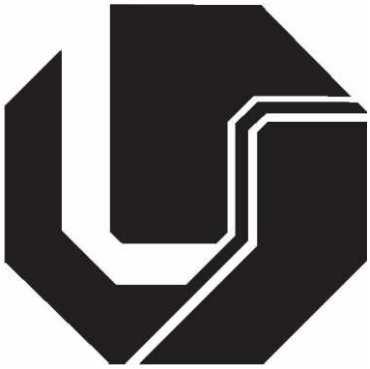
FEDERAL UNIVERSITY OF UBERLANDIA
FACULTY OF ELECTRICAL ENGINEERING



A NEW CHANNEL AND QOS-AWARE SCHEDULING ALGORITHM FOR REAL-TIME AND NON-REAL-TIME TRAFFIC IN 5G HETEROGENEOUS NETWORKS

GABRIEL ANDRADE QUEIROZ

UBERLÂNDIA – 2024

GABRIEL ANDRADE QUEIROZ

A NEW CHANNEL AND QOS AWARE SCHEDULING ALGORITHM FOR REAL-TIME AND NON-REAL-TIME TRAFFIC IN 5G HETEROGENEOUS NETWORKS

DISSERTATION SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCES TO THE POST-GRADUATE PROGRAM OF THE FACULTY OF ELECTRICAL ENGINEERING AT THE FEDERAL UNIVERSITY OF UBERLÂNDIA.

MEMBERS OF THE COMMITTEE

PROF. DR. ÉDERSON ROSA DA SILVA (ADVISOR) – UFU

PROF. DR. ANDRÉ LUIZ AGUIAR DA COSTA – UFU

PROF. DR. MÁRCIO ANDREY TEIXEIRA – IFSP

Uberlândia - 2024

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

Q3
2024

Queiroz, Gabriel Andrade, 1999-
A new channel and QoS-aware scheduling algorithm for real-time and non-real-time traffic in 5G heterogeneous networks [recurso eletrônico] / Gabriel Andrade Queiroz. - 2024.

Orientador: Éderson Rosa da Silva.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Engenharia Elétrica.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2024.445
Inclui bibliografia.
Inclui ilustrações.

1. Engenharia elétrica. I. Silva, Éderson Rosa da, 1984-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Engenharia Elétrica. III. Título.

CDU: 621.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica
Av. João Naves de Ávila, 2121, Bloco 3N - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4707 - www.posgrad.feelt.ufu.br - copel@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Engenharia Elétrica				
Defesa de:	Dissertação de Mestrado, 794, PPGEELT				
Data:	Vinte e seis de julho de dois mil e vinte e quatro	Hora de início:	09:00	Hora de encerramento:	11:30
Matrícula do Discente:	12222EEL006				
Nome do Discente:	Gabriel Andrade Queiroz				
Título do Trabalho:	A new Channel and QoS Aware Scheduling algorithm for real-time and non-real-time traffic in 5G heterogeneous networks				
Área de concentração:	Processamento da Informação				
Linha de pesquisa:	Processamento Digital de Sinais e Redes de Comunicação				
Projeto de Pesquisa de vinculação:	Coordenador do projeto: Éderson Rosa da Silva. Título do projeto: Desenvolvimento e simulação de técnicas de alocação de recursos em redes de comunicação. Agência financiadora: não se aplica. Número do processo na agência financiadora: não se aplica. Vigência do projeto: 2018 – atual.				

Reuniu-se no formato híbrido, no Anfiteatro do Bloco 1E, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Engenharia Elétrica, assim composta:

Professores Doutores: André Luiz Aguiar da Costa (UFU), Márcio Andrey Teixeira (IFSP) e Éderson Rosa da Silva, orientador do discente. Os professores André Luiz Aguiar, Éderson Rosa da Silva e o candidato Gabriel Andrade Queiroz participaram presencialmente. O professor Márcio Andrey Teixeira participou de forma remota.

Iniciando os trabalhos o presidente da mesa, Dr. Éderson Rosa da Silva, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra para a exposição do seu trabalho. A duração da apresentação do discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir, o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

APROVADO.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre. O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme, foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Ederson Rosa da Silva**, **Professor(a) do Magistério Superior**, em 26/07/2024, às 11:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Andre Luiz Aguiar da Costa**, **Professor(a) do Magistério Superior**, em 26/07/2024, às 11:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **MARCIO ANDREY TEIXEIRA**, **Usuário Externo**, em 26/07/2024, às 11:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5521207** e o código CRC **40B0104C**.

Referência: Processo nº 23117.043910/2024-51 SEI nº 5521207

Acknowledgments

First, I would like to thank my parents Marta and Celso (*in memorian*) for teaching me the value of knowledge. To my mother, for her unconditional dedication to me and for being my most precious foundation. Today I am aware of her sacrifices and where her love has taken me. To my father, for unconsciously introducing me to one of the first algorithms I fell in love with: the Tower of Hanoi. I suspect I now have more practice at this game than you, Dad.

To my tutor, Prof. Dr. Éderson Rosa da Silva, for the opportunity to conduct this work, for guiding me all the way from the beginning of my graduation and for his suggestions and contributions to the development of this research. I thank him for his comprehension, partnership, and his compassionate view of the academic world.

I thank my brother Celso Galeno for going through many of my experiences with me during this time. Thank you for often supporting me in my research and publications as if they were your own work. I would also like to thank my sister Cintia for her enthusiasm and encouragement regarding my academic decisions.

To the Queiroz Alvarenga family for welcoming me during my master's journey. Thank you, Aunt Eliane, Uncle Ronaldo, Isabella, and Octávio. I would especially like to thank my aunt Eliane for reinforcing that my knowledge should not be confined to my mind but should be passed on.

I thank my girlfriend, Marina, for her unconditional support for my decisions. I met you during this stage of my life and discovered that through love, we inspire our loved ones and get inspired by them. I fell in love with you because I learned about myself from your eyes. After all, you are a heart that stirs but also calms, shelters, and shares. It turns out that sometimes you know me better than I know myself.

To my friends, because life is more joyful and less complicated when you have such good company. Thank you, Rafael, Guilherme and Manu, the group of friends affectionately known as the "M&A", Marcelo, and Lucas.

To my cousin Mateus, who influenced me positively in many of my interests and decisions, as well as being my gateway to the computer networks world. To my aunt Márcia, who is the major influence of education in my life. Thank you, auntie.

Finally, I would like to thank the Department of Electrical Engineering of the Federal University of Uberlândia (UFU) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their financial support.

“Those who have the privilege to know have the duty to act, and in that action are the seed of new knowledge.”

Albert Einstein

Abstract

Queiroz, A. G., *A NEW CHANNEL AND QOS AWARE SCHEDULING ALGORITHM FOR REAL-TIME AND NON-REAL-TIME TRAFFIC IN 5G HETEROGENEOUS NETWORKS*, UFU, Uberlândia, Brazil, 2024, 81p.

5G mobile communication systems present growing demands aimed at Quality of Service (QoS) requirements combined with the high densification of users in a heterogeneous network infrastructure (HetNet). Thus, 5G networks are expected to accommodate various applications and services. Therefore, it is necessary to develop scheduling methods that can benefit both real-time (RT) and non-real-time (NRT) services. In this sense, we propose two new scheduling algorithms, the Weighted QoS Aware Scheduler (WQAS) and the Channel and QoS Aware Scheduler (CQAS), in a HetNet scenario characterized by the diversity of traffic models: full buffer (IoT), HTTP, vehicular, VoIP, gaming, and video. System-level simulations are carried out to analyze the performance of WQAS and CQAS compared to the Round Robin (RR), best Channel Quality Indicator (CQI), and QoS Aware Scheduler (QAS) algorithms, considering the variation in the number of users as a way of stress testing the network. The results indicate that WQAS and CQAS show similar results for RT applications, with WQAS standing out for reliability. Furthermore, CQAS shows significant overall throughput gains except for HTTP when compared with QAS and WQAS. Both WQAS and CQAS perform well in adapting to the delay constraints related to the QoS requirements of each 5G RT application, in addition to good reliability performance. Finally, CQAS achieved the results described above for RT applications while generating the lowest impact on NRT applications, demonstrating the importance of a Channel and QoS-aware algorithm.

Keywords: Channel and QoS Aware Scheduler, Weighted QoS Aware Scheduler, Scheduling Algorithm, Real-Time Traffic, Non-Real-Time Traffic, 5G Heterogeneous Networks.

Contents

List of Figures

List of Tables

List of Abbreviations

16QAM	16 Quadrature Amplitude Modulation
1G	First Generation
256QAM	256 Quadrature Amplitude Modulation
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
5QI	5G QoS Identifier
64QAM	64 Quadrature Amplitude Modulation
AFTOS	Advanced Fair Throughput Optimized Scheduler
ALWDF	Adjusted Largest Weighted Delay First
AMBR	Aggregate Maximum Bit Rate
AMC	Adaptive Modulation and Coding
AMF	Access and Mobility Management Function
AMPS	Analogue Mobile Phone System
AMR	Adaptive Multi-Rate
AP	Access Point
APIs	Application Programming Interfaces
ARP	Allocation and Retention Priority
ARQ	hybrid-automatic Repeat Request
AS	Access Stratum
AUC	Authentication Center
AUSF	Authentication Server Function
BBU	Baseband Unit
BCCH	Broadcast Control Channel
BCH	Broadcast Channel
BDMA	Beam Division Multiple Access
BLER	Block Error Ratio
BSR	Buffer Status Reports
BSs	Base Stations
BSS	Base Station Subsystem
BTSs	Base Transceiver Stations
CAOPF	Channel Aware Optimized Proportional Fair
CCCH	Common Control Channel
CCI	Co-Channel Interference
CDMA	Code Division Multiple Access
CIS	Channel-Independent Scheduling
CN	Core Network
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CoMP	Coordinated Multipoint
CQAS	Channel and QoS-aware Scheduler
CQIs	Channel Quality Indicators
C-RAN	Centralized/Cloud Radio Access Network

CS	Circuit Switched
CSI	Channel State Information
D2D	Device-to-Device
DC	Delay Constraint
DCCH	Dedicated Control Channel
DL-SCH	Downlink Shared Channel
DOS	Denial of Service
DRB	Data Radio Bearer
DTCH	Dedicated Traffic Channel
ECDF	Empirical Cumulative Distribution Function
eCOMP	enhanced COMP
ECSD	Enhanced Circuit Switched Data
EDGE	Enhanced Data Rates for GSM Evolution
EGPRS	Enhanced General Packet Radio Service
eIMT-A	enhanced Interference Mitigation and Traffic Adaptation
EIR	Equipment Identity Register
eMBB	enhanced Mobile Broadband
eNodeB	evolved Node Base station
EPC	Evolved Packet Core
EPS	Evolved Packet System
E-UTRA	Evolved Universal Terrestrial Radio Access
EV-DO	Evolution Data Optimized
FBSs	femto BSs
FDD	Frequency Division Duplexing
FDM	Frequency Division Multiplexing
FDMA	Frequency Division Multiple Access
FD-MIMO	Full Dimensional MIMO
FEC	Forward Error Correction
FIFO	First-In, First-Out
FM	Frequency Modulation
FSK	Frequency Shift Keying
FTOS	Fair Throughput Optimized Scheduler
FUs	Femtocell users
GBR	Guaranteed Bit Rate
GFBR	Guaranteed Flow Bit Rate
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Switching Center
GMSK	Gaussian Minimum Shift Keying
gNodeBs	5G Node Base Stations
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GTP-U	General Packet Radio Service Tunneling Protocol User Plane
HARQ	Hybrid Automatic Repeat Request
H-CRANs	Heterogeneous-Cloud Radio Access Networks
HetNet	Heterogeneous Network
HLR	Home Location Register
HomNets	Homogeneous Networks
HRPD	High Rate Packet Data
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access

HSPA+	Evolved HSPA
HSPDA	High-Speed Downlink Packet Access
HSUPA	High-Speed Uplink Packet Access
HTTP	Hypertext Transfer Protocol
IMT	International Mobile Telecommunications
IMT-2000	International Mobile Telecommunications-2000
IoT	Internet of Things
IP	Internet Protocol
IRS	Intelligent Reflecting Surface
ISDN	Integrated Services Digital Network
ITU	International Telecommunications Union
J-TACS	Japanese Total Access Communication System
LOS	Line-Of-Sight
LTE	Long-Term Evolution
LTE-A	LTE-Advanced
LTE-LAA	LTE-A Pro
LTE-LWA	LTE WLAN Aggregation
LTE-M	LTE Machine Type Communication
M2M	Machine-to-Machine
MAC	Medium-Access Control
MBSs	macro BSs
MCS	Modulation Coding Scheme
MDBV	Maximum Data Burst Volume
ME	mobile equipment
MFBR	Maximum Flow Bit Rate
MIMO	Multiple Input Multiple Output
M-LWDF	Modified Largest Weighted Delay First
MME	Mobile Management Entity
MMS	Multimedia Messaging Service
mMTC	massive Machine Type Communications
MR	Maximum Rate
MSS	mobile stations
MSC	Mobile Switching Center
MT	Maximum Throughput
MTC	Machine Type Communication
MTSO	Mobile Telephone Switching Office
MU-MIMO	Multiple User-MIMO
MUs	Macrocell Users
NAICS	Network Assisted Interference Cancellation and Suppression
NAS	Non-Access Stratum
NA-TDMA	North American TDMA
ng-eNodeB	next-generation evolved Node Base station
NG-RAN	next-generation RAN
NLOS	Non Line-Of-Sight
NLP	nonlinear programming
NMT	Nordic Mobile Telephone System
node B	node Base
NOMA	Non-Orthogonal Multiple Access
NR	New Radio
NRT	non-real-time

NSA	Non-Standalone
NSS	Network Switching Subsystem
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OMC	Operation and Maintenance Center
PCCH	Paging Control Channel
PCH	Paging Channel
PDB	Packet Delay Budget
PDC	Personal Digital Cellular
PDCP	Packet Data Convergence Protocol
PDP	Power Delay Profile
PDU	Protocol Data Unit
PER	Packet Erros Rate
PF	Proportional Fair
P-GW	Packet Data Network Gateway
PHY	Physical Layer
PS	Packet Switched
PSTN	Public Switched Telephone Network
QAS	QoS Aware Scheduler
QCI	Quality of Service Identifier
QFI	QoS Flow ID
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAN	Radio-Access Network
RATs	Radio Access Technologies
RBs	Resource Blocks
RE	Resource Element
RL	Reinforcement Learning
RLC	Radio-Link Control
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RQ	Reflective QoS
RQA	Reflective QoS Attribute
RR	Round Robin
RRC	Radio Resource Control
RRHs	Remote Radio Heads
RRM	Radio Resource Management
RT	real-time
RUEs	Remote User Equipments
SA	Standalone
SBA	Service-Based Architecture
SC-FDMA	Single Carrier Frequency Division Multiple Access
SCTP	Stream Control Transmission Protocol
SDAP	Service Data Application Protocol
SDOs	Standards Development Organizations
SDU	Service Data Unit
SFG	Simplified Fine Granularity
SFN	System Frame Number
SGSN	Serving GPRS Support Node

S-GW	Serving Gateway
SIC	Successive Interference Cancellation
SID	Silence Insertion Descriptor
SIM	Subscriber Identity Module
SMF	Session Management Function
SMS	Short Message Service
SMSF	Short Message Service Function
SPSK	Space Polarization Shift Keying
SRBs	Signalling Radio Bearers
TACS	Total Access Communications System
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
TF	Transport Format
TTI	Transmission Time Interval
UDMF	Unified Data Management Function
UDNs	Ultradense Networks
UDP	User Datagram Protocol
UDR	Unified Data Repository
UE-AMBR	UE Aggregate Maximum Bit Rate
UEMR	UE-based Maximum Rate
UEs	User Equipments
UFU	Universidade Federal de Uberlândia
UL-SCH	Uplink Shared Channel
UMTS	Universal Mobile Telecommunications Service
UPF	User Plane Function
URLLC	Ultra-Reliable, Low-Latency Communications
USA	United States of America
USIM	Universal SIM
UTRAN	UMTS Terrestrial Radio Access Network
V2X	Vehicle-To-Everything
VLR	Visitor Location Register
VoIP	Voice over IP
WCDMA	Wideband Code Division Multiple Access
WQAS	Weighted QoS-aware Scheduler
WRR	Weighted Round Robin
WWW	World Wide Web

"No matter where you go, everyone is connected."

- Ryutaro Nakamura

1

INTRODUCTION

DEMAND FOR MOBILE COMMUNICATION SERVICES is increasing in parallel with new technologies, applications, and services that integrate more users and devices into operators' mobile network infrastructure. In this sense, 5G systems create a highly connected world and encompass a variety of technologies such as Device-to-Device (D2D), Machine-to-Machine (M2M), Internet of Things (IoT), and Vehicle-To-Everything (V2X) [1].

5G and beyond-5G systems must be able to cope with the significant increase in the number of users, variety of services, and applications through networks with greater capacity, better spectral and energy efficiency, higher data rates, and lower latency values [2]. Thus, the study of the capacity of mobile communications systems stands out, which can be linked to three main factors: the exploitation of the channel in search of greater efficiency, the increased use of the spectrum, and the increase in the number of nodes in the mobile network [3].

Cellular communication systems have developed in the sense of an ecosystem of base stations (BSs) that configure different cell sizes (i.e. macrocells, microcells, picocells, and femtocells) operating at distinct powers and supporting a variety of medium access technologies [4]. Thus, this description of a non-homogeneous network infrastructure is commonly referred to as a heterogeneous network (HetNet) and is a crucial feature for managing the diversification of demands arising from the variety of devices connected in the 5G communication system.

The factors described above can therefore be explained using the concept of network densification, which is divided into spatial densification and spectrum aggregation [5]. Spatial densification seeks to ensure a uniform distribution of users between base stations and then resorts to increasing the number of antennas per node and the density of BSs per region. In turn, spectrum aggregation involves intensifying the use of the electromagnetic spectrum by using millimeter wave bands (30-300 GHz).

In addition to the complex structure of HetNets, 5G systems must deal with the heterogeneity of applications and services offered to users, which involves analyzing the typical traffic model in each situation. These traffic models have throughput, data loss, and latency criteria related to Quality of Service (QoS) requirements. In this way, the effective use of radio resources is a critical factor in network configuration and operation, highlighting the importance of studying Radio Resource Management (RRM) techniques. In particular, techniques related to scheduling algorithms work to configure the efficient use of resources, including antenna power and bandwidth, while mitigating interference between cells and users, ensuring that QoS requisites are met and increasing the Quality of Experience (QoE) of heterogeneous users [3].

The following sections describe the main challenges and contributions proposed by researchers in the field. A study base is then developed to guide this research in proposing a new scheduling algorithm for the downlink of 5G networks, focusing on channel conditions and quality of service requirements.

1.1 Problem Definition

In 5G systems, User Equipments (UEs) access the network through base stations called 5G Node Base Stations (gNodeBs), the network elements responsible for allocating resources to the users attached to it. The gNodeBs must take into account the UEs current channel conditions reported by the UE through the Channel Quality Indicators (CQIs) to allocate resources effectively. Thus, based on channel quality information, the system can use Adaptive Modulation and Coding (AMC) techniques to perform link adaptation, i.e. spectral efficiency can be increased by selecting a better Modulation Coding Scheme (MCS) in the face of worse channel conditions.

Considering the great diversity of applications and use cases for modern mobile networks, developing methods that take into account the QoS requirements for each traffic model is another influential factor in decision-making during the resource allocation procedure. Thus, the presence of real-time (RT) and non-real-time (NRT) applications stands out. In this

sense, the search for channel-aware and QoS-aware solutions for 5G networks is noteworthy, as schedulers can exploit variations in both channel quality and user-specific requirements when assigning priority to each user.

Finally, it is visible that designing a scheduling algorithm for 5G networks is a complex task that depends on several factors to achieve greater spectral efficiency, exploit the capabilities of the network infrastructure, and provide QoE to the users. In addition, it is important to analyze metrics such as throughput, Block Error Ratio (BLER), fairness index, delay constraints (DCs) related to QoS requirements, and computational performance. Thus, channel and QoS-aware algorithms appear as an effective method for the problem of resource allocation.

In the next section, the most relevant works on resource allocation, and heterogeneous networks with multiple traffic models are presented along with solutions to the problem mentioned above.

1.2 State of the Art

Resource allocation techniques for mobile networks have been a field of intense study so some literature reviews stand out by highlighting the evolution of methods and discussing the main proposals. Thus, such research is of great importance in understanding schedulers and analyzing the objectives of research in the field, for example, the optimization criteria that can be adopted. Therefore, some literature reviews about RRM and scheduling algorithms are highlighted below.

In this sense, the study presented in [6] provides an overview of the techniques used by scheduling algorithms, including a classification of the different methods used. It also provides details on the architecture of Long-Term Evolution (LTE) networks, the precursor to 5G technology, and resource allocation, as well as evaluating the performance of algorithms via simulation. In [7], there is an analysis of RRM techniques focusing on multi-objective optimization techniques to achieve desired levels of QoS in LTE-Advanced (LTE-A) systems. There is a classification of multi-objective optimization implementation methods, highlighting algorithms according to their throughput and latency performance.

The research [8] focuses on content-aware scheduling algorithms for video streaming and proposes a review of algorithms in the LTE downlink, summarizing the techniques and parameters used by each algorithm. It also simulates some of the algorithms considered to be the best fit for the study: the Proportional Fair (PF) scheduling rule as a QoS unaware strategy,

the Modified Largest Weighted Delay First (M-LWDF) scheduling procedure as a QoS aware strategy, a video quality driven and delay blind scheduling through Simplified Fine Granularity (SFG) scheduling, a video quality driven and delay aware scheduling, and a proxy driven radio resource allocation.

In [9], the authors analyze resource allocation schemes revolving around UDNs in 5G and beyond-5G systems, based on optimization approaches, methods, and criteria, as well as describing the main characteristics, and technical and research challenges of the allocation process concerning emerging technologies. The following optimization criteria are highlighted in the paper: energy efficiency, interference, throughput, QoS, spectrum efficiency, fairness, and computational complexity.

The authors of [10] propose a review focused on packet scheduling algorithms in URLLC for 5G and beyond-5G, in which the decentralized, centralized, and joint scheduling techniques stand out. The authors then analyze the performance of some algorithms and highlight the main challenges in the area: latency and reliability, flexible technology development, heterogeneous services, scheduling in multi-connected networks, resource scheduling in a cloud environment, puncturing the eMBB packet, processing time, and Intelligent Reflecting Surface (IRS) and holographic MIMO.

Finally, the study [1] stands out in its research into the operation of scheduling algorithms, determining the factors that influence the scheduler's decision (channel quality indicator, average data rate, head of the line packet delay, queue status, and buffer levels, and quality of service identifier) and the performance evaluation parameters (delay, throughput, goodput, fairness, spectral efficiency, and packet loss ratio). It also classifies existing techniques according to their evaluation parameters, performance, and ability to meet the needs of 5G services. Finally, the positive and negative points of the main algorithms analyzed are highlighted.

Considering the implementation of heterogeneous networks, the research [11] presents a detailed overview of network characteristics, different network densification scenarios, and the challenges of implementing resource allocation in HetNets of future systems. Although there is no classification of scheduling algorithms, the authors propose a classification of resource allocation models, as well as providing possible 6G system structure solutions for dealing with resource allocation in heterogeneous networks.

Although there is no classification of scheduling algorithms, the authors of [12] analyzed various approaches to RRM schemes in Ultradense Networks (UDNs). The study of resource allocation schemes highlights the importance of research into spectrum allocation, i.e.

the implementation of methods that emphasize system throughput, spectrum efficiency, QoS and QoE, and fairness. It also highlights interference management, power allocation through energy harvesting and energy efficiency, user association, and the implementational and computational complexities.

The authors of [13] highlight how the study of RRM methods has become critical in the face of increasing network densification and the expectation of 5G systems to accommodate multiple innovative services based on QoS and user QoE. Resource allocation is attributed to more efficient use of spectrum, load balancing, and efficient use of network energy resources. The study analyzes the significance of the challenges of interference, user association, and power allocation for heterogeneous 5G networks.

In [14], the authors propose a state-of-the-art investigation on the coexistence mechanisms between enhanced mobile broadband (eMBB) and ultra-reliable, low-latency communications (URLLC) traffic for resource allocation in 5G systems. The paper presents a classification of proposals based on the following approaches: multiplexing, QoS provisioning, network slicing, machine learning, and Centralized/Cloud Radio Access Network (C-RAN). In the case of approaches based on QoS provisioning, the importance of scheduling algorithms considering both QoS framework policy and user requirements is highlighted. In addition, the paper emphasizes fairness as one of the most important metrics when designing a QoS-aware strategy for resource allocation.

About the multiple traffic models expected for HetNets, the survey [15] associates existing traffic models with the use cases of 5G systems and carries out performance analysis. To this end, it takes into account attributes such as traffic volume, network deployments, and the main performance objectives. In addition, this work brings together the main references from Standards Development Organizations (SDOs) and industry associations on the subject.

Thus, some proposals for scheduling algorithms stand out in the context of this work and are outlined below.

The study in [16] considers algorithms with strategies driven by network metrics, highlighting the need for schedulers to consider QoS requirements to adapt to the heterogeneous traffic of 5G networks. Thus, it evaluates the Maximum Rate (MR), Round Robin (RR), PF, and UE-based Maximum Rate (UEMR) scheduler algorithms based on throughput and fairness.

In [17], a new scheduling policy called Channel Aware Optimized Proportional Fair (CAOPF) is proposed to optimize the channel behavior based on CQI. The scheduler's performance is analyzed and compared to algorithms such as RR and PF. The simulation results indicate that CAOPF has better QoS performance since there is no pending data of users in

good channel conditions. It also has higher average cellular throughput, higher user throughput for users in good and average channel conditions, and an optimized fairness index. Despite studying several parameters and being a recent work, the algorithm was proposed based on LTE networks.

The authors of [18] take channel conditions into account to improve the throughput performance of guaranteed bit rate and non-guaranteed bit rate traffic models. In this sense, a delay control mechanism on QoS requirements of the multiple traffic classes is introduced to improve a system's average throughput. However, the proposal, namely the Prioritized QoS-Aware downlink scheduling algorithm is evaluated on LTE networks and does not consider packet drop ratio and delay. In turn, [19] proposes a 5G eMBB scheduling algorithm that is aware of throughput and CQI and is metric-based. It highlights the gains in throughput and fairness compared to the best CQI and PF schedulers.

In [20], a joint QoS-Aware component carrier selection and resource allocation scheme for carrier aggregation in 5G systems is proposed. The research addresses component carrier selection and resource allocation according to 5G QoS identifiers, to maximize average throughput and satisfy QoS requirements regarding delay. To this end, three index classes are considered according to packet loss ratio and packet delay budget. Thus, the proposed scheme maximizes proportionally fair average throughput for different classes of services in the face of rate and delay requirements. Moreover, this research indicates the importance of studying resource allocation schemes considering delay constraints for 5G networks.

The authors of [21] propose two new policies for RT and NRT traffic models: the Adjusted Largest Weighted Delay First (ALWDF) and the Fair Throughput Optimized Scheduler (FTOS). Based on these two ideas, they created the Advanced Fair Throughput Optimized Scheduler (AFTOS), which seeks to maximize spectral efficiency and throughput based on the metrics of fairness, delay, and packet loss ratio. The study is based on a wide range of performance metrics and the results indicate that AFTOS performs better than the Maximum Throughput (MT), PF, and MLWDF algorithms, but is implemented for LTE networks.

Finally, we highlight the research carried out in [22], which considers a network with mixed traffic models to propose a QoS Aware Scheduler (QAS) aimed at achieving the QoS requirements for 5G systems. Thus, simulations are carried out to evaluate the performance of the scheduler related to sum throughput, average throughput, BLER, and latency and compare it to the RR and best CQI algorithms.

1.3 Justification

Based on what was presented in the previous section, the complexity of resource allocation in the downlink of 5G networks stands out. Despite the existence of various proposals for scheduling algorithms, there is a lack of work on techniques for scheduling algorithms in heterogeneous networks considering multiple traffic models, highlighting both RT and NRT applications.

Thus, considering the actual history of scheduling algorithms for resource allocation and the need to study more complex environments that more faithfully represent 5G systems, there is a great opportunity for research in the design of a 5G downlink channel and QoS-aware scheduler for real-time and non-real-time traffic in heterogeneous networks.

1.4 Research Scope and Objectives

This work aims to propose a solution to the problem of resource allocation in heterogeneous networks in the 5G downlink, considering RT and NRT traffic models. To this end, the implementation of a channel and QoS-aware scheduler is chosen considering the following specific objectives:

- implement a QoS-aware scheduling algorithm that prioritizes RT application users when allocating resources to improve the QAS algorithm presented in [22];
- implement a channel and QoS-aware scheduling algorithm by studying the CQI parameters and delay constraints of RT traffic models;
- compare the proposed algorithms with relevant schedulers in the literature, namely RR, best CQI, and QAS, as to throughput, BLER, fairness index, latency, and simulation time.

To evaluate the proposed objectives, it is necessary to define a simulation environment, as there are a substantial number of parameters involved. Therefore, the MATLAB tool was chosen as the simulation environment since it has features for modeling heterogeneous networks using the Vienna 5G System Level Simulator [23]. In this context, the analyses are conducted through computer simulation runs to reproduce the conditions of the proposed scenario and meet the expectations of users and network operators. Finally, we highlight the analysis of a highly densified scenario through the configuration of macrocells, picocells, and femtocells to increase network capacity.

1.5 Novelty and Contributions

The main contribution of this dissertation is the development of two algorithms, named Weighted QoS-aware Scheduler (WQAS) and Channel and QoS-aware Scheduler (CQAS).

Considering Section 1.2, these proposed algorithms are based on the work [22]. The three strategies consider QoS requirements and use the same optimization problem called mixed binary integer programming. However, WQAS allocates more resource blocks (RBs) to RT users, while CQAS adds the CQI parameter to the scheduling solution. This work therefore presents the following contributions:

- the algorithms are evaluated for six traffic models: full buffer, Hypertext Transfer Protocol (HTTP), vehicular, Voice over IP (VoIP), gaming, and video. Unlike [22], full buffer traffic is also applied to IoT users arranged in clusters.
- the simulation environment is more complex than the other studies due to the heterogeneous nature of the network, both in terms of infrastructure (macrocells, picocells, and femtocells) and the variety of users.
- the proposed WQAS algorithm shows improvements for RT traffic models in throughput and latency.
- the CQAS algorithm presents general improvements in throughput and latency, as well as taking into account both channel conditions through the implementation of a tuning parameter related to CQI and QoS requirements, highlighting the delay constraints of RT traffic.

Finally, the novelties and contributions made during the development of this research are listed in the following publications:

- Queiroz, A. G., Silva R. E. **A new Channel and QoS Aware Scheduler algorithm for real-time and non-real-time traffic in 5G heterogeneous networks.** Accepted by IEEE Latin America Transactions.
- Queiroz, A. G., Silva R. E. **A Weighted QoS Aware Scheduler algorithm for multiple traffic models in 5G heterogeneous networks.** Accepted by XLII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais.

1.6 The structure of the chapters

This dissertation is organized as follows.

Chapter 2 introduces the main concepts about 5G networks, which are essential for understanding the research carried out in this dissertation.

Chapter 3 provides a theoretical background on resource allocation techniques for 5G systems, initially highlighting the structure of heterogeneous networks, traffic models, and how these issues influence the design of resource allocation methods.

Chapter 4 is dedicated to presenting the requirements and operation of scheduling algorithms, the proposed algorithms, and the algorithms adopted for comparison.

Chapter 5 describes the simulation environment used to evaluate the algorithms and then presents the analysis of the results.

Finally, Chapter 6 presents the conclusions and future developments of the research.

”Mr. Watson, come here. I want to see you.”

- Alexander Graham Bell

2

THE 5G NETWORK

THE 5G NETWORK is a very complex mobile communication system based on the operation of various network entities and their different technologies. This chapter presents the structure of the network, its main elements, and the essential concepts for understanding the research carried out in this dissertation.

2.1 Historical Evolution of Mobile Networks

5G is defined as the mobile network of the fifth generation. Each generation of mobile communications systems has been marked by an increase in the bandwidth and throughput provided to end-user consumers as the core focus. However, 5G networks differ from previous generations in that they indicate a change in the main consumer, becoming industries rather than end-users. Thus, there is innovation in both technology and the business model [4]. Table 2.1 illustrates the standards, releases, modulation schemes, Radio Access Technologies (RATs), and Forward Error Correction (FEC) for each generation.

Table 2.1: 1G to 5G – Timeline, standards, releases, RAT and FEC. (Adapted from [24]).

Generation	**Development period**	**Standard/ Technology**	**Release**	**Modulation scheme**	**RAT**	**FEC**
1G	1980s	AMPS NMT TACS/JTACS PSTN	IS-95 TIA-EIA 95 CDMAOne	FM FSK	FDMA	-
2G	1992-1997	GSM IS-95A IS-136 PDC	CDMAOne NA-TDMA	GMSK	FDMA TDMA CDMA	-
2.5G	1998	GPRS	3GPP Release 97	GMSK	TDMA CDMA	-
2.75G/2.9G	1999	EDGE EGPRS ECSD	3GPP Release 98	8PSK	TDMA CDMA WCDMA EV-DO Rev. A EV-DO Rev. B TD-SCDMA	-
3G	2000-2001	UMTS	CDMA2000 IMT-2000 3GPP Release 99 3GPP Release 4	QPSK	EV-DO	Turbo codes
3.5G	2002-2007	HSPA: HSDPA HSUPA	3GPP Release 5 3GPP Release 6	DL: 16QAM or QPSK (HSDPA)	EV-DO	Turbo codes
3.75G	2007-2008	HSPA+	3GPP Release 7	DL: 64QAM or 2x2 MIMO stream UL: 16QAM	DL: OFDMA UL: SC-FDMA	Concatenated codes
3.9G	2009-2010	LTE	3GPP Release 8 3GPP Release 9	DL: 4x4 MIMO UL: 64QAM SISO	DL: OFDMA UL: SC-FDMA	Concatenated codes
4G	2011-2015	LTE-A	3GPP Release 10 IMT-Advanced 3GPP Release 11	DL: MU-MIMO (8x8) UL: SU-MIMO (4x4) TDD FDD	DL: OFDMA UL: SC-FDMA	Turbo codes
4.5G	2015-2016	LTE-A Evolution LTE-M LTE-U comprising LTE-A Pro (LTE-LAA) LTE-LWA LTE-MulteFire	3GPP Release 12 3GPP Release 13 eIMT-A	eMIMO FD-MIMO Elevation beam forming eCOMP UL: SU/UM MIMO DL: 3D-MIMO FDD-TDD CA 256QAM	BDMA Multi-RAT NOMA	Turbo codes
5G	2016-2020	SA NR NSA NR	3GPP Release 14 3GPP Release 15 3GPP Release 16 3GPP Release 17 ITU/IMT-2020	FQAM FBMC Massive MIMO Advanced MIMO		Low-density parity check codes

2.1.1 First-Generation Mobile Networks (1G)

1G represents analog transmission systems focused on basic voice services. Although digital signals are used through frequency modulation (FM) to connect radio towers and the rest of the telephone system, the signals are analog and have no data capabilities. Frequency Division Multiplexing (FDM) is thus used to divide the bandwidth into specific frequencies associated with different connections. In addition, the cells are between 2 and 20 km in size [24].

There have been some independent projects on 1G systems: the Analogue Mobile Phone System (AMPS) was implemented in the United States of America (USA), while the Nordic Mobile Telephone System (NMT) and the Total Access Communications System (TACS) were used in Europe, and finally, the Japanese Total Access Communication System (J-TACS) was used in Japan and Hong Kong.

As 1G networks use analog signals, the main problem regarding the quality of connections was susceptibility to interference. Furthermore, there was a lack of security since analog signals do not allow advanced encryption methods to be implemented. Thus, 1G systems suffer from limited user capacity, poor voice quality, short battery life, and frequent call dropping [25]. Figure 2.1 shows the architecture of 1G AMPS. The telephones connect to the Public Switched Telephone Network (PSTN), and the telephone landline connects the PSTN to the Mobile Telephone Switching Office (MTSO), which performs call routing for the mobile users (MUs) through their mobile stations (MSs) connected to the Base Transceiver Stations (BTSs). The MTSO also controls the cell connections to the Mobile Switching Center (MSC).

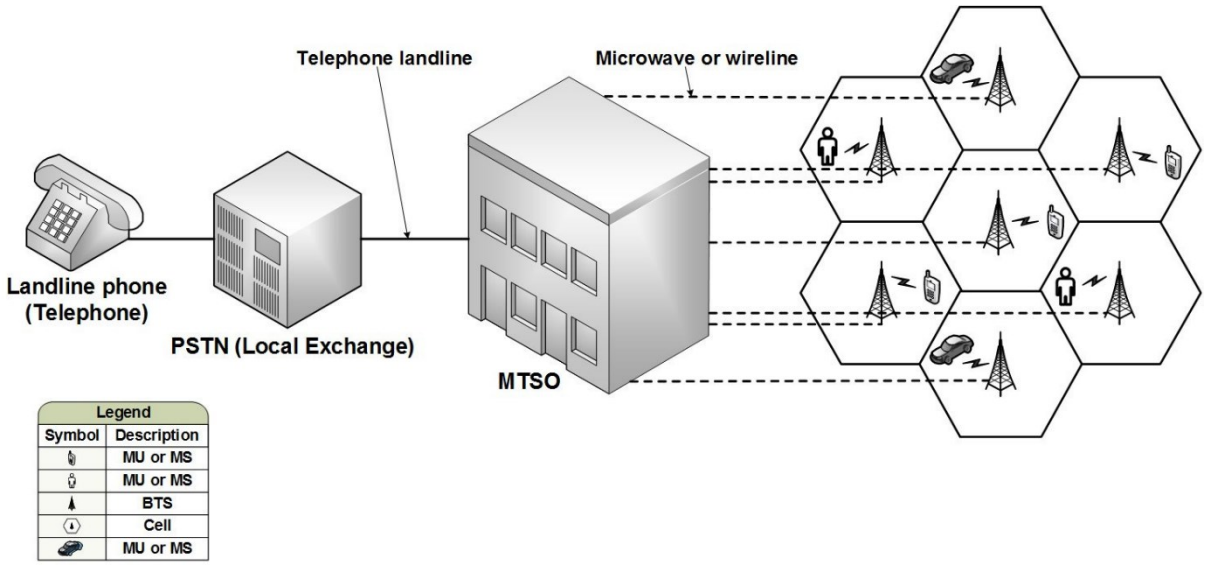


Figure 2.1: 1G AMPS architecture. (Adapted from [24]).

2.1.2 Second-Generation Mobile Networks (2G)

2G systems incorporated digitalization into mobile networks, achieving better voice quality and the first data service. Voice applications were improved through better use of spectrum allocation due to multiple access schemes: Frequency Multiple Access (FDMA), Time Division Multiple Access (TDMA), and Code Division Multiple Access (CDMA). Short Message Service (SMS), Multimedia Messaging Service (MMS), and a semi-global roaming system were also developed [24].

The 2G Global System for Mobile Communications (GSM) uses cells up to 35km, including macrocells, microcells, picocells, and femtocells [25]. Figure 2.2 illustrates the architecture of 2G GSM. The mobile stations (MSS) are made up of mobile equipment (ME), which is linked to the Subscriber Identity Module (SIM), which is responsible for identifying the user. Interface Um is the air interface between the MSS and the Base Station Subsystem (BSS), where the BTSs and Base Station Controllers (BSCs) are configured.

In turn, the BSCs connect to the central element of the Network Switching Subsystem (NSS), the MSC, enabling the allocation of channels and timeslots, as well as connection to PSTN, Integrated Services Digital Network (ISDN), and data networks. The NSS is also made up of user registration elements such as the Visitor Location Register (VLR), Home Location Register (HLR), and Equipment Identity Register (EIR), authentication in the case of the Authentication Center (AUC), and operation and maintenance in the case of the Operation and Maintenance Center (OMC).

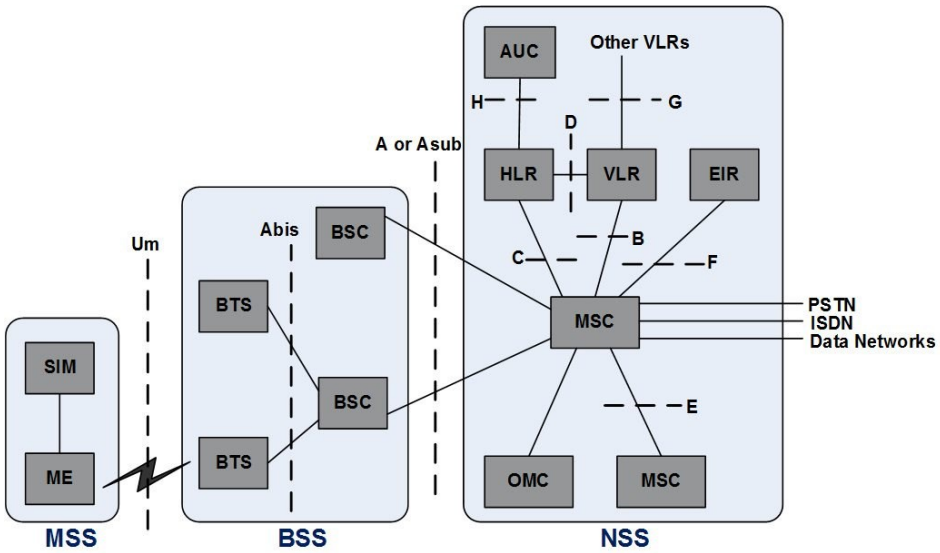


Figure 2.2: 2G GSM architecture (Adapted from [24]).

GSM then evolved to General Packet Radio Service (GPRS), considered 2.5G, and then to pre-3G systems called 2.75G or 2.9G, called Enhanced Data Rates for GSM Evolution (EDGE). EDGE technology includes two switches: circuit switching, called Enhanced Circuit Switched Data (ECSD), and packet switching, called Enhanced General Packet Radio Service (EGPRS) [24]. There were also other 2G standards, such as TIA/EIA-136, also known as the North American TDMA (NA-TDMA) standard, TIA/EIA IS-95A, also known as CDMAOne, based on CDMA and used in the USA, and Personal Digital Cellular (PDC), implemented exclusively in Japan [4].

Concerning 2G problems, the following stand out: Co-Channel Interference (CCI) due to frequency reuse, link failures due to TDMA pulse decay under certain circumstances, as well as security flaws related to false base stations, Denial of Service (DOS), eavesdropping and Subscriber Identity Module (SIM) attacks [26].

2.1.3 Third-Generation Mobile Networks (3G)

3G systems deliver multimedia services via dedicated digital networks. Because they were developed with the advancement of the Internet and IP networks in mind, 3G systems support improvements in throughput and QoS. Compared to its predecessors, it has improved both voice quality and the development of global roaming, but its downside is a slight deterioration in energy efficiency, as the devices consume more energy and have higher installation and operating costs [24].

During the development of 3G systems, two working groups stood out: the 3GPP and 3GPP2. Both groups developed proposals for 3G technology based on the International Mobile Telecommunications-2000 (IMT-2000) program initiated by the International Telecommunications Union (ITU) [27]. 3GPP developed 3G based on GSM technology, based on Wideband Code Division Multiple Access (WCDMA) specifications, while 3GPP2 focused its efforts on IS-95 technology, based on cdma2000.

Figure 2.3 illustrates the architecture of the Universal Mobile Telecommunications Service (UMTS) developed by 3GPP. UMTS is compatible with previous generations because it can exist heterogeneously with legacy GSM or AMPS. In addition, UMTS has evolved through High-Speed Packet Access (HSPA) and Evolved HSPA (HSPA+), achieving better end-to-end network performance.

In Figure 2.3, the MSS is the evolution of the SIM into the Universal SIM (USIM) and is connected to the UMTS Terrestrial Radio Access Network (UTRAN) via the Uu air interface. In the UTRAN, the base stations are called node Bases (node Bs) and the Radio Network

Subsystem (RNS) controls each node B via the Radio Network Controller (RNC). In turn, the Core Network (CN) is connected to the UTRAN via two interfaces: Iu Circuit Switched (CS) and Iu Packet Switched (PS). Thus, call routing is carried out through the Gateway Mobile Switching Center (GMSC), while packet routing goes through the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN) to the Internet.

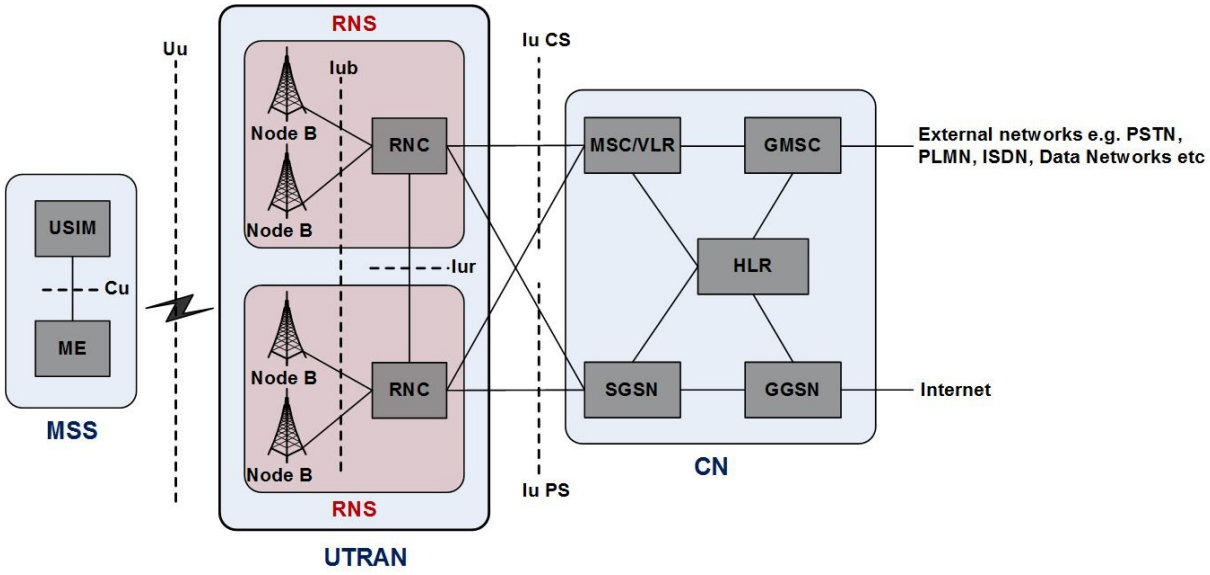


Figure 2.3: 3G UMTS architecture (Adapted from [24]).

In turn, 3GPP2 went on to develop High Rate Packet Data (HRPD) or Evolution Data Optimized (EV-DO), which was later improved to CDMA EV-DO Revision A and CDMA EV-DO Revision B. Among the differences between WCDMA and cdma2000 are the bandwidth, which is 5 MHz for WCDMA and 1.25 MHz for cdma2000, and the fact that WCDMA allows the same subcarrier to share data and voice traffic [28].

2.1.4 Fourth-Generation Mobile Networks (4G)

The development of the 4G LTE standard involves the search for packet-switched traffic with seamless mobility, QoS, and reduced latency. Thus, data, voice, and multimedia services are offered through packet-switched. The architecture of LTE networks is illustrated in Figure 2.4.

In Figure 2.4, the UEs connect to eNodeB, which manages all the radio protocols, mobility management, header compression, ciphering, reliable delivery of packets, and packet retransmissions via the Radio Network Controller (RNC). About the control plane, the eNodeB is responsible for admission control and RRM functions. The eNodeBs can communicate with each other via the X2 interface and communicate with the Mobile Management Entity (MME) via the S1 interface. The core of the LTE network is the Evolved Packet Core (EPC), while the

combination of the EPS and the access network constitutes the entire system and is called the Evolved Packet System (EPS) [24]. The EPC has experienced the evolution of the SGSN into the Serving Gateway (S-GW) and the GGSN into the Packet Data Network Gateway (P-GW).

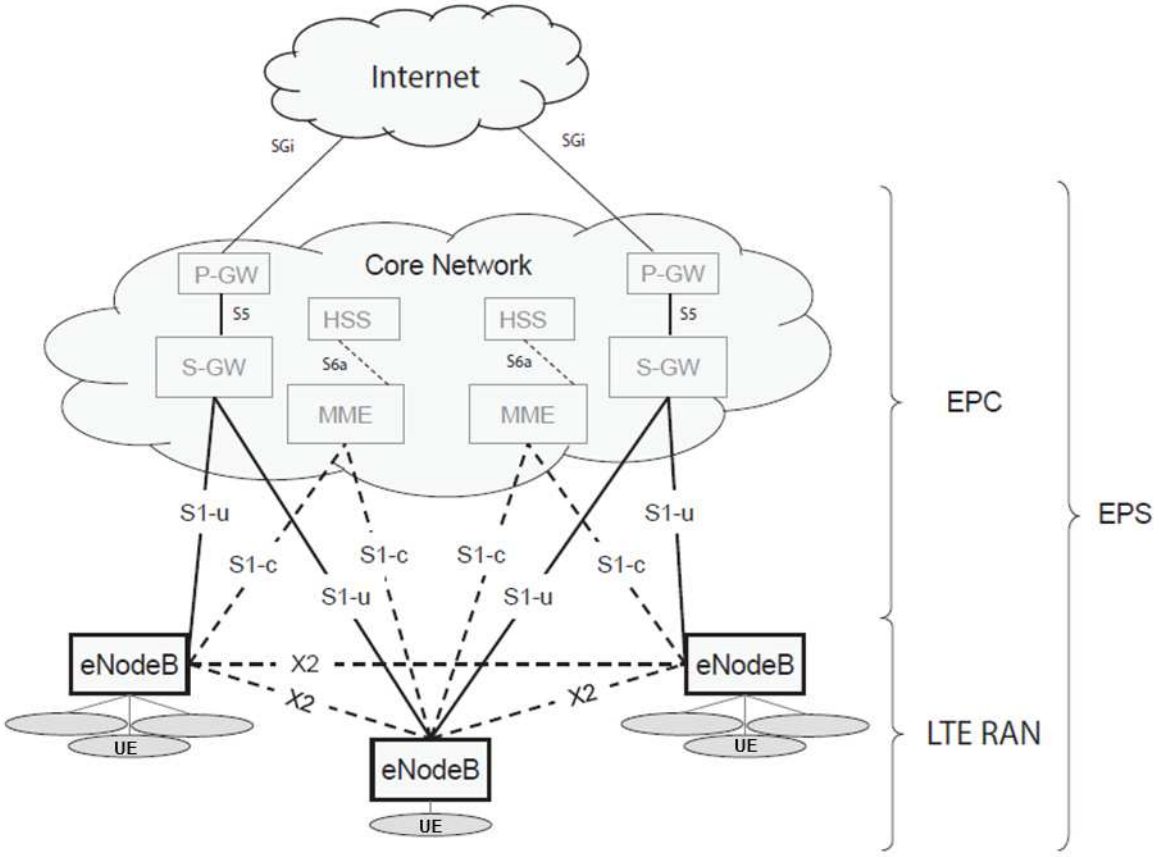


Figure 2.4: 4G LTE architecture (Adapted from [3]).

Among the features of 4G technology is the advance of Multiple Input Multiple Output (MIMO) techniques in the face of multiple radio access technologies such as Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier Frequency Division Multiple Access (SC-FDMA). In addition, higher data rates, support for heterogeneous networks, and spectrum flexibility are the result of techniques such as carrier aggregation, relaying, and Coordinated Multipoint (CoMP) transmission and reception [24].

In addition, 4G technology brings advances in multicasting and interference mitigation, as well as the development of services communications such as Machine Type Communication (MTC), M2M, and IoT. In addition, some implementations, such as Network Assisted Interference Cancellation and Suppression (NAICS), have been implemented to improve the performance of intracellular and intercellular interference [24]. Finally, the LTE standard evolved into LTE-A, which now includes techniques such as Multiple User-MIMO (MU-MIMO) and error correction codes based on turbo codes.

2.1.5 Fifth-Generation Mobile Networks (5G)

The 5G New Radio (NR) standard has two versions, 5G Standalone (SA) NR and 5G Non-Standalone (NSA) NR. In general, the SA NR architecture refers to implementing the 5G system based on the NR access network and the 5G core. On the other hand, the NSA NR architecture characterizes systems that use LTE as a control plane anchor for NR [29].

5G communication systems are designed in line with the demand for data volume arising from the higher level of connectivity. According to Ericsson's mobility report [30], 5G mobile subscriptions are set to reach nearly 5.6 billion in 2029. It also highlights that measurements from a leading service provider show that 97 percent of all user activities on 5G mid-band achieved a time-to-content of less than 1.5 seconds.

According to the ITU, International Mobile Telecommunications (IMT) systems, such as 5G systems, must support [31]:

- **Very low latency and high reliability human-centric communication**: enables users to experience instant connectivity, i.e. system responses are perceived as if they were instantaneous. This allows the development of applications aimed at, for example, health, security, and entertainment.
- **Very low latency and high reliability machine-centric communication:** accounts for improved quality of life based on services provided by M2M communication systems, such as driverless cars, real-time optimization of traffic control, solutions for emergencies and disasters, smart grid, and e-health.
- **Supporting high user density:** users expect a satisfactory experience even in the presence of many active users, which means a high density of traffic and devices per area. This includes, for example, users of public transport systems, and professionals such as police officers, firefighters, and healthcare workers in crowded zones.
- **Maintaining high quality at high mobility:** the user experience in a state of mobility should be similar to the static scenario, enabling robust and reliable communication while maintaining QoS.
- **Enhanced multimedia services:** there is a greater demand from devices for high-definition multimedia in areas such as entertainment, medical support, and security, so increasing consumption capacity involves developing services such as Ultra-High Definition, 3D projection, immersive videoconferencing, and augmented reality.
- **Internet of Things:** the number of connected devices grows fast as a greater diversity of connected devices is explored, such as UEs, sensors, and vehicles, among others. Thus, 5G systems encompass everything from less complex to advanced devices. For this reason, there is great concern about the capacity of mobile networks, which must support a greater number of connections, as well as address critical issues such as energy consumption, transmission power, and latency requirements. Some fields related to IoT are smart energy distribution systems, agriculture, healthcare, and V2X.

- **Convergence of applications:** the variety of applications running on IMT systems, such as multimedia content, government services, security, and health, make it necessary to analyze requirements that support the plurality of applications.
- **Ultra-accurate positioning applications:** the accuracy of location services has grown so that the number of applications based on navigation technology has increased and requires attention from mobile network systems.

Based on these support trends offered by IMT-2020 systems, 5G systems should contribute to the development of three main usage scenarios: enhanced Mobile Broadband (eMBB), Ultra-Reliable And Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC) [31].

Firstly, the eMBB scenario stems from the increased demand for mobile broadband services, which provide access to multimedia content, services, and data. Therefore, it considers coverage in large areas and access points. In the case of large-area coverage, there is a need for uninterrupted coverage, a medium to a high level of mobility, and a high data rate. In the case of access points, there is a need for high user density, high traffic capacity, a low level of mobility, and a higher data rate than in the previous case. In the case of URLLC, there are strict requirements regarding throughput, latency, and availability. Some examples related to this use case are remote medical procedures and the automation of smart grid systems. Finally, the mMTC use case is characterized by the large number of connected devices, which, in general, transmit a low volume of data that is not sensitive to delay. The requirements for these devices are low cost and long battery life.

Figure 2.5 illustrates some examples of applications for the usage scenarios for IMT for 2020 and beyond [31].

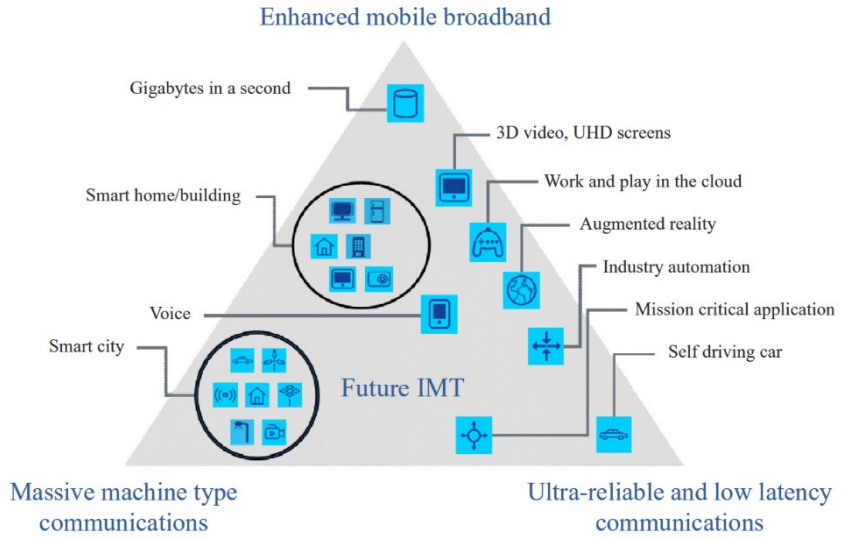


Figure 2.5: The usage scenarios of IMT for 2020 and beyond (Adapted from [31]).

Regarding the usage scenarios presented, there are eight key capabilities considered essential for IMT-2020 systems [31] as listed below. Figure 2.6 illustrates the relationship between these capacities in different usage scenarios.

- **Peak data rate:** the maximum data rate achieved under ideal conditions per user or device (in Gbits/s).
- **User experienced data rate:** the data rate available given a coverage area for the user or device (in Mbits/s or Gbits/s).
- **Latency:** the time between the sender sending a packet and the recipient receiving it (in ms).
- **Mobility:** is characterized by the maximum speed achieved given an established level of QoS and continuous transfer between radio interface nodes that may belong to multilayers and/or different access technologies (in km/h).
- **Connection density:** the total number of connected and/or accessible devices per unit area (per km^2).
- **Energy efficiency:** considering the network, energy efficiency refers to the number of bits of information transmitted to or received from users per unit of energy consumed by the Radio Access Network (RAN) (in bits/J). From the point of view of the device, it refers to the number of bits per unit of energy consumed by the communication module (in bits/J).
- **Spectrum efficiency:** the average data throughput per unit of spectral resource and cell (in bits/s/Hz).
- **Area traffic capacity:** it is the total throughput served per area (in $Mbits/m^2$).

There are also parameters related to system flexibility, as well as reliability and security:

- **Spectrum and bandwidth flexibility:** refers to the system's flexibility to operate at different frequencies and channel bandwidths.
- **Reliability:** it is the ability to provide service with a high degree of availability.
- **Resilience:** refers to the network's ability to continue operating during natural or man-made interference, such as loss of power.
- **Security and privacy:** takes into account encryption, user data integrity and signaling (control functions), user privacy from unauthorized tracking, and network protection from possible attacks.
- **Operational lifetime:** this is the relationship between operating time and energy capacity to, for example, the use of batteries. It deals mainly with M2M device scenarios, which require a long battery life and have fewer maintenance incidents.

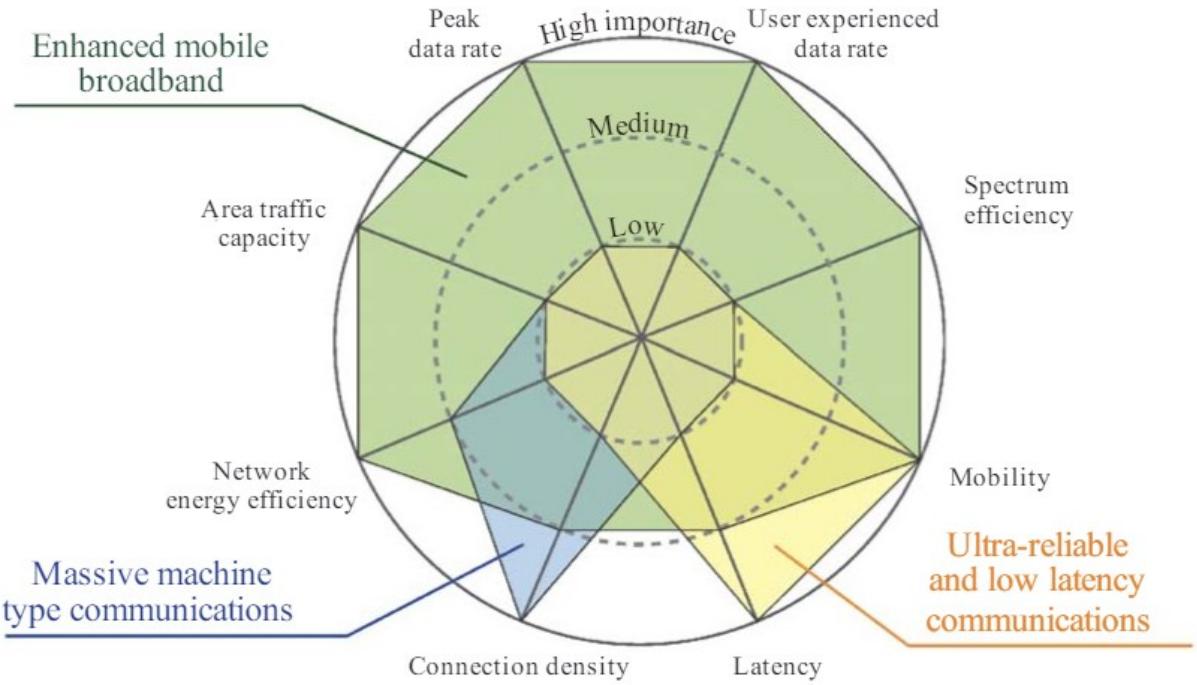


Figure 2.6: The impact of key capabilities in different 5G usage scenarios (Adapted from [31]).

2.2 Network Architecture

In the architecture of 5G networks, the RAN is responsible for radio-related functions, which include radio-resource handling, scheduling, retransmission protocols, coding, and multi-antenna schemes. On the other hand, the Core Network performs functions not related to radio access but which complement the operation of the network, for example, authentication, charging functionality, and setup of end-to-end connections. Managing these non-RAN tasks allows the core network to support multiple medium access technologies.

2.2.1 Network core

The architecture of 5G networks presents the virtualization of network functions as the biggest difference compared to previous network standards. This makes it possible to use resources in a more optimized and dynamic way through network slicing, i.e. it is possible to configure logical networks with different characteristics and capacities [32].

Based on the virtualization of network functions, the core of the 5G network is represented by the use of service-based interactions between network functions rather than in the form of nodes or network elements. In this way, each network function offers one or more services to the other network functions. The services are made available via network function interfaces connected to a Service-Based Architecture (SBA), i.e. the functionalities of the services are accessed via Application Programming Interfaces (APIs) [33].

The functionalities of the network core include establishing sessions securely and routing data from users to or from cellular devices, guaranteeing connectivity [33]. The main network functions that make up the core of the 5G network are shown in Figure 2.7.

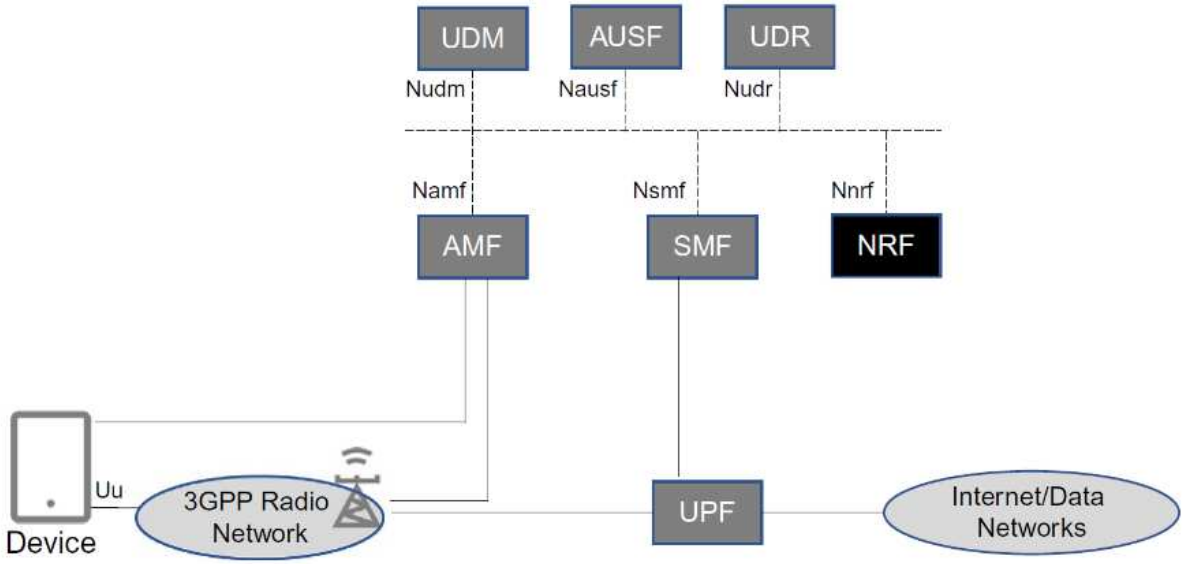


Figure 2.7: Mandatory components of the 5G network. (Adapted from [33]).

The Access and Mobility Management Function (AMF) interacts with the access network and devices, respectively, via the N2 and N1 interfaces, and is responsible for signaling call flows, supporting encrypted signaling to devices, which enables user registration, authentication, and mobility between cells on the network. The AMF supports the activation of idle devices. Unlike the core of the 4G network, AMF does not perform session management functions like the MME [33]. It also provides transport of SMS messages between UEs and the Short Message Service Function (SMSF) and performs the UE location service [32].

In turn, the Session Management Function (SMF) is responsible for managing the UEs' sessions. This includes establishing, modifying, and terminating individual sessions, as well as allocating Internet Protocol (IP) addresses for each session. The SMF communicates with the devices indirectly, as the AMF forwards session-related messages between these entities [33]. It also controls the User Plane Function (UPF), performing routing, directing traffic, and establishing control policies. Also related to the UPF, the SMF controls the charging data. Finally, it stands out for its roaming functions. Finally, the SMF works together with the UPF to replace the S-GW and P-GW elements of 4G networks [32].

The UPF's main task is to process and forward user data. It establishes connections with external IP networks without the other networks needing to know about mobility. So even when a device moves around the network, the packets destined for it will always be routed to the UPF

responsible. Regarding data processing, it inspects data packets and reports traffic usage statistics. Finally, it applies QoS markings to packets for both the access network and external networks, making it possible, for example, for the transport layer to determine the priority for each packet in the event of network congestion [33].

The Unified Data Management Function (UDMF) acts as a front-end for user subscription data stored in the Unified Data Repository (UDR), which also stores user access policies. The UDM function generates the authentication data for the UEs associated with the network, authorizing services based on each subscriber's profile. Finally, the Authentication Server Function (AUSF) not only provides an encryption service for roaming information and other device parameters but also authenticates the devices using the credentials generated by the UDMF [33].

2.2.2 Radio Access Network (RAN)

The access network architecture is based on multiple base stations connected to the network core and to each other, as shown in Figure 2.8.

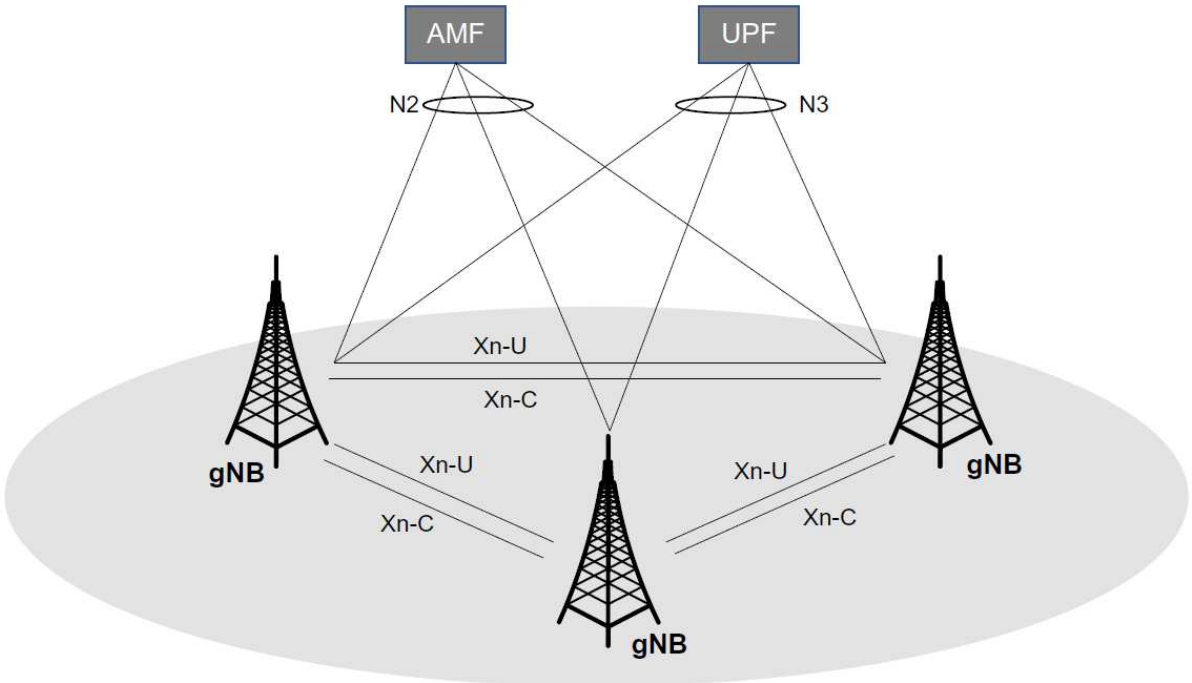


Figure 2.8: 5G access network architecture overview. (Adapted from [33]).

The next-generation RAN (NG-RAN) has two types of nodes: a gNodeB, which provides the control and user plans to the UEs via the New Radio (NR) air interface, or a next-generation evolved Node Base station (ng-eNodeB), which provides the control and user plans to the UEs based on the Evolved Universal Terrestrial Radio Access (E-UTRA) interface. The

gNodeBs and ng-eNodeBs are interconnected via Xn interfaces, in addition to connecting to the core of the 5G network, specifically the AMF, and UPF, via the NG-C and NG-U interfaces respectively [34]. Both the gNodeB and ng-eNodeB are responsible for all functions related to the radio interface of one or more cells, for example, RRM, admission control, connection establishment, QoS flow management, and data routing and signaling [3].

User data is sent between the base stations and the UPF via the IP network so that IP packets are encapsulated and transported using the General Packet Radio Service Tunneling Protocol User Plane (GTP-U) protocol. In turn, GTP-U is loaded onto a User Datagram Protocol (UDP)/IP stack, running over network protocols such as Ethernet [33].

Control or signaling information is sent between base stations or between base stations and AMFs in the network core, based on the Stream Control Transmission Protocol (SCTP), which replaces UDP and guarantees message delivery and security mechanisms. Regarding interfaces, NG provides functionalities related to mobility signaling, transport of Non-Access Stratum (NAS) messages between devices and the network core, as well as paging signaling to UEs that are in an idle state. In turn, the Xn interface offers mobility signaling functions and dual connectivity, i.e. the use of more than one access technology, for example, NR and LTE on different frequencies [33].

2.3 User-plane and control-plane

Figure 2.9 illustrates the protocol stack that constitutes the user-plane and control-plane structure.

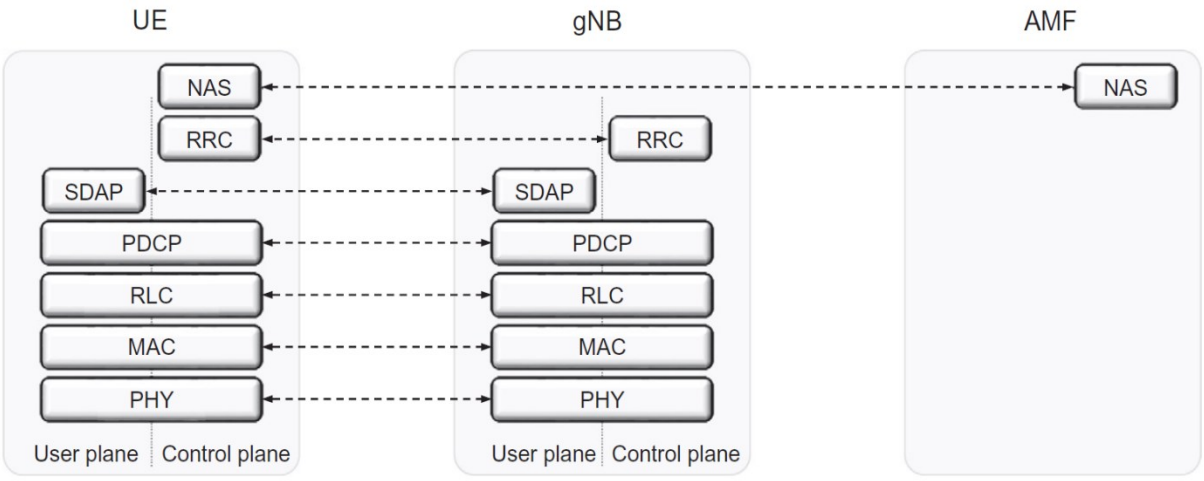


Figure 2.9: User-plane and control-plane protocol stack. (Adapted from [3]).

The entities that characterize the user plane are [3]:

- **Service Data Application Protocol (SDAP):** responsible for mapping QoS bearers onto radio bearers, i.e., it guarantees QoS requirements for the different data flows.
- **Packet Data Convergence Protocol (PDCP):** performs IP header compression, encryption, and integrity protection, as well as retransmission functions, delivering data in the correct sequence and removing replicated data. Finally, it handles routing in the event of dual connectivity.

- **Radio-Link Control (RLC):** handles data segmentation and retransmission. The RLC provides services to the PDCP via RLC channels, so there is one RLC entity per RLC channel or one RLC per radio bearer configured per device.
- **Medium-Access Control (MAC):** is responsible for channel multiplexing, retransmissions based on hybrid-automatic Repeat Request (ARQ), and functions related to scheduling. The base stations perform scheduling in the downlink and uplink. Compared to the RLC layer, the MAC layer offers services via logical channels.
- **Physical Layer (PHY):** deals with coding and decoding, modulation and demodulation, antenna mapping, and other physical layer functions. It provides services to the MAC layer via transport channels

Data flow that considers the protocol layer of the downlink data plane begins with mapping IP packets to different radio bearers as an action of the SDAP protocol. Data related to a higher protocol layer in the stack is called a Service Data Unit (SDU), and data related to a lower layer is called a Protocol Data Unit (PDU). Thus, the output of the SDAP mapping is an SDAP PDU equivalent to a PDCP SDU. The PDCP protocol can perform IP header compression with encryption for each radio bearer. The PDCP header provides information needed for decryption in the UE and the sequence number used for retransmission and delivery sequence. The output of the PDCP is then forwarded to the RLC.

In turn, the RLC protocol generates the segmentation of the PDCP PDU when necessary and adds an RLC header containing the sequence number used in retransmissions. The RLC PDUs are forwarded to the MAC layer, which multiplexes some PDUs and appends a MAC header to generate a transport block.

As for the structure of the control plane, which focuses on connection establishment, mobility, and security, there are two functionalities: the NAS control plane and Radio Resource Control (RRC). The former acts between the MFA in the network core and the device, and is responsible for authentication, security, registry control, and mobility, as well as assigning IP addresses to UEs.

In turn, RRC operates between the RRC located in the gNodeB and the UE and is responsible for RAN control procedures: broadcasting system information necessary for the device to be able to communicate with the cell; transmitting paging messages to notify an idle

device of upcoming requests; connection control, i.e. setting parameters for communication between the UE and the RAN by establishing an RRC context; mobility functions such as changing cell selection; setting and reporting metrics.

RRC messages are transmitted to the UE via Signalling Radio Bearers (SRBs) using the protocol layers of the data layer. An SRB is mapped to a Common Control Channel (CCCH) during connection establishment and once the connection is established, it is mapped to a Dedicated Control Channel (DCCH). Data from the control and user planes can be multiplexed at the MAC layer and transmitted to the device in the same Transmission Time Interval (TTI).

2.4 Medium Access Control (MAC)

The MAC layer is responsible for logical-channel multiplexing, Hybrid Automatic Repeat Request (HARQ), scheduling, and scheduling-related functions [3].

2.4.1 Logical Channels and Transport Channels

The MAC layer provides services to the RLC via logical channels, which are defined according to the type of information carried and are generally classified as control channels or traffic channels. Control channels transmit control data and configuration information, while traffic channels are used for user data. Table 2.2 shows the set of logical channel types specified for 5G.

Table 2.2: Logical-channel specified for 5G.

Channel name	Transported information
Broadcast Control Channel (BCCH)	System information for the user to access the system
Paging Control Channel (PCCH)	Paging messages
Common Control Channel (CCCH)	Control information in conjunction with random access
Dedicated Control Channel (DCCH)	Control information for individual device configuration
Dedicated Traffic Channel (DTCH)	Data transmission

From the physical layer, the MAC layer uses services in the form of transport channels, defined according to how and with what characteristics the information is transmitted on the radio interface. Thus, data in the transport layer is organized as transport blocks, and in each TTI, a maximum of one transport block of dynamic size is transmitted on the radio interface. In addition, associated with each transport block is a transport format (TF), which specifies how a transport block should be transmitted and includes information on the size of the transport block, the MCS, and the antenna mapping. The 5G transport channel types are shown in Table 2.3. Finally, Figure 2.10 illustrates the mapping between logical-channel types and transport-channel types, which is a specific feature of the MAC [3].

Table 2.3: Transport-channel specified for 5G.

Channel name	Transported information
Broadcast Channel (BCH)	BCCH system information – Master Information Block (MIB)
Paging Channel (PCH)	Paging information
Downlink Shared Channel (DL-SCH)	Transmission of downlink data
Uplink Shared Channel (UL-SCH)	Transmission of uplink data

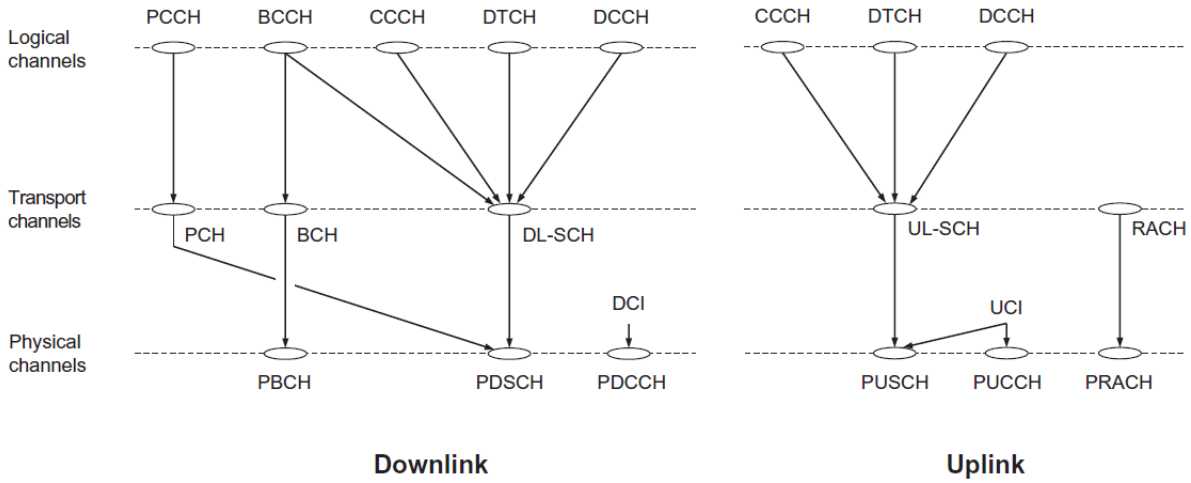


Figure 2.10: Mapping between logical, transport, and physical channels. (Adapted from [3]).

2.5 Summary

This chapter presents a historical overview of the development of mobile communication systems from 1G to 5G. Next, the network structure and the main concepts of the 5G network were highlighted. It can be seen that 5G systems are complex and involve various network elements and protocols. Therefore, understanding this basic structure and its evolution over the years is essential for developing new technologies, applications, and services. The next chapter looks at how gNodeB performs the resource allocation procedure, the concepts of QoS, heterogeneous networks, and traffic models.

"Information is the resolution of uncertainty."

- Claude Elwood Shannon

3

TRANSMISSION STRUCTURE, HETNETS, AND TRAFFIC MODELS

MOBILE COMMUNICATION SYSTEMS are developed to serve a variety of users with a limited amount of resources. This procedure is called resource allocation. This chapter presents a description of the structure of resource transmission, as well as QoS requirements, the structure of a HetNet, and its traffic models.

3.1 Transmission Scheme

The Orthogonal Frequency-Division Multiplexing (OFDM) defines the suitable waveform for 5G systems due to its robustness to time dispersion and ease of simultaneously exploiting the time and frequency domains in determining the structure of various channels and signals. Therefore, it is the transmission scheme used for both the downlink and the uplink [3].

OFDM presents the selection of numerology, in particular, the subcarrier spacing and the cyclic prefix length. While LTE works with a subcarrier spacing of 15 kHz and a cyclic prefix of approximately 4.7μs, 5G systems are designed to support a wider range of scenarios, from large cells with sub-1GHz carrier frequency to mm-wave scenarios with wide spectrum allocation. Therefore, it would not be interesting to present a single numerology. Due to the coexistence of LTE and 5G, the 15 kHz subcarrier spacing was adopted as the basis for 5G and

the other values range from 15 kHz to 240 kHz, which means that the cyclic prefix duration varies proportionally between 4.7μs and 0.29μs. Table 3.1 shows the subcarrier spacing supported by 5G. The OFDM Symbol time is the sum of the cyclic prefix and the useful symbol time [3].

Table 3.1: Subcarrier spacing supported by 5G (Adapted from [3]).

Subcarrier spacing (kHz)	Useful Symbol time (μs)	Cyclic prefix
15	66.7	4.7
30	33.3	2.3
60	16.7	1.2
120	8.33	0.59
240	4.17	0.29

3.1.1 Time-Domain Structure

In the time domain, transmissions are structured based on frames of length 10 ms, which are divided into subframes of length 1 ms. The subframes are split into slots of 14 OFDM symbols. At a higher level, each frame is identified as a System Frame Number (SFN), so transmission cycles longer than one frame are defined based on SFNs.

The time-domain structure is derived from the 15 kHz base by powers of two. In this sense, an OFDM symbol is divided into two OFDM symbols of the next higher numerology, so that 14 consecutive symbols form a slot. Thus, scaling by powers of two allows the existence of symbol boundaries between numerologies, which makes it possible to mix multiple numerologies on the same carrier. Regardless of the numerology, a subframe lasts 1 ms and consists of 2^{μ} slots, where μ indicates the subcarrier spacing configuration [3]. Figure 3.1 illustrates the structure of 5G frames, subframes, and slots.

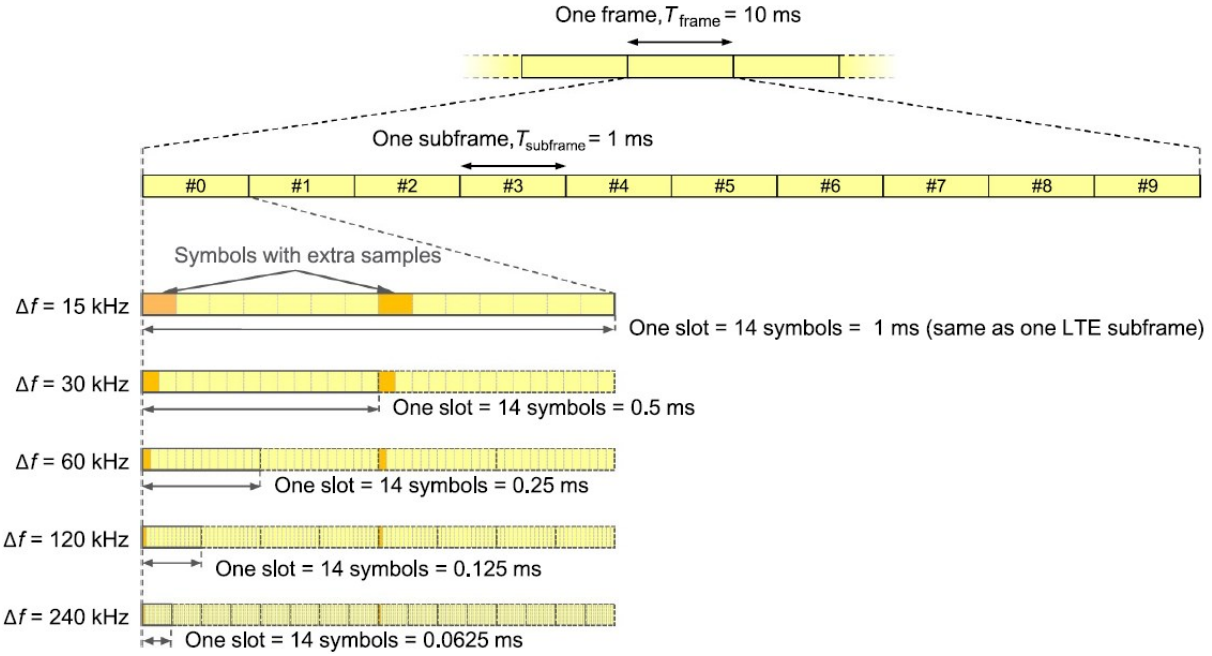


Figure 3.1: Frames, subframes, and slots in 5G systems. (Adapted from [3]).

3.1.2 Frequency-Domain Structure

Concerning the frequency-domain structure, the resource element (RE) is the lowest physical structure and consists of a subcarrier during an OFDM Symbol. In addition, 12 consecutive subcarriers form a resource block in the frequency domain. Unlike LTE networks, a resource block in the 5G network is a one-dimensional measure that encompasses only the frequency domain due to the flexibility in the duration of time slots for each transmission [3]. Figure 3.2 illustrates the relationship between resource element and resource block for the 5G system.

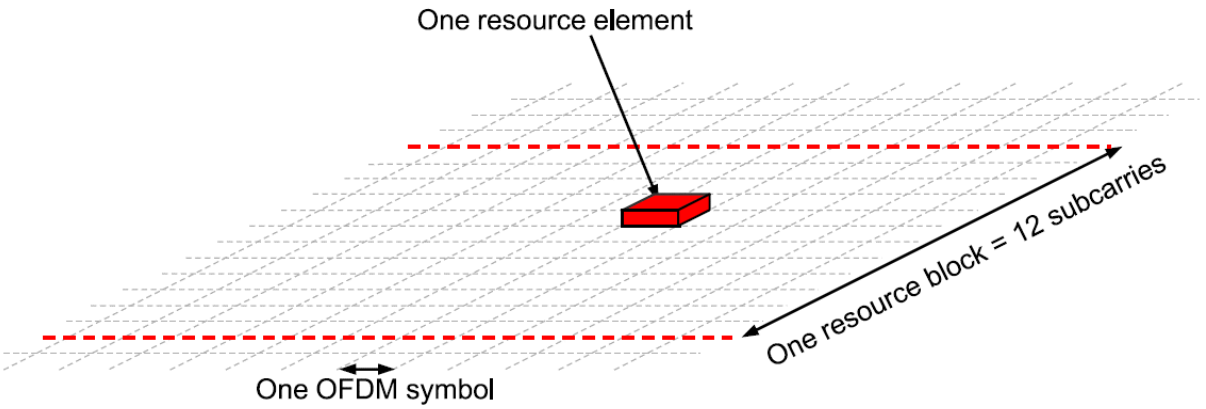


Figure 3.2: Resource element and resource block structures. (Adapted from [3]).

3.2 Quality of Service (QoS)

Quality of Service is the ability to provide differentiated treatment in the resource allocation to different types of users, applications, services, or media within the same application. This

distinctive treatment can be related, for example, to prioritizing the data or guaranteeing a performance level for a data flow [33]. As discussed in Chapter 2, 5G seeks to address various use cases based on eMBB, URLLC, and mMTC.

Considering an EPS, QoS is implemented by the EPC through the classification of data and its association with EPS bearers, the application of QoS parameters, and the execution of packet forwarding through schedulers in the downlink and uplink of the access network [35].

Regarding QoS, 5G systems take flexibility into account and the network core must support any access type. In addition, the access network must completely deal with QoS requirements, i.e. the complete separation of this function between the core and the access network is defined. Finally, there is a reduction in the signaling required for QoS establishment and modifications [33].

In 5G systems, the QoS framework is based on QoS Flows identified by a unique QoS Flow ID (QFI) within the PDU Session. A 5G-RAN can establish a Data Radio Bearer (DRB) per QoS Flow or combine more than one QoS Flow in a single data radio bearer [33]. Figure 3.3 shows the 5G QoS framework. Thus, Table 3.2 shows the 5G QoS characteristics, and Table 3.3 identifies some QoS parameters.

Table 3.2: 5G QoS characteristics (Adapted from [33]).

5G QoS characteristics	**Description**
Resource Type	Guaranteed Bit Rate (GBR), Delay Critical GBR or Non-GBR
Priority level	indicates the priority level in scheduling
Packet Delay Budget (PDB)	upper bound for the time that a packet might be delayed between UE and UPF
Packet Error Rate (PER)	upper bound for the rate of non-congestion related packet losses
Averaging Window	it is the duration over which the bitrate is calculated
Maximum Data Burst Volume (MDBV)	the largest amount of data that the access network is required to serve within the period of the 5G access network part of the PDB

Table 3.3: 5G QoS parameters (Adapted from [33]).

5G QoS parameter	**Description**	
Per QoS Flow	5G QoS Identifier (5QI)	a scalar reference to the 5G QoS characteristics
	Allocation and Retention Priority (ARP)	It contains the following information: • priority level: 1-15 values. • pre-emption capability: it determines when a service data flow can receive resources that have already been associated with another service data flow with a lower ARP priority level. • pre-emption vulnerability: it indicates when a service data flow can lose resources associated with it to admit a service data flow with a higher ARP priority level.
	Reflective QoS Attribute (RQA)	indicates that the traffic is subject to Reflective QoS
	Notification control	indicates when notifications are requested by the 5G-RAN when a Guaranteed Flow Bit Rate (GFBR) cannot or must again be guaranteed for a QoS flow
	Flow Bit Rates	Guaranteed Flow Bit Rate (GFBR) – separate for downlink and uplink. Maximum Flow Bit Rate (MFBR) – separate for downlink and uplink.
	Maximum Packet Loss Rate	indicates the maximum rate for lost packets tolerated on both the downlink and uplink
Additional QoS parameters	Aggregate Bit Rates	PDU sessions are associated with per session Aggregate Maximum Bit Rate (Session-AMBR) which limits the aggregate bit rate for Non-GBR QoS flows given a PDU Session. On the other hand, each UE is associated with per UE Aggregate Maximum Bit Rate (UE-AMBR) which limits the aggregate bit rate for Non-GBR QoS flows for a UE.

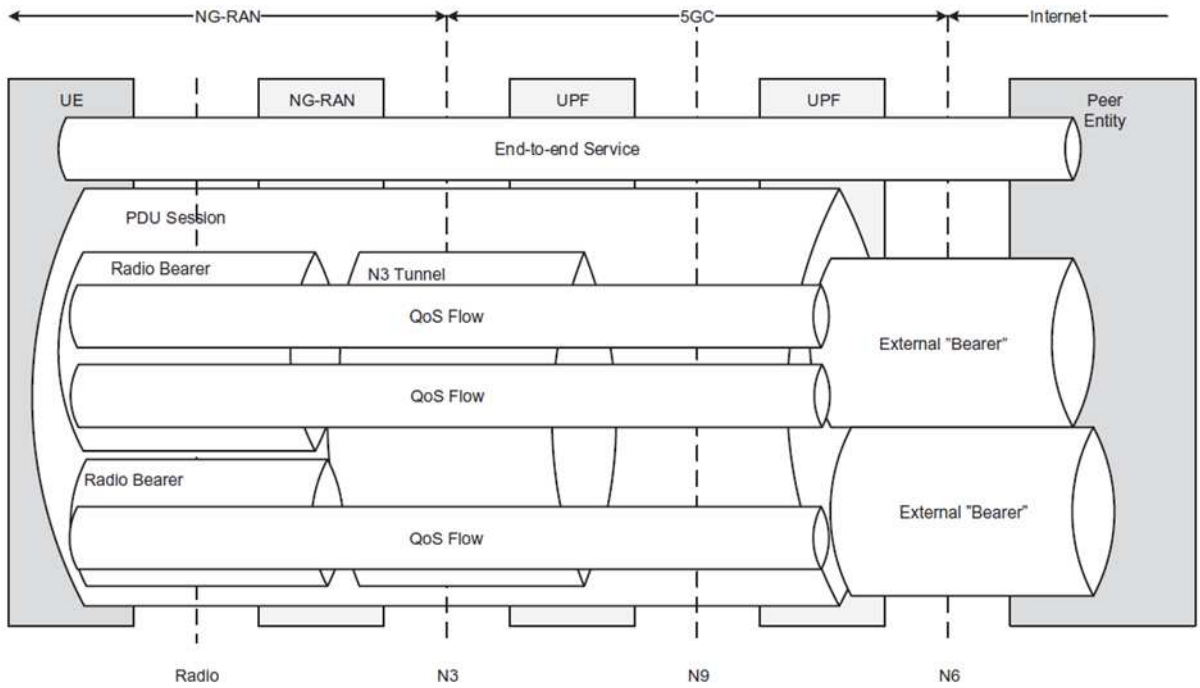


Figure 3.3: 5G QoS framework. (Adapted from [33]).

Based on this QoS framework, Figure 3.4 illustrates the QoS Flow to DRB mapping for an example of five IP flows being classified into three QoS flows. Following the downlink direction, the RAN decides how to map the QoS Flows into DRBs based on the QFI marking and QFI QoS Profile received. The SDAP allows multiplexing of more than one QoS Flow sent

in the same DRB. For QFI5, the RAN decides to use a dedicated DRB, while QFI2 and QFI3 are multiplexed in the same DRB. Thus, the data packets are sent to the UE via the DRBs and are routed to the application layer's socket interfaces in the UE as IP packets [33].

On the uplink, the UE's application layer generates the data packets which are compared with the Packet Filter Sets in the UE, and when a match is found the data packet is associated with a QFI. Thus, the assigned QFI and the data packet are sent through the UE's Access Stratum (AS) SDAP layer, which is responsible for performing QFI to DRP mapping. When there is a match, the data packet is sent to the corresponding DRB and if there is no match then the data packet is sent to a default DRB. In Figure 3.4 the QFI5 is sent on DRB1 and QFI2 and QFI3 are sent on DRB2, so the SDAP header indicates the QFI of each data packet. Finally, the UPF resolves the data packets into IP flows [33].

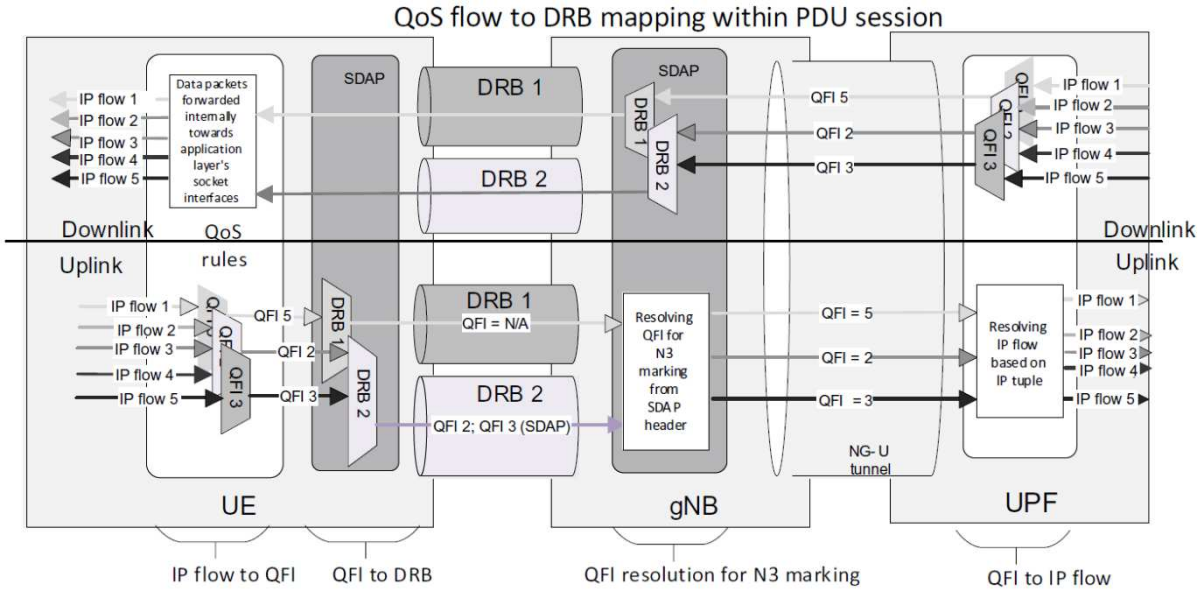


Figure 3.4: QoS Flow to DRB mapping. (Adapted from [33]).

A concept introduced in 5G systems is Reflective QoS (RQ), which minimizes signaling between the UE and the network core so that the decision on which QoS will be offered is made by reflecting the previously received QoS. In other words, the mirrored data packet receives the same QoS treatment as the received data packet. Thus, when RQ is active for a QFI, the UE creates a derived QoS rule for data classification based on the received DL data packet. Hence, when the UE sends a UL data packet the UE checks the QoS rules including the derived QoS rule, and applies the matched QoS Rule's QFI to the UL data packet [33].

3.3 Heterogeneous Networks (HetNets)

Another factor impacting the resource allocation configuration is the evolution of mobile communications systems from homogeneous networks (HomNets) to heterogeneous networks (HetNets). The 3GPP introduced heterogeneous networks in Release 12 [36]. HetNets allow different types of small cells to coexist alongside macrocells by sharing the same spectral resources, making it possible to increase spectral efficiency, reduce areas without coverage, increase system density, improve network capacity, and reduce link loss and delays. Thus, three spectrum-sharing strategies in HetNets stand out [11]:

- **Overlay Spectrum Sharing:** the small cells are allowed to use the resources not used by the macrocells.
- **Underlay Spectrum Sharing:** the small cells share the same resources with the macrocells simultaneously, with the introduction of an interference control for the small cells' transmissions concerning each macrocell, i.e. there is a cross-tier interference power constraint.
- **Hybrid Spectrum Sharing:** resources are divided into two types. The first is used only by the small cells and supports high data rates. The second is used by both macrocells and small cells, aiming for high-spectrum utilization.

In terms of access mode, Open Access mode and Closed Subscriber Groups mode stand out. In Open Access mode, users can join small cells or macrocells according to their coverage areas. For example, if a UE is in a small cell coverage area, it can access this network preferentially. On the other hand, in Closed Subscriber Groups mode, UEs associated with small cells can connect exclusively to them [11].

Handoff schemes in HetNets are essential for achieving QoS requirements, allowing seamless mobility between networks without users experiencing performance degradation. These include vertical handoff and horizontal handoff. Vertical handoff is also called intra-system handoff because the UE moves in an overlapping area covered by different networks, i.e. there are multiple access technologies. In a horizontal handoff, the user moves between networks with the same access technology, i.e. the process takes place within the system itself [37].

3.3.1 Cell Types

Figure 3.5 shows the structure of a HetNet with multiple small cells. Thus, the different cell types can be categorized into macrocells, microcells, picocells, and femtocells, according to the coverage area and application scenario.

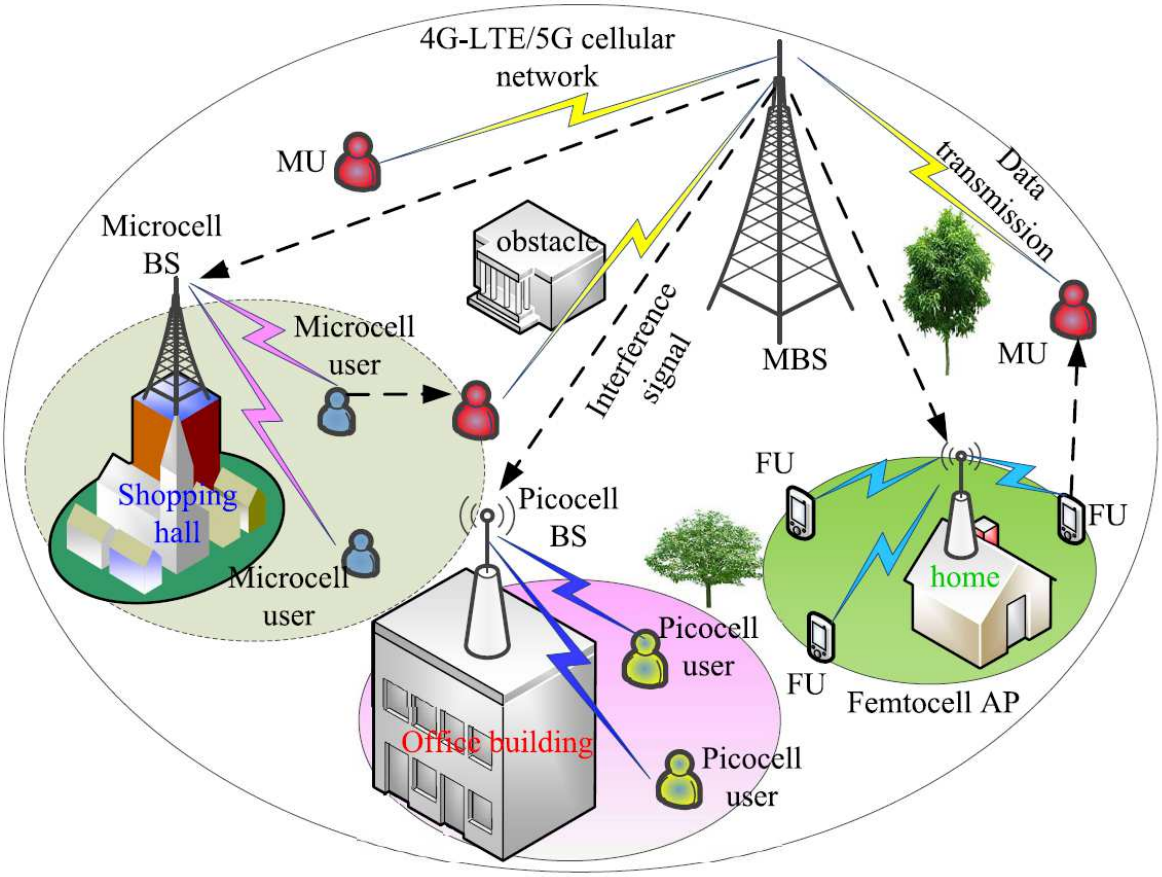


Figure 3.5: HetNet structure with multiple small cells. (Adapted from [11]).

Macrocells are located high up so there is a clear view of buildings and obstacles around their coverage area. Thus, they have high-power BSs and their transmission distances encompass a vast region, reaching from 1 km to 25 km of coverage. The QoS of macrocells is affected due to shadow fading and multipath interference, negatively highlighting indoor users due to uncovered spots [11].

In turn, microcells have low-power BSs and therefore cover regions of 200 m to 1 km, for example, highly populated areas such as shopping centers. The number of channels and traffic density are increased by reducing the frequency reuse distance of the low-power BS. Picocells, on the other hand, cover smaller coverage areas than microcells, reaching between 100 m and 200 m, for example, commercial buildings. In this way, picocells stand out when it comes to extending coverage to indoor areas, reducing the number of uncovered sites [38].

Finally, femtocells are called Home gNodeBs and have small and low-power BSs focused on improving communication conditions in homes and small buildings. They thus improve the QoS of indoor users connected to Home gNodeBs through simple installation. In addition, they can work in conjunction with femtocells to eliminate signal loss in buildings by

filling in areas without coverage [39]. Table 3.4 provides a summary comparison of the different cell types.

Table 3.4: Summarized comparison between different cell types (Adapted from [11]).

Cell	Radius (km)	Power (W)	Scenario
Macro	1 to 25	20 to 160	mountaintop
Micro	0.2 to 1	2 to 20	shopping malls, railway station
Pico	0.1 to 0.2	0.25 to 2	office building, underground parking
Femto	0.01 to 0.05	0.01 to 0.2	home, small enterprises

3.3.2 HetNet Models

Given the different types of cells, various communication scenarios have been designed for HetNets, including Traditional HetNets, OFDMA-Based HetNets, NOMA-Based HetNets, Relay-Based HetNets, H-CRANs, and Multi-Antenna HetNets.

The structure of a traditional HetNet is shown in Figure 3.6. In this network model, there are two types of cells, such as Macro BSs (MBSs), which act as primary cells in resource allocation, and Femto BSs (FBSs), which are secondary and occupy the same spectrum resources as macrocells. As can be seen in the Figure 3.6, users communicate with BSs without relay nodes. Femtocell Users (FUs) have their transmission power adjusted to reduce cross-tier interference compared to Macrocell Users (MUs), as well as achieving higher throughput values to satisfy QoS levels [11].

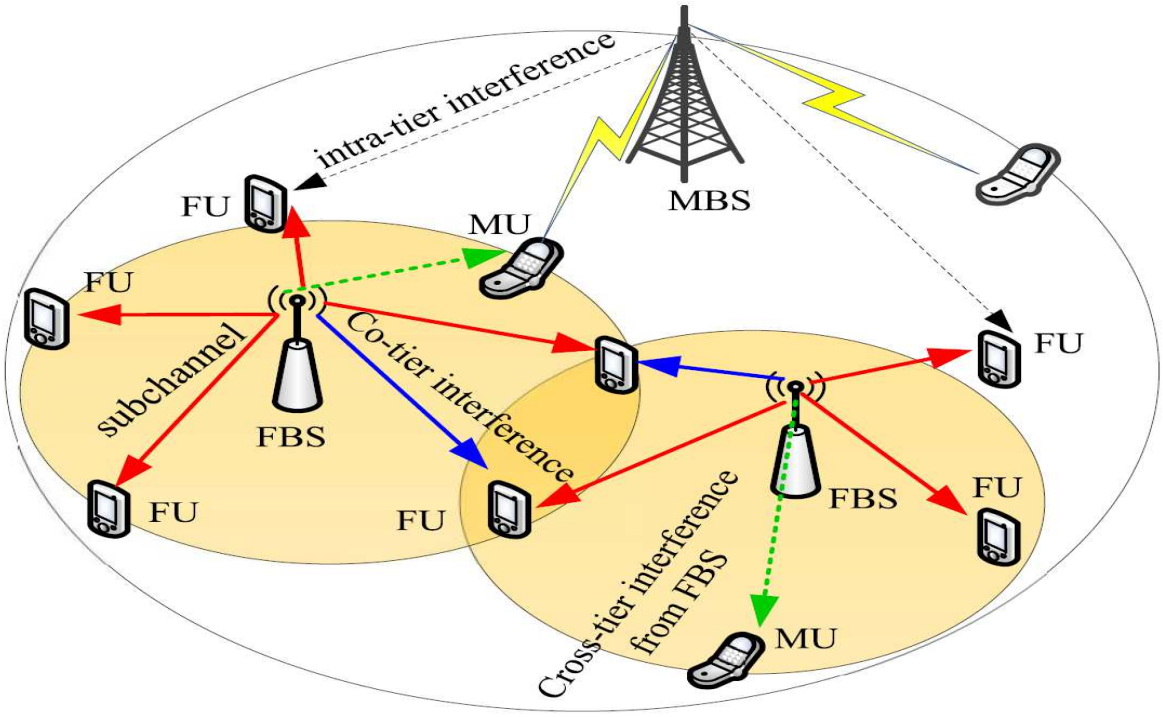


Figure 3.6: A traditional heterogeneous macro-femto network. (Adapted from [11]).

OFDM was introduced in HetNets to divide resources into multiple orthogonalized subcarriers, thus reducing mutual interference between subchannels and increasing the number of connected users. Hence, the subcarriers satisfy the communication requirements of the users through dynamic subcarrier assignment. This makes OFDMA-based HetNets important models for increasing the capacity of mobile communication systems by increasing the number of connected users [11]. Figure 3.7 illustrates the OFDMA-based heterogeneous macro-femto network.

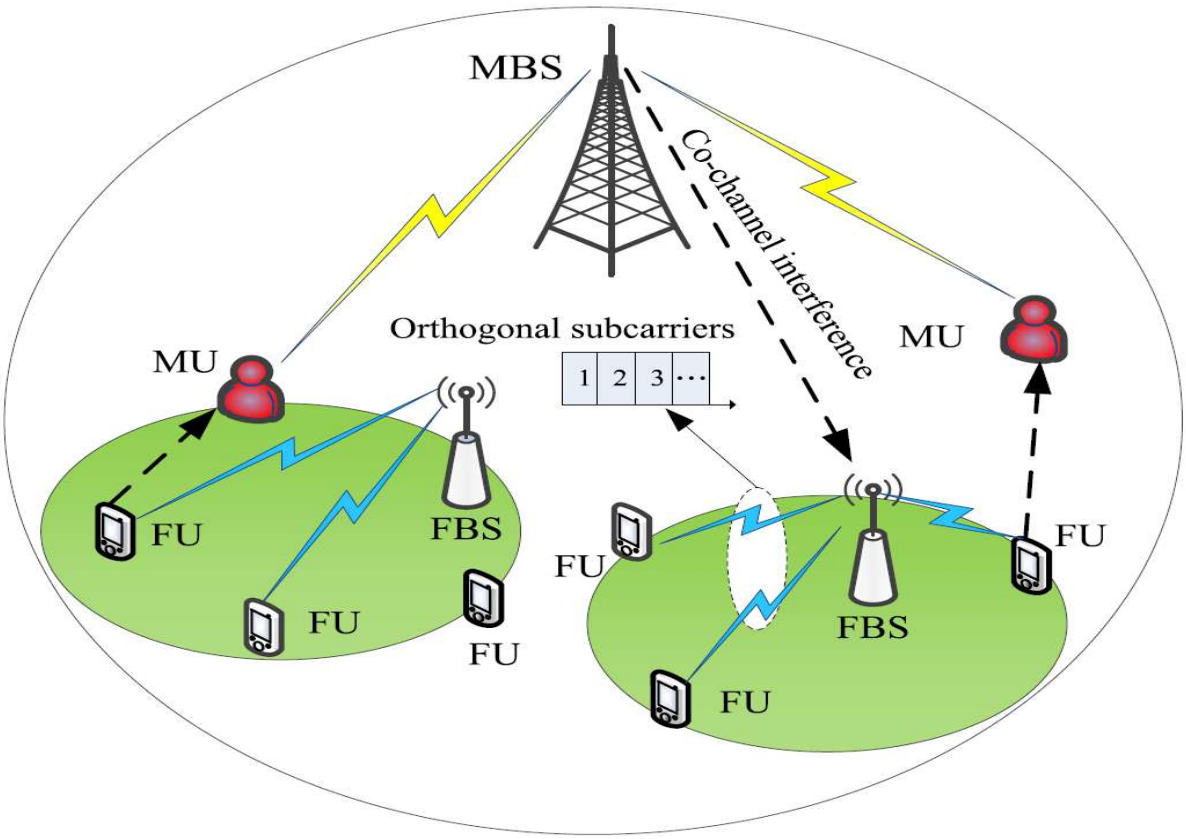


Figure 3.7: An OFDMA-based heterogeneous macro-femto network. (Adapted from [11]).

In turn, NOMA-Based HetNets use Non-Orthogonal Multiple Access (NOMA) to achieve non-orthogonal resource allocation between multiple users at the cost of increasing the complexity of the receiver hardware [40]. In a power-domain NOMA network, NOMA users are allocated different powers by the BSs according to the channel quality. As far as the receiver is concerned, the signals from multiple users are decoded using Successive Interference Cancellation (SIC). Thus, the same time and frequency resources can be shared between several users.

Figure 3.8 illustrates the downlink of a NOMA-based HetNet. Taking into account macrocells and femtocells as shown in the Figure 3.8, the process of using resources owned by macrocells by femtocells takes place as follows: femtocell users with poor channel conditions

detect their signals from other femtocell users with good channel conditions as the interference power. Users with good channel conditions detect the signals of those with poor channel conditions and then subtract these signals from the received signals. In this way, users with poor channel conditions can detect their signals.

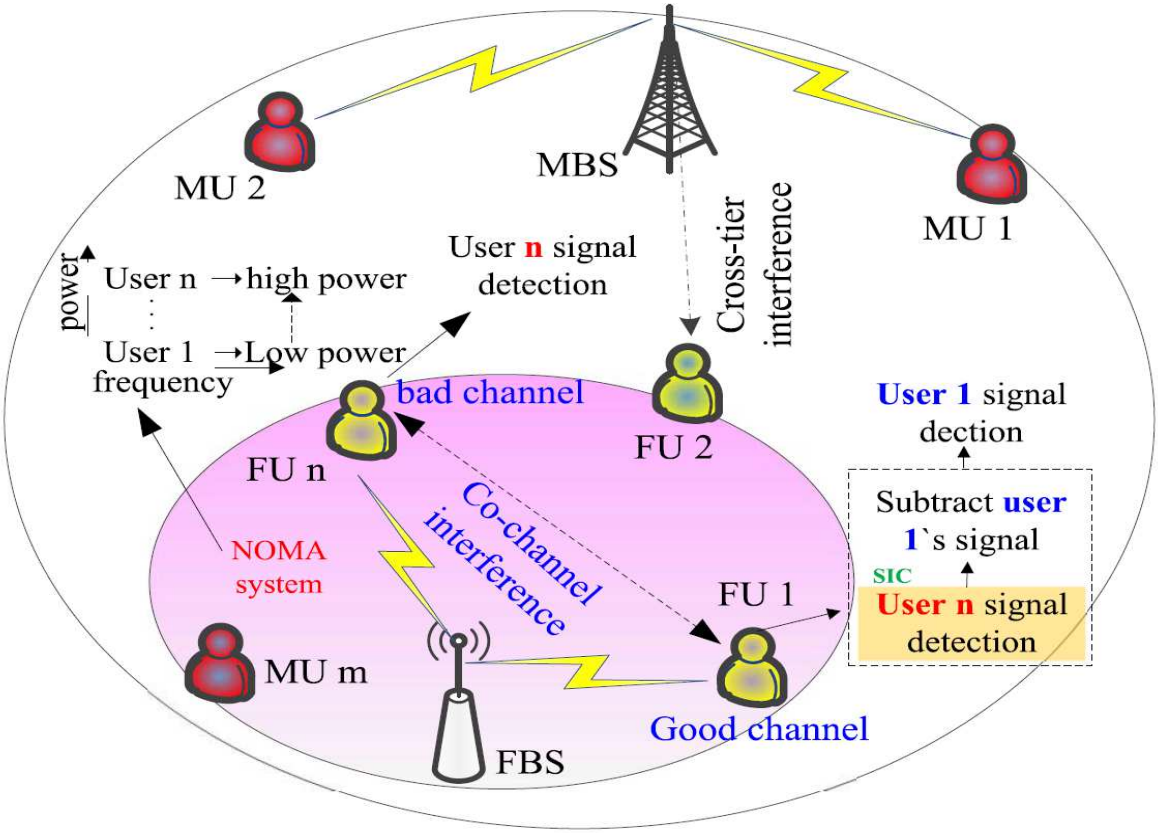


Figure 3.8: A downlink NOMA-based heterogeneous macro-femto network. (Adapted from [11]).

Relay-based HetNets use relay communication to increase the network's coverage area, as illustrated in Figure 3.9. In this sense, the transmission path is more complex and the transmission model is more flexible, allowing users outside the coverage zone of macrocells to be associated with macro BSs through femto BSs relay Access Point (AP) [11].

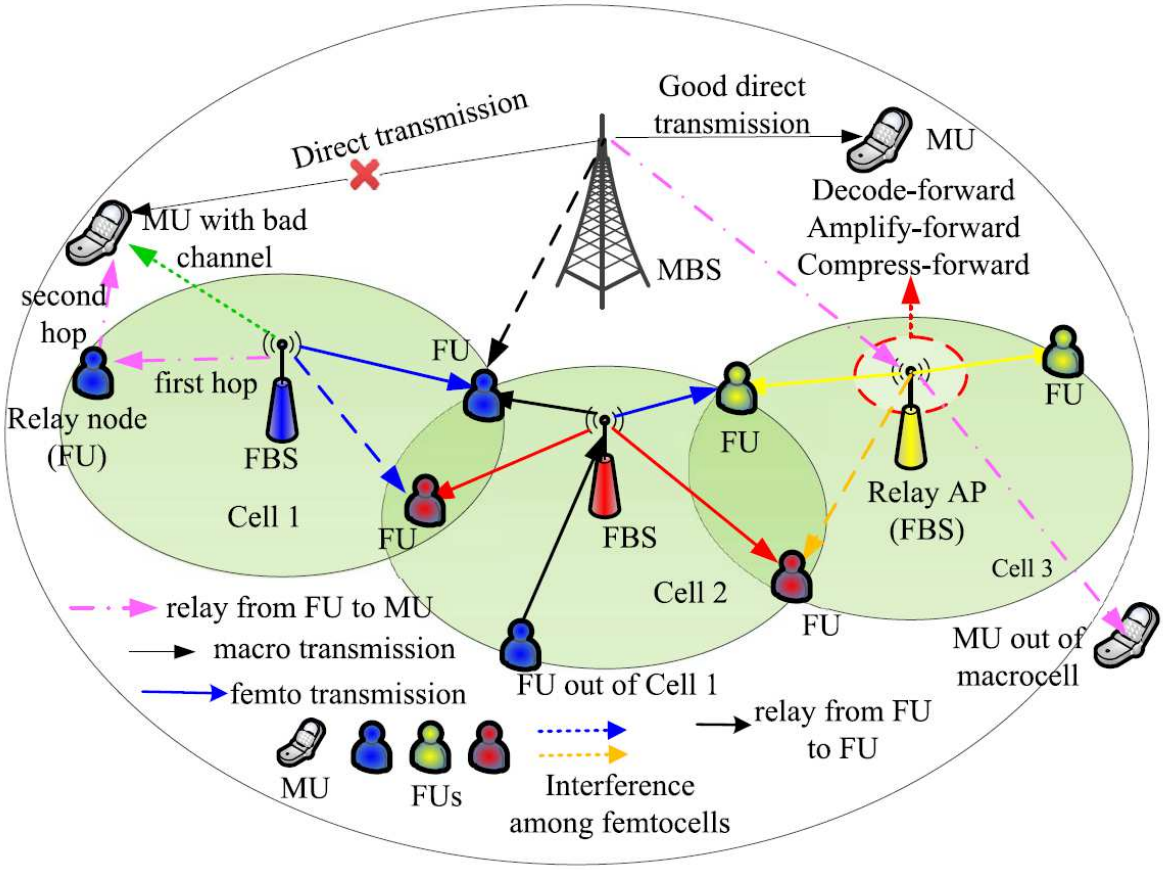


Figure 3.9: A relay-based heterogeneous macro-femto network. (Adapted from [11]).

Heterogeneous-Cloud Radio Access Networks (H-CRANs) introduce cloud computing into HetNets to achieve centralized network processing. Therefore, computational capacities are increased, achieving high throughput, and acting efficiently in large-scale data processing and control. Figure 3.10 illustrates how the user plane is decoupled from the control plane. Thus, macro BSs offer large-scale coverage and control signals, while Remote Radio Heads (RRHs) are configured in hot spots to provide high-speed data services to Remote User Equipments (RUEs). Finally, the Baseband Unit (BBU) pool in the cloud coordinates network resources and is connected to the RRHs via fronthaul links [11].

Finally, Multi-Antenna HetNets were developed based on spatial multiplexing to increase system capacity and spectral efficiency due to the installation of multiple antennas at the BSs and UEs. Thus, MIMO channels perform spatial division multiple access for multiple users through beamforming. Figure 3.11 illustrates a MIMO-based heterogeneous macro-pico network.

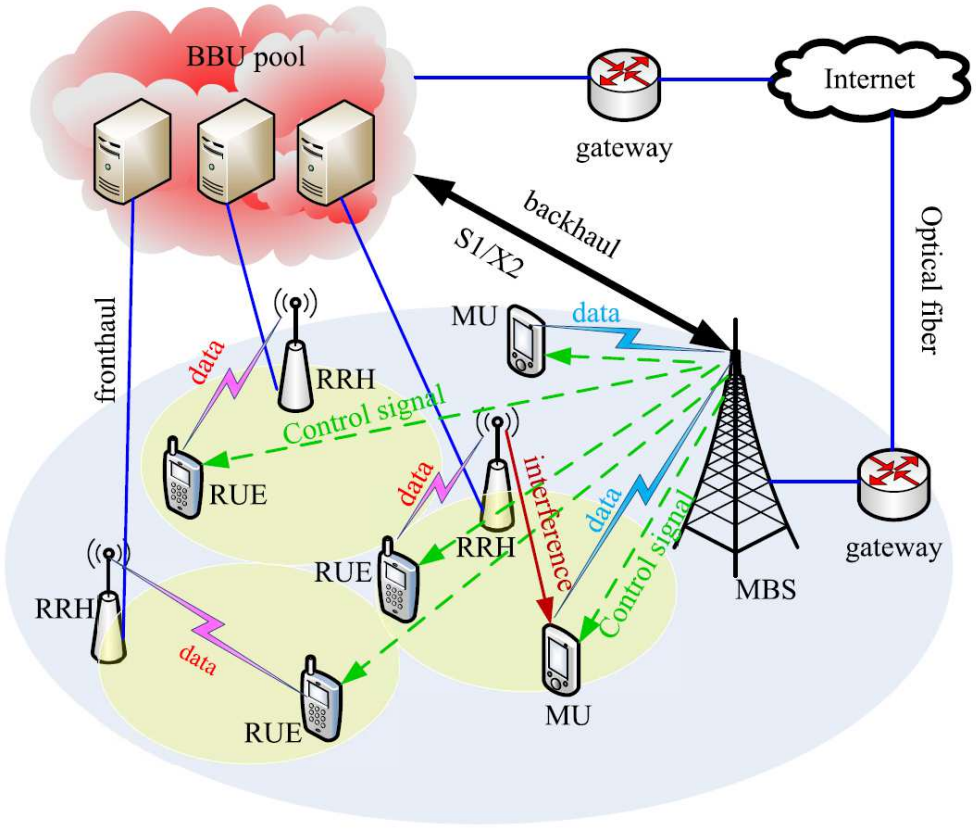


Figure 3.10: A downlink H-CRAN. (Adapted from [11]).

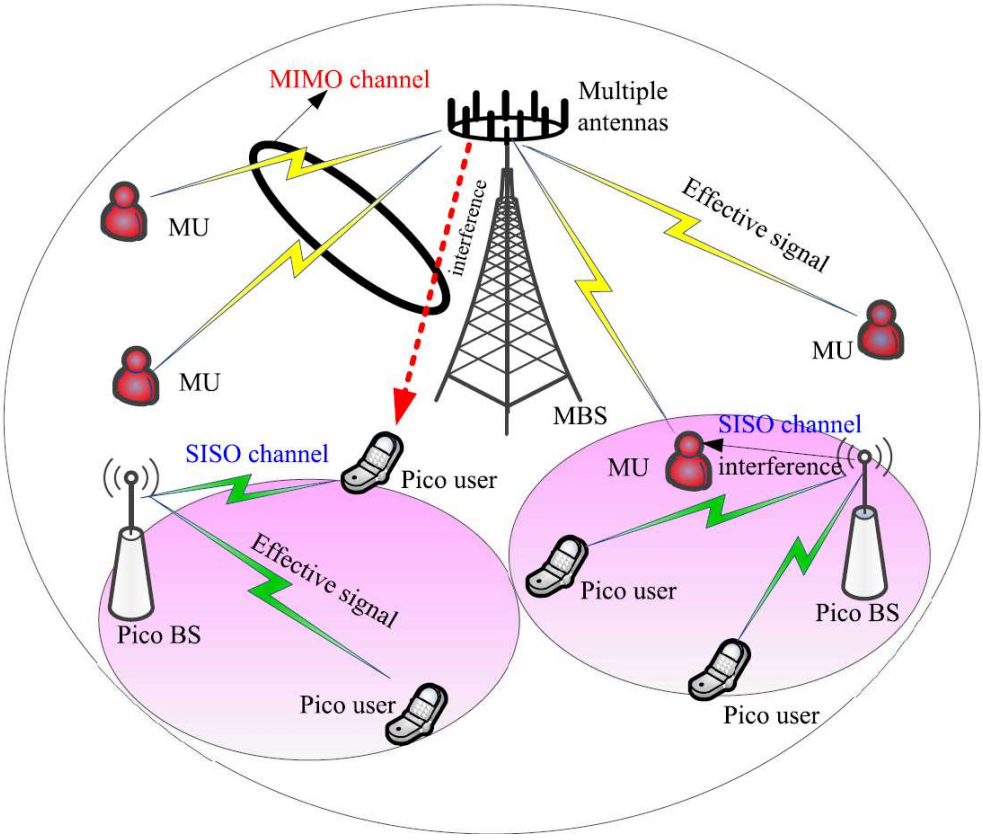


Figure 3.11: A MIMO-based heterogeneous macro-pico network. (Adapted from [11]).

3.4 Traffic Models

Heterogeneous 5G networks encompass a diversity of users, which makes it necessary to study traffic models that are consistent with the use cases presented in 2.1.5, namely eMBB, URLLC, and mMTC. According to [15], the main factors influencing the traffic of 5G systems are the demand for video, which will account for around two-thirds of all mobile network traffic; the proliferation of devices, since it is estimated that there will be an increase of around 1.4 billion smartphones and tablets from 2020 to 2030; and application uptake, since it is expected that more than 270 billion applications will be downloaded during the same period.

In addition, applications related to VoIP and real-time gaming tend to take up a lot of the network's consumption. Also noteworthy is access to web pages, which involves HTTP traffic usage. Regarding the proliferation of devices, the demand is due to its increase in IoT applications in various sectors such as industry and the incorporation of smart devices in homes. Finally, vehicular traffic encompasses mobile network users on the move or even IoT applications related to monitoring and cars' intