice call, by comparison to a regular voice call using the Public Switched Telephone Network (PSTN). Jitter also needs to be controlled, as packets do not arrive at the same rate as they were transmitted due to the delay introduced

by the network. The control of delay and jitter has to be ensured, while guaranteeing that spectral efficiency is not compromised. Table 2.3 shows a list of common International Telecommunication Union-Telecommunication (ITU-T) standard codecs, including the required rates and one-way delay.

Table 2.3. Common ITU-T Codecs and their defaults (extracted from [AhIl08]).

Codec	Data Rate [kbps]	Datagram Size [ms]	A/D Conversion Delay [ms]	Combined Bandwidth (Bidirectional) [kbps]
G.711u	64.0	20	1.0	180.8
G.711a	64.0		1.0	180.8
G.729	8.0		25.0	68.8
G.723.1 (MPMLQ)	6.3	30	67.5	47.8
G.723.1 (ACELP)	5.3		67.5	45.8

Adaptive Multi Rate (AMR) is the speech codec adopted by 3GPP, which allows multiple source rates by adapting to the radio channel conditions, from 4.75 to 12.2 kbps, and includes Voice Activity Detection (VAD), a comfort noise generation system and an error concealment mechanism. AMR-wideband (AMR-WB) is an extension of this codec, with rates from 6.60 to 23.85 kbps. GSMA specifies AMR as the mandatory voice codec for VoLTE, stating that AMR-WB can be used optionally to provide better quality [GSMA16a]. Most of the operators deploying the service are opting for this codec in order to provide improved call quality. Moreover, the Enhanced Voice Services (EVS) codec has been standardised by 3GPP as the next-generation to provide high quality voice [3GPP17c].

Figure 2.6 shows how VoIP traffic is generated over time using the AMR codec, where voice frames are generated periodically every 20 ms. Discontinuous transmission (DTX) is used during silent periods where the VAD algorithm does not detect that there is voice to be transmitted by sending Silence Insertion Descriptor (SID) frames every 160 ms to provide information about the acoustic background noise.

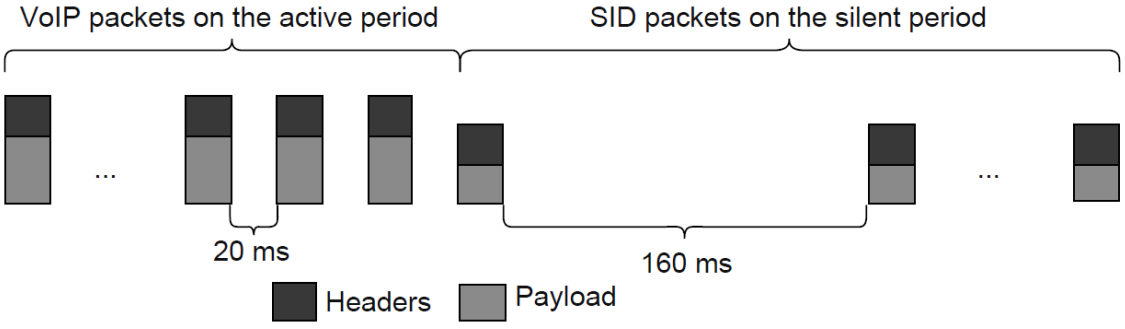


Figure 2.6. Illustration of VoIP traffic (extracted from [AhIl08]).

SIP is currently the main VoIP application-layer protocol, being the Internet Engineering Task Force (IETF) standard defined for multimedia conferencing over IP, used to establish, maintain, and terminate calls between multiple end points. It provides signalling functions, to carry call information across the

network, as well as session management functions to control the attributes of the call with an agreement between the UEs. After a session is established using SIP, the exchange of packets is guaranteed through the Real-time Transport Protocol (RTP). User Datagram Protocol (UDP) is the transport-layer protocol used to carry VoIP packets. As it is the case for real-time applications, delay is a priority over the reliability of the delivery of packets and, thus, UDP is preferable over Transmission Control Protocol (TCP) as there is no need to establish a connection. With UDP the data stream keeps being transmitted, not depending on the level of the existing packet losses.

As VoIP operates using these protocols, packets include an overhead from IP, UDP, and RTP headers, which leads to an excessive amount of bytes to be transmitted compared to the smaller size of the actual data. To avoid that, header overheads are compressed using Robust Header Compression (RoHC), achieving a compression from 46 bytes, for Internet Protocol version 4 (IPv4), down to 3 bytes in the best case.

The transmission of voice through packets has also an impact in the required power for transmission. As packets need the inclusion of extra overheads, more power is needed to send the same amount of information, resulting in a lower coverage. Moreover, the transmission of packets through bursts leads to a higher PAPR, also requiring more transmitted power.

2.2.2 Functionality and performance of VoLTE

This section provides the underlying aspects of the implementation of VoLTE, as well as how its performance can be measured. It is based on [Cox14], [HoTo11], [PoHo12] and [Ahll08].

VoLTE generically stands for the delivery of VoIP calls over LTE, using IMS to provide IP connectivity and acts as the service control architecture, operating as an overlay layer on top of EPS. IMS uses the SIP protocol for registration and control of the service sessions, not being only used between the terminal and the IMS but also among multiple IMS internal nodes. IMS is fundamental for the implementation of VoLTE, as it allows not only to provide IP connectivity and service control but also QoS and IP policy control, interworking with other networks, charging of services and also the execution of secure communications. It is through IMS that the UE specifies its characteristics and sets its QoS requirements by specifying parameters, like media type, direction of traffic, packet size and rate, bit rate for each media type and bandwidth specifications. With this information, the PCRF defines the required policy rules to implement which allow the P-GW to set bearers for a given application. The setup of bearers for an end-to-end service like VoLTE is done by the UE using IMS on top of the default bearer by signalling to an AS in the first place.

As LTE has a network architecture that is entirely packet-switch oriented, there is a need to provide voice calls, and also text messages, over IP while guaranteeing compatibility with previous circuit-switched systems. To ensure mobility to circuit-switched voice, especially while VoLTE is not massively available, two main functions were defined: CSFB and Single Radio Voice Call Continuity (SR-VCC). CSFB is currently the most popular solution among network operators and it consists basically of using their GSM/UMTS network to cope with voice calls, by moving to a GSM or an UMTS cell. This move is

done by MME, which interfaces with the Mobile Switching Centre (MSC) server, a common core network element of both the GSM and UMTS architectures, which is responsible for ensuring traffic forwarding. This method has disadvantages, such as an additional call setup time and the need to have coverage areas with LTE and GSM/UMTS simultaneously. SR-VCC uses a similar concept, but it is used to handover calls from a VoLTE call to the circuit-switched domain. Without this technology, operators cannot provide the user experience offered by VoLTE if they are not able to provide LTE coverage in their entire geographic range of service subscribers.

Regarding its radio functionalities, the implementation of VoLTE deals with one major aspect, which is packet scheduling. The packet scheduling algorithm, implemented at the eNodeB, deals with the network capacity by performing user selection and the respective resource allocation, a feature that does not exist in regular VoIP. This algorithm takes the QoS parameters of each bearer into account, as suggested in Figure 2.7, and it can be fully dynamic or semi-persistent. LTE uses dynamic scheduling by default, where each voice packet that arrives every 20 ms is scheduled, allowing total flexibility for optimising radio resources based on channel conditions. Its main constraint is that it requires capacity at the control channel, more specifically in the PDCCH. Semi-persistent scheduling avoids the need for control channel capacity by pre-allocating resources for the initial voice packets, every 20 ms, based on previously allocated transmission resources instead. An alternative to deal with the limited control channel capacity, but still using dynamic scheduling, is packet bundling, which consists of merging two voice packets before the transmission of a user, allowing a reduction of the required signalling overhead at the expense of the possibility of increasing the delay for future transmissions.

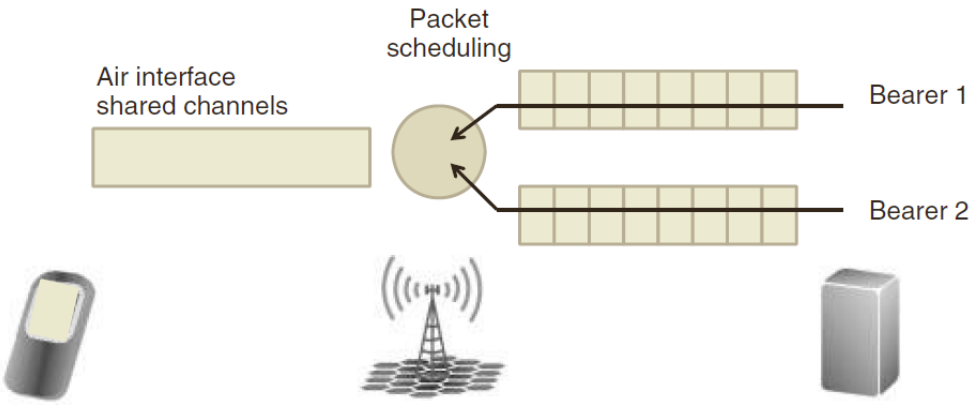


Figure 2.7. QoS differentiation provided by packet scheduling (extracted from [PoHo12]).

LTE uses DTX and discontinuous reception (DRX) for power saving. DRX is based on the concept that the UE does not monitor the PDCCH continuously, but instead it only receives control information after certain periods. Even though packets are received every 20 ms for voice, this method can be used efficiently for this purpose, as LTE's TTI is considerably smaller, allowing the UE to go to a power saving mode in between this time interval. Nevertheless, in what refers to UL coverage, one voice packet can be transmitted in a single TTI (1 ms), which is fairly inefficient as the terminal power amplifier only transmits in 5% of the time between the reception of consecutive packets. LTE uses TTI bundling, eventually combined with retransmissions, to overcome this fact as depicted in Figure 2.8. With TTI bundling, the same data is transmitted in up to four consecutive TTIs, introducing redundancy and

allowing consequently a reduction of the required overheads which does not happen with regular VoIP. Retransmission of these packets can also occur as long as it does not violate the maximum transmission delay requirement.

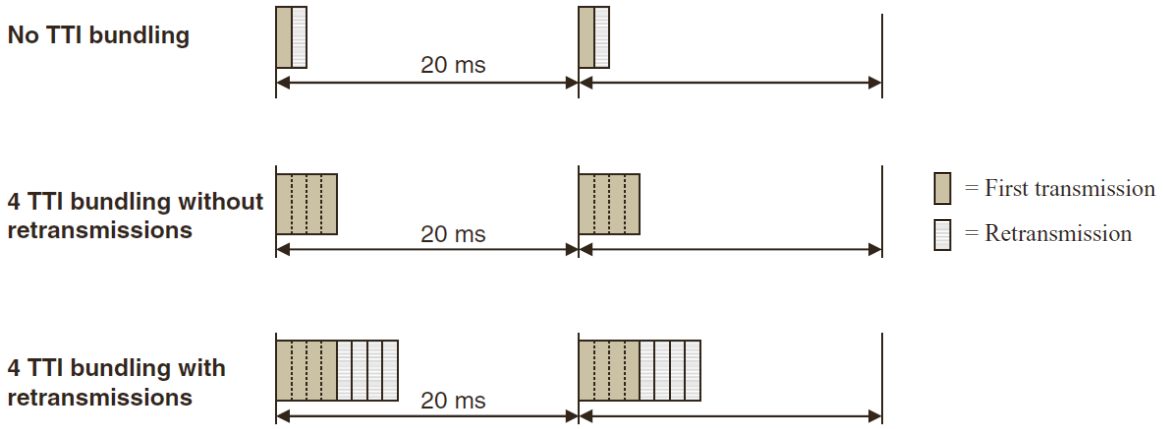


Figure 2.8. TTI bundling for enhancing VoIP coverage in uplink (adapted from [HoTo11]).

The impact of mobility on the implementation of VoLTE can also be analysed as handover introduces delay in transmission. For LTE, there is typically a delay around 25 ms for the break in the physical connection. Figure 2.9 illustrates how LTE handles the intra-frequency mobility between eNodeBs. This process is only possible because the UE measures continuously the signal coming from neighbour cells, allowing it to signal to the source eNodeB if a strong enough signal was found. The handover impact over VoLTE calls is not considered in this study, as it focuses on aspects at the single cell level.

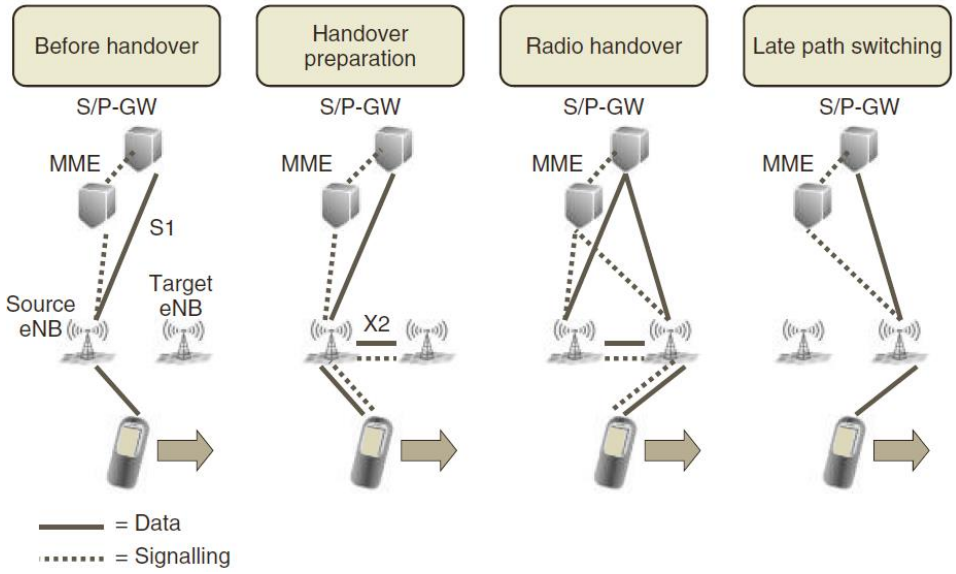


Figure 2.9. Intra-frequency mobility (extracted from [PoHo12]).

Performance measurement of VoLTE can be approached not only from the service perspective that is provided by network operators, but also in terms of the quality perceived by the end user. To evaluate the performance of the voice packet traffic in terms of the provided QoS, several Key Performance Indicators (KPIs) should be considered, just like in regular over-the-top VoIP, namely end-to-end delay,

jitter and packet loss rate. Besides these, the performance of signalling and call set up can also be evaluated, as well as the bit rate available for data in parallel with the voice call.

As it was referred in Section 2.2.1, the target for the maximum end-to-end delay is 200 ms. For jitter, a variable delay of 20 ms is usually recommended. Jitter is controlled by using a jitter buffer, which size has to be properly dimensioned as a function of a trade-off between delay and packet loss. If jitter exceeds the size of the buffer, voice packets are lost, while if the buffer size is too big, delay is increased as the packets take longer to be processed. In what refers to packet loss, voice is relatively tolerant to it, as humans are not sensitive to short drops in received speech. Therefore, a maximum packet loss rate of 2% is a typical value to consider if a user is in outage.

To obtain a measurement of the QoE of a VoLTE service, the more traditional method is MOS, which is a subjective method where a set of users classify the quality of a voice call with scores from 1 (bad) to 5 (excellent). According to [RaCZ13], there are two different types of algorithms used to replicate these results: perceptual/subjective and network based. In the former, Perceptual Objective Listening Quality Analysis (POLQA) is the most up to date method to measure QoE in LTE networks, being a successor of the Perceptual Evaluation of Speech Quality (PESQ) method, both defined as ITU-T standards in [ITUT14a] and [ITUT01] respectively. These methods require an analysis of the speech signal on both transmission ends, meaning that they require a sample of the signal on the listener's side.

Based on network packet transmission measurements, the E-model, defined in [ITUT14b], outputs a scalar factor named R, which reflects the level of user satisfaction, similarly to MOS, by receiving as input parameters jitter, packet loss rate and delay among others. This method is fairly more appealing from a network operator perspective, compared to perceptual methods, as it only requires parameters that are intrinsic to the network and are actually used to perform a QoS evaluation. The drawback is that the E-Model might hide some aspects that POLQA and PESQ methods can address, like voice activity detection, echo cancelation, among others. Nevertheless, perceptual/subjective methods like these are computationally intensive and are not adequate for measurements on high-scale networks.

2.2.3 VoLTE implementation

The deployment of VoLTE raises several challenges for operators, which in most cases need to upgrade their existing LTE networks. The IMS infrastructure must be settled in order to support VoLTE, which is something that in some cases has been a source of problems for operators, as they find out that they have to replace previously deployed equipment that did not support VoLTE [Qunh11]. However, the complexity of the VoLTE implementation is far beyond deploying an IMS, as it involves almost all core network elements.

The full interoperability between circuit-switched and VoLTE networks has to be ensured, as these two are expected to coexist during a long period with solutions like CSFB and SR-VCC. According to [Qunh11], two approaches can be taken, which consist of deploying or not the IMS Centralised Service (ICS). The architecture for ICS has been standardised by 3GPP [3GPP17b] as a way to apply services to users independently of the AN they use. This means that it works not only with IMS devices but also

with GSM or UMTS ones, without the need for any additional software upgrade. However, MSC servers must be upgraded to act as IMS clients, which most of the times proves to be a costly solution for operators. The need for a large-scale network upgrade is a synonym of high costs, causing operators to avoid this technology during the early stages of their VoLTE deployment. Not deploying ICS means that the service is completely migrated to the circuit-switch domain, without any intervention of IMS, and consequently no need to upgrade existing MSC servers.

Besides the network component, implementing VoLTE also means having compatible terminals. A VoLTE UE has to connect with LTE's RAN and EPC, meaning that they need an embedded IMS stack and a VoLTE application to operate with the LTE stack. [Aric15] lists the most relevant implementation aspects in terms of software for VoLTE. One of the most important aspects that influence the delivered QoE is power optimisation. Three major aspects influence power consumption on a UE performing VoLTE: battery life optimisation and power management at the hardware level; software architecture and the related multi-core processor architecture; and radio bearers and transmission protocols.

Several options exist in terms of test and simulation of VoLTE solutions. Simulators for generic network performance management can be used, like OPNET [Rive17] or OMNet++ [OMNe17]. In the case of OPNET, the simulator is basically a tool that allows to simulate and analyse the performance of any kind of network. In order to simulate VoLTE, the most common method is to consider a LTE network model and to configure an application with a voice profile. OMNet++ is yet another solution intended for networks simulations. Several projects exist as external extensions of this simulator, namely the INET framework aimed at the simulation of wired and mobile networks, and SimuLTE specifically focused on the analysis and performance of LTE and LTE Advanced networks. A similar approach, also focused on LTE systems only, is offered by the open source framework LTE-Sim [PGBC11], which allows the complete simulation of LTE networks, being able to simulate UL and DL scheduling algorithms in scenarios with multiples cells and users. Through its application layer it is also possible to generate VoIP traffic in the network. By default, the traffic generator considers G.729 voice flows, but all the parameters relative to traffic generation can be modified to suit a VoLTE scenario. MATLAB is also an alternative to test the physical layer of LTE systems, using its LTE System Toolbox [Math17] and implementing a VoIP service to perform end to end simulations.

In addition to simulation tools, many solutions for VoLTE performance testing currently exist in industry. Manufacturers, like Rohde & Schwarz or Spirent, provide tools to test all the infrastructure associated with the deployment of VoLTE, from core network elements to devices and call quality assessment [RoSc17] [Spir17]. These tools include for example audio analysis using PESQ or POLQA algorithms, battery performance tests correlated with IP traffic analysis or IMS equipment emulation.

2.3 Services and applications

This section describes the most relevant aspects that characterise the vast amount of services and

applications available over modern communication systems, being based on [3GPP17a], [HoTo11], [Cox14] and [Corr06].

Mobile data usage in recent years has largely risen as a result of multiple factors. The availability of mobile communication systems (namely since the UMTS era, which allow users to access Internet data services almost in every place at every time, led users to a whole new range of possibilities in terms of the variety of services and applications they have at their disposal. The evolution in smartphone technology also helped this trend, becoming more attractive to the end user and allowing the development of new applications. This growth is, in great measure, one of the main motivations for the study of the impact of the implementation of VoLTE over an existing LTE network. During a long time, most of the mobile traffic was originated from voice calls, which was by far the service with the biggest demand. Even though voice remains as the most popular service, this reality is changing and operators have to guarantee QoS for the more and more demanding data services, while ensuring that voice performance is not compromised.

To account for this situation, 3GPP established that UMTS services can be grouped into four distinct QoS classes. This service classification was maintained to characterise services in LTE too. These classes are defined according to service specific characteristics and requirements:

- Conversational refers to symmetric real-time services, which have strict delay requirements in order to allow simultaneous interaction between users.
- Streaming includes services that are mostly associated with unidirectional flows of information in real-time. The unidirectional nature of these services makes them more tolerant to delay than Conversational ones.
- Interactive corresponds to services like web browsing that have a request-response behaviour where the user requests data from a server.
- Background comprises services where the delivery time of information is not relevant as data is usually stored to be accessed later.

A comparison between the main characteristics of each of these classes is shown in Table 2.4. These characteristics are based on their most relevant performance requirements, which vary in terms of real-time requirements, symmetry on the traffic flows, need to guarantee a given bit rate, delay, buffer usage and traffic burstiness. One of the main distinction lies between the first two classes, Conversational and Streaming, which are associated with real-time services and the last two, Interactive and Background, which are not. This implies totally opposite approaches when dealing with the delay requirements of each service. In the case of real-time services, they have much stricter delay requirements, as users expect real-time responsivity, while for non-real-time ones this is not the case, as they are associated with services like file transfer or e-mail where users simply gather information like files from remote servers.

Table 2.4. QoS service classes' characteristics (adapted from [3GPP17a]).

	Service class			
	Conversational	**Streaming**	**Interactive**	**Background**
Real-time	✓	✓	×	×
Symmetric	✓	×	×	×
Bit Rate	Guaranteed	Guaranteed	Non-Guaranteed	Non-Guaranteed
Delay	Minimum fixed	Minimum variable	Moderate variable	High variable
Buffer	×	✓	✓	✓
Bursty	×	×	✓	✓
Example	VoLTE	Video Streaming	Web Browsing	E-mail

LTE natively includes signalling at the EPS level to control and monitor the experienced QoS. Signalling includes specified QoS parameters or only a simple indication to the service. For each EPS bearer, a set of QoS parameters are specified:

- QoS Class Identifier (QCI). It is an index used to refer to a set of pre-configured classes in terms of three QoS attributes: Priority, Delay and Loss Rate. These classes are divided in two categories, Guaranteed Bit Rate (GBR) and non-GBR bearers.
- Allocation and retention priority (ARP). It defines the priority of the bearer compared to others, allowing to decide upon the admission or modification of a given bearer. This also allows to decide if a bearer should be rejected in cases of congestion.
- GBR. It refers to the expected bit rate for the case of a GBR bearer.
- Maximum Bit Rate (MBR). It defines the maximum limit for the bit rate provided by the bearer.

For sets of EPS bearers, two QoS parameters are also defined:

- APN-Aggregate Maximum Bit Rate (AMBR). It defines the maximum bit rate that can be provided to all non-GBR bearers associated with the same APN.
- UE-AMBR. It defines the maximum bit rate that can be provided to all non-GBR bearers of the same UE and it is specified at the eNodeB level.

Bearers are essentially distinguished according to their assigned QCI and its corresponding QoS attributes. Priority determines the packets priority for scheduling purposes at the radio interface in a scale of 0.5 to 9, where 0.5 represents the highest priority. The Delay Budget is a reference maximum delay value useful for the packet scheduler to cope with delay requirements of each bearer. Loss Rate is a short notation for packet error loss rate, which is defined as the percentage of higher layer packets lost during uncongested periods of the network and it is useful for management of Radio Link Control (RLC) settings, like the number of re-transmissions, for example. Table 2.5 shows the correspondence between each QCI and the parameters standardised by 3GPP.

Table 2.5. QoS parameters for QCI (adapted from [3GPP16c]).

QCI	Resource type	Priority	Delay budget [ms]	Loss rate	Example application
1	GBR	2	100	10^{-2}	VoIP
2		4	150	10^{-3}	Video call (live streaming)
3		3	50		Real-time gaming
4		5	300	10^{-6}	Non-conversational video (streaming)
65		0.7	75	10^{-2}	Mission Critical user plane Push To Talk voice
66		2	100	10^{-2}	Non-Mission-Critical user plane Push To Talk voice
75		2.5	50	10^{-2}	Vehicle-to-everything (V2X) messages
5	Non-GBR	1	100	10^{-6}	IMS signalling
6		7		10^{-3}	Interactive gaming
7		6	300	10^{-6}	Video (buffered streaming) TCP-based
8		8			Application with TCP: browsing, email, file download, etc.
9		9			
69		0.5	60		Mission Critical delay sensitive signalling
70		5.5	200		Mission Critical Data
79		6.5	50	10^{-2}	V2X messages

2.4 State of the Art

An overview of the current state of the art regarding the subject of the thesis is presented in this section. The implementation of VoLTE in the context of already deployed LTE networks increased the interest of the scientific community in aspects related to DL scheduling algorithms, as these heavily influence how resources are distributed among network users. In this section, one highlights some of the most relevant studies related to radio resource allocation mechanisms in the presence of VoIP traffic as it is the case for VoLTE. Besides this subject, this section also analyses several studies in the literature that address the performance of VoLTE in terms of the variety of implementation options that network operators have at their disposal.

In [SiWa08], the authors present results that show the importance of service prioritisation for the performance of delay-critical services like VoIP, in the presence of concurrent Best Effort traffic like web browsing. For that purpose, three scenarios with VoIP traffic combined with traffic from three distinct services are considered: real-time video, mobile television and web browsing. Simulations are performed using two distinct approaches. The first one considers no service prioritisation and each user,

which can perform more than one service simultaneously, has a single queue. The later assumes one queue per distinct service and different scheduling priorities among each of them. The study concludes that capacity gains of 105 to 710% can be obtained in terms of VoIP capacity at the cost of small capacity losses in the other services. Gain variation is essentially justified by the different channel conditions of users. The Best Effort scenario is more favourable for users with good radio conditions.

In [BoBa14], a Channel and QoS Aware (CQA) scheduler is proposed to enhance the capacity of VoLTE systems and a comparison with schedulers of the same nature is presented. They start by referring that the most common scheduling algorithms, like Round Robin, Maximum Throughput or Proportional Fair, are unsuited for VoLTE as they are not QoS-aware. They propose a scheduler based on both time frequency domains. In the time domain, a metric that prioritises users with the highest value of Head of Line (HOL) delay. In the frequency domain, the proposed metric has the purpose of providing to all data flows the same level of QoS in terms of delay and bit rate by prioritising flows with higher HOL delay and a larger ratio of achievable throughput and past throughput. The authors compare this scheduler with the Priority Set Scheduler (PSS), which is similar to the CQA scheduler, but does not consider HOL delay, and the HOL scheduler, which only considers HOL delay as the scheduling metric. They conclude that the proposed scheduler outperforms PSS and HOL schedulers in 20 and 100% in terms of VoLTE capacity, for a realistic pedestrian scenario.

In what refers to VoLTE performance, in [Vizz14a], the author performs simulations to measure several KPIs for four different scenarios with different network congestion conditions, which reflects in different values for the available link bit rate. Simulations were performed using the OPNET Modeler 17.5 PL6 software tool, [Rive17], considering a typical campus area with 10×10 km^2, with the main results presented in Table 2.6. It is concluded that for link utilisations above 75% performance is still acceptable, as the obtained MOS reflects a speech quality perceived by the end user that is “not annoying”. If this percentage lies under 50%, speech quality becomes “slightly annoying”, with an increase of the end-to-end delay.

Table 2.6. VoLTE performance comparison (extracted from [Vizz14a]).

Scenario ID	Link utilisation [%]	Link bit rate [Mbps]	MOS	End-to-end packet delay [ms]	Mean voice packets sent	Mean voice packets received
1	100	44.73	4.3	140	30	19
2	75	33.55	4.0	140	26	16
3	50	22.36	3.7	150	22	13
4	25	11.18	3.6	150	20	11

In [Vizz14b], another study from the same author, a similar approach is considered to analyse the influence of different voice codecs in the performance of VoLTE by performing simulations for G.711, GSM Enhanced Full Rate (GSM EFR), AMR (12.2 kbps), IS 641 and G.729A codecs. The author concludes that in terms of the obtained MOS, G.711 and GSM EFR achieve better performances as a direct consequence of using higher transmission bit rates. On the other hand, AMR (12.2 kbps), IS 641

and G.729A are more susceptible to network transmission factors like delay for example, registering lower values of MOS.

In [KaSa14], the performance of VoLTE is evaluated in conditions where the transport network is congested with data traffic and no QoS prioritisation is considered, using the OPNET software to perform simulations. The considered scenario divides into a first case where only voice traffic is generated, compared to a second case where both voice and File Transfer Protocol (FTP) traffic are generated. End-to-end delay and jitter remain constant in the first case as the peak traffic is never reached, even for a worst case with 100 users. The same is not true for the second case where the network becomes congested and VoLTE packets are queued while FTP packets are processed. In this case, delay can reach values around 350 ms in the worst conditions, which is not tolerable.

In [RiDM13], the authors intend to propose a more realistic approach to the analysis of QoS and QoE in LTE networks, using VoLTE as a use case, by experimenting in a real context with real devices, services, and radio configurations. Moreover, they perform cross-layer measurements by correlating radio configurations with the QoS parameters measured at the application layer. The authors presented an experimental testbed to be used for multiple scenarios that included an LTE test base station able to emulate features like channel propagation for example, and that supported connection to real LTE devices. The use of realistic impairments, like fading and noise, produces noticeably variable results, which cannot be obtained through non-simulated results, like many of those that are available in the literature. Correlations below 0.8 were obtained when low level parameters and IP parameters are compared.

Among many other parameters already mentioned, the authors show the relationship between the received Signal-to-Noise Ratio (SNR) and the packet loss rate, concluding that the radio conditions in the receiver's side can impact on the amount of losses, as a variation of 5 dB in the SNR can result in a variation of the packet loss rate of approximately 1%. The authors also show that a linear estimation can be obtained to relate MOS to packet losses, and that for packet losses close to 0% the PESQ algorithm outputs the maximum quality that corresponds to a MOS equal to 4.5.

In [OzVa13], a study of DL performance of VoLTE is conducted in macro-cell homogeneous and heterogeneous networks, considering macro- and small-cells, where both voice and data traffic are generated using dynamic and semi-persistent scheduling. Simulations were held and statistics for VoLTE packets, total user throughput, number of PDCCH resources and RB usage were collected.

For the case of considering only macro-cells, the higher priority VoLTE results in the fact that the only impact of data users on VoLTE is through inter-cell interference. On the other hand, voice traffic has an impact in the data users' performance as the total throughput decreases almost linearly with the increase of the number of VoLTE users. The inclusion of a limit to the number of PDCCH resources also has an impact on data users, resulting in unused RBs. To avoid this, the use of semi-persistent scheduling shows similar gains to a scenario where there is no PDCCH limit.

Heterogeneous networks are considered with the intent of providing gains in terms of performance, mostly due to the offloading of users to pico-cells and to the reduction of control channel limitation, which

increases control channel capacity. The authors registered very good performances for pico-cell users compared to macro-cell ones and, furthermore, there is actually an improvement in performance with the increase of VoLTE users due to reduction of the interference in pico-cells originated by macro-cells. In terms of capacity, the authors conclude that it can be increased significantly with dynamic scheduling when there is no limitation on PDCCH, and emphasise that the VoLTE capacity would be higher if no data users were considered in the system.

In order for one to understand the current context in most of the current LTE deployments where VoLTE is still not available, in [TuPe13], the authors provide a study of the drawbacks of CSFB and its interaction with packet-switched data services. They conducted experiments in two major US LTE operators, named symbolically OP-I and OP-II, using six different phone models and collecting traces for further analysis using the Wireshark network protocol analyser. Four mains aspects are analysed: the data performance degradation when voice calls occur, the impact of the voice calls on the data session, additional performance impacts of voice in data traffic beyond throughput degradation, and the impact of packet-switch data on circuit-switch voice. Table 2.7 shows the main conclusions of this study where, to the authors' surprise, they verify that voice and data have a mutual impact on each other.

Table 2.7. Summary of the impact of CSFB on LTE systems (extracted from [TuPe13]).

Finding	Detail	Root cause
Throughput slump	Data throughput decreases (up to 83.4% observed) OP-I: only during the call OP-II: during and after the call	Handoffs triggered by CSFB and speed gap between 3G and 4G
Losing 4G connectivity	Never returns to 4G after the CSFB call under certain data traffic OP-I: when the call fails to be established OP-II: any CSFB call	State machine "loophole" in 3G to 4G transition
Application aborts	Application aborts occasionally (3.4% for OPI and 5.7% for OP-II) after the call	Network state changed by circuit-switch domain operation (here, network detach caused by CSFB voice calls)
Missing incoming call	Misses all incoming calls temporally (for several seconds) while enabling the packet-switch service	Network state changed by packet-switch domain operations

Chapter 3

Models and Simulator

This chapter provides a description of the developed model for DL resource allocation on an LTE cell and the metrics used to evaluate the performance of VoLTE and other services. The implementation of this model on a simulator and its assessment are also presented in detail.

3.1 Model overview

This section describes the model developed to evaluate the impact of the implementation of VoLTE over other services, focusing on aspects related to the overall capacity of the network and the degree of satisfaction of the served users. One considers a single-cell model where all users are connected to a single base station, each of these performing a given service. This model is based on the resource allocation of the DL traffic generated in the network, describing its time evolution for every TTI at the packet level for different types and classes of services. Interference and handover issues are not taken into account for simplification purposes, as they do not constitute the main focus of this thesis, which is aimed at the cell level and not at the cellular network one. From a high-level perspective, the model is composed of three core parts as illustrated in Figure 3.1.

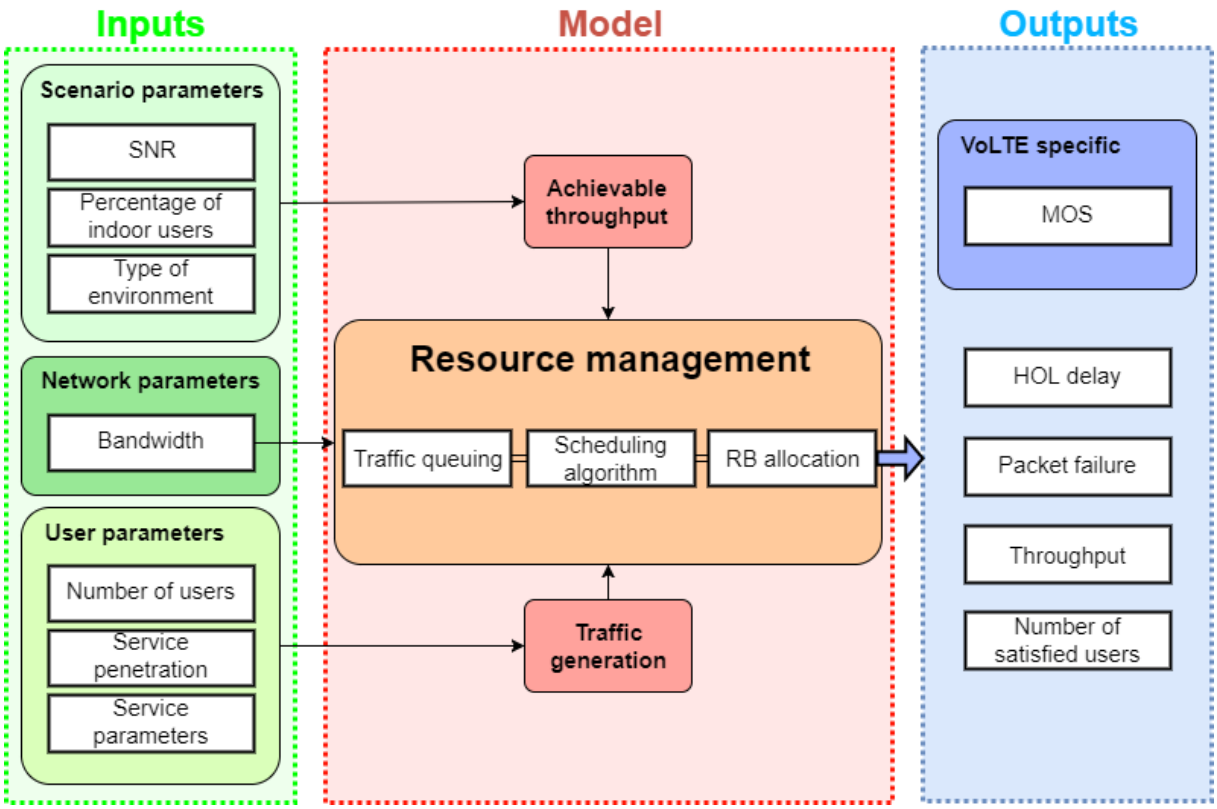


Figure 3.1. Model architecture.

Input parameters are classified as Scenario, Network and User ones. Scenario parameters refer to the characterisation of the cell environment, namely the type of environment, the associated SNR parameters and the percentage of indoor users. For this work, three types of environment are considered: rural, suburban and urban. Network parameters refer in this context to the configured radio channel bandwidth, as this is one of the main aspects that determines the system's capacity together with SNR. User parameters refer to all the aspects involved in traffic generation. The number of users determines the average rate of user arrivals. Each user is associated with the service he/she is

requesting based on service penetration, which defines the number of users performing each service. The service parameters characterise the behaviour of the network traffic generated by them.

The proposed model consists of three main steps: calculation of the achievable throughput, traffic generation, and resource management. The achievable throughput per RB in every TTI is estimated based on each user's experienced SNR, which is statistically generated. For traffic generation, users are associated with the network according to a Poisson process with a specified average rate of users. Traffic for each user and the corresponding service is modelled through appropriate traffic source models, presented in Annex B. The core step of the model deals with resource management in order to schedule users at the radio interface level. In a first instance, network generated traffic is queued for each user and allocated according to a scheduling algorithm that manages RB allocation according to the system's capacity. Finally, the model outputs are evaluated by using several metrics for the performance of all the services and VoLTE specifically. The performance of each service is assessed based on the experienced throughput, queuing delay, packet failure associated with the DL resource allocation and the number of satisfied users. Besides these parameters, QoE for VoLTE is evaluated based on the calculation of the MOS using the E-model with estimations of the end-to-end delay and packet loss.

3.2 Model description

This section describes in detail all the elements that compose the proposed model for traffic generation and resource allocation.

3.2.1 Achievable throughput

A statistical approach to characterise the radio conditions of each UE is considered, where it is assumed that SNR is described by a Log-Normal Distribution that depends on the considered type of environment, e.g., rural, suburban or urban. Each environment is characterised by its SNR with given values of mean and standard deviation of the SNR, which provides statistical variance in the time domain. Assuming that the considered reference values for the SNR refer to outdoor environments, an extra attenuation $L_{p\ ind}$ is added to account for indoor users. To fully model the time variation of SNR, the channel coherence time is also considered. The coherence time corresponds to the time interval in which the channel conditions are considered to be invariant, meaning that SNR, and consequently the achievable throughput, keep constant. An estimation of the coherence time for 50% correlation is obtained by [Corr16]:

$$T_{c\,[\mathrm{s}]} \cong \frac{9}{16\pi \frac{v_{[\mathrm{m/s}]}}{\lambda_{[\mathrm{m}]}}} \tag{3.1}$$

where:

- v: User speed.
- λ: Wavelength (dependent of the operating frequency band).

Even though mobility is out of the scope of this thesis, since a single cell is considered, the speed of each user inside the cell is modelled to account for propagation changes. For this purpose, user speeds are updated at regular time intervals, using a triangular distribution density model as detailed in Annex C. The estimated value of SNR is used to compute the achievable transmission rate as described in Annex A, where a mathematical formulation developed by [Alme13] is presented to approximate the throughput of a single RB. This formulation considers a network using MIMO 2x2 for QPSK, 16-QAM and 64-QAM modulations, based on recent measurements done by 3GPP. One considers that the serving eNodeB has full knowledge of the instantaneous DL SNR value reported by each user, therefore allowing to choose the modulation that provides the best throughput. The chosen Modulation and Coding Scheme (MCS) is the same for all the RBs allocated to a user, during a given TTI. The algorithm for the calculation of the achievable throughput is detailed in Figure 3.2. For a better understanding of all the variables involved in this algorithm, the variable not defined up to this point has the following notation:

- $t_{last,SNR}$: Last instant where the SNR for a given user changed.

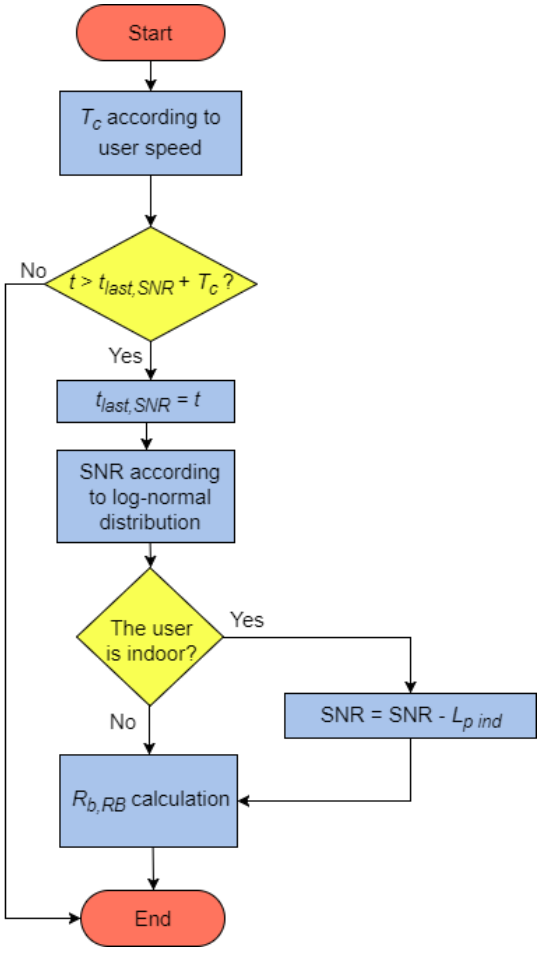


Figure 3.2. Achievable throughput algorithm.

3.2.2 Traffic generation

Several traffic source models are considered for traffic generation, benefiting from the work developed in [Agui03], [Khat14] and [Serr12], where a complete description of the considered models is provided as well as their statistical validation. Using these models, network traffic is generated at the packet level considering the specificities of each service. For the purpose of this work, seven services are considered, namely: VoLTE, video calling, video streaming, music streaming, web browsing, file transfer and e-mail. This list of services is sufficiently detailed to take conclusions on the impact of the implementation of VoLTE over an LTE network, as it includes the main sources of traffic over current networks and covers all the existing service classes [3GPP17a].

One considers an ON-OFF model for VoLTE calls as described in Annex B.1. Streaming services, namely video and music streaming, are also implemented using this model for convenience of implementation, as they share common characteristics with the behaviour of voice traffic, especially the fact that both use a fixed framerate. Figure 3.3 shows the ON-OFF model used for VoLTE and streaming services. The variables involved in the ON-OFF model that where not defined up to this point have the following notation:

- $t_{arrival}$: Instant of arrival of a user to the network.
- T_{call}: Call/session duration.
- t_{frame}: Frame duration.
- T_{ON}: Active state duration.
- T_{OFF}: Silent state duration.
- t_{state}: Duration of the ongoing state.
- t_{last}: Last instant where a packet was generated.
- $t_{activity}$: Instant where the current ON or OFF state started.
- p_{gen}: Signal that a packet was generated during the current frame.

For the video calling service, the Gamma Beta Auto-Regressive (GBAR) video source model presented in Annex B.2 is used, as it is based in real traces of video traffic generated from video conferencing services. The algorithm used to model the video calling service traffic using the GBAR model is detailed in Figure 3.4.

Several approaches exist in the literature to model video conferencing traffic with different video codecs. The mandatory codec to provide ViLTE is H.264 but, nevertheless, as the service is still not massively available over the existing LTE networks, one considers that for a short-term realistic scenario, video conferencing traffic is mostly generated from older codecs, like the H.263, which is the mandatory codec for 3G video telephony. For that reason, one considers the GBAR model that is based on statistical features observed in H.261 and H.263 codecs. Moreover, this model is also appealing from the implementation point of view, as simulating it only requires generating random values from a stationary stochastic process based on Gamma and Beta Distributions.

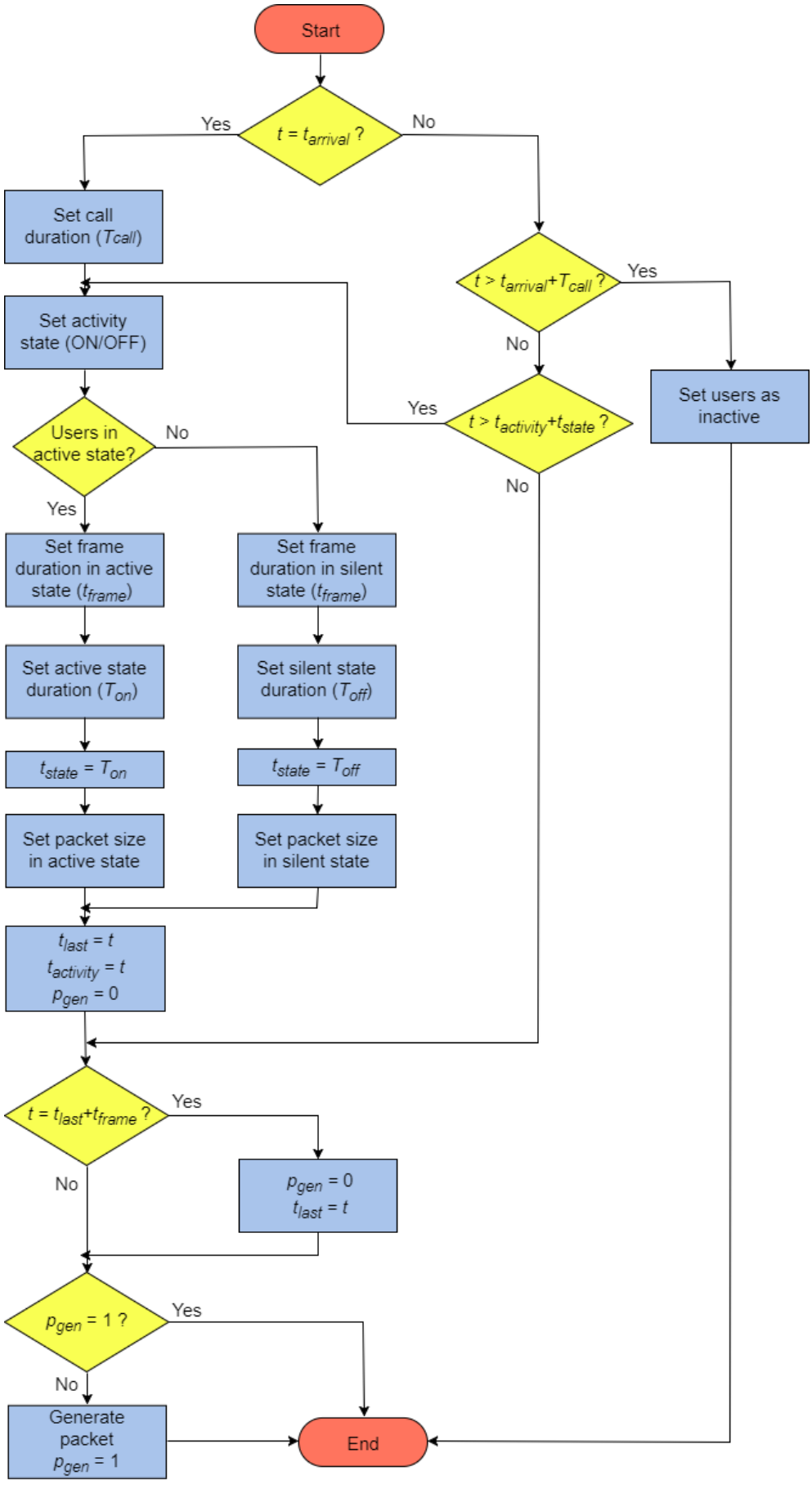


Figure 3.3. ON-OFF model algorithm.

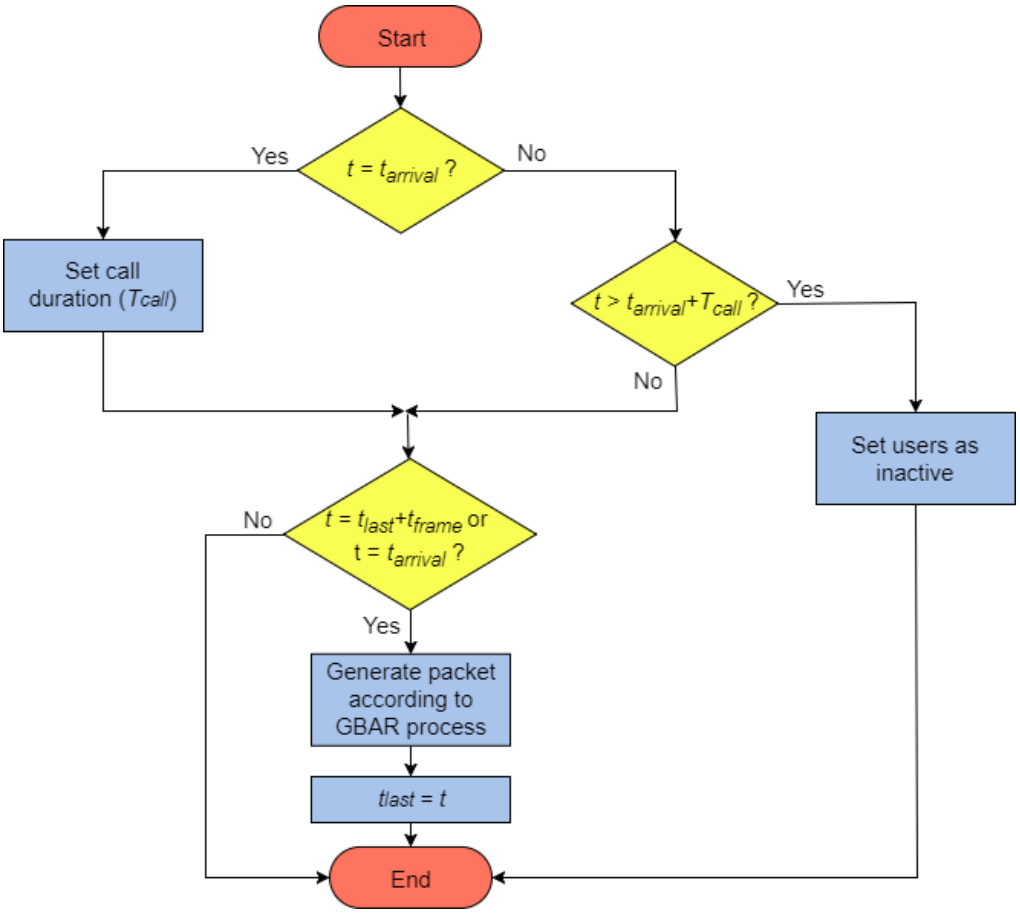


Figure 3.4. GBAR model algorithm.

Regarding non-conversational applications, the considered models are described in Annex B.4, for web browsing, and Annex B.5, for file transfer and e-mail. Figure 3.5 shows the non-conversational algorithm used for web browsing, file transfer and e-mail. This algorithm is applied to the three services, as they present a similar behaviour. For web browsing, packet calls correspond to the objects that are downloaded, namely the webpages and all the embedded content associated with them. In file transfer and e-mail, packet calls correspond directly to the downloaded files or e-mails, respectively.

The variables involved in this algorithm that where not defined up to this point have the following notation:

- N_{pc}: Number of packet calls.
- V_D: Object data volume.
- P_S: Packet size.
- D_d: Parsing time between two consecutive packets.
- D_{pc}: Reading time between two consecutive packet calls.
- t_{silent}: Instant where the silent state started.

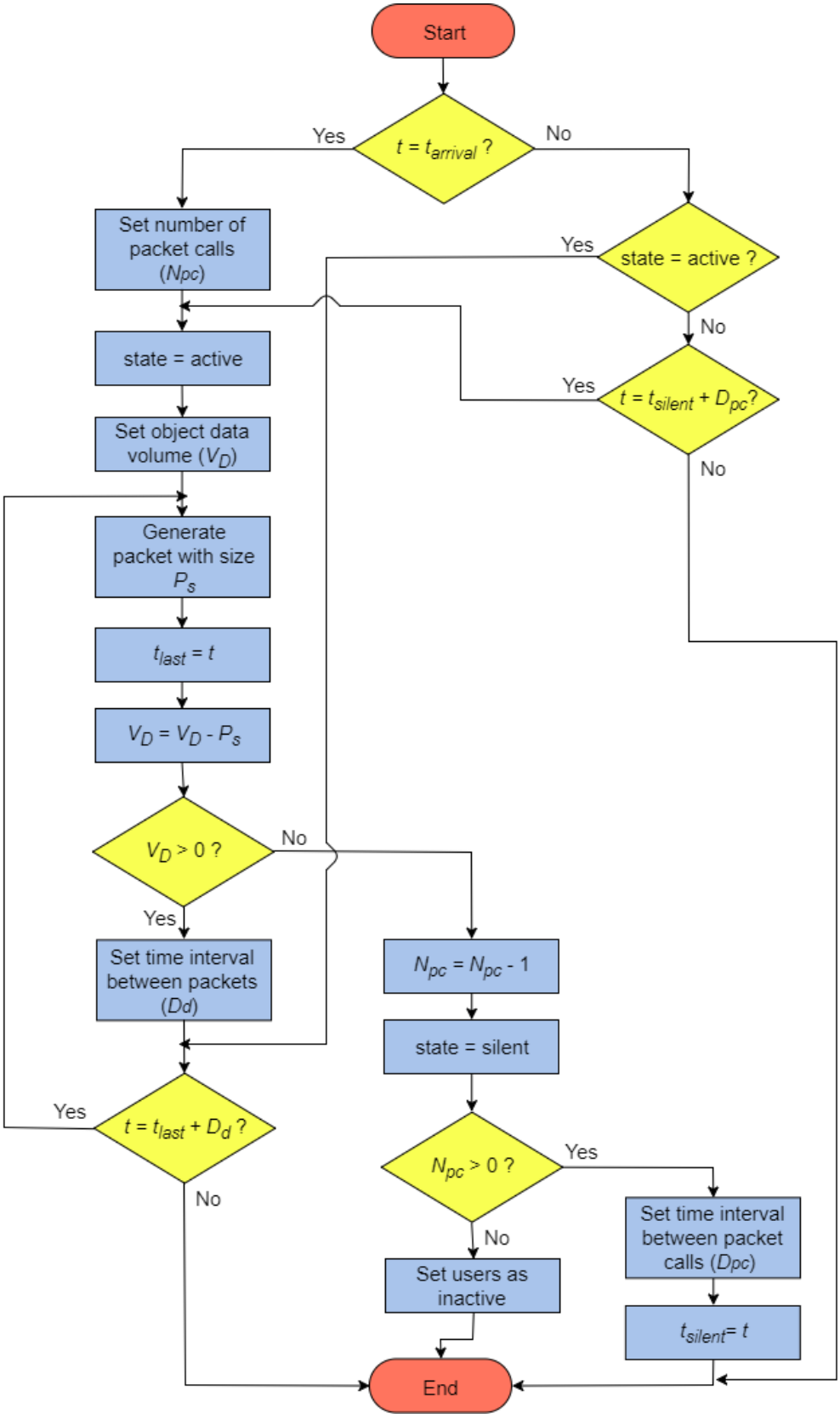


Figure 3.5. Non-conversational model algorithm.

3.2.3 Resource management

The most crucial aspect of the developed model refers to resource management, as it basically determines how the generated traffic is assigned to each user and how the trade-off between cell capacity and QoS is handled. Figure 3.6 illustrates the processes involved in resource management, which are divided in three main stages: traffic queuing, the scheduling algorithm, and RB allocation.

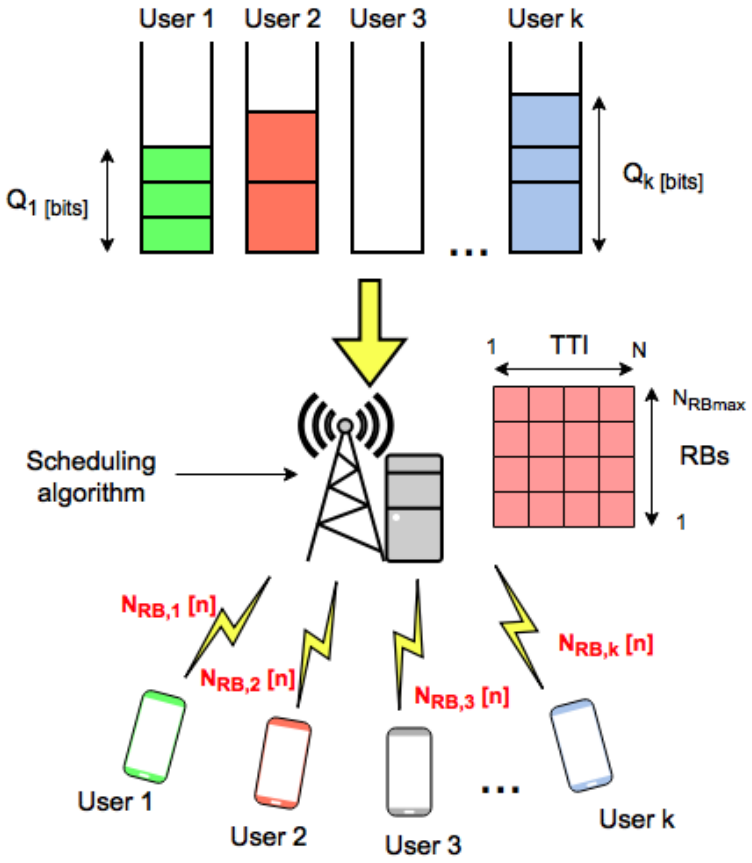


Figure 3.6. Resource management model.

The incoming traffic for each user is handled through an individual First-In-First-Out (FIFO) queuing system. The model for queue management at the eNodeB level consists of one queue buffer per user, with $Q_k[n]$ representing the total number of bits for all the packets in queue for user k at the time instant n. Each queue is updated whenever a packet arrives at the eNodeB in order to add its corresponding number of bits. In parallel with this process, a timer is started for each packet received in the eNodeB. This timer is updated every TTI n, allowing the computation of the HOL delay $\tau_{HOL}[n]$, which represents the amount of time spent by the first packet to be transmitted. This information is used to monitor the time spent by each packet in the queue.

A maximum delay for each service is defined with the purpose of preventing that packets remain in the queues for too long. The adopted limits correspond to the delay budget specified by 3GPP recommendations as presented in Table 2.5, Section 2.3. Figure 3.7 describes the algorithm to manage user queues, including the mechanisms to insert packets, and to discard packets exceeding the maximum delay.

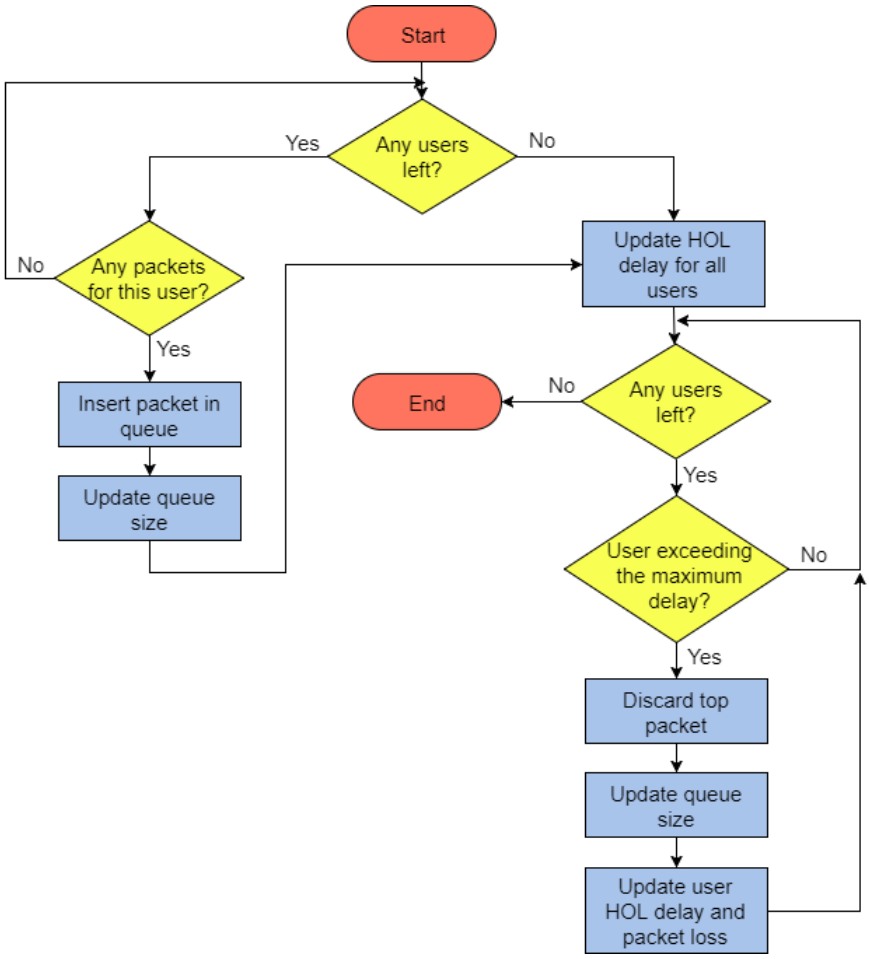


Figure 3.7. Queue management algorithm.

RBs are assigned at every time instant n in order to ensure that the generated DL traffic is efficiently allocated in the radio interface. As previously described in Section 2.1.2 (and shown in Table 2.2), the number of available RBs is determined by the radio channel bandwidth. In a first approach, the total number of RBs required to allocate all the traffic in queue for each user is computed. The number of RBs needed to accommodate all the queued traffic for each user k is estimated as:

$$\widetilde{N}_{RB,k}[n] = \left\lceil \frac{Q_{k\,[\text{bits}]}[n]}{R_{b,RB\,[\text{kbps}]}[n] \times T_{TTI\,[\text{ms}]}} \right\rceil \tag{3.2}$$

where:

- $R_{b,RB}$: Achievable RB throughput.
- T_{TTI}: TTI (equal to 1 ms).

After the total number of requested RBs for each user is computed, network capacity is checked, as the number of RBs requested by all users cannot exceed the maximum number of RBs available at each TTI. To perform this comparison, the cell load at instant n is computed as follows:

$$L_{cell}[n]_{[\%]} = \frac{\sum_{k=1}^{N_u} N_{RB,k}[n]}{N_{RB,max}} \times 100 \quad (3.3)$$

where:

- $N_{RB,k}$: Number of RBs allocated to user k.
- $N_{RB,max}$: Maximum number of RBs per TTI (corresponds to the network capacity which is directly related to the system's radio channel bandwidth).
- N_u: Total number of active users in the cell.

If the requested cell load is above 100%, the total number of requested RBs is larger than the network capacity and a reduction must occur. To manage this need for optimising RBs allocation, a scheduling algorithm composed of two levels of optimisation runs iteratively until network capacity is not exceeded. At a higher level, users requesting for RBs are distinguished according to the priority of the service they are performing. This means that the set of users performing a service that is associated with the lowest priority of all existing services are the first to be reduced. Then, for each of these sets of users performing a given service, further optimisation must occur in order to distinguish among these users. At this point, a concept inspired by the Proportional Fair algorithm is introduced [HoTo11]. The main assumption is that the number of RBs assigned to each user is proportionally reduced among each set of users. By using this approach, fairness is maintained among users, based on their radio conditions and the amount of traffic they have in queue. Users experiencing a low achievable throughput per RB and that have more traffic in queue are more penalised because they require more RBs.

Once it is guaranteed that the network capacity is not exceeded, RBs are assigned to the requesting users. After this assignment, the state of the traffic queues is updated coherently. For that purpose, for every TTI n that a user k is successfully scheduled, the total number of bits in the queue of user k is updated as:

$$Q_k[n]_{[\text{bits}]} = Q_k[n-1]_{[\text{bits}]} - (N_{RB,k}[n] \times R_{b,RB\ [\text{kbps}]}[n] \times T_{TTI\ [\text{ms}]}) \quad (3.4)$$

Figure 3.8 shows the developed algorithm to allocate RBs in every TTI. The variables involved in this algorithm that where not defined up to this point have the following notation:

- P_{max}: Higher Priority level.
- $N_{RB,reduce}$: Sum of all the RBs that must be reduced.
- $N_{RB,usersub}$: Number of RBs that must be reduced for each user.

Other scheduling strategies could be applied at this point. For the purpose of this thesis, increasing the complexity of this algorithm would mean significantly increasing simulation times. Considering a scheduling algorithm also in the frequency domain assumes that scheduling decisions are taken for each RB in the same time slot, which could increase the computational effort up to 100 times, corresponding to the number of available RBs per TTI for a maximum 20 MHz bandwidth.

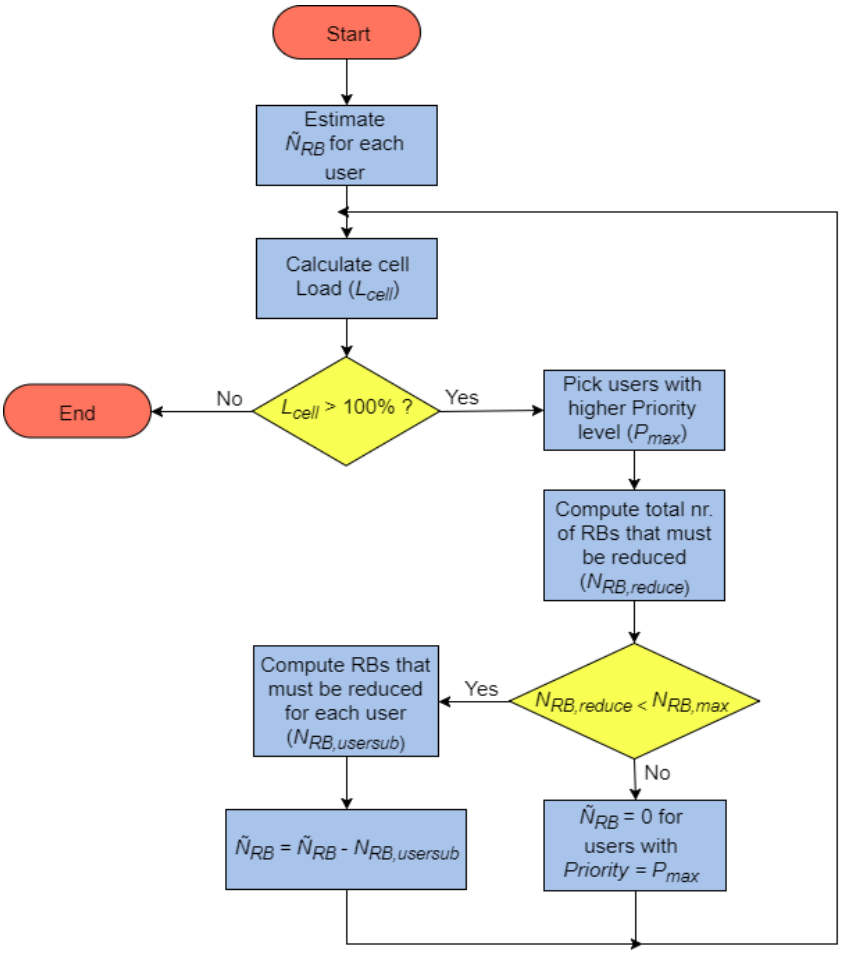


Figure 3.8. RB allocation algorithm.

3.3 QoS and QoE Metrics

This section describes the main performance metrics considered to evaluate the developed model. To approach the measurement of QoS and QoE, a distinction between VoLTE and other services is made, in the sense that the considered metrics are divided among those that allow a qualitative characterisation of VoLTE and those that provides a measure of the overall performance of any service and the corresponding degree of satisfaction from the user viewpoint.

As the quality of VoLTE as a real-time application is deeply affected by network delay, one considers the HOL delay of each voice packet to assess the queuing delay due to the DL resource allocation. To have a global metric of the delay experienced by all users, one considers in first place the average HOL delay for all users performing a service s, taking the average HOL delay of each user into account:

$$\overline{\tau_{HOL,s}}\ [\text{ms}] = \frac{1}{N_{u,s}} \sum_{k=1}^{N_{u,s}} \overline{\tau_{HOL,k}}\ [\text{ms}] \tag{3.5}$$

where:

- $N_{u,s}$: Number of users performing service s.
- $\overline{\tau_{HOL,k}}$: Average HOL delay of the packets received by user k.

This parameter allows estimating the end-to-end delay in order to have a complete understanding of the total delay involved in a voice call and moreover to obtain a measurement of QoE using an estimation of MOS. As the entire end-to-end network path between users performing a voice call is not addressed in this work, several factors must be estimated. Using the delay budget analysis considered in [HoTo11], an estimation of the end-to-end delay of a user k for a VoLTE call using the AMR-WB codec is given by:

$$\tau_{k\,[ms]} = \tau_{enc\,[\mathrm{ms}]} + \tau_{UL\,[\mathrm{ms}]} + \tau_{trans\,[\mathrm{ms}]} + \overline{\tau_{HOL,k}}_{\,[\mathrm{ms}]} + \tau_{dec\,[\mathrm{ms}]} + \tau_{proc\,[\mathrm{ms}]} \quad (3.6)$$

where:

- τ_{enc}: Voice encoding delay (equal to 30 ms corresponding to a 20 ms frame size, a 5 ms look ahead delay common to all codec modes and 5 ms processing time).
- τ_{UL}: Radio interface delay in the UL (50 ms is assumed).
- τ_{trans}: Transport delay (assumed to be 10 ms).
- τ_{dec}: Voice decoding delay (equal to 5 ms).
- τ_{proc}: Processing delay in the UE and in eNodeB (equal to 10 ms).

For the time being, one does not acknowledge the existence of a similar analysis to estimate delay for the EVS codec. Similarly, an estimation of MOS cannot be obtained using a simplified model as the one presented throughout this section. For this reason, this thesis is focused on the study of QoE for the AMR-WB codec. Nevertheless, queuing delay among other parameters can still be used to address the performance of both voice codecs at the radio interface level.

Packet loss is also considered, as it is not only a requirement for a good voice service, but it is also an important QoS metric for every other service. Two main aspects lead to the existence of packet loss in the radio interface, namely losses due to congestion and transmission errors. In terms of congestion issues, packets are discarded when they exceed their time limit to stay in queue to be scheduled. Therefore, a maximum queueing delay is defined for each service, corresponding to the delay budget associated with the corresponding QCI's (as listed in Table 2.5, Section 2.2.2). Regarding transmission errors, as the study of the radio link falls out of the scope of this thesis, one analyses the behaviour of the network in terms of the packet loss due to lack of cell capacity. It is crucial to mention that no packet retransmission mechanisms or rate adaptation algorithms were implemented, and therefore the values obtained for packet loss are far above what is usually obtained in the context of real network measurements. The main goal in this case is to characterise the degradation of the services' throughputs, characterising it through both delay and the number of lost packets. For this purpose, one introduces packet failure as the metric under analysis in this study to measure the percentage of transmitted packets that had transmission problems due to excessive delay and would require a retransmission. Thus, the packet failure ratio for a given time interval is computed as:

$$\Gamma_{failure\ [\%]} = \frac{N_{sent} - N_{received}}{N_{sent}} \times 100 \tag{3.7}$$

where:

- N_{sent}: Number of packets sent during the time interval being considered.
- $N_{received}$: Number of packets received during the time interval being considered.

To achieve an estimation of the QoE at the user end, one adopts the E-model defined in [ITUT14b] by converting its output, the transmission rating factor R, to MOS, which reflects the level of user satisfaction in terms of the perceived voice call quality. As the original E-model accounts for a high number of variables, several assumptions can be made to simplify it. The model can be applied to voice calls that use the AMR-WB codec, as it is the case for VoLTE, where R is simplified as [Nguy16]:

$$R = 129 - I_{d,wb} - I_{e,eff,wb} \tag{3.8}$$

where:

- $I_{d,wb}$: Delay impairment factor, representing the impact of delay over voice signals.
- $I_{e,eff,wb}$: Equipment impairment factor, which measures the impact of the signal distortion caused by low codec bit rates and packet losses of random distribution.

With the simplified model, $I_{d,wb}$ depends only on the experienced end-to-end delay and can be obtained as [Nguy16]:

$$I_{d,wb} = \begin{cases} 0.024\tau_{[\mathrm{ms}]} & ,\tau < 177.3\ \mathrm{ms} \\ 0.024\tau_{[\mathrm{ms}]} + 0.11 \times (\tau_{[\mathrm{ms}]} - 177.3) & ,\tau \geq 177.3\ \mathrm{ms} \end{cases} \tag{3.9}$$

The value of $I_{e,eff,wb}$ depends on packet loss and on the AMR-WB codec source bit rate. One considers the packet failure ratio as an estimation of packet loss for this parameter, which is given by [Nguy16]:

$$I_{e,eff,wb} = I_{e,wb} + (129 - I_{e,wb}) \frac{\Gamma_{failure\ [\%]}}{\Gamma_{failure\ [\%]} + B_{pl}} \tag{3.10}$$

where:

- $I_{e,wb}$: Equipment impairment factor without any packet loss (11.0 is the value proposed by [MoRa06] for the AMR-WB 12.65 codec).
- B_{pl}: A codec-specific factor which characterises its robustness against packet loss (13.0 is the value proposed by [MoRa06] for the AMR-WB 12.65 codec).

Once all factors are defined, the R factor is obtained and it can be converted to MOS. MOS is associated with a subjective classification of call quality as stated in Table 3.1. The conversion to MOS is simply given by [Nguy16]:

$$\mathrm{MOS} = \begin{cases} 1 & ,R_x < 0 \\ 1 + 0.035R_x + R_x(R_x - 60)(100 - R_x) \times 7 \times 10^{-6} & ,0 \leq R_x \leq 100 \\ 4.5 & ,R_x > 100 \end{cases} \quad (3.11)$$

where R_x stands for a correction of the original transmission factor given by [Nguy16]:

$$R_x = \frac{R}{1.29} \quad (3.12)$$

Table 3.1. Relation between R-value and user satisfaction (extracted from [ITUT14b]).

R-value (lower limit)	MOS	User satisfaction
90	>4.34	Very satisfied
80	>4.03	Satisfied
70	>3.60	Some users dissatisfied
60	>3.10	Many users dissatisfied
50	>2.58	Nearly all users dissatisfied
-	<2.58	Not recommended

The performance parameters for the VoLTE service were described up to this point. In what refers to all the other services, performance is measured by considering user throughputs and the level of satisfaction associated with them. As this is a time-based model, and thus one has access to the total number of transmitted bits, to obtain the total cell throughput during a given period of time one computes:

$$R_{b,eNodeB\,[\mathrm{Mbps}]} = \frac{N_{bits}}{T_{sim\,[\mathrm{s}]} \times 10^6} \quad (3.13)$$

where:

- N_{bits}: Total number of bits transmitted.
- T_{sim}: Simulation time.

In order to monitor the network throughput, the average throughput per user for a given service is obtained from:

$$\overline{R_{b,s}}_{\,[\mathrm{Mbps}]} = \frac{1}{N_{u,s}} \sum_{k=1}^{N_{u,s}} \overline{R_{b,k}}_{\,[\mathrm{Mbps}]} \quad (3.14)$$

where:

- $\overline{R_{b,k}}$: Average throughput of user *k*.

Finally, the number of satisfied users served by the cell is also a relevant metric. A user is considered

to be satisfied if its minimum throughput can be guaranteed. This parameter is used in multiple contexts like, for example, the number of satisfied users associated with a given service or service class.

3.4 Model implementation

Towards the implementation of the models and metrics described in Sections 3.1 and 3.2, a time-based simulator was developed to allow the analysis of the network during a given period of time. This simulator was implemented using MATLAB, and it operates with a resolution of 1 ms, corresponding to the TTI in LTE. Figure 3.9 represents the architecture of the simulator where three main scripts are used: *inputs.m*, *simulation_setup.m* and *main.m*.

inputs.m corresponds to the script where all input parameters and simulation environment are specified. This file must be run before any simulation whenever one wants to change the input setup. *simulation_setup.m* is responsible for reading the input parameters from *inputs.m* and scheduling one or more simulations with different input configurations. This allows the user to perform multiple simulations in a single run, while performing a sweep of input values, considering for example different environments or a different number of users. Therefore, *simulation_setup.m* and *inputs.m* are the scripts that the user interacts directly with.

main.m, as the name suggests, is the main script of the entire simulator where the whole network simulation is performed and runs in a loop inside the *simulation_setup.m* script. After the initialisation of variables and data structures, the program consists of a loop that runs with a 1 ms resolution until the simulation time is reached. Note that except when clearly mentioned, all the algorithms involved in network simulation apply simultaneous to all users or sets of users by benefiting from MATLAB capabilities in terms of array structures and logical indexing. The first step in the simulation loop corresponds to the update of all variables related to the environment. For every second, the speed of each user is updated and the corresponding coherence time is estimated. The coherence time will determine how fast SNR will change and consequently the achievable throughput per RB.

After all the scenario variables are updated, traffic is generated by using the three models used to implement the services for this study (see Section 3.2). The ON-OFF model consists of intermittent sequences of active states, where packets are generated at fixed framerates and silent states where no packets are generated (for the streaming case) or SID frames are generated (for the VoLTE case). Each session is limited by the time duration T_{call}, which is counted since the arrival of the user to the network. The GBAR model used for the video calling service is simply implemented by generating for each user, at a fixed framerate, a packet with a size given by the stochastic process described in Annex B.2. The non-conversational algorithm consists of generating sequences of Hypertext Transfer Protocol (HTTP) objects in the web browsing case and files in the file transfer and e-mail cases. These files are separated by reading times, which correspond to the time needed by the users to process the information. By opposition to the streaming services, in these services each session is limited by the volume and the

number of the files to be transferred, which has a totally different behaviour when the session duration has great variability. A detailed description of traffic source models is presented in Annex B.

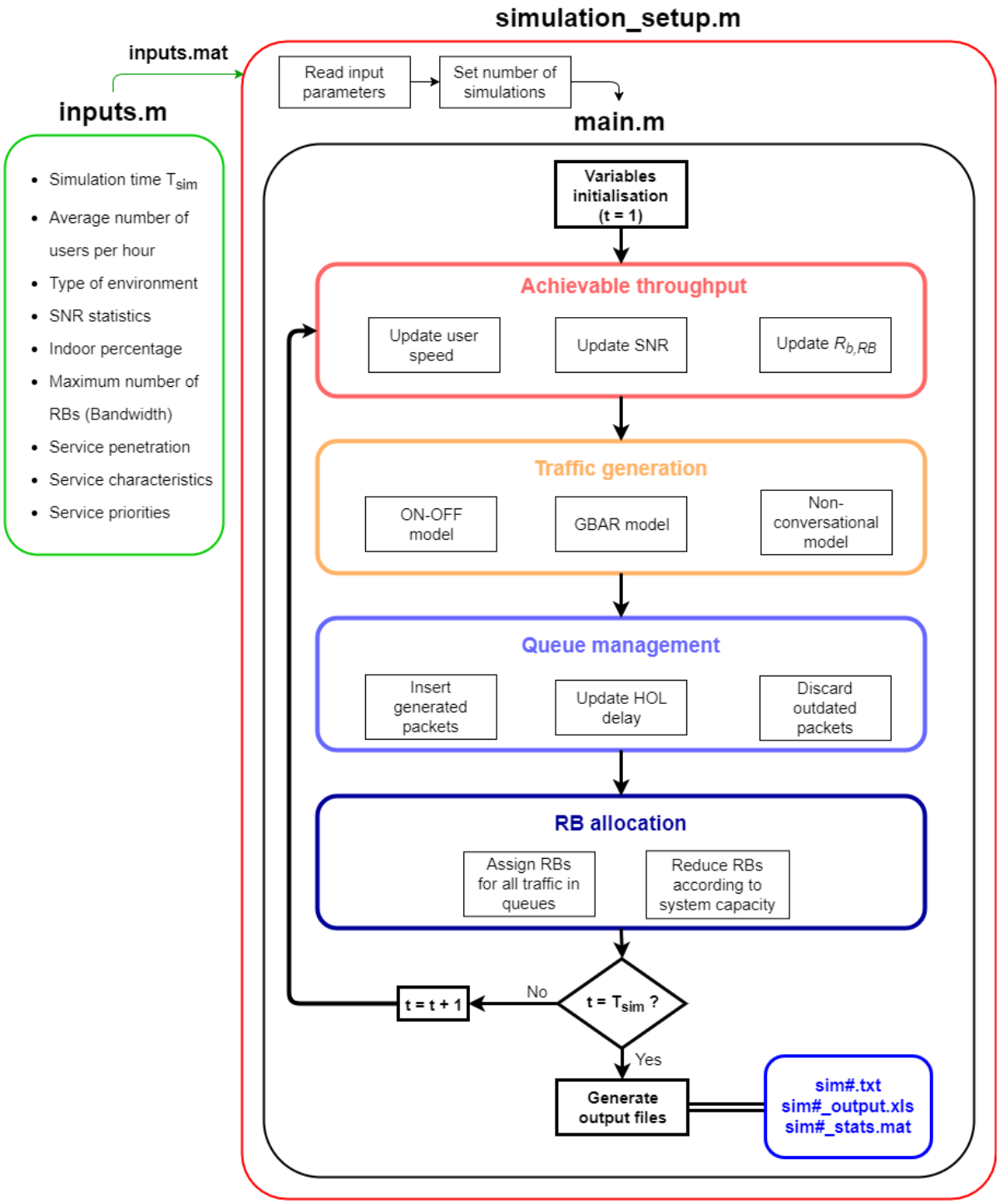


Figure 3.9. Overall simulator architecture.

After packets were generated according to the multiple traffic source models, the queue management block is responsible for updating the state of user queues. Generated packets are inserted in the end of each queue, and the HOL delay is updated according to the current simulation time. Finally, this block searches for packets that are exceeding their maximum allowed queuing delay and discards them. At this point, packet failure statistics are updated to account for the number of lost packets.

Once all the generated traffic is queued, it is ready to be scheduled through RB allocation. In a first stage, the estimation of cell load is based on the number of RBs needed to allocate all the queued traffic. When capacity is exceeded, the algorithm performs RB reduction, which corresponds to the strategy described in Section 3.2. After the number of RBs was determined taking cell capacity into consideration, the RBs are allocated to the corresponding users and all the output statistics are updated, namely the user throughput, average delay, and the number of allocated packets. Queues are updated to account for the allocated traffic.

At the end of each simulation, output files are generated with different information gathered from the data structures used in the program. Post-processing calculations were performed using Microsoft Excel for easier statistical and graphical analysis. Three output files are directly generated in *main.m* as follows:

- *sim#.txt*. It is mostly used for debugging and simulator assessment purposes as it includes a register in every time instant of information related to the instantaneous number of active users for each service, the experienced throughputs and the size of the user queues.
- *sim#_output.xls*. It contains the most relevant statistics of each simulation distributed by four separate sheets. The first one provides details about the input configurations of the simulation and also information related to its computational effort, namely the total simulation time, the total number of users and the average number of active users. The second one shows the main results for the output parameters for each service, such as throughput or average delay among others. The third sheet has information for each user namely throughput, queuing delay, packet failure and, in the VoLTE case, end-to-end delay and MOS. It also includes information about the users' environment, speed and service as well as the total cell throughput. Finally, the fourth sheet refers to information about traffic generators, namely the total generated traffic and the users' bit rate and average experienced SNR.
- *sim#_stats.mat*. It corresponds to the MATLAB workspace in the end of each simulation, including all the data structures and parameters used. This file includes additional results that can be post processed for more details on the statistics obtained in the other output files.

3.5 Simulator assessment

This section describes the procedures taken in order to assess the simulator before starting with the generation of output results through simulations, and the corresponding results analysis that are presented in Chapter 4. Several tests were done beforehand to guarantee that the input configurations are working properly. First, aspects related to the user environment and the calculation of the achievable throughput were assessed. One checked that SNR follows the desired Log-Normal Distributions, and that the achievable RB throughput in each TTI is correctly computed and changes at each coherence time, while keeping a constant value throughout this period. Speed is also periodically updated while changing the coherence times accordingly. Figure 3.10 shows a time sample of the variation of the

achievable throughput per RB, for a user with pedestrian speed in an outdoor urban environment, where one can observe the random behaviour and the intervals in which throughput is constant, which correspond to the coherence time in that interval. The validation of the random number generators used for SNR Distributions and for user speeds according to the mobility model can be found in Annexes A.2 and C.1, respectively.

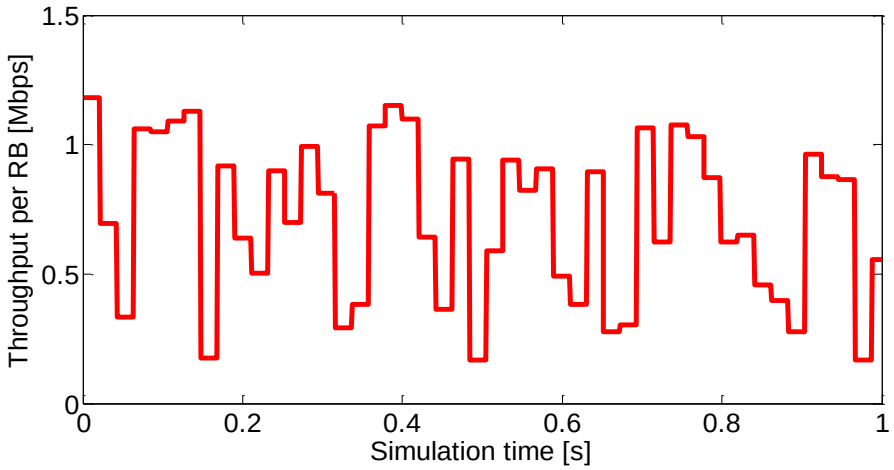


Figure 3.10. Time sample of the achievable throughput per RB variation.

Traffic generators were also assessed to ensure that the three traffic source models are working properly. For that purpose, simulations with large numbers of users from each service were done in order to check the distributions for the multiple parameters that characterise the services, from the call durations of VoLTE users to the file volumes of the file transfer users, for example. All these results can be found in Annex B.6.

Queues for DL traffic were tested by generating traffic and defining a fixed value for the global throughput of the queues. One has observed that if the throughput is larger than the bit rate of the network generated traffic the queues maintain a stable behaviour by filling up and emptying without any sort of overload. On the opposite, if the value for the fixed throughput is lower than the bit rate of the generated traffic the capacity of the queues is exceeded.

In terms of resource allocation, one verified that the radio channel bandwidth allows the allocation of the expected number of RBs per TTI, namely 50 RBs for a 10 MHz bandwidth and 100 RBs for a 20 MHz one. The LTE peak data rate was also tested by considering a scenario with a single static user in very good radio conditions with a fixed SNR of 40 dB and 100 RBs during the entire simulation period. Under these conditions, the achieved throughput is 119.6 Mbps which is coherent with the empirical expressions described in Annex A.1 for a 64-QAM modulation with a 3/4 coding rate. With a fully functional simulator, simulator parameters are configured in order to establish the statistical relevance of the simulation outputs. As users are considered to be arriving at an average arrival rate described by a Poisson process, the number of simultaneous active users is monitored to check its behaviour. Figure 3.11 shows the system behaviour in terms of the simultaneous number of users during a 75 minutes period.

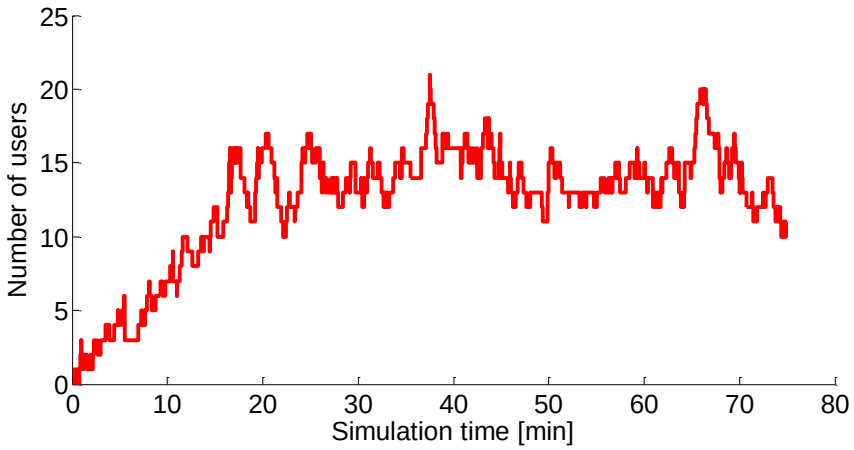


Figure 3.11. Time evolution of the number of simultaneous active users.

The minimum simulation time is set to one hour. This value allows enough time to achieve a statistically relevant sample of the system time evolution. As there is no traffic in the system during the transient simulation period, the initial instants (corresponding to the transient interval) must be neglected until it stabilises. According to the observed results, one decided to neglect the first 15 minutes of simulation to fully filter all the transient effects. In conclusion, the simulation time is set to 75 minutes from which 15 minutes are excluded and 60 minutes (after the transient effect) are used for results analysis. In order to estimate the adequate numbers of simulations needed to provide results with statistical relevance, a set of preliminary simulations considering the reference scenario were performed to identify the best compromise between the range of results variation and simulation time. It was expected in advance that the required number of simulations would be low as each simulation already covers a statistically significant time window. Figure 3.12 shows how the average number of simultaneous users during a one hour simulation changes for five different simulations. One observes that the average values for the average number of simultaneous active users per simulation do not have a deviation from the mean value larger than 5% for all the simulations.

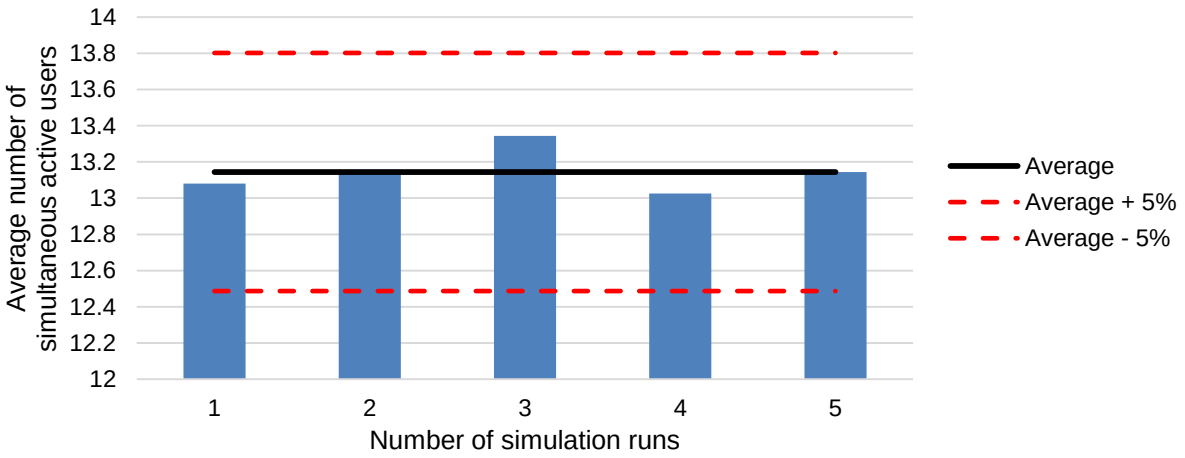


Figure 3.12. Cumulative average number of simultaneous users in the system.

In terms of the global throughput during each simulation, the average values have a larger variation as depicted in Figure 3.13. As some variation is experienced in the average results for the first three

simulation runs, one assumes that three simulations are enough to correctly describe a simulation scenario, as they ensure an error lower than 10% relative to the average value.

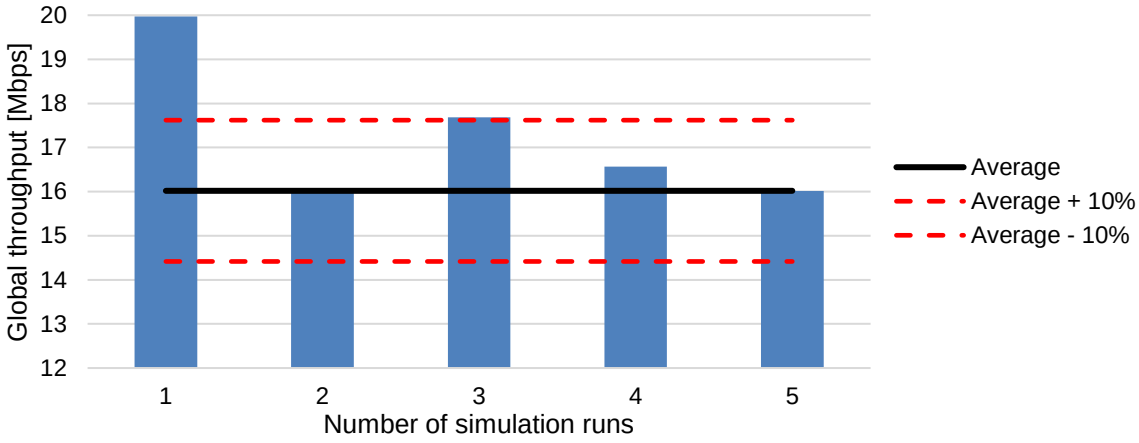


Figure 3.13. Cumulative average global throughput.

With the complete setup for the simulator, a set of simulations for the reference scenario with different numbers of users was done in order to assess the services priorities and the percentage of satisfied users for each service, as shown in Figure 3.14. Results for VoLTE, video calling and video streaming are overlapped as approximately all users are satisfied for this range of user arrival rates. A detailed description of the reference scenario considered for these simulations can be found in Section 4.1.

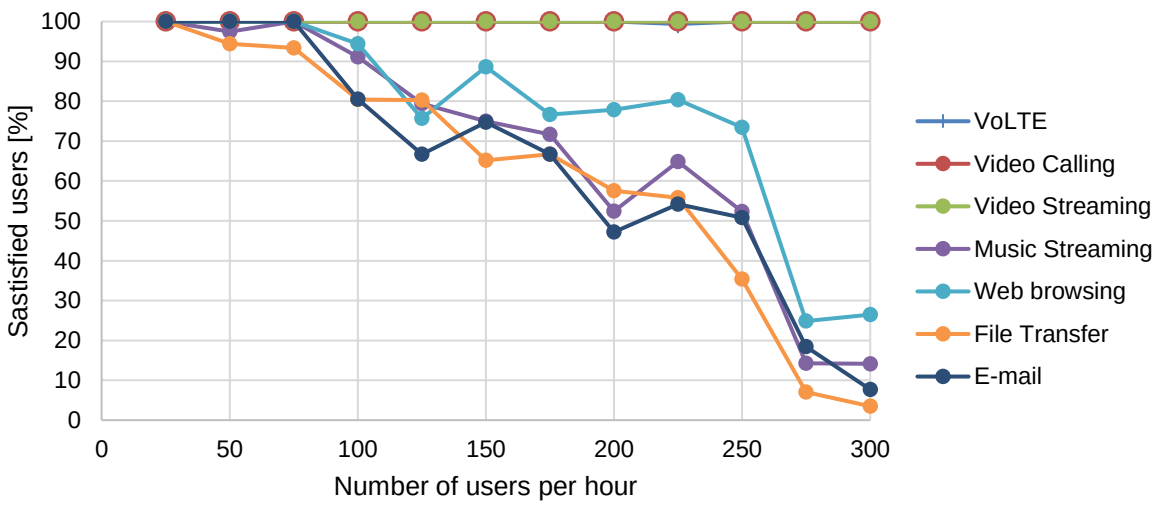


Figure 3.14. Percentage of satisfied users per service for different numbers of users.

The percentage of satisfied users clearly decreases for the lowest priority services, more specifically for the services with a priority lower than video streaming, which is the most demanding one in terms of the required bit rate. Voice and video calling, which are the highest priority services, have a 100% level of satisfaction, appearing alongside with the video streaming service for all the average numbers of users tested. The large variation verified for a number of users higher than 150 is related to the stabilisation of the cell throughput, which is fully discussed in Section 4.3.

The required processing time is also a crucial factor in order to schedule the 168 simulations needed

for this work. The complexity of each simulation depends fundamentally of the number of users and the resulting need for network optimisation, which is also related to the considered radio channel bandwidth as it basically determines network capacity. To understand the involved computation times, Figure 3.15 shows the relationship between the number of users and the corresponding average values of real-time consumption, where one observes an approximately linear behaviour. The total computational effort required for all simulations consisted of approximately 288 continuous hours (about 12 days) of processing for a single computer with a 2.3 GHz dual core processor and 6 GB of Random Access Memory (RAM).

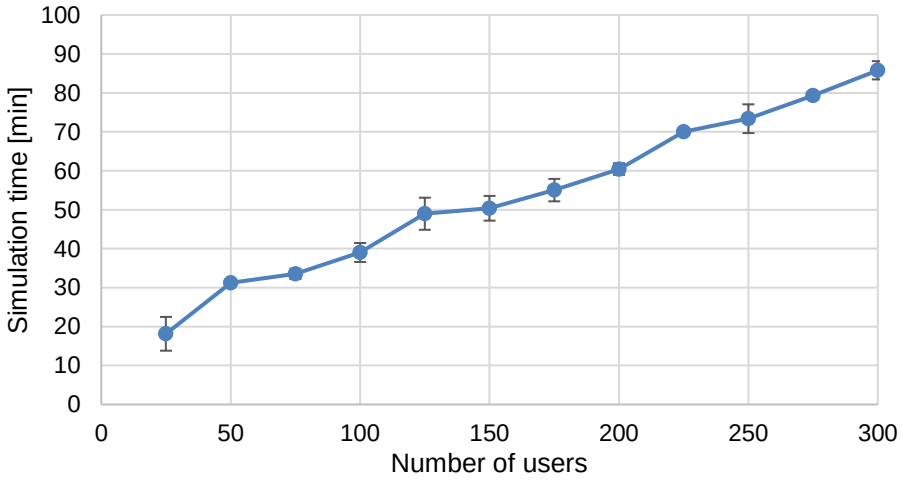


Figure 3.15. Simulation time in terms of the number of users in the cell.

Chapter 4

Results Analysis

This chapter starts with a description of the reference scenario, characterised in terms of all its input configurations and service parameters. The simulation strategy followed for the thesis is also described and, after that, the results obtained are presented and the corresponding analysis is made.

4.1 Reference scenario

The reference scenario for this study is a generic cell operating in an urban environment, assuming that 80% of the users are indoor. This approach intends to be as generic as possible, regardless of the geographic location of the cell, allowing this work to be adaptable to any case where the SNR statistical behaviour is available. All input parameters used for the simulation of the reference scenario are detailed in Table 4.1.

Table 4.1. Input configuration for the reference scenario.

Input parameter	Description/Value
Average number of users	150/hour
Type of environment	Urban
Percentage of indoor users	80%
Bandwidth	20 MHz
VoLTE codec	AMR-WB 12.65

Table 4.2 includes the characterisation of the seven different types of services being considered for the reference scenario, where all the four service classes can be found. Reference values for the minimum, average and maximum throughputs are also specified as reference guides to define QoS levels for these services. Each service has also the associated QoS priorities, with VoLTE having absolute priority above any other service. Two variations of the VoLTE service are presented as they represent the two voice codecs under analysis: AMR-WB and EVS. The service penetration for the reference scenario is also presented, where it is assumed that the majority of the users, 70%, performs voice and streaming services. Table 4.3 includes all the main service parameters considered for the reference scenario.

Table 4.2. Services characterisation (based on [Khat14] and [Guit16]).

Service	Service Class	Bit Rate [kbps]			QoS Priority	Service Penetration [%]
		Min	Avg.	Max		
VoLTE 12.65	Conversational	5.30	12.65	64.00	1	22
VoLTE 23.85			23.85			
Video Calling	Conversational	64.00	384.00	2 048.00	2	8
Video Streaming	Streaming	500.00	5 120.00	13 000.00	3	28
Music Streaming	Streaming	16.00	64.00	160.00	4	20
Web Browsing	Interactive	30.50	500.00	-	5	10
File Transfer	Interactive	384.00	1 024.00	-	6	8
E-Mail	Background	10.00	100.00	-	7	4

Table 4.3. Reference scenario service parameters (adapted from [Khat14] and [Dout15]).

Service	Parameter	Value
VoLTE	Average call duration [s]	60
Video Calling	Average call duration [s]	60
Video Streaming	Average video duration [s]	150
	Average number of videos per session	3
Music Streaming	Average music duration [s]	150
	Average number of songs per session	3
Web Browsing	Average main object size [kB]	10.71
	Average embedded object size [kB]	7.758
	Average number of embedded objects	5.64
	Average reading time [s]	10
File Transfer	Average file size [MB]	2
	Average number of files per session	3
E-Mail	Average file size [kB]	100
	Average number of e-mails per session	1

The simulation strategy followed in this thesis consists of varying each input parameter and observing the corresponding impact on the output parameters. As the impact of the other services on VoLTE is negligible in terms of allocation capacity as it is the service with the highest priority, one defines a scenario with only VoLTE users in order to analyse the VoLTE call quality degradation. This analysis focuses on the impact of the number of users and bandwidth on the level of satisfaction of VoLTE users, allowing to take conclusions on the fact that capacity for voice users only is not an issue.

The analysis focuses then on the results obtained for the performance of the other services in the presence of VoLTE traffic. One analyses the impact of the variation of the number of users on the performance of the multiple services, by comparing the reference scenario with an additional scenario where VoLTE is the highest priority service and there is no distinction among the remaining data services. The main goal is to emulate conditions where a network operator deals with encrypted traffic in their networks for all its data services. There is a rise trend in encrypted mobile traffic [Cisc17] due to the wide adoption of Hyper Text Transfer Protocol Secure (HTTPS) with the objective of providing secure communication between web servers and clients. Finally, one also performs a variation of the number of VoLTE users, while fixing the number of data users in the reference scenario, to evaluate the degradation of the other services. The influence of the two voice codecs AMR-WB and EVS is also analysed.

An analysis of the reference scenario configuration is carried through rural, suburban and urban environments. The influence of the number of indoor users is also evaluated for three situations: 20, 50 and 80% of the number of users in the reference scenario. The impact of the service parameters on system performance is analysed. Users streaming videos with a longer average duration or downloading bigger files will certainly influence the load of the system. Service penetration is expected to have a

similar effect, as different numbers of users performing each service might mean for example more users performing a highly demanding service like video streaming and consequently an increased load over the network. Finally, different scenarios in terms of service penetration are also tested, considering two additional scenarios: the first one, which is Video centric and assumes a higher percentage of video streaming users, and a second one, VoLTE centric, with more VoLTE users.

4.2 VoLTE quality

This section presents the analysis of VoLTE's call quality and how it degrades as a function of network capacity. VoLTE requires a low throughput and as it is considered as the highest priority service for scheduling purposes, it is beforehand expected that capacity for VoLTE users is not an issue under usual conditions. For the average SNR associated with each environment, VoLTE requires in the active state roughly between one and two RBs every 20 ms, which means that an LTE cell with, e.g., a 10 MHz bandwidth, can theoretically support up to 1 000 simultaneous active users without significant quality degradation. This is obviously an unrealistic situation in most cases, unless specific conditions with heavily crowded events are considered, like for example big stadiums or festivals.

For that purpose, one analyses a scenario with 100% penetration of VoLTE users, where the remaining scenario parameters are the same as for the reference scenario in an urban environment with 80% of indoor users. Figure 4.1 shows the variation of MOS for scenarios with average user arrivals rates ranging from 2 500 to 17 500 users per hour, for 10 and 20 MHz bandwidths.

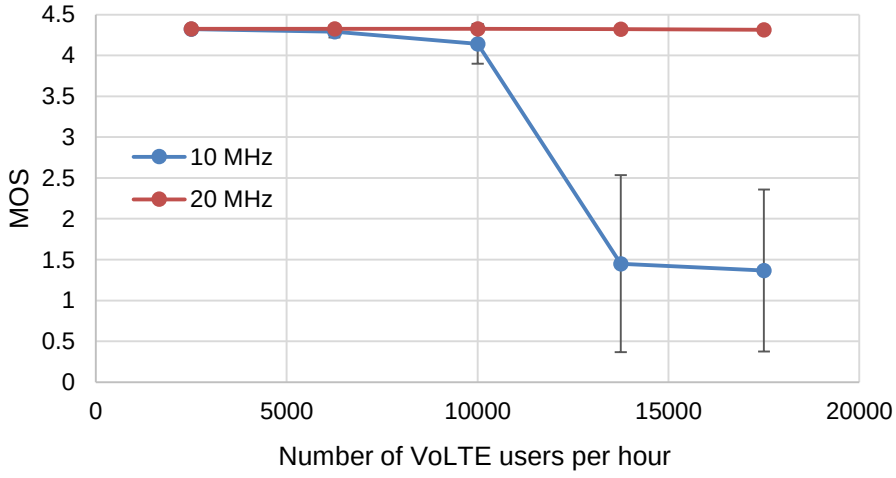


Figure 4.1. Average user MOS depending on the number of users in the cell.

From Figure 4.1, one concludes that for the 10 MHz bandwidth, MOS is abruptly reduced for a number of users per hour above 10 000, which, according to simulation results, corresponds to an average number of simultaneous users of 164. For the 20 MHz bandwidth, the value of MOS is practically independent of the number of users for the case being studied. To get further in deep on the analysis of satisfaction of VoLTE users, Figure 4.2 shows the distribution of the numbers of users in each MOS

category, reflecting the end user perceived quality. One presents the results for the 10 MHz bandwidth as the conclusions for the 20 MHz one are basically similar, but for a larger number of VoLTE users.

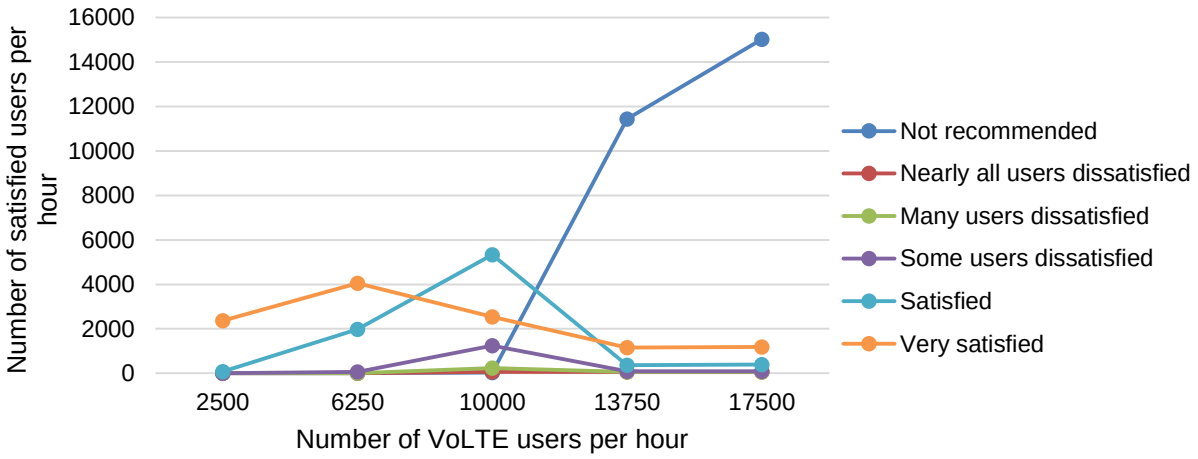


Figure 4.2. Call satisfaction according to MOS for a 10 MHz bandwidth.

Up to 10 000 users per hour, the vast majority of the users have a MOS that reflects that they are “Satisfied” or “Very satisfied”, and only a small portion, 1 565 users more specifically, reports ”Some users dissatisfied”. For a larger number of users, the call quality quickly becomes “Not recommended” going along with the degradation in terms of average MOS. Basically, the degradation is associated with the fact that as the network capacity is reached, the queuing delay of the packets increases and with it also the packet failure. For a matter of a better understanding, a variation on the number of users from 10 000 to 13 750 users implies that the average estimated end-to-end delay goes from 108.8 to 164.3 ms, which gets significantly closer to the advisable maximum value of 200 ms for voice communications.

4.3 Number of users

In this section, one analyses the influence of the variation of the total number of users in the cell. One starts by characterising the overall behaviour of the reference scenario. Figure 4.3 shows the variation of the total cell throughput as a function of the total number of users. One studies throughput evolution for both 10 and 20 MHz bandwidths in a range of 25 to 300 users, which, as one can observe, is enough to reach a condition of stabilisation in terms of total throughput for both bandwidths.

Results suggest that in terms of total cell throughput, the reference scenario has a similar behaviour for both bandwidths for a number of users up to 125. At this point, the throughput stabilises for the 10 MHz bandwidth, while for the 20 MHz one it only stabilises at approximately 225 users. The stabilisation of the offered throughput for a scenario with a fixed configuration is an expected behaviour as network capacity is finite. This implies that as users are arriving at a constant rate at the cell, performing the same set of services according to a fixed service penetration, the maximum throughput the cell can offer

is a constant value. It is interesting to notice that this maximum throughput for the 10 and 20 MHz bandwidths is about 10 and 20 Mbps, respectively, corresponding to a spectral efficiency of one. That is not surprising since 80% of users are indoor ones, hence, experiencing low SNR values. An important aspect is to check how the reference scenario fits in this analysis. The reference scenario with 150 users and a total cell throughput of approximately 16 Mbps is not at the stabilisation point of the cell, which means that it does not correspond to a high load scenario.

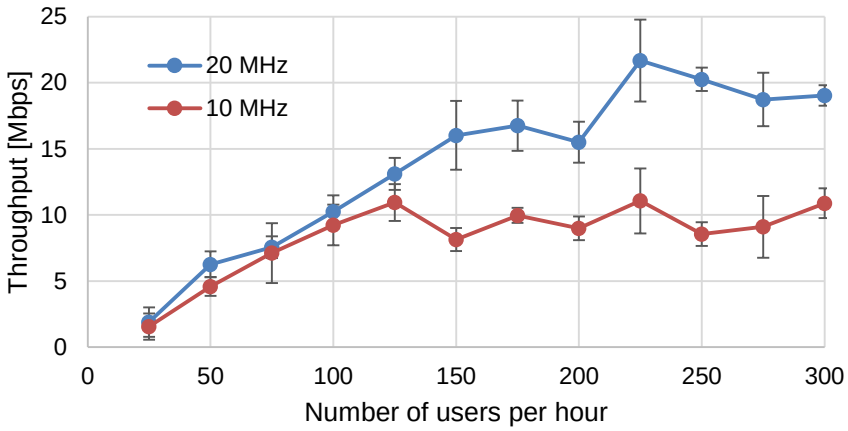
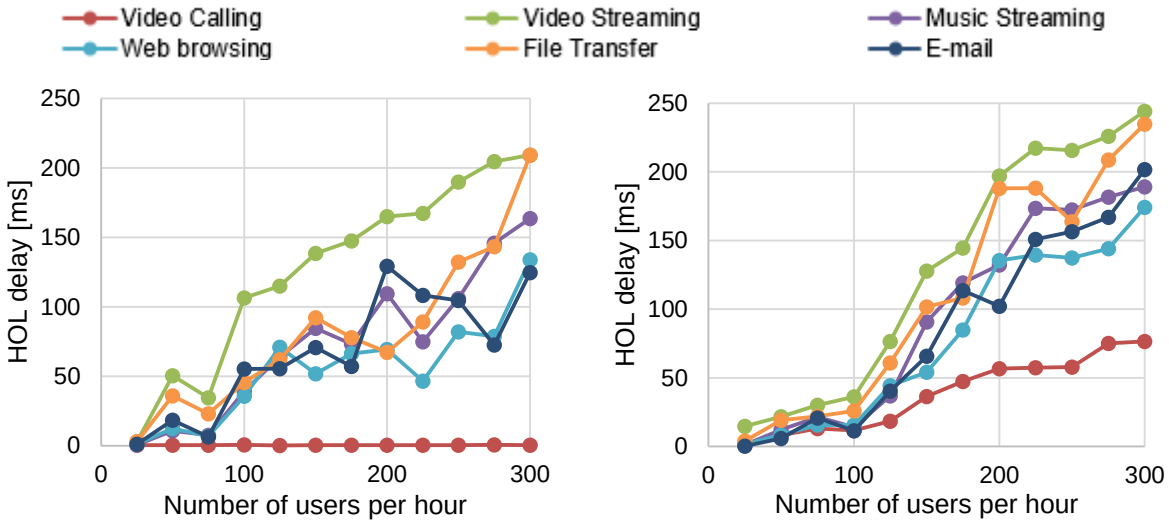


Figure 4.3. Total cell throughput for different numbers of users.

The analysis is further taken to the performance of the various services. As the system throughput tends to stabilise, this means that with the increase of the number of users, it cannot provide the expected average throughputs, resulting in delayed packets that in some situations get discarded. The way this affects the performance of services mostly depends on the nature of the service itself. For services that are non-GBR, like web browsing, file transfer and e-mail, their behaviour is characterised by high fluctuations of the average throughput and allow bigger values of delay. For GBR services, like video calling, video and music streaming or VoLTE itself, they commonly have a much lower variation of the average throughput, but are more delay sensitive. To analyse the evolution of the system in terms of delay for the multiple services, Figure 4.4 shows the variation of the average HOL delay for each service, comparing the reference scenario with a scenario where no priorities are defined among data services. As one has concluded that the performance of VoLTE is barely affected for the considered range of the number of users in the cell, one discards its analysis and focuses on its impact on the performance of other services.

For the scenario with prioritised data services, one observes that video calling, similarly to what happens with VoLTE, is not affected for these amounts of users. This is mostly due to the fact that both services have the highest priority and are shadowed by video streaming, which is the service with the highest priority besides them and the biggest share of users. Video streaming is also the service that demands more resources, because of its high average throughput. Together with the fact that it is a GBR service, it makes it more susceptible to delay as the high amount of generated traffic means a bigger delay for packets queued due to congestion. From 275 users onwards, video streaming reaches a HOL delay of more than 200 ms apart from small fluctuations, which becomes close to the recommended delay budget of 300 ms for non-conversational video. Music has a similar behaviour, but the much lower

throughput makes it less susceptible to delay. Non-GBR services all have similar values of delay, which reach close to 200 ms in the worst case for file transfer. These values of delay do not have a significant impact in the performance of these services, as they are not aimed at providing information in real-time.



(a) With priorities among data services.

(b) Without priorities among data services.

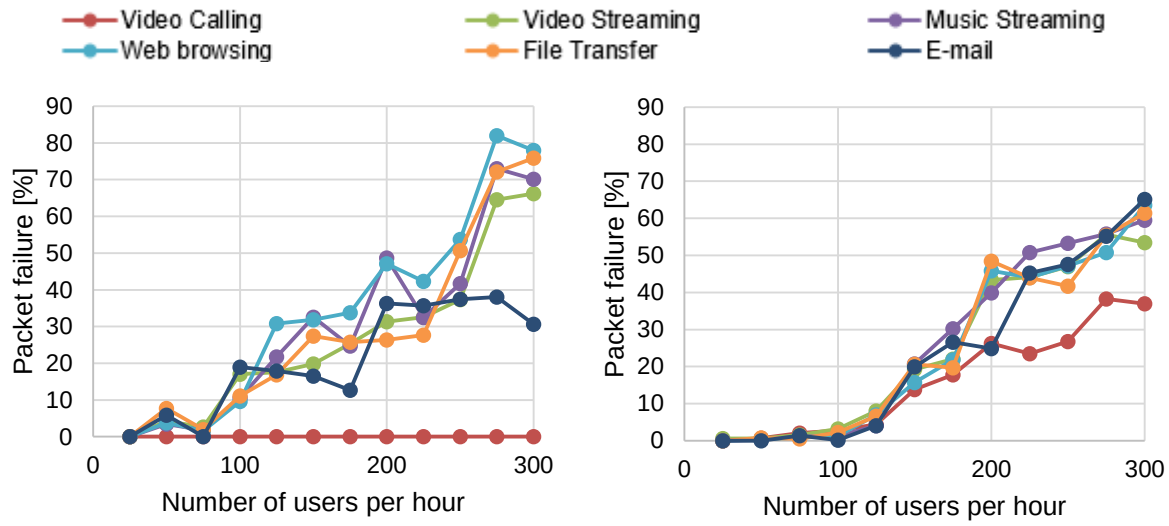
Figure 4.4. Average HOL delay for different numbers of users.

When no priorities are considered among data services, video calling is naturally affected, but due to the relatively low number of users performing the service, HOL delay does not exceed 80 ms in the worst case. In terms of video streaming, it would be expected that without priority over other services, video packets would have a higher average delay. This is verified when the network load increases due to the increase on the number of users per hour. For lower loads, resources are distributed more evenly among all services, resulting in a lower delay and a better exploitation of the achievable throughputs. The remaining services also tend to have a similar behaviour, with higher HOL delay for higher loads.

It is relevant to mention that considering priorities among data services causes HOL delay to degrade faster at lower loads. Up to 75 users, all the services have low HOL delay values, below 50 ms and bigger variations are noticeable between 75 and 100. With no priorities, HOL delay values above 50 ms are only verified for more than 125 users per hour. Figure 4.5 shows the results obtained in terms of packet failure per service for the scenarios with and without priorities among data services.

Packet failure tends to follow the growth trend of HOL delay, as the increase of the average packet delay means that more packets exceed their delay budgets and get discarded. However, even with video streaming being the more susceptible service to delay, higher packet failure percentages are verified for lower priority services when priorities among data services are considered. As these services generate less data and have lower priorities, the probability of losing packets due to excessive packet delay increases compared with video streaming, which allocates a higher percentage of packets due to the higher priority. Interestingly, when no priorities are considered among data services, most of them have similar values of packet failure and they are considerably lower than in the scenario with service priorities. Even though packets stay in queue for a longer time, the fact that video streaming is no longer

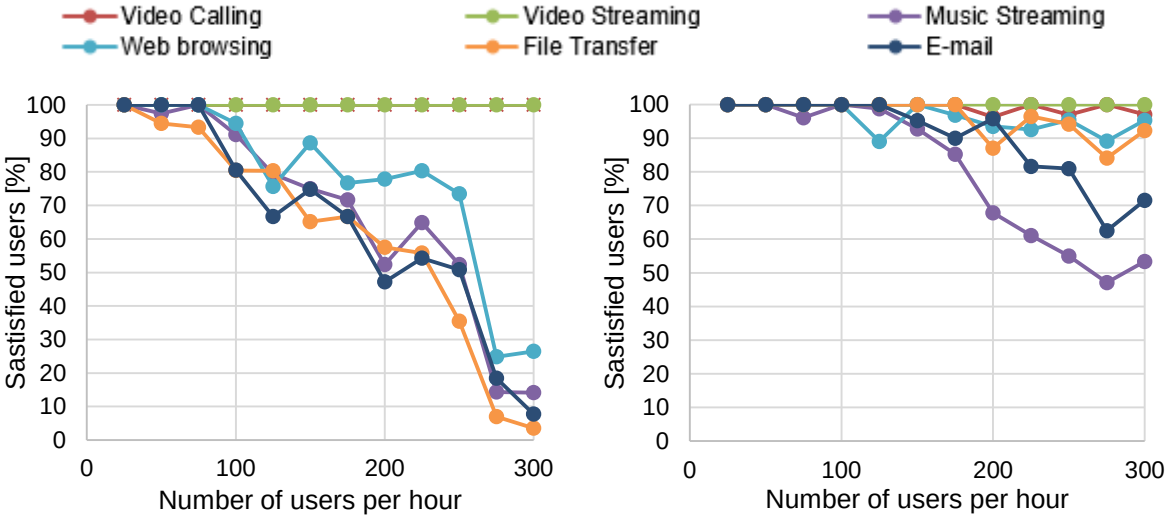
a high priority service means that the capacity for packets from other services increases. Regarding user satisfaction, Figure 4.6 shows the variation of the percentage of satisfied users for each service, for the scenarios with and without priorities among data services.



(a) With priorities among data services.

(b) Without priorities among data services.

Figure 4.5. Packet failure per service for different numbers of users.



(a) With priorities among data services.

(b) Without priorities among data services.

Figure 4.6. Satisfied users per service for different numbers of users.

The results show that guaranteeing a satisfaction of throughout requirements for all services of at least 90% is only possible for an arrival rate of 75 users per hour when priorities are considered. After that, user satisfaction degrades with the lowest priority services being the first to suffer in terms of throughput. The exception is music streaming where the results show a more emphasised decrease on user satisfaction than in web browsing which has a lower priority. This is basically justified by the fact that the

music service has much stricter throughput requirements compared to web browsing. It is a direct consequence of the totally distinct types of traffic of both services. Music streaming requires an approximately constant bit rate that should not suffer much degradation, while web browsing only requires the delivery of data objects with low losses and the experienced throughput is not a major issue.

For the scenario with no priorities, the results reflect the low levels of packet failure for higher loads. It is possible to ensure a 90% throughput satisfaction in all services up to an arrival rate of 150 users per hour. In this scenario, capacity for higher priority services is essentially taken by lower priority ones. Because of this, services like video calling and video streaming suffer a bigger impact in terms of throughput. However, results show that for the simulated range of users, this is not enough to significantly degrade the number of satisfied users for these services. Some impact is verified in the case of video calling but the percentage of satisfied users never drops below 90%. Video streaming is not affected due to the large difference between its average and minimum throughput values.

Another approach to analyse the influence of the number of users is to fix the number of users performing all services in the reference scenario and to change only the number of VoLTE ones. This makes it clear on how service's performance is degraded with the introduction of VoLTE in the context of an LTE network with no native voice service and priorities among data services. Table 4.4 shows the conversion between service penetration percentages and the actual average number of users arriving at the cell for each service.

Table 4.4. Average number of data users in the reference scenario.

Service	Service Penetration [%]	Average number of users/hour
VoLTE	22	33
Video Calling	8	12
Video Streaming	28	42
Music Streaming	20	30
Web Browsing	10	15
File Transfer	8	12
E-Mail	4	6

These are the values that were fixed apart from the VoLTE case. The number of VoLTE users was varied from the reference scenario with 33 users up until 5 000 users in order to show the high number of users that is required to observe significant degradation on services' QoS. Figure 4.7 shows the percentage of the total generated traffic during one hour that corresponds to VoLTE traffic. One compares results using both AMR-WB and EVS codecs. As the considered mode for the EVS codec has a source bit rate that is roughly two times the one for AMR-WB, network generated traffic for EVS is also approximately the double compared to AMR-WB, for the tested range of users.

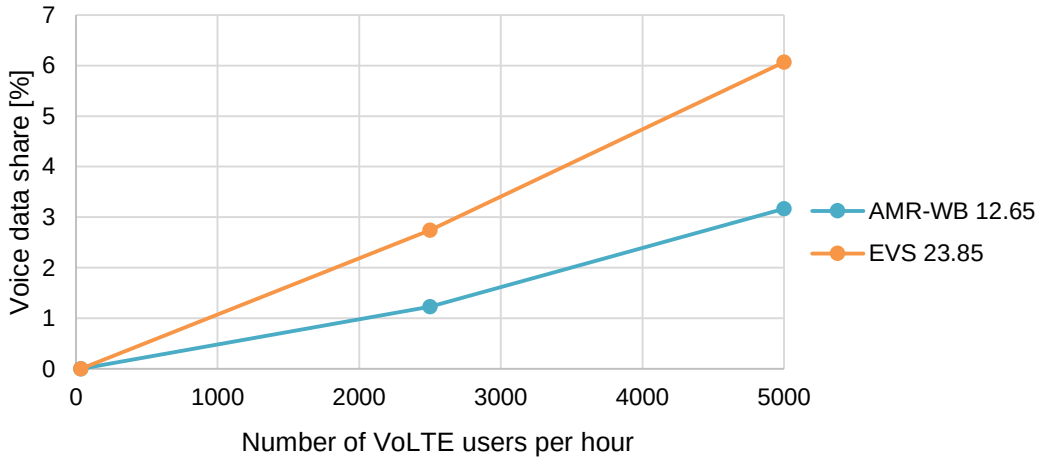
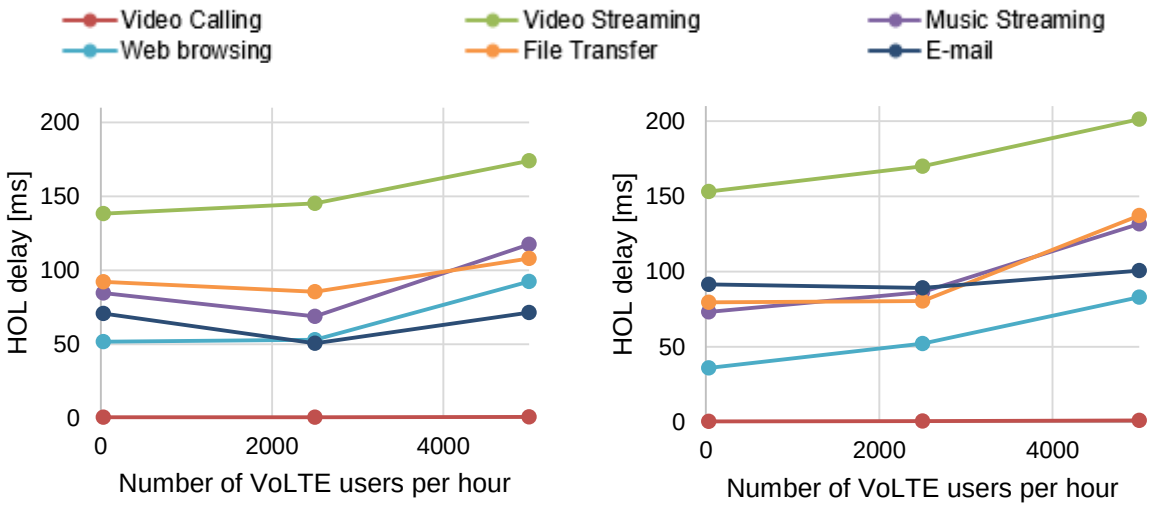


Figure 4.7. Percentage of voice traffic for different numbers of VoLTE users.

The most important aspect to observe is that even for a very high number of users and with a high quality voice codec, the percentage of voice traffic does not exceed 6% of the total network traffic under a regular scenario in terms of data services penetration. Assuming a situation where VoLTE is deployed over an existing LTE network, one does not expect a significant performance impact if the service usage pattern remains similar to the one observed in older voice technologies. The remaining contents of this section describe the performance parameters for the various services in order to assess their performance degradation. Figure 4.8 shows the behaviour of the system in terms of queuing delay for different numbers of VoLTE users considering the two codecs under analysis in this study.



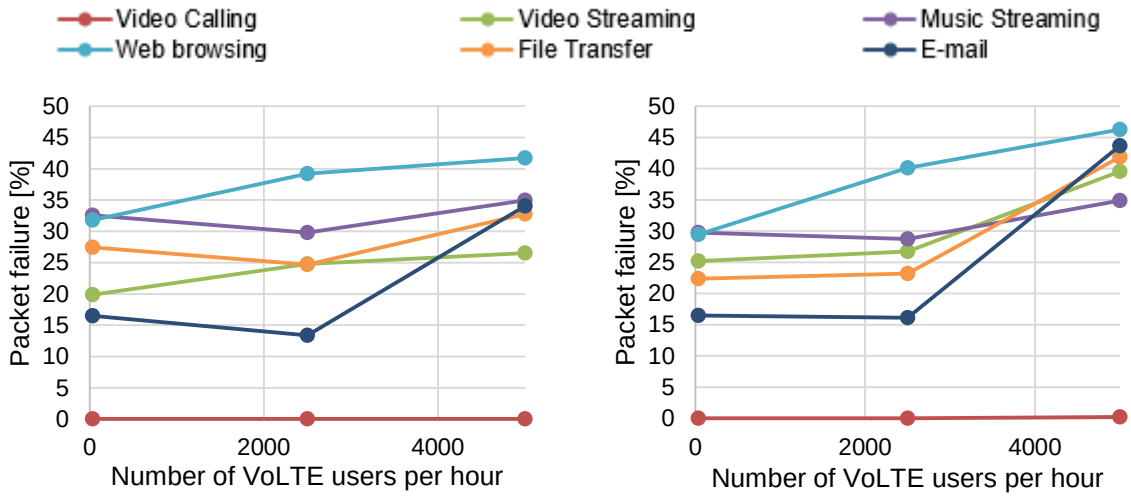
(a) AMR-WB codec.

(b) EVS codec.

Figure 4.8. Average HOL delay for different numbers of VoLTE users.

As one can observe, the impact in terms of delay is only noticeable between 2 500 and 5 000 users per hour, for both codecs. In some cases, namely in music streaming and e-mail, a slight decrease in delay is verified between 33 and 2 500 VoLTE users per hour. As there is no significant effect on the average HOL delay for these services, this variation on the obtained results falls within the simulator's margin of error. Naturally, the EVS codec due to its higher bitrate has a bigger impact in the other services. Due

to its high throughput and priority, video streaming shows the most noticeable impact. For this service, the difference between codecs reaches 27 ms, for a rate of 5 000 VoLTE users per hour. For this user rate, file transfer and music streaming are the services with biggest differences between the two codecs. Figure 4.9 shows the behaviour of the system in terms of packet failure for different numbers of VoLTE users considering the two codecs under analysis in this study. As all services have HOL delay values that are not close to their respective delay budgets, packet failure does not present big variations.



(a) AMR-WB codec.

(b) EVS codec.

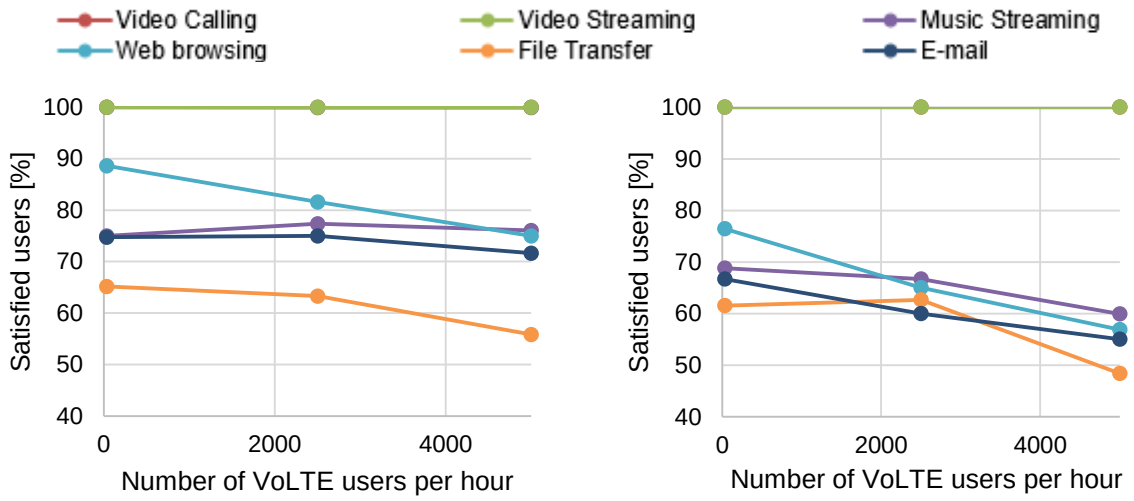
Figure 4.9. Packet failure for different numbers of VoLTE users.

Packet failure barely changes when the number of VoLTE users increases to 2 500 per hour, for both codecs. As the amount of voice traffic is low, this metric presents high variability especially noticeable in low priority services like e-mail. A growth trend is observed between 2 500 and 5 000 VoLTE users per hour for all services except video calling as a consequence of the slight increases observed in queuing delay. It is also observed that this trend is higher for the EVS codec.

Finally, user throughput satisfaction is analysed in order to assess the real impact that voice traffic has on services' throughputs. For that purpose, Figure 4.10 shows the behaviour of the system in terms of satisfied users for different numbers of VoLTE users considering the two codecs. Video calling appears superimposed with video streaming, as both these services are not affected in these conditions. One observes for this metric that significant changes only arise between 2 500 and 5 000 VoLTE users per hour, with the exception of web browsing. The high number of generated packets contributes to a high packet failure ratio and, consequently, to a reduction of the experienced throughput.

Video calling and video streaming are not affected at all, as they are the highest priority services. The percentage of satisfied users decreases in general for all the other services. Comparing the results between the 33 VoLTE users' case and the 5 000 one, web browsing shows the worst impact with a 13% reduction in terms of satisfied users for the AMR-WB code and close to 20% for the EVS codec. This means that the impact on service performance is not very significant, even when a high number of VoLTE users is assumed, as the worst case implies a reduction of 20% on the number of satisfied users, for web browsing, which is a service that is not severely conditioned by throughput requirements.

Figure 4.11 shows how the average number of simultaneous VoLTE and data users varies in the range of 33 to 5000 VoLTE users per hour. The main goal is to understand how many concurrent users can perform VoLTE calls without significantly degrading other services.



(a) AMR-WB codec.

(b) EVS codec.

Figure 4.10. Satisfied users for different numbers of VoLTE users.

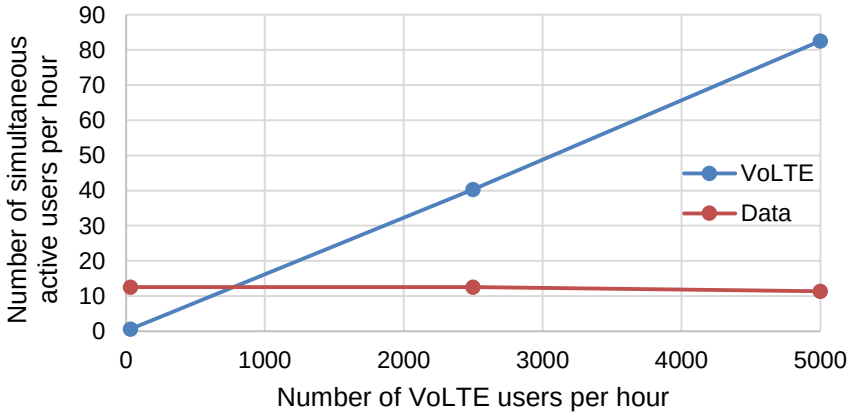


Figure 4.11. Number of simultaneous VoLTE and data users for different arrival rates.

Assuming a target of 10% as the maximum reduction on the percentage of satisfied users for all services, one can estimate a maximum number of simultaneous VoLTE users that the system can support. Results show that for the AMR-WB codec, the web browsing service suffers the higher degradation. In this case, one verifies that a 10% reduction of the satisfied users occurs for 3 800 VoLTE users per hour. Therefore, according to Figure 4.11, an approximate maximum of 60 simultaneous VoLTE users can be supported. For the EVS codec, web browsing is also the critical service. In this case, the 10% reduction in terms of satisfied users occurs approximately for 2 200 VoLTE users per hour, which corresponds to an approximate maximum of 36 simultaneous VoLTE users.

For a matter of comparison, if a higher margin, e.g., 20% for the reduction on the number of satisfied users was considered, more than 82 simultaneous VoLTE users could be supported for both codecs. It

is important to mention that these values correspond to a scenario with approximately 12 users (Figure 4.12) performing data services simultaneously and should not be interpreted as maximum theoretical values but rather a typical urban scenario.

4.4 Environment

The scenarios under analysis are based on a single cell operating on rural, suburban and urban single-cell environments. Table 4.5 shows the values used for the Log-Normal Distributions that characterise each of these environments, considering the average, μ, and standard deviation, σ, as defined in Annex A.2. These values were based on trial measurements performed by [Carr11] in different environments within an LTE cluster in the city of Porto. Besides the type of environment, each user in a given scenario can be characterised as indoor or outdoor. As the default values considered for SNR are for outdoor environments, indoor users are assigned an additional attenuation between 12 and 20 dB.

Table 4.5. Statistical parameters for SNR distributions.

Environment	**SNR**	
	μ **[dB]**	σ **[dB]**
Rural	13	6
Suburban	16	
Urban	18	

The type of environment heavily influences network capacity, as radio conditions directly define the achievable cell throughput. In this section, one splits the analysis into two main aspects. The first one refers to the characterisation of the cell environment, which can be rural, suburban or urban. This classification depends essentially on the average SNR value in each scenario. The second aspect deals with the number of indoor users in each scenario, as they get their experienced SNR heavily reduced due to indoor attenuation. Figure 4.12 shows how the total throughput offered by the cell changes for the three considered types of environment. While the difference between urban and suburban environments has only a decrease close to 11%, between urban and rural one observes a difference of approximately 6 Mbps, which corresponds to a 38% decrease.

The obtained results are directly related to how the achievable throughput for each RB changes according to the experienced SNR. According to the mapping of SNR to the achievable LTE throughout per RB presented in Annex A, if one considers the average SNR values for each environment, a reduction of 54% on the achievable throughput per RB is expected between urban and rural. As the values of SNR for the rural environment are more prone to using lower order modulations, like 16-QAM and QPSK, the difference between both environments in terms of throughput is attenuated, since for lower order modulations the achievable throughput decreases more slowly with the reduction of SNR,

compared to a higher order modulation, like 64-QAM. Once again, to get further detail into the performance of each of the services, Figure 4.13 shows the behaviour of the system in terms of queuing delay for the three types of environment.

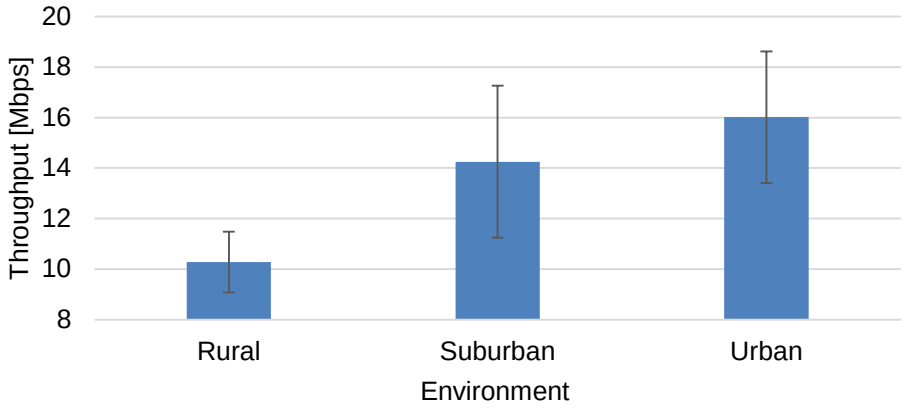


Figure 4.12. Total throughput for different types of environment.

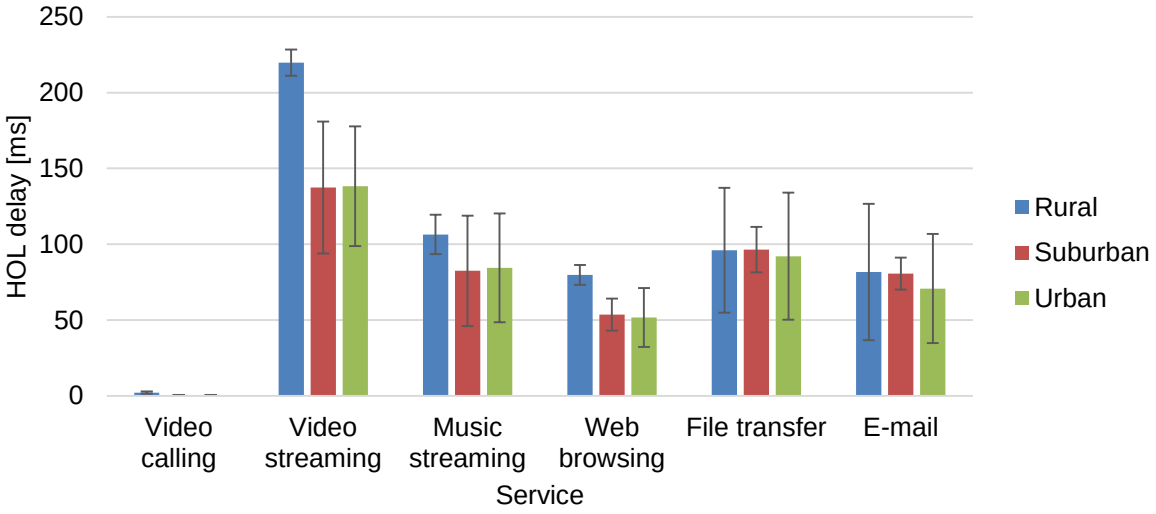


Figure 4.13. Average HOL delay per service for the different environments.

Even with a significant decrease of the achievable throughputs, non-GBR services, namely web browsing, file transfer and e-mail, are not deeply affected in terms of delay and all of them have HOL delays below 100 ms for the three environments. For GBR services, namely video calling, video streaming and music streaming, results do not show any significant distinctions between suburban and urban environments. Differences are more noticeable between the video and music streaming services on the rural case. Results show an increase of over 20 ms on the music streaming service but it is video streaming that suffers a greater impact. When one compares the rural environment with the suburban and urban ones, an increase of about 80 ms on the average HOL delay is verified. Figure 4.14 shows the behaviour of the system in terms of packet failure for the three types of environment.

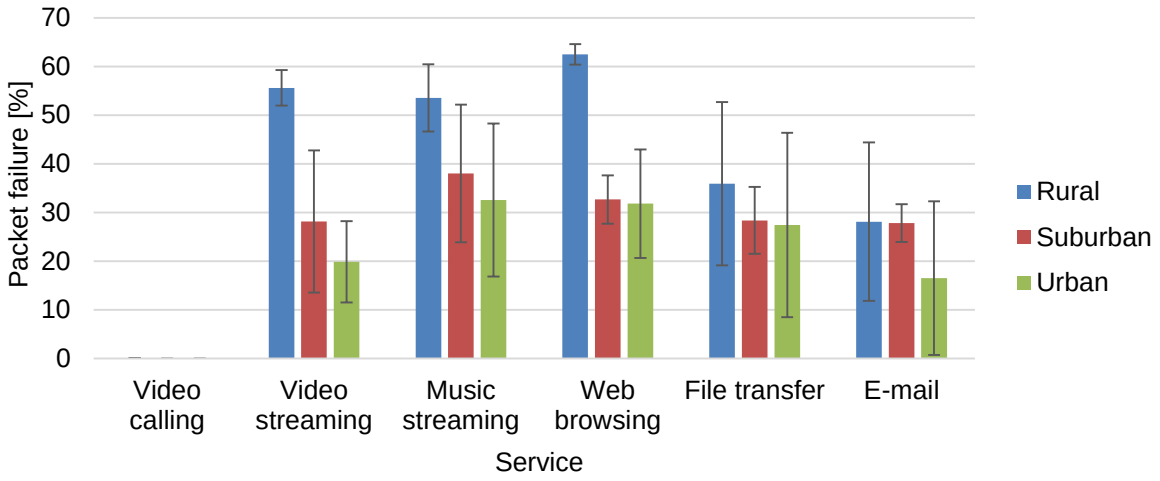


Figure 4.14. Packet failure per service for different environments.

In terms of packet failure, results do not show any significant difference between suburban and urban environments as observed for HOL delay, except for the case of video streaming and e-mail, where the observed difference is close to 10%. This is caused by the large amount of traffic generated by video streaming and the low priority of e-mail. The situation for the rural environment shows additional problems. Unlike delay, where only video streaming is more affected, packet failure for music streaming and web browsing increases 26% and 40% from the suburban to the rural environments. In the case of music streaming, the high number of users performing the service means a high number of generated packets, which, even if their size is small compared to video, are discarded due to the lower priority. A similar effect occurs with web browsing, but which is mostly due to the high number of generated packets as a consequence of the high number of requested data objects during a typical session. Figure 4.15 shows the behaviour of the system in terms of the satisfied users for the three types of environment.

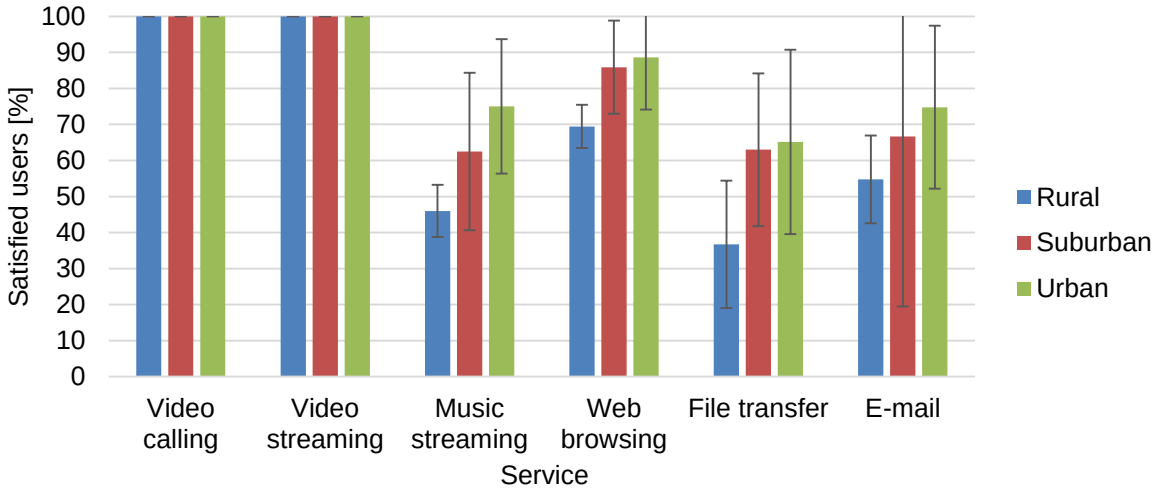


Figure 4.15. Satisfied users for different environments.

In terms of user satisfaction, no relevant differences exist between urban and suburban environments, unless for music streaming where the reduction of satisfied users reaches about 13%. For rural environments, the level of satisfaction is significantly lower in all cases, with the exception of video calling and video streaming services, which in these scenarios are not affected in terms of throughput

satisfaction. In the worst case, the percentage of satisfied file transfer users has a reduction of approximately 28% from the urban to the rural scenarios.

Similarly to the type of environment, one analyses the results when the percentage of indoor users is changed in each scenario. Figure 4.16 shows how the total offered throughput changes for the three different values of the percentage of indoor users, including the 80% that corresponds to the reference scenario.

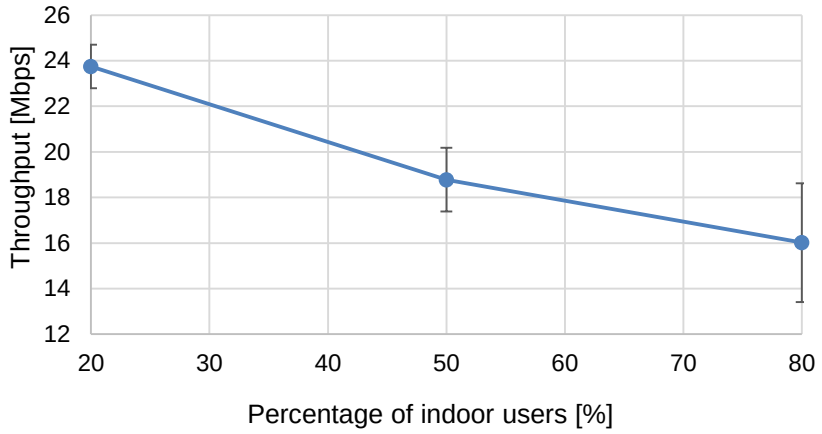


Figure 4.16. Total throughput for different percentages of indoor users.

One observes that there is an approximately linear variation of the total throughput in terms of the percentage of indoor users. There is almost a 50% increase on the total throughput when the percentage of indoor users is reduced from 80% to 20%. Figure 4.17 shows the behaviour of the system in terms of queuing delay for the three different values of the percentage of indoor users.

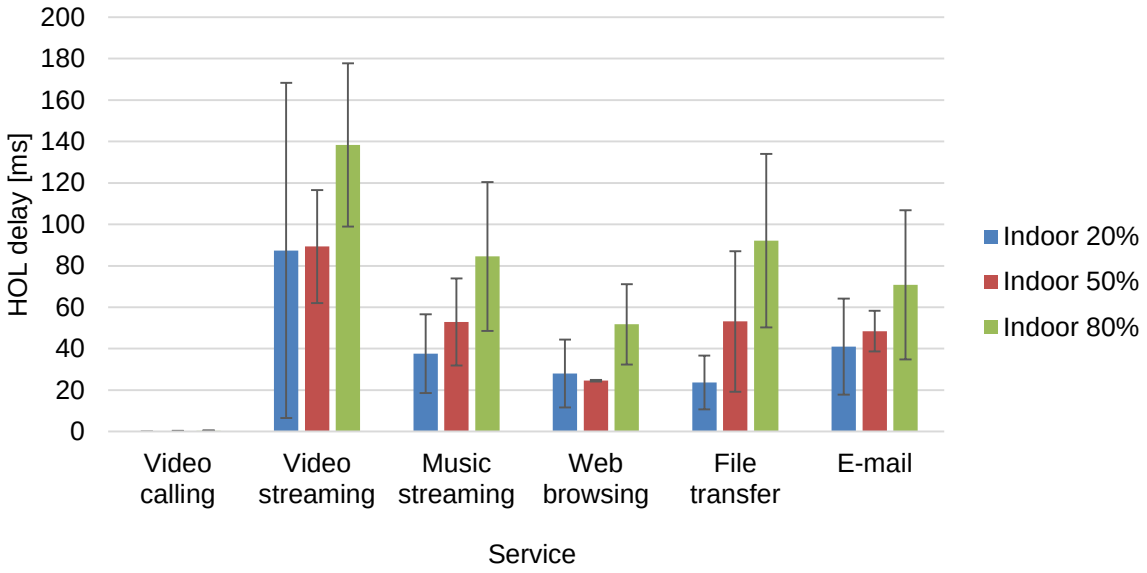


Figure 4.17. Average HOL delay per service for different percentages of indoor users.

The results in Figure 4.17 show that the variation of the percentage of indoor users impacts the average HOL delay in an unproportioned way. Big differences are verified between the 80% and 50% scenarios with special evidence, as expected, for the music and video streaming services. For instance, video and

music have a reduction of about 50 ms and 32 ms, respectively. However, this trend softens when one considers the 50% and 20% scenarios. The biggest difference is verified for file transfer, where the difference in between scenarios reaches 30 ms. Figure 4.18 shows the behaviour of the system in terms of packet failure for the three different values of the percentage of indoor users.

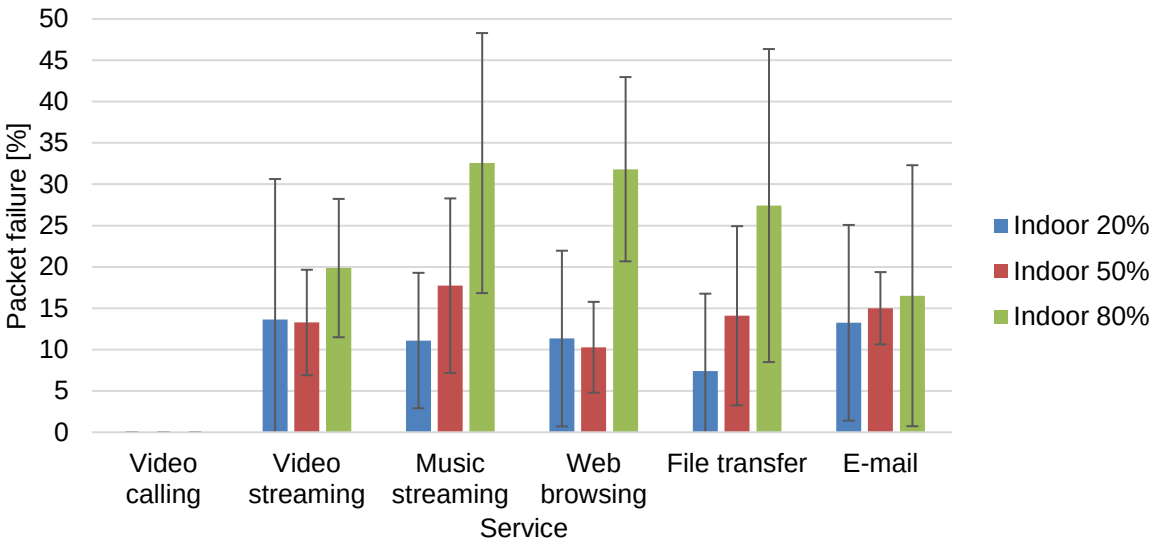


Figure 4.18. Packet failure per service for different percentages of indoor users.

Similar conclusions to the ones mentioned for HOL delay are taken for packet failure as big drops are verified between the 80% and 50% scenarios, and no significant changes are observed between the 50% and 20% scenarios. However, the most relevant aspect lies on the packet failure variation among services for the case of 80% indoor users; in this scenario, music streaming and web browsing have the highest packet failure. In general, having a majority of users performing services in indoor conditions means that the majority of users have increased probability of experiencing very low SNR values caused by indoor attenuation. This implies that these users experience significantly low achievable throughputs, which makes it harder for lower priority services to get all their packets delivered.

To understand how this behaviour translates in terms of satisfied users, Figure 4.19 shows the behaviour of the system in terms of the satisfied users for the three different values of the percentage of indoor users. Even with the results for the HOL delay showing that the packets' delay tends to have a significant decrease with the decrease of the number of indoor users, the same behaviour does not translate in terms of user throughput satisfaction. This suggests that, for the reference scenario configuration, the number of indoor users does not significantly impact services' performance. Nevertheless, the reduction of the number of indoor users theoretically suggests a performance improvement as a consequence of improving the achievable throughput of the users. This improvement is noticeable when one compares the 80% and 50% scenarios but becomes irrelevant between the 50% and 20% ones. For the case of web browsing, a slight decrease on the percentage of satisfied users is verified with the reduction of the number of percentage of indoor users. It must be stressed that for these cases one can say that the results are similar if one considers the simulator's margin of error as represented by the black vertical lines.

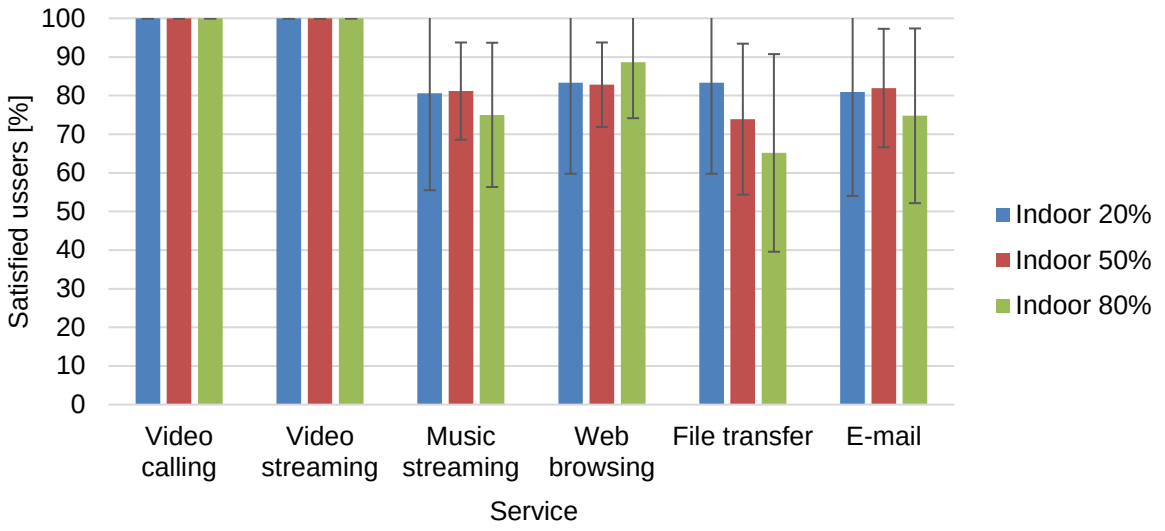


Figure 4.19. Satisfied users for different percentages of indoor users.

4.5 Bandwidth

In this section, one analyses the influence of the two considered radio channel bandwidths on services performance. Switching from the 20 MHz bandwidth, as in the reference scenario, to the 10 MHz one directly reduces the instantaneous available capacity to a half as exactly a half of the RBs are available for traffic allocation. As it is illustrated in Figure 4.3 (see Section 4.3), the total cell throughput is reduced to a half, from 16 to 8 Mbps for the reference scenario. This shows how all the allocated cell traffic is affected by this parameter. However, this should not be taken as a general case, as the cell throughput reduction depends on the network load. If the network is on a low load scenario, with for example a low number of users, the network performance tends not to be affected by the bandwidth, as shown in Section 4.3. To analyse the bandwidth influence for the reference scenario, Figure 4.20 shows the behaviour of the system in terms of queuing delay for the two bandwidths considered in this study.

Results show that for most services, only slight increases on the HOL delay are verified. The exception in this case is video streaming, which has an increase of 94 ms. Even though the service has a high priority, the amount of traffic that it generates in the network due to its high throughput makes it vulnerable to delay issues. However, as lower priority services do not have significant degradation in terms of HOL delay, this suggests a significant increase on the number of failed packets to make room for the incoming video streaming traffic. Figure 4.21 shows the behaviour of the system in terms of packet failure for the two bandwidths considered in this study.

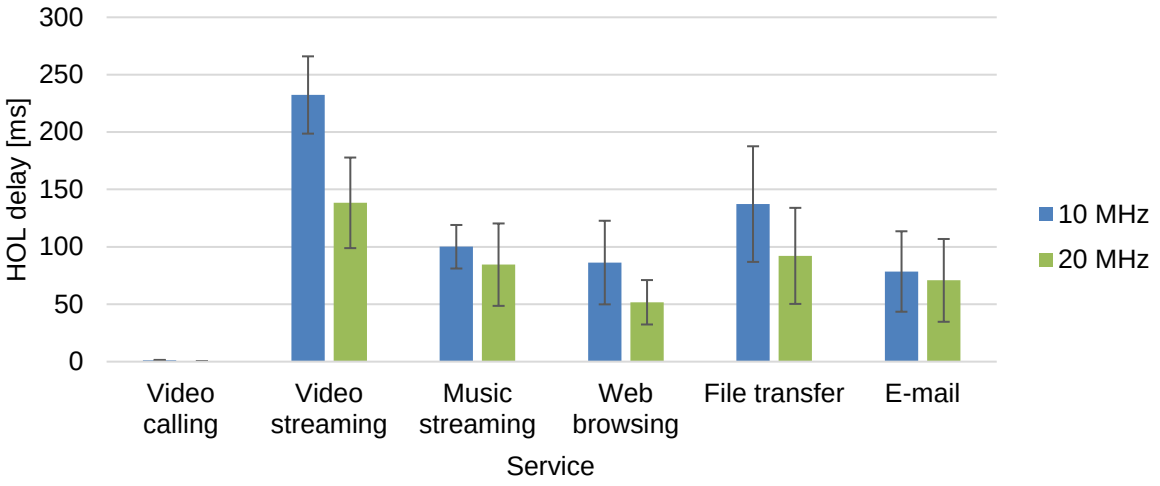


Figure 4.20. Average HOL delay per service for the different bandwidths.

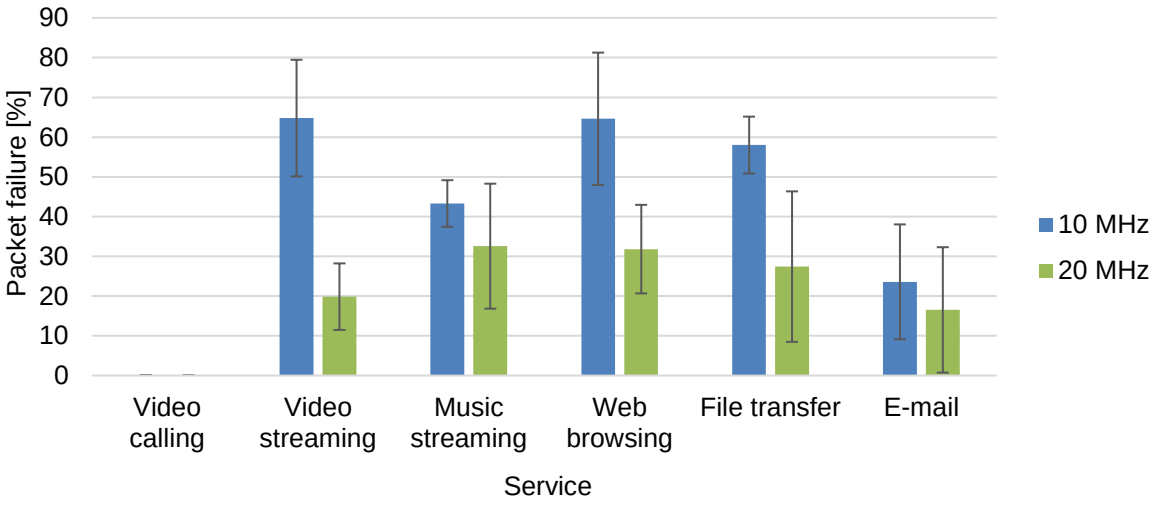


Figure 4.21. Packet failure per service for the different bandwidths.

Results are coherent with the analysis made for HOL delay values. Lower priority services, specially web browsing and file transfer, have an increase of 33% and 31% in packet failure, respectively, which happens because these are the most demanding non-GBR services, requiring higher amounts of data. Video streaming has an increase of 45% in packet failure due to the high average values of HOL delay which raises the number of packets close to the delay budget specified for streaming packets. Once again, this will seriously impact the provided QoS as it will require a significant rate reduction to control the number of lost packets. Figure 4.22 shows the behaviour of the system in terms of user satisfaction for the two bandwidths considered in this study.

While the 10 MHz bandwidth still does not impact the higher priority services, namely video calling and video streaming, results show that the remaining services suffer severe reductions in terms of throughput satisfaction. In the worst case, file transfer shows a reduction of around 40% in the percentage of satisfied users as a consequence of the high packet failure. Music streaming and web browsing are also similarly affected. E-mail, even though it is the service with the lowest priority, is not very affected, because it does not require as much data as the other services.

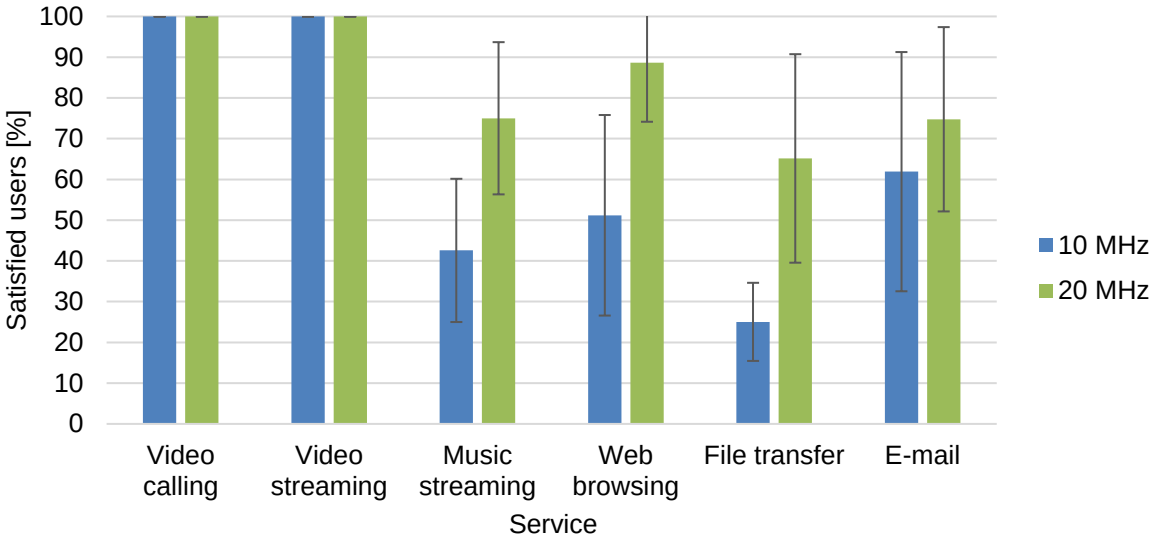


Figure 4.22. Satisfied users for different bandwidths.

4.6 Service parameters

This section describes the analysis of the results obtained for scenarios where the service parameters are changed. The main goal is to show how differences on the services' usage patterns influence system performance. It is predictable that for critical services like video streaming, considering different parameters like the average video duration will affect the performance of other services. Table 4.6 shows the service parameters that were changed from the reference scenario, and Figure 4.23 shows the system behaviour in terms of queuing delay for the four scenarios with different service parameters.

Table 4.6. Service parameters for alternative scenarios.

Service	Parameter	Reference scenario	Other scenarios
Video Streaming	Average video duration [s]	150	300, 600
Web Browsing	Average main object size [kB]	10.71	100
	Average embedded object size [kB]	7.758	80
File Transfer	Average file size [MB]	2	20

Increasing the average video duration causes an abrupt increase in terms of HOL delay for all services with no exception. This is an expected result as the amount of generated traffic by video streaming is two to four times higher compared to the reference scenario, for the 300 s and 600 s scenarios respectively. HOL delay for video streaming comes close to its delay budget, while for the others delay is lower as most packets fail to be transmitted due to congestion as depicted in Figure 4.24, which shows the behaviour of the system in terms of packet failure for the four alternative scenarios considered.

For the scenario where the average object size of the web browsing service was increased to 100 kB, results in terms of delay are similar to the ones obtained for the reference scenario. In the file transfer case with 20 Mbyte files, the biggest impact occurs for the e-mail service, as its average delay increases about 88 ms. This is a direct consequence of having a lower priority, which in this situation means less resources to allocate e-mail packets.

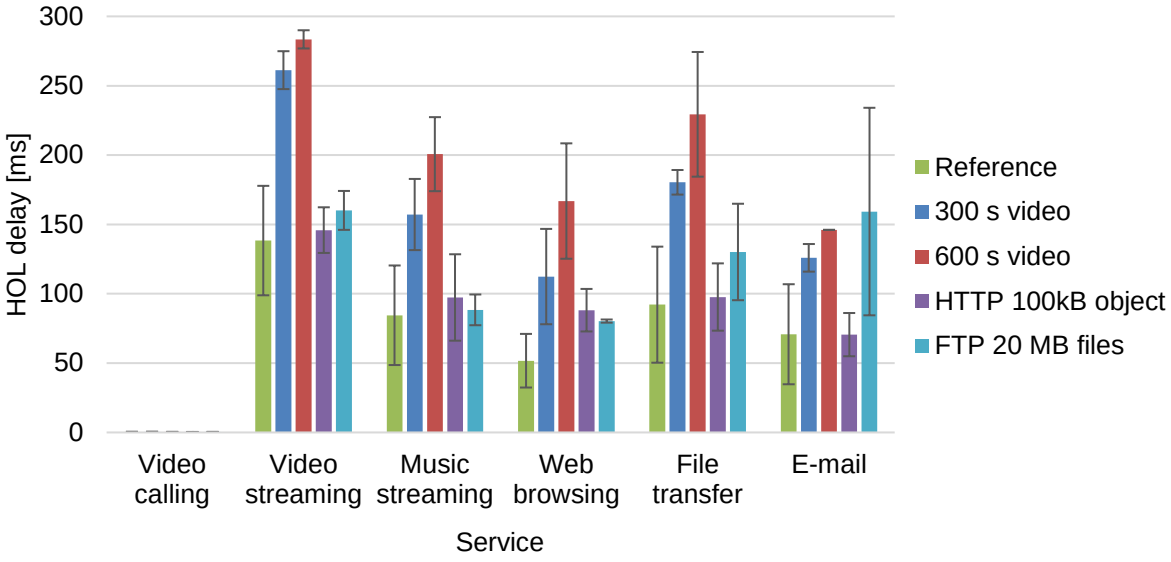


Figure 4.23. Average HOL delay per service for scenarios with different service parameters.

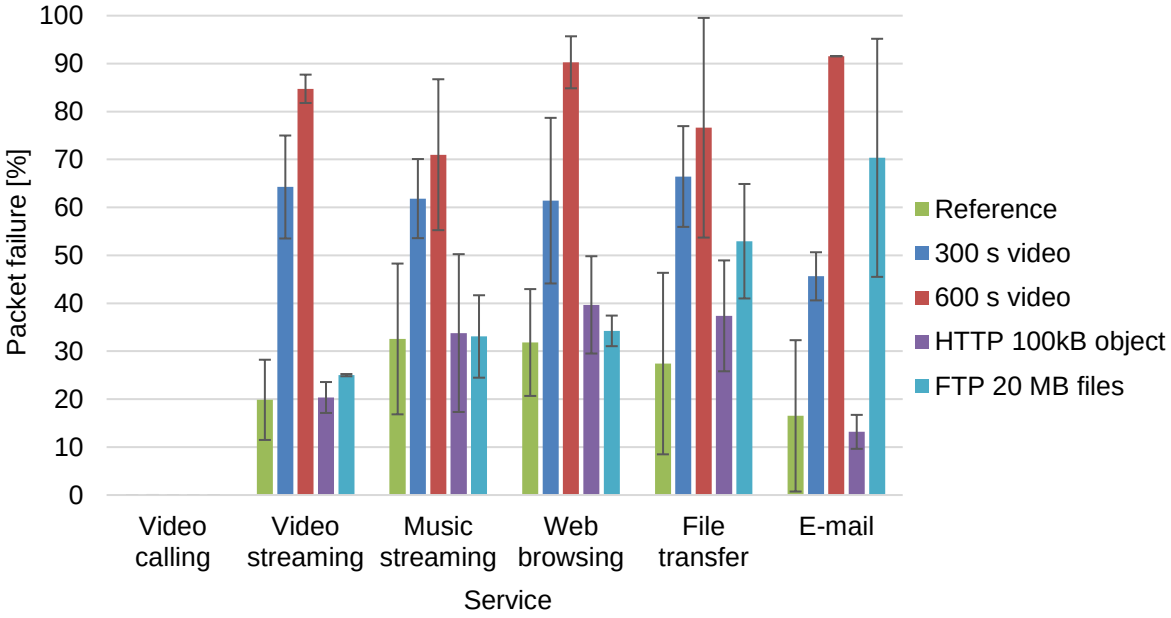


Figure 4.24. Packet failure per service for scenarios with different service parameters.

Packet failure results are coherent with the ones obtained for delay. Increasing the average video duration causes the system to be severely congested and packet failure increases to unacceptable values for video streaming and all services with a lower priority. For the scenario with bigger HTTP objects, results stay similar to the reference scenario. For the scenario with bigger FTP files, results are also similar to the reference, with the exception of file transfer and e-mail services. Increasing the

average file size for file transfer causes packet failure to increase 26% for file transfer and most noticeably 54% for e-mail. Figure 4.25 shows the behaviour of the system in terms of user satisfaction for the four scenarios with different service parameters considered in this study.

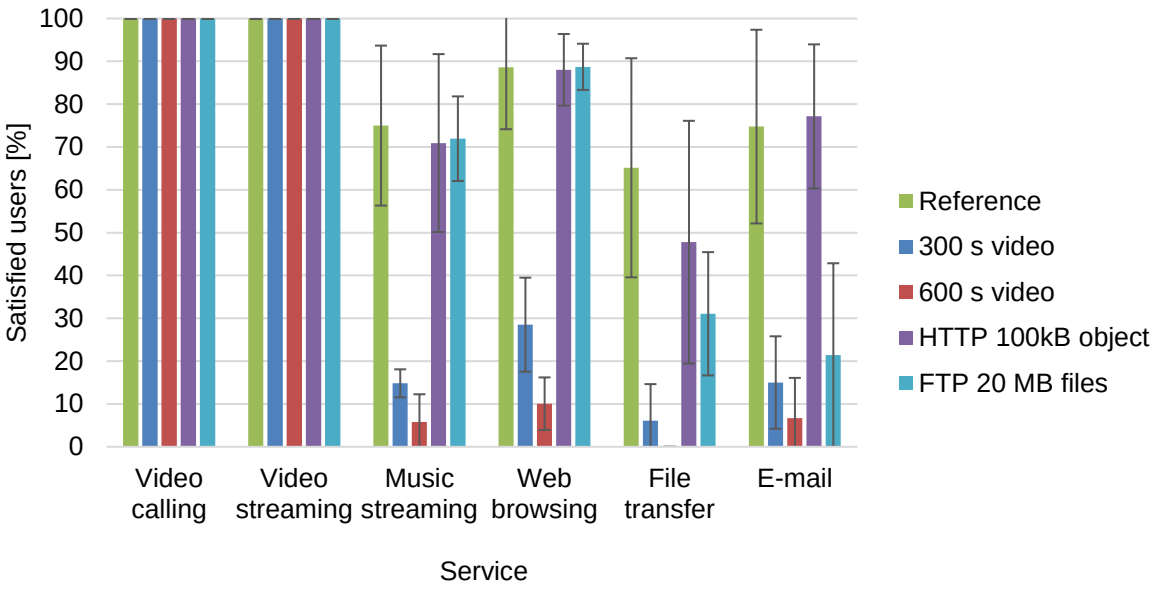


Figure 4.25. Satisfied users for scenarios with different service parameters.

All services with a lower priority than video streaming suffer severe degradation in the scenarios with 300 s and 600 s video durations, as all of them get throughput satisfaction levels below 30%. This is consistent with the results of total congestion with highly delayed packets and a high number of failed packets. For the scenario of varying the average size of the web browsing objects it becomes clear that the number of satisfied users follows the results of the reference scenario. Web browsing does not generate great amounts of traffic due to low object sizes and large reading times. Following the increase in HOL delay and packet failure, the percentage of satisfied e-mail users drops 53% as a consequence of changing the file size in file transfer.

4.7 Service penetration

This section describes the results obtained when scenarios with different service penetrations are considered, apart from the reference one used in the other simulations. For that purpose, one defined two additional scenarios that aim to characterise two distinct situations as summarised in Table 4.7. The first is Video centric as one assumes that 50% of the users perform video streaming. This scenario is motivated by the fact that video is becoming the dominant source of mobile data traffic. According to [Eric17], 75% of the mobile data traffic will come from video sources in 2022. The other scenario is VoLTE centric with a 50% penetration which reflects situations where users perform voice calls more frequently than usual. It is also important to mention that introducing VoLTE may lead to a growth of the users' interest in voice communications, thus increasing its service penetration.

Table 4.7. Service penetrations for additional scenarios.

Service	Reference [%]	Video centric [%]	VoLTE centric [%]
VoLTE	22	8	50
Video Calling	8	6	8
Video Streaming	28	50	16
Music Streaming	20	18	8
Web Browsing	10	8	8
File Transfer	8	6	6
E-Mail	4	4	4

The obtained results show how different these situations can be in terms of their impact on the network. Figure 4.26 shows the results for the variation of the total cell throughput for these scenarios. As one can observe, having a VoLTE centric scenario decreases the total cell throughput by approximately 46%. Having more VoLTE users basically requires less resources and more capacity is available for other services to reach higher throughputs. On the other hand, having more video streaming users makes the total throughput increase about 14% for values close to throughput stabilisation, for this scenario. This emphasises how important it is to accurately monitor the consumption of highly demanding services like video streaming.

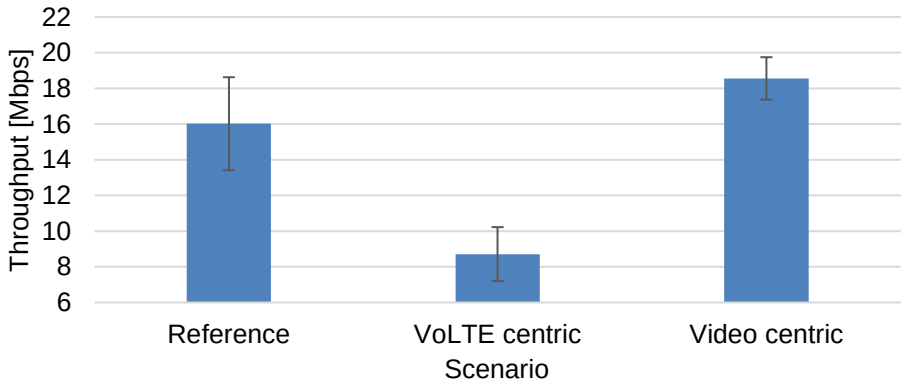


Figure 4.26. Total throughput for different service penetrations.

In what refers to services' performance, Figure 4.27 shows the results obtained in terms of queuing delay for the three different scenarios being considered. The results for the VoLTE centric scenario clearly enhance how switching the number of video users to a service with low throughput requirements like VoLTE reduces the load of the system. For this scenario, the HOL delay becomes basically inconsequent as it stays below 50 ms for all services, including video streaming. The same does not happen with the Video centric scenario. In this case, the majority of services registers a queuing delay similar to the one obtained in the reference scenario. However, HOL delay for video streaming users increases about 86% which is directly caused by the larger number of users performing the service, achieving an average delay value close to the maximum delay budget of 300 ms. Figure 4.28 shows the results obtained in terms of packet failure for the three defined scenarios.

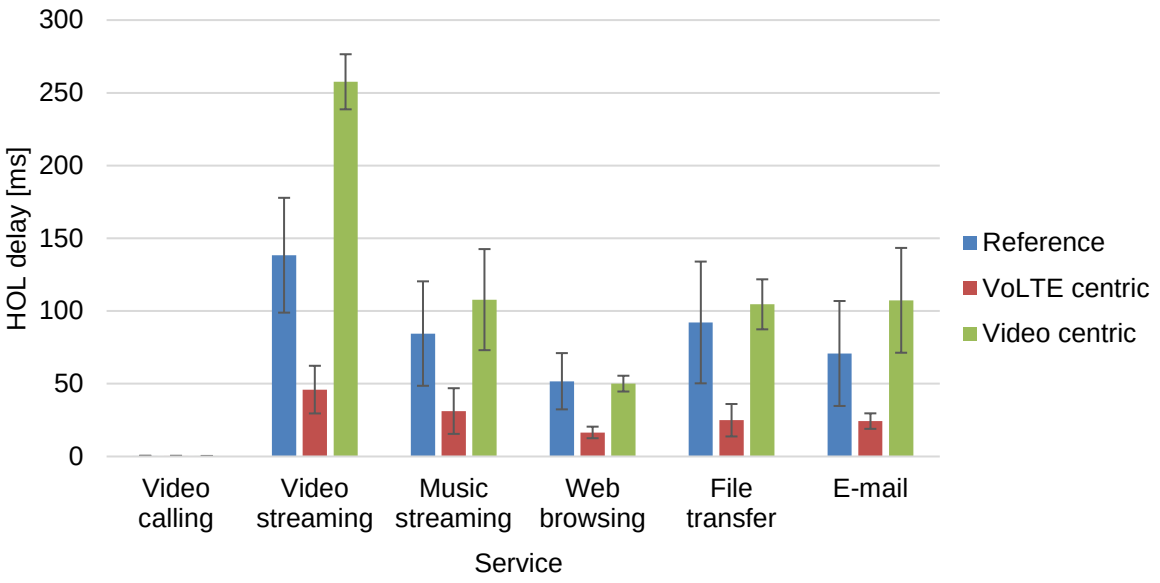


Figure 4.27. Average HOL delay per service for different service penetrations.

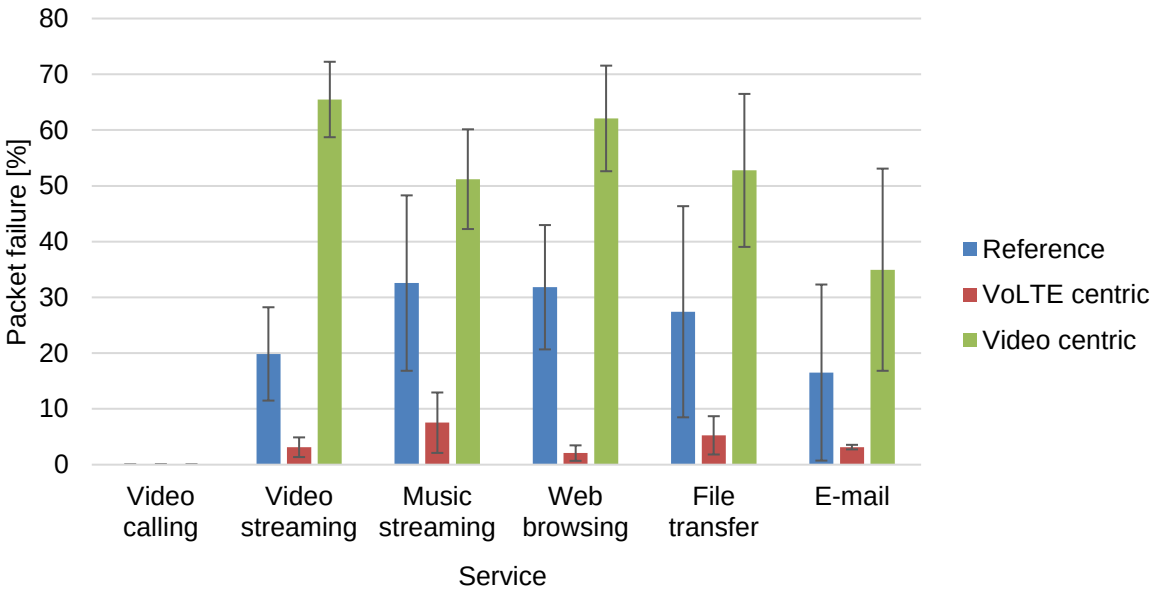


Figure 4.28. Packet failure per service for the different service penetrations.

The results in Figure 4.28 show that in the VoLTE centric scenario services' performance improves as expected. Packet failure does not exceed 7.5%, which is a result that highlights how service throughputs are barely affected as most of the requested packets are delivered. The opposite happens for the Video centric scenario. The high average value of delay experienced by video streaming shows that the service loads the system with high amounts of video traffic to be allocated, leaving less capacity for traffic from other services and causing high levels of packet failure in all of them.

Figure 4.29 shows how this behaviour translates in terms of user satisfaction for the three defined scenarios. The results for the percentage of satisfied users highlight the impact of increasing the amount of video traffic in the system. For the Video centric scenario, most of the services with a lower priority than video streaming and video calling have unacceptable percentages of satisfied users, all of them

below 50%. For the VoLTE centric scenario, results show the improvements due to significantly reducing the cell load. Satisfaction levels are all above 90%, representing the conditions to which the cell is expected to work to provide good levels of QoS to all of its users.

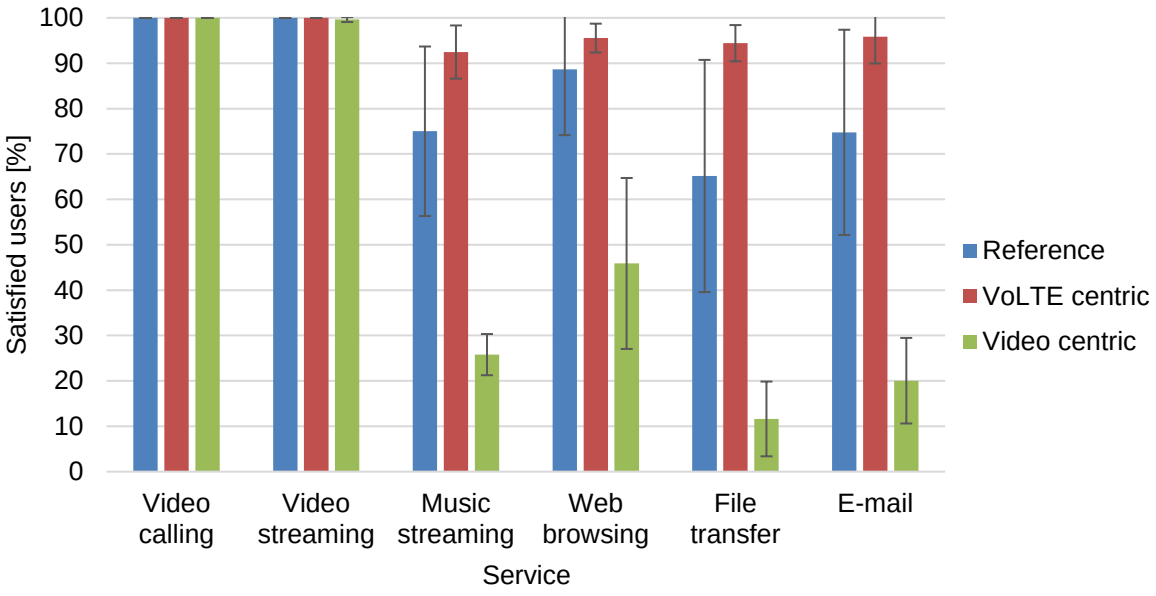


Figure 4.29. Satisfied users for different service penetrations.

Chapter 5

Conclusions

This chapter summarises all the work developed for this thesis and highlights its main conclusions. The chapter closes by pointing out perspectives for the development of future work.

The main goal of this thesis was to evaluate the impact of the implementation of VoLTE in the provided QoS of other services in the context of an already deployed LTE network, while monitoring the performance of the VoLTE service. In order to accomplish this goal, a single-cell model for the DL resource allocation at the radio interface in LTE was proposed and implemented in a simulator from which several simulations were performed. The results from these simulations allowed the analysis of the influence of several parameters, namely, type of environment, percentage of indoor users, bandwidths, average rate of user arrivals, service penetration and service-specific parameters.

The first chapter presents the underlying context to the study of VoLTE systems as part of the evolution over time of mobile communications systems. It also provides information regarding the current state of the technology in terms of industry and predictions for the growing demand of VoLTE in the upcoming years. After that, the motivation that lead to the development of this thesis was presented as well as its structure and contents.

Chapter 2 starts by introducing the fundamental aspects regarding LTE, describing their network architecture and radio interface. The network architecture is described in terms of its main layers and the most important network elements, such as eNodeB, MME, PCRF, among others. The radio interface is described in terms of FDD, TDD, multiple access techniques, frequency bands allocation, resource allocation, bandwidth and physical channels. After that, it presents a study of the main aspects of VoIP services and describes how this type of service is implemented in the context of an LTE network to provide VoLTE calls. VoIP traffic is constituted by periodically generated voice frames during active instants and several codecs exist to support this service. For VoLTE, AMR-WB and EVS were standardised by 3GPP and are foreseen as the main codecs to support this technology while providing high quality voice. IMS has a fundamental importance in enabling service functionality at the application layer and two main solutions exist to support the operator's transition to full VoLTE networks: CSFB and SR-VCC. In terms of assessing QoE, there are essentially three methods to consider that are basically applicable to any type of voice communications. MOS allows a subjective measurement using a set of users as sample, perceptual methods like PESQ/POLQA that allow an analysis based on the received voice signal, and network based methods like the E-model that predicts user satisfaction based on network packet transmission measurements. A description of services and applications' characteristics is also provided in this chapter in order to show the different classes of traffic that can be generated. At the end of the chapter, a brief state of the art is presented, where aspects related to scheduling algorithms in the presence of VoIP traffic and performance assessment of VoLTE are discussed.

In Chapter 3, the model developed for this thesis is fully described as well as its implementation on a simulator and the corresponding assessment. The developed single-cell model aims at describing the processes involved in DL resource allocation in LTE, in a time-based approach at the IP packet level. The model assumes that all users are connected to a single base station, performing a given service. It considers several aspects that influence system performance, namely the cell environment described through each user's SNR and indoor attenuation, number of users, bandwidth and several other service-related parameters. The user's SNR is statistically generated for three distinct types of environment

and the corresponding achievable throughput for each UE is computed through the expressions presented in Annex A. Traffic for each user is modelled through the traffic source models presented in Annex B. Then, a resource management block deals with the traffic queuing and RB allocation. System performance is evaluated through several metrics aimed at measuring QoS and QoE of all services. A time-based simulator was developed from scratch in MATLAB to implement these models and allow a network analysis over a fixed time window. Finally, the simulator assessment is described, including the validation of all the blocks that compose the simulator and the statistical relevance of the results in terms of simulation time and number of simulations.

Chapter 4 starts by presenting the description of the reference scenario, the configuration of all the parameters of the simulator and the simulation strategy considered in this thesis. The reference scenario for this thesis is an urban one with 80% of users subject to indoor attenuation. Seven services are considered, namely: VoLTE, video calling, video streaming, music streaming, web browsing, file transfer and e-mail. Two source bit rate modes were considered for VoLTE to analyse the influence of two voice codecs: AMR-WB 12.65 and EVS 23.85. The results analysis is organised in terms of VoLTE call quality, number of users, type of environment, different bandwidths, different service parameters and different service penetration scenarios.

For the analysis of VoLTE call quality, a scenario with a full 100% penetration of VoLTE users was tested. As the developed model assumes that VoLTE has the highest priority, the remaining services do not influence capacity for VoLTE packets. Therefore, it is observed that the call quality is not affected for realistic values of the amount of users in the cell. The average MOS remains fairly constant up to a point where the network capacity is exceeded causing MOS to quickly drop from values close to 4.3 to 1.4. At this point, the average packet delay increases suddenly and with it the number of failed packets. For the 10 MHz bandwidth, until approximately 200 simultaneous active users only 15% present some sort of dissatisfaction towards the call quality. The remaining users have a satisfaction level of “Satisfied” or “Very satisfied” with an average MOS above 4.0. These results highlight that capacity for VoLTE is not a major issue as it is a service with a low priority, requiring under common radio conditions no more than one or two RBs every 20 ms to ensure its throughput. After that, the analysis focused on the influence of model’s input parameters on other services performance in the presence of VoLTE traffic.

For the variation of the total number of users using the cell, both 10 and 20 MHz bandwidths allow a similar cell throughput until a rate of 125 users per hour. After that, throughput stabilises for the 10 MHz bandwidth while the 20 MHz stabilises at approximately 225 users per hour. Lower user throughputs are observed when the number of users approaches these values, as the network can no longer allocate capacity to them. In what refers to services’ performance, one compares the QoS metrics for two scenarios, with and without scheduling priorities among data services. Results show that for a lower number of users per hour, considering no priorities allows a higher number of satisfied users. To guarantee that at least 90% of the users is satisfied for each service, up to a rate of 150 users per hour is supported for the scenario with no priorities while only 75 users per hour are supported when priorities are considered. For higher cell loads, low priority services like web browsing, file transfer or e-mail,

benefit in terms of user satisfaction for the scenario with no priorities among the services. On the other hand, the capacity for higher priority services is reduced. This is especially noticeable in the case of video calling, which is rarely affected for the scenario with services priorities as a consequence of having a higher priority than video streaming.

To analyse the impact of the number of VoLTE users over the performance of the other services, the number of data users in the reference scenario was fixed and the rate of VoLTE user arrivals was varied between 33, corresponding to the reference scenario, and 5000 users per hour. Results were compared for both AMR-WB and EVS codecs. It is observed that VoLTE generated traffic does not exceed 6% of the total cell traffic in the worst case with the EVS codec. To check its influence on the remaining services in terms of QoS, it is observed that for the HOL delay and the packet failure ratio, a significant impact is only verified for more than 2 500 users per hour, for both codecs. In terms of user satisfaction, this has no effect in video calling and video streaming due to their high priorities. For the remaining services, the number of satisfied users reduces with more impact caused by the EVS codec. One concludes that to guarantee that the reduction on the number of satisfied users for each service stays below 10%, up to 60 and 36 simultaneous VoLTE users are supported for the AMR-WB and EVS codecs, respectively.

Three types of environment were analysed to assess their influence on the system's performance: rural, suburban and urban. The cell throughput for the urban scenario is 38% above the rural one. No significant degradation is caused when comparing the urban with the suburban environment. In the worst case, the reduction of satisfied users reaches about 13% in the music streaming case. For the rural environment this reduction reaches up to 28%, in the file transfer case. This result is the consequence of an 80 ms increase in the average HOL delay for video streaming which causes the system to get more congested with the generated video traffic. Packet failure for lower priority services increases up to 40%, in the web browsing case, from the suburban to the rural environment.

The percentage of indoor users was changed to assess how indoor attenuation affects the achievable user throughputs due to a reduction of the average SNR. The total cell throughput in the reference scenario decreases about 50% when the percentage of indoor users increases from 20% to 80%. However, impact in terms of performance is more noticeable between the 80% and 50% scenarios. In terms of HOL delay, the biggest reductions occur for video and music streaming, with 50 ms and 32 ms respectively. Results show that there is no significant degradation in terms of the number of satisfied users for the load conditions of the reference scenario.

Regarding the radio channel bandwidths, a comparison between the reference 20 MHz and the 10 MHz one was made. For the defined reference scenario, reducing the available bandwidth results in a scenario where the total cell throughput is reduced from 16 Mbps to a point where it stabilises at 8 Mbps. As a consequence of decreasing the available capacity to a half, services with a priority lower than video streaming have a reduction that can go up to 40% in the percentage of satisfied users, as in the file transfer case. Higher priority services, namely video calling and video streaming, do not show any degradation in terms of satisfied users. However, HOL delay for video streaming has an increase of 94 ms, reaching an average value of 232 ms that is close to its delay budget.

Several service parameters were changed in order to assess their influence in the system performance. Increasing the average video duration to 300 and 600 s has a serious impact in all the services with a lower priority, as the percentage of satisfied users for these services drops below 30%. This is a consequence of getting the network loaded with video traffic, which causes the average HOL delay of video streaming to get close to its delay budget and to increase packet failure. Increasing the web browsing object sizes by a factor of 10 barely had any influence in the observed QoS metrics as results were similar to those of the reference scenario. This shows that relatively to other services, web browsing does not account for as much traffic due to small object sizes and low throughput requirements. Increasing the file sizes in file transfer shows that the biggest impact occurs for e-mail which is the only service with lower priority. The percentage of satisfied e-mail users dropped 53% following an 88 ms increase in the average HOL packet delay.

Finally, the influence of the service penetration was analysed by defining two alternative scenarios, one of them being VoLTE centric, with 50% VoLTE users, and the other being Video centric, with 50% video streaming users. Increasing the percentage of VoLTE users from 22% to 50% causes a decrease on the total cell throughput of 46%, which highlights the effect of reducing the number of users using other services. For the VoLTE centric scenario, HOL delay stays below 50 ms and packet failure does not exceed 7.5% for all services, highlighting how the system performs under low load conditions. For the Video centric scenario, the majority of the services registers a queuing delay similar to the one obtained in the reference scenario. The exception is video streaming that increases close to 120 ms, causing an increase in packet failure for this service and for those with a lower priority. This shows how in these conditions capacity for the lower priority services is severely reduced. For these services, less than 50% of the users get their throughput satisfied while for the VoLTE centric scenario, user satisfaction levels are above 90% for all services.

Several aspects can be studied in future work regarding the subject of this thesis. The proposed model could be improved to encompass the study of a real cellular network, considering the analysis of a specific geographical area for example. Assuming a multi-cell environment would allow studying aspects related to mobility like, for example, the impact of handover on the performance of all services including VoLTE. Real SNR measurements could be used to improve the accuracy of the characterisation of the radio conditions of the UEs.

Other aspects related to the modelling of resource management in LTE could be further studied. Several scheduling algorithms are proposed in the literature to improve VoLTE capacity. A comparison between the most relevant ones would be interesting. In this thesis, it is assumed that all the LTE capacity is used for traffic allocation which in practice is not entirely true. Signalling and control functions may occupy part of the available RBs per TTI and modelling of these mechanisms would increase the results accuracy. At the packet level, retransmission mechanisms could be implemented to allow realistic measurements of packet loss. The obtained results for LTE could also be compared with the voice capacity of currently deployed GSM and UMTS networks. The main goal would be to evaluate how many voice users should VoLTE support, coming from circuit-switched services.

Annex A

SNR and Throughput

This annex provides the analytical models used to relate SNR and the obtainable throughput in LTE for a given set of configurations. The simulator assessment of these models is also shown.

A.1 SNR and throughput

The mapping of SNR to throughput presented in this annex is based on the formulas described by [Guit16] and which are the result of the previous work provided by [Alme13]. For this purpose three expressions were considered for three modulation types in the DL: QPSK, 16QAM and 64QAM. These expressions are the logistic functions which provide the best-fit approach to a set of values collected by 3GPP based on throughput performance tests done by manufacturers. Three MCSs and the corresponding average coding rates were chosen with the intent of achieving a more realistic approach to the behaviour of a real network:

- QPSK with a coding rate of 1/3.
- 16QAM with a coding rate of 1/2.
- 64QAM with a coding rate of 3/4.

For 2x2 MIMO, QPSK and a coding rate of 1/3, throughput in the DL is approximated by:

$$R_{b,RB}\,[\text{bps}] = \frac{2.34201 \times 10^{6}}{14.0051 + e^{-0.577897\rho_{N}\,[\text{dB}]}} \tag{A.1}$$

where:

- ρ_N: SNR.

For 2x2 MIMO, 16-QAM and a coding rate of 1/2, throughput in the DL is approximated by:

$$R_{b,RB}\,[\text{bps}] = \frac{47613.1}{0.0926275 + e^{-0.295838\rho_{N}\,[\text{dB}]}} \tag{A.2}$$

For 2x2 MIMO, 64-QAM and a coding rate of 3/4, throughput in the DL is approximated by:

$$R_{b,RB}\,[\text{bps}] = \frac{26405.8}{0.0220186 + e^{-0.24491\rho_{N}\,[\text{dB}]}} \tag{A.3}$$

For the sake of a better understanding, Figure A.1 shows the graphical representation of the described curves where the solid lines corresponds to the best achievable throughput for every SNR value from the three considered modulations.

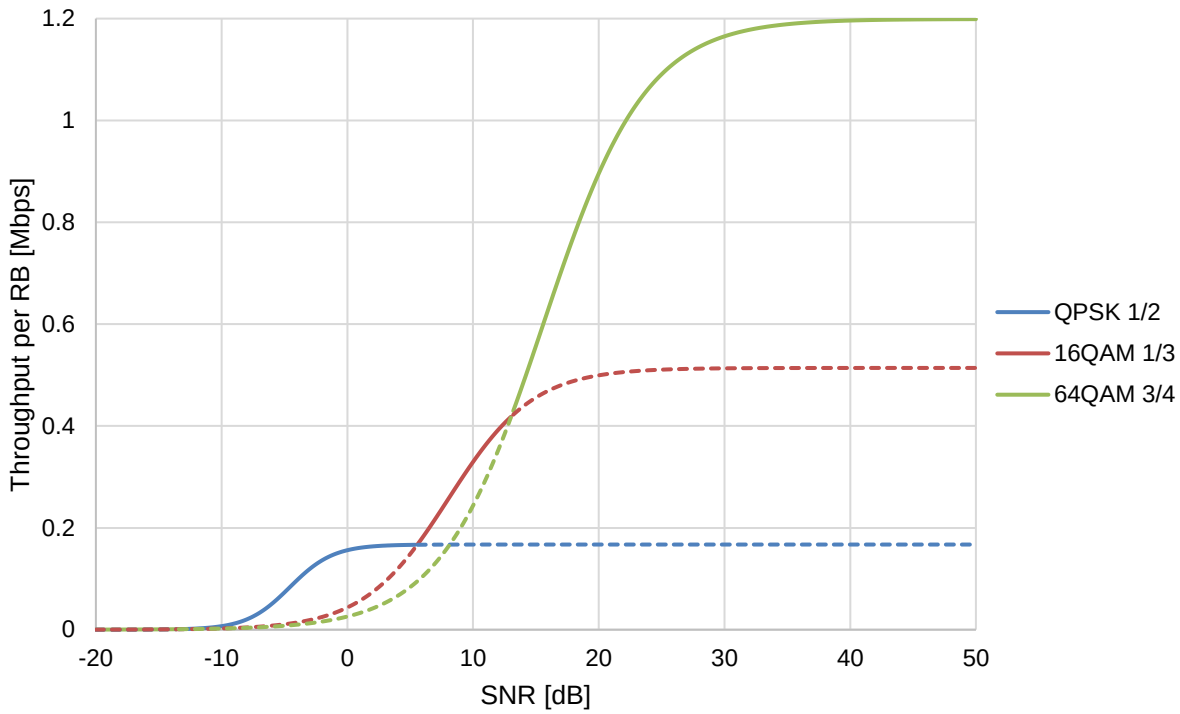


Figure A.1. Throughput per RB versus SNR for three different MCSs.

A.2 Simulator assessment

This annex shows the assessment of the implementation of the analytical models used to relate SNR and the obtainable throughput in LTE. Statistical distribution parameters, average and standard deviation, are evaluated from (A.4) and (A.5), respectively.

$$\mu = \frac{\sum_{i=1}^{N_Z} z_i}{N_Z} \tag{A.4}$$

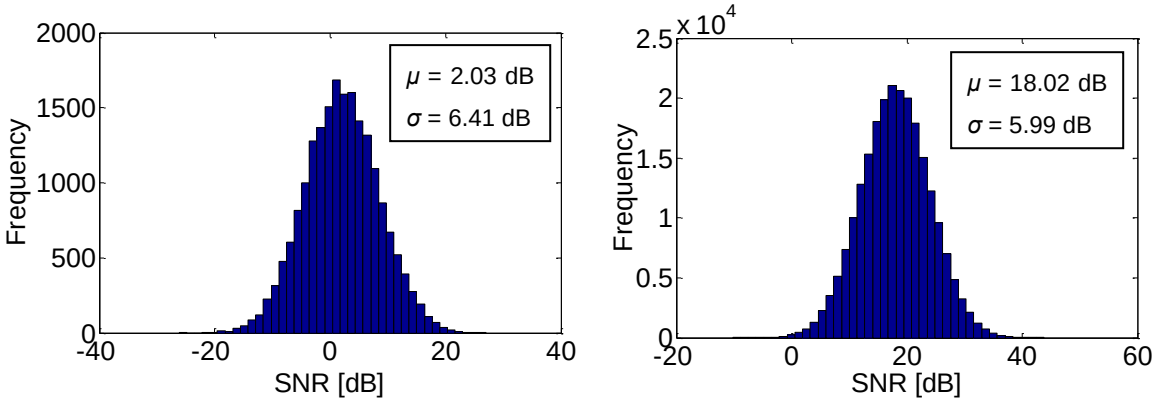
where:

- N_Z: Number of samples.
- z_i: Value of sample i.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N_Z} (z_i - \mu)^2}{N_Z}} \tag{A.5}$$

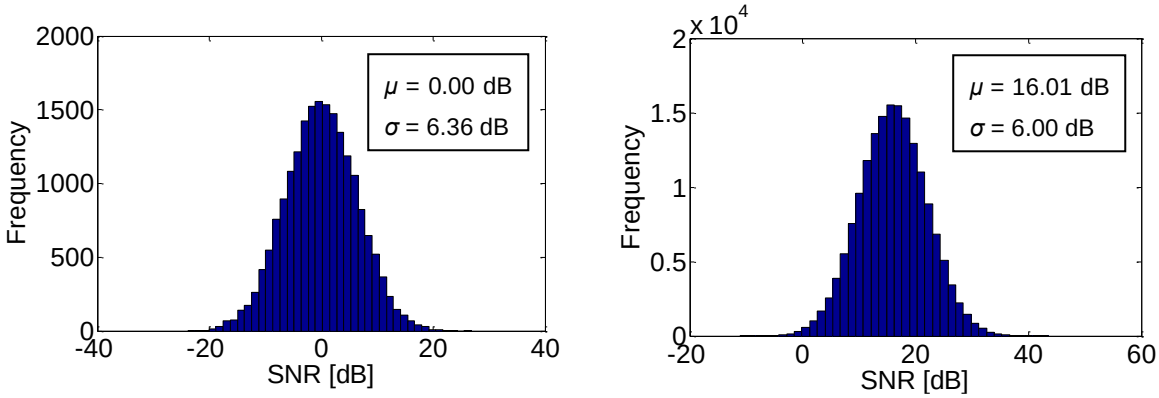
From Figure A.2 to Figure A.4, the obtained SNR distributions for three types on environment considered in this thesis are presented. From Figure A.5 to Figure A.7, the correspondent distributions for the

achievable throughput per RB are shown.



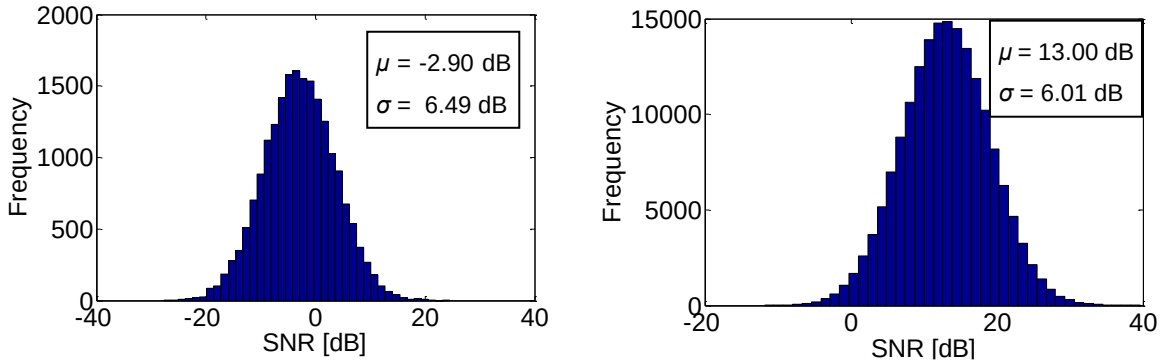
(a) Indoor. (b) Outdoor.

Figure A.2. SNR distributions for the urban environment.



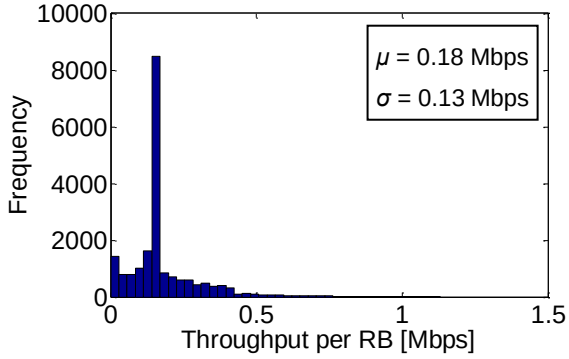
(a) Indoor. (b) Outdoor.

Figure A.3. SNR distributions for the suburban environment.

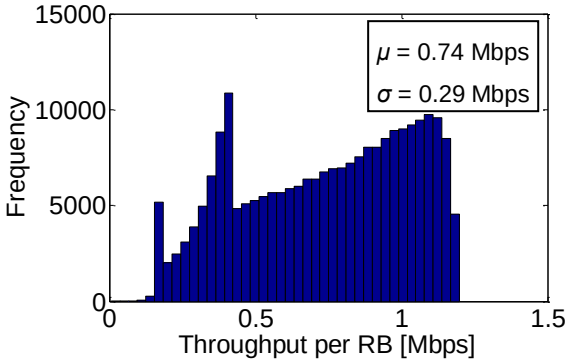


(a) Indoor. (b) Outdoor.

Figure A.4. SNR distributions for the rural environment.

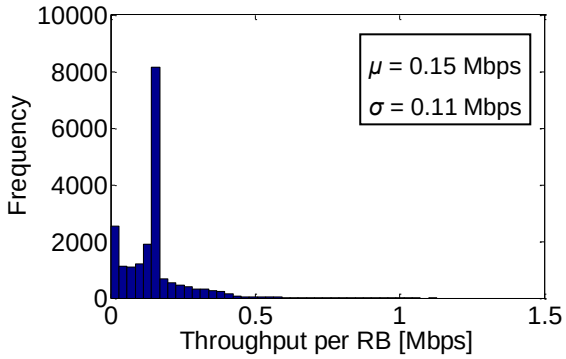


(a) Indoor.

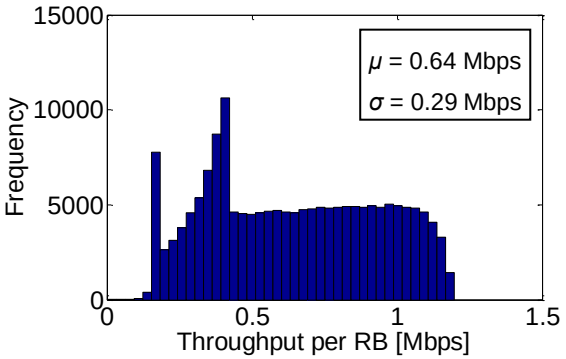


(b) Outdoor.

Figure A.5. Achievable throughput per RB for the urban environment.

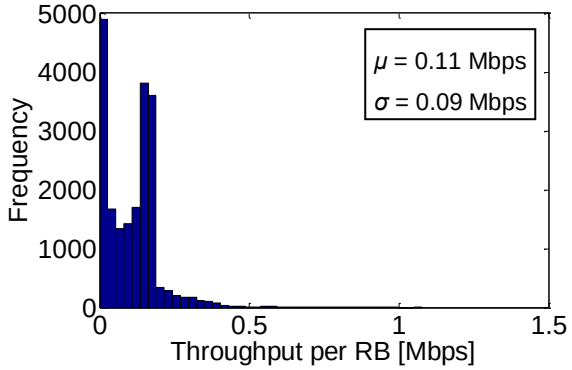


(a) Indoor.

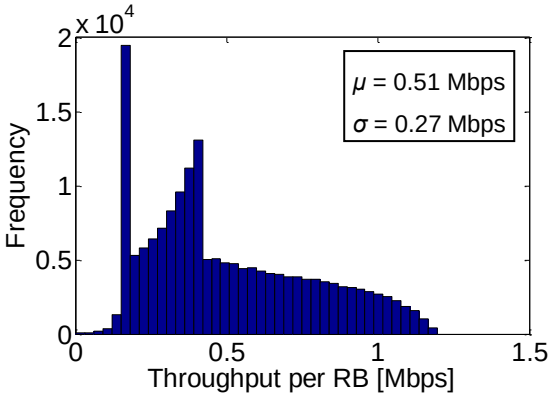


(b) Outdoor.

Figure A.6. Achievable throughput per RB for the suburban environment.



(a) Indoor.



(b) Outdoor.

Figure A.7. Achievable throughput per RB for the rural environment.

Annex B

Traffic source models

This annex presents the traffic source models used for the services considered in this study. The statistical validation of the models is also presented.

B.1 VoLTE (VoIP) model

VoLTE, as for every other VoIP services, is traditionally characterised by an ON-OFF behaviour composed of intercalated speech and silent bursts. According to [Khan09], this behaviour can be modelled as a two-state Markov model where the two states are silence or inactive state (State 0) and talking or active state (State 1), as shown in Figure B.1. An important parameter that is usually used to characterise the model is the voice activity factor (VAF) which corresponds to the probability of being in the active state:

$$VAF = \frac{\alpha}{\alpha + \beta} \tag{B.1}$$

where:

- α : Probability of transition from the silent state to the talking state
- β : Probability of transition from the talking state to silence state.

Both activity and silent periods are generated by an exponentially distributed random variable with mean values t_{ON} and t_{OFF}, respectively. The state update is done every time these periods are exceeded.

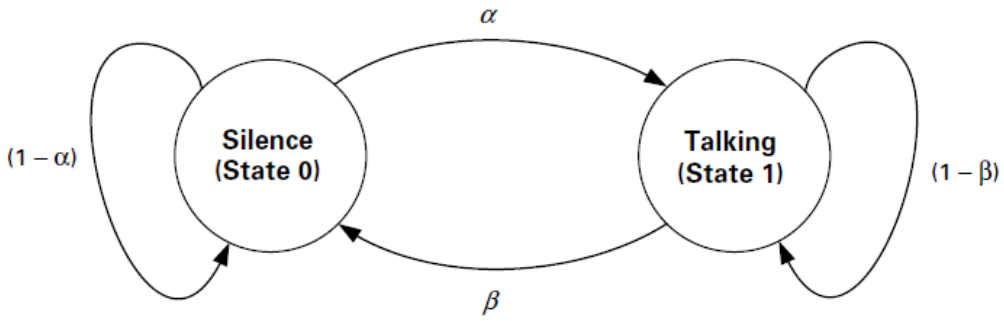


Figure B.1. Two-state voice activity model (extracted from [Khan09]).

The considered speech codec and the corresponding packet rate determine the payload size of the IP packets carrying speech frames. Besides these, SID frames with a frame size of 56 bits are also periodically inserted every 160 ms, during silent periods. Table B.1 shows the main characteristics of the codecs and the corresponding modes considered for this thesis, namely AMR-WB 12.65 and EVS 23.85. Additionally to the payload size, each voice packet also includes a header overhead from the fact that VoIP uses UDP and RTP at the transport layer, along with the IP overhead. The use of RoHC tackles the problem of reducing the size of this overhead.

Table B.1. VoLTE codec characteristics (based on [Cox14]).

Codec	Payload size [bits]	Frame size [bits]	Source rate [kbps]
AMR-WB 12.65	253	264	12.65
EVS 23.85	477	488	23.85

In what refers to the voice calls generation process, it is described by a Poisson process and the duration of the calls is well modelled by an exponential distribution. The complete parameter set considered to model VoLTE traffic is presented in Table B.2.

Table B.2. Parameter set for VoLTE traffic.

Voice Activity Factor [%]	50
Mean Active Phase, t_{ON} [s]	3
Mean Silent Phase, t_{OFF} [s]	3
Transmission Time Interval [ms]	20
Mean Call Duration [s]	60

B.2 Video calling model

The approach considered to model video conferencing traffic is based on the video source model proposed by [Heym97] and extensively described in [Agui03]. This model is named the Gamma Beta Auto-Regressive (GBAR) model which is a first order auto-regressive process based on statistical features observed in H.261 and H.263 Variable Bit Rate (VBR) video conferencing traffic. This model is based on the definition of a Gamma distributed stationary stochastic variable $\{X_n\}$ given by:

$$X_n = A_n X_{n-1} + B_n \tag{B.2}$$

where:

- A_n: Auto-correlation function.
- B_n: Marginal distribution.

By assuming that $Ga(\beta, \lambda)$ denotes a random variable with a Gamma distribution with shape parameter β and scale parameter λ and that $Be(p, q)$ denotes a Beta distributed random variable with parameters p and q, $\{X_n\}$ has a marginal $Ga(\beta, \lambda)$ distribution if A_n is $Be(\alpha, \beta - \alpha)$ and B_n is $Ga(\beta - \alpha, \lambda)$. Therefore, simulating the GBAR model consists of generating random values from the $\{X_n\}$ process through Gamma and Beta distributions defined by three parameters β, λ and α. This process is used as a source model by rounding the obtained values to the nearest integers to assume them as packet sizes. As described in [Heym97], assuming that the mean μ and variance v of the sample data are

known, β and λ are given by:

$$\beta = \frac{\mu^2}{v} \text{ and } \lambda = \frac{\mu}{v} \quad \text{(B.3)}$$

The parameter α is calculated by:

$$\alpha = \rho \cdot \beta \quad \text{(B.4)}$$

where:

- ρ: Lag 1 autocorrelation coefficient of the sample data (estimated from the data statistics).

B.3 Streaming model

The model considered for streaming services is based in [Khan09] which assumes that each frame of data arrives at a regular time interval defined by the number of frames per second. Each video frame is decomposed into a fixed number of slices, each transmitted as a single packet. The size of these packets/slices is modelled as a truncated Pareto distribution. The video encoder introduces encoding delay intervals between the packets of a frame. These intervals are also modelled by a truncated Pareto distribution. The video streaming traffic model parameters are given in Table B.3. In this model, the video source rate is assumed at 64 kbps.

Table B.3. Video streaming traffic model parameters (extracted from [Khan09]).

Parameter	Statistical characterisation
Inter-arrival time between the beginning of each frame	Deterministic at 100 ms (10 frames per second)
Number of packets (slices) in a frame	Deterministic, 8 packets per frame
Packet (slice) size	Truncated Pareto distribution, Mean = 10 bytes, Maximum = 250 bytes (before truncation), PDF: $f_X = \frac{\alpha_k^\alpha}{\alpha+1}, k \le x < m, f_X = \left(\frac{k}{m}\right)^\alpha, x = m,$ α = 1.2, k = 20 bytes, m = 250 bytes
Inter-arrival time between packets (slices) in a frame	Truncated Pareto distribution, Mean = 6 ms, Maximum = 12.5 ms (before truncation), PDF: $f_X = \frac{\alpha_k^\alpha}{\alpha+1}, k \le x < m, f_X = \left(\frac{k}{m}\right)^\alpha, x = m,$ α = 1.2, k = 2.5 ms, m = 12.5 ms

B.4 Web browsing HTTP model

Web browsing corresponds to a non-conversational application and therefore exhibits an asymmetrical behaviour which corresponds to the requests for information by users to remote servers. As stated in [Khan09] and further described in [Khat14], each session is divided into active and inactive periods representing webpage downloads and the intermediate reading times. The webpage downloads such as web pages with images, text, etc. are name packet calls. The active and inactive periods are the result of human interaction where the packet call represents a web user's request for information and the reading time corresponds to the time required to process the webpage content. Figure B.2 shows a packet trace of a typical HTTP web browsing session.

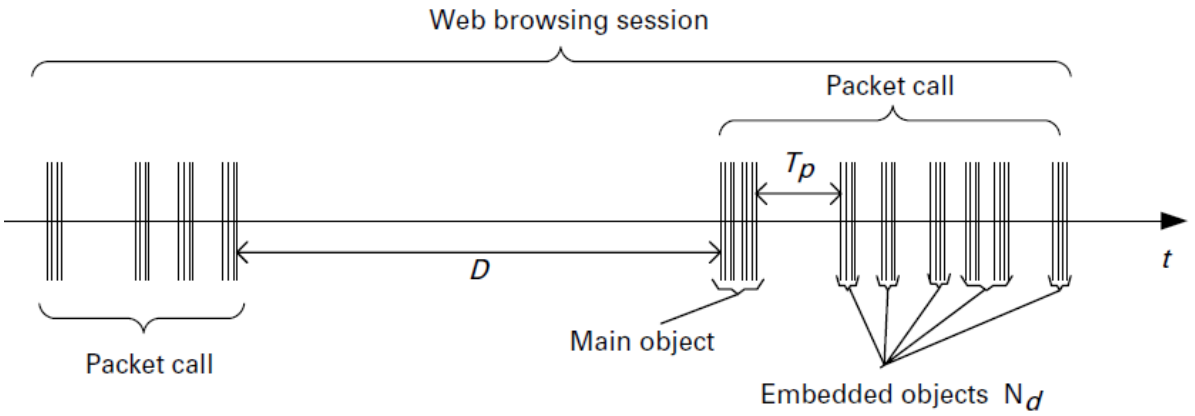


Figure B.2. Typical packet trace of a WWW session.

Due to the fact that web traffic has a self-similar behaviour where the traffic statistics on different timescales are similar, packet calls, similarly to packet sessions, are also divided into active and inactive periods. Unlike packet sessions, these periods are attributed to machine interaction rather than human interaction. A packet session may include one or more packet calls, depending on the application. A web browser serves a user's request by requesting the initial HTML page which corresponds to the main object, using an HTTP GET request. The retrieval of the initial page and each of the embedded objects (e.g. pictures, videos, advertisements, etc.) is represented by the active period within the packet call while the parsing time and protocol overhead are represented by the inactive periods within a packet call. The parsing time refers to the time the browser spends in parsing for the embedded objects in the packet call or the web page.

Table B.4 shows the key parameters used to characterise the web browsing traffic. These are the main object size, S_M, the size of an embedded object in a web page, S_E, the number of embedded objects, N_D, the reading time, D and the parsing time T_p.

Table B.4. Web browsing traffic model parameters (extracted from [Khan09]).

Parameter	Statistical characterisation
Main object size S_M	Truncated lognormal distribution, Mean = 10 710 bytes, Standard deviation = 25 032 bytes, Minimum = 100 bytes, Maximum = 2 Mbytes (before truncation), PDF: $f_X = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0,$ σ = 1.37, μ = 8.37
Embedded object size S_E	Truncated lognormal distribution, Mean = 7 758 bytes, Standard deviation = 126 168 bytes, Minimum = 50 bytes, Maximum = 2 Mbytes (before truncation), PDF: $f_X = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0,$ σ = 2.36, μ = 6.17
Number of embedded objects per page N_D	Truncated Pareto distribution, Mean = 5.64, Maximum = 53 (before truncation), PDF: $f_X = \frac{\alpha_k^\alpha}{\alpha+1}, k \le x < m, f_X = \left(\frac{k}{m}\right)^\alpha, x = m,$ α = 1.1, k = 2, m = 55
Reading time D	Exponential distribution, Mean = 30 s, PDF: $f_X = \lambda e^{-\lambda x}, x \ge 0$, λ = 0.033
Parsing Time T_p	Exponential distribution, Mean = 0.13 s, PDF: $f_X = \lambda e^{-\lambda x}, x \ge 0$, λ = 7.69

B.5 FTP and E-mail traffic model

[Khan09] describes a model for a FTP session as a sequence of file transfers separated by reading times. This model basically consists of a simplified version of the web browsing model presented in Annex B.4 where the packet calls simply correspond to the files requested by the end user. The two key FTP session parameters are S, the size of the file to be transferred, and D, the reading time which is the time interval between the end of a file download and the request for the next file. The parameters considered for the FTP traffic model are presented in Table B.5. For the e-mail service, the same model used for FTP traffic is considered with a different configuration based on the work developed in [Seba08] where the sizes of all the emails sent and received in a real network on a given working day, allowing their statistical characterisation as presented in Table B.6.

Table B.5. FTP traffic model parameters (extracted from [Khan09]).

Parameter	Statistical characterisation
File size S	Truncated lognormal distribution, Mean = 2 Mbytes, Standard deviation = 0.722 Mbytes, Maximum = 5 Mbytes (before truncation), PDF: $f_X = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0,$ σ = 0.35, μ = 14.45
Reading time D	Exponential distribution, Mean = 180 s, PDF: $f_X = \lambda e^{-\lambda x}, x \geq 0$, λ = 0.006

Table B.6. E-mail traffic model parameters (adapted from [Seba08]).

Parameter	Statistical characterisation
E-mail size *S*	Lognormal distribution, Mean = 100 kbytes, Standard deviation = 812 bytes PDF: $f_X = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0,$ σ = 0.0081, μ = 11.51
Reading time *D*	Exponential distribution, Mean = 600 s, PDF: $f_X = \lambda e^{-\lambda x}, x \geq 0$, λ = 0.0017

B.6 Traffic source models validation

This annex shows the most relevant examples of the validation of the implemented traffic source models by providing the obtained distributions for the multiple parameters that describe each of the services. Figure B.3 shows the histogram of the instants of arrival for a one hour simulation with 10000 users arriving to the cell. As expected, the instants of arrival follow a uniform distribution during this period. Calls duration are generated for the VoLTE, video calling and streaming services using exponential distributions which follow the targeted behaviour as shown in Figure B.4. Active and silent states are also generated using this distribution with different average values. Packet sizes in streaming are assumed to follow truncated Pareto distributions. Figure B.5 shows the expected distribution where the effect of the truncation is evident as the packet sizes are limited at a maximum of 6700 bytes.

For the non-conversational services, the objects and file sizes are distributed according to Log-Normal distributions. Figure B.6 shows the distribution of file sizes for 10000 users requesting a file transfer service. Reading times for non-GBR services use an exponential distribution just like the one considered for the call duration of the GBR services.

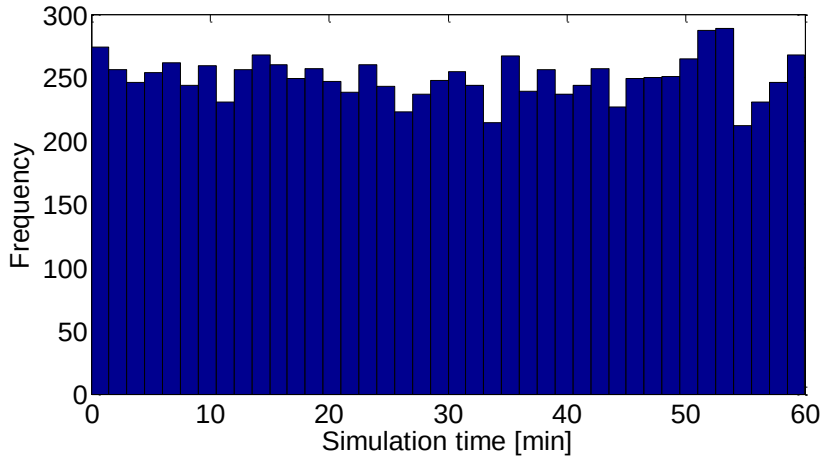


Figure B.3. User's instants of arrival to the cell.

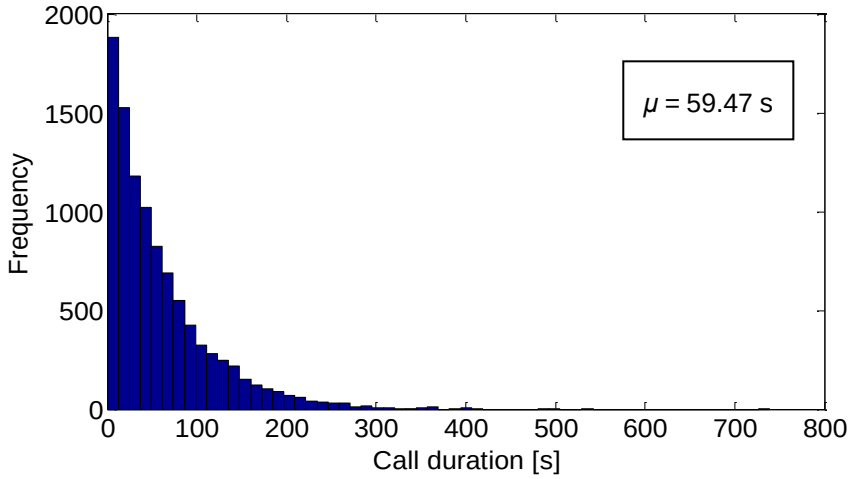


Figure B.4. Call duration distribution for 10000 VoLTE users.

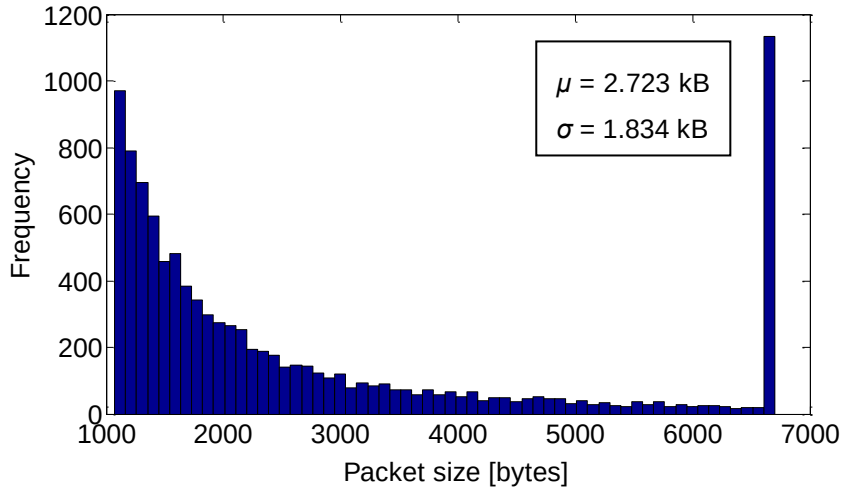


Figure B.5. Packet sizes for 10000 packet samples from video streaming users.

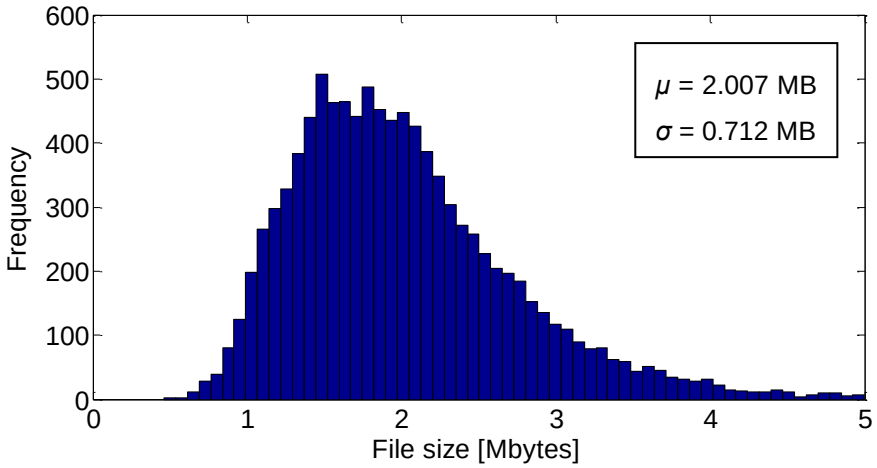


Figure B.6. File sizes for 10000 samples from file transfer users.

Annex C

Mobility Model

This annex describes the mobility model considered in this thesis which consists of a random generator for user speeds.

C.1 Triangular Distribution Mobility Model

The model presented in [ChLu95] and described in detail by [Serr12] considers a triangular distribution for speed, as shown in Figure C.1.

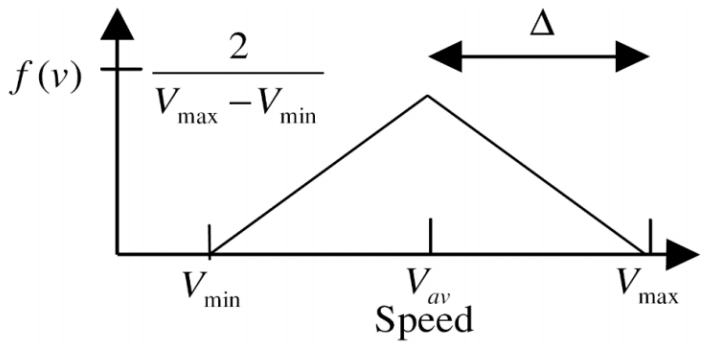


Figure C.1. Velocity probability density function (extracted from [ChLu95]).

This distribution is used with average user speed and deviation as given in (C.1) and (C.2), respectively.

$$V_{av} = \frac{V_{max} + V_{min}}{2} \tag{C.1}$$

where:

- V_{max}: Maximum user speed.
- V_{min}: Minimum user speed.

$$\Delta = \frac{V_{max} - V_{min}}{2} \tag{C.2}$$

The density function for the triangular distribution is given by:

$$f(v) = \begin{cases} \frac{1}{\Delta^2}[v - (V_{av} - \Delta)] & , V_{av} - \Delta \leq v \leq V_{av} \\ -\frac{1}{\Delta^2}[v - (V_{av} + \Delta)] & , V_{av} \leq v \leq V_{av} + \Delta \\ 0 & , otherwise \end{cases} \tag{C.3}$$

Five different mobility types are considered as given in Table C.1. Figure C.2 shows the validation of the mobility model by showing the results of the obtained speeds during a two hour simulation (with a total of 7200 samples) for a single urban user, where the histogram clearly shows the triangular shape that characterises the density function.

Table C.1. Mobility type speed characteristics (extracted from [ChLu95]).

Mobility type	V_{av} **[m/s]**	Δ **[m/s]**
Static	0.0	0.0
Pedestrian	1.0	1.0
Urban	10.0	10.0
Main Roads	15.0	15.0
Highways	22.5	12.5

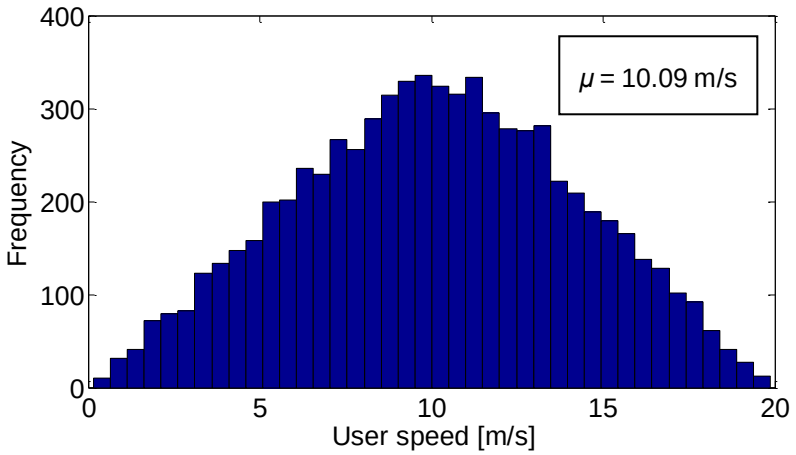


Figure C.2. Triangular distribution validation for a user with an urban mobility type.

References

[3GPP16a] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (Release 14),* TS 36.101, Ver. 14.0.0, Jun. 2016 (http://www.3gpp.org).

[3GPP16b] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 14)*, TS 36.300, Ver. 14.0.0, Sep. 2016 (http://www.3gpp.org).

[3GPP16c] 3GPP, *Technical Specification Group Services and System Aspects*; *Policy and Charging Control Architecture (Release 14)*, Report TS 23.203, Ver. 14.2.0, Dec. 2016 (http://www.3gpp.org/).

[3GPP17a] 3GPP, *Technical Specification Group Services and System Aspects; Quality of Service (QoS) concept and architecture (Release 14)*, Report TS 23.107, Ver. 14.0.0, Mar. 2017 (http://www.3gpp.org/).

[3GPP17b] 3GPP, *Technical Specification Group Services and System Aspects; IP Multimedia Subsystem (IMS) centralised services; Stage 2 (Release 14)*, Report TS 23.292, Ver. 14.2.0, Mar. 2017 (http://www.3gpp.org/).

[3GPP17c] 3GPP, *Technical Specification Group Services and System Aspects; Codec for Enhanced Voice Services (EVS); General Overview (Release 14)*, Report TS 26.441, Ver. 14.0.0, Mar. 2017 (http://www.3gpp.org/).

[Agui03] J. Aguiar, *Traffic Analysis at the Radio Interface in Converging Mobile and Wireless Communication Systems*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2003.

[AhIl08] S. Ahson and M. Ilyas, *VoIP Handbook: Applications, Technologies, Reliability, and Security*, First Edition, CRC Press, Boca Raton, FL, USA, Dec. 2008.

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

[ANAC12a] ANACOM, *Final Report of the Auction*, Public Consultation, Lisbon, Portugal, Jan. 2012 (http://www.anacom.pt/streaming/Final_Report_Auction.pdf?contentId=1115304&field=ATTACHED_FILE).

[ANAC12b] ANACOM, *Decision of homologation of the agreement regarding the location of spectrum in the 1800 MHz band* (in Portuguese), Public Consultation, Lisbon, Portugal, Mar. 2012 (http://www.anacom.pt/streaming/Decisao_ReshufflingMarco2012.pdf?contentId=1120288&field=ATTACHED_FILE).

[AnLS12] M. Anehill, M. Larsson, G. Strömberg and E. Parsons, “Validating voice over LTE end-to-end”, *Ericsson Review*, Vol. 91, No. 1, Jan. 2012, pp. 10-15.

[Aric15] Aricent, *Voice over LTE (VoLTE) implementation for User Equipment (UE)*, White Paper, July 2015.

[BoBa14] B. Bojovic and N. Baldo, “A new channel and QoS aware scheduler to enhance the capacity of voice over LTE systems”, in *Proc. of SSD14 2014 - 11th IEEE International Multi-Conference on Systems, Signals & Devices*, Barcelona, Spain, Feb. 2014.

[Carr11] P. Carreira, *Data Rate Performance Gains in UMTS Evolution to LTE at the Cellular Level*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2011.

[ChLu95] E. Chlebus and W. Ludwin, “Is handoff traffic really Poissonean?”, in *Proc. of ICUPC’95 – 4th IEEE International Conference on Universal Personal Communications*, Tokyo, Japan, Nov. 1995.

[Cox14] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications*, Second Edition, John Wiley & Sons, Chichester, UK, July 2014.

[Cisc16] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020*, White Paper, Feb. 2016 (https://www.cisco.com/c/dam/m/en_in/innovation/enterprise/assets/mobile-white-paper-c11-520862.pdf).

[Cisc17] Cisco, *Encrypted Traffic Analytics*, White Paper, June 2017 (https://www.cisco.com/c/dam/en/us/solutions/collateral/enterprise-networks/enterprise-network-security/nb-09-encrytd-traf-anlytcs-wp-cte-en.pdf).

[Corr06] L.M. Correia, *Mobile Broadband Multimedia Networks: Techniques, models and tools for 4G*, First Edition, Elsevier, London, UK, May 2006.

[Corr16] L. M. Correia, *Mobile Communication Systems - Course Notes*, Instituto Superior Técnico, 2016.

[DPBK06] J. Davidson, J. Peters, M. Bhatia, S. Kalidindi and S. Mukherjee, *Voice over IP Fundamentals*, Second Edition, Cisco Press, Indianapolis, IN, USA, July 2006.

[Dout15] D. Doutor, *Load balancing between LTE and WiFi*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2015.

[Eric17] Ericsson, *Ericsson Mobility Report*, Public Consultation, Stockholm, Sweden, June 2017.

[GSA17] GSA, *VoLTE & ViLTE Market Status Snapshot*, Industry Report, Farnham, UK, Aug. 2017 (https://gsacom.com/paper/volte-vilte-market-status-snapshot/).

[GSMA16a] GSMA, *GSMA IR.92: IMS Profile for Voice and SMS*, Ver. 10.0, May 2016 (http://www.gsma.com/newsroom/wp-content/uploads/IR.92-v10.0.pdf).

[GSMA16b] GSMA, *GSMA IR.51 – IMS Profile for Voice, Video and SMS over Wi-Fi*, Ver. 4.0, May 2016 (http://www.gsma.com/newsroom/wp-content/uploads/IR.51-v4.0.pdf).

[Guit16] J. Guita, *Balancing the load in LTE urban networks via inter-frequency handovers*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2016.

[Heym97] D. Heyman, "The GBAR Source Model for VBR Video conferences", *IEEE/ACM Transactions on Networking*, Vol. 5, No. 4, Aug. 1997, pp. 554-560.

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

[ITUT01] ITU-T, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,* Recommendation P.862, Geneva, Switzerland, Feb. 2001.

[ITUT14a] ITU-T, *Perceptual objective listening quality assessment,* Recommendation P.863, Geneva, Switzerland,Sep. 2014.

[ITUT14b] ITU-T, *The E-model: a computational model for use in transmission planning,* Recommendation G.107, Geneva, Switzerland, Feb. 2014.

[KaSa14] O. Kadatskaya and S. Saburova, "Research of Requirements to QoS for Voice over LTE", in *Proc of 2014 First International Scientific-Practical Conference: Problems of Infocommunications, Science and Technology*, Kharkiv, Ukraine, Oct. 2014.

[Khan09] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*, Cambridge University Press, Cambridge, UK, 2009.

[Khat14] S. Khatibi, *Radio Resource Management Strategies in Virtual Networks*, Ph.D. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2014.

[Math17] Mathworks, LTE System Toolbox, https://www.mathworks.com/products/lte-system.html, Sep. 2017.

[MoRa06] S. Moller, A. Raake, N. Kitawaki, A. Takahashi and M. Waltermann, "Impairment factor framework for wideband speech codecs", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No.6, Nov. 2006, pp. 1969–1976.

[MRum08] M. Rumney, "3GPP LTE: Introducing Single-Carrier FDMA", *Agilent Measurement Journal*, Issue 4, 2008, pp. 18-27.

[Nguy16] D. Nguyen and H. Nguyen, "A dynamic rate adaptation algorithm using WB E-model for voice traffic over LTE network", in *Proc. of 2016 Wireless Days*, Toulouse, France, May 2016.

[Noki17] Nokia, Evolve to richer voice with Voice over LTE (VoLTE), https://resources.ext.nokia.com /asset/200306, Sep. 2017.

[OMNe17] OMNet++, Discrete Event Simulator, https://omnetpp.org/, Sep. 2017.

[OzVa13] O. Ozturk, and M. Vajapeyam, "Performance of VoLTE and Data Traffic in LTE Heterogeneous Networks", in *Proc. of GLOBECOM'13 – IEEE Global Communications*

Conference, Atlanta, GA, USA, Dec. 2013.

[PGBC11] G. Piro, L. Grieco, G. Boggia, F. Capozzi and P. Camarda, “Simulating LTE Cellular Systems: Na Open-Source Framework”, *IEEE Transactions on vehicular technology*, Vol. 60, No. 2, Feb. 2011, pp. 498-512.

[PoHo12] M. Poikselkä, H. Holma, J. Hongisto, J. Kallio and A. Toskala, *Voice over LTE (VoLTE)*, First Edition, John Wiley & Sons, Chichester, UK, Feb. 2012.

[Qunh11] C. Qunhui, *Evolution and deployment of VoLTE*, Huawei Communicate, Issue 61, Shenzhen, China, Sep. 2011, pp. 52-55 (http://www.huawei.com/mediafiles/CORPORATE/PDF/Magazine/communicate/61/HW_094164.pdf).

[RaCZ13] J. Rankin, A. Costaiche and J. Zeto, *Validating VoLTE: A Definitive Guide to Successful Deployments*, First Edition, IXIA, Calabasas, CA, USA, Aug. 2013.

[RiDM13] F. J. Rivas, A. Díaz, and P. Merino, “Obtaining More Realistic Cross-Layer QoS Measurements: A VoIP over LTE Use Case”, *Journal of Computer Networks and Communications*, Vol. 2013, Article ID 405858, Aug. 2013.

[Rive17] Riverbed Modeler (former OPNET Modeler Suite), https://www.riverbed.com/gb/products/steelcentral/opnet.html, Sep. 2017.

[RoSc17] Rohde & Schwarz, VoLTE Test Solutions, https://www.rohde-schwarz.com/us/solutions/wireless-communications/lte/in-focus/lte_voice_solutions_71515.html, Sep. 2017.

[Seba08] D. Sebastião, *Algorithms for Quality of Service in a WiFi Network*, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2008.

[Serr12] A. Serrador, *Joint Radio Resource Management in Heterogeneous Networks*, Ph.D. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2012.

[SeTB11] S. Sesia, I. Toufik and I. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice (2nd Edition)*, John Wiley & Sons, Chichester, UK, Aug. 2011.

[SiWa08] I. Siomina and S. Wanstedt, “The impact of QoS support on the end user satisfaction in LTE networks with mixed traffic”, in *Proc. of PIMR 2008 - 19th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Cannes, France, Sep. 2008.

[Spir17] Spirent, VoLTE & RCS Testing, https://www.spirent.com/Solutions/Volte, Sep. 2017.

[TaKo11] I. Tanaka and T. Koshimizu, “Overview of GSMA VoLTE Profile”, *NTT DOCOMO Technical Journal*, Vol. 13, No. 4, Mar. 2011, pp. 45-51.

[TuPe13] G. Tu, C. Peng, H.Wang, C. Li and S. Lu, “How Voice Calls Affect Data in Operational LTE Networks”, in Proc. of *MobiCom’13 - 19th Annual International Conference on Mobile computing & networking*, Miami, FL, USA, Sep. 2013.

[Vizz14a] A. Vizzarri, “Analysis of VoIP Over LTE End-To-End Performances in Congested

Scenarios", in *Proc. of 2014 Second International Conference on Artificial Intelligence, Modelling and Simulation*, Madrid, Spain, Nov. 2014.

[Vizz14b] A. Vizzarri, "Analysis of VoLTE End-To-End Quality of Service using OPNET", in *Proc. of UKSim 2014 - AMSS 8th European Modelling Symposium*, Pisa, Italy, Oct. 2014.

[YiCh12] S. Yi, S. Chun, Y. Lee, S. Park and S. Jung, *Radio Protocols for LTE and LTE-Advanced (First Edition)*, John Wiley & Sons, Chichester, UK, Sep. 2012.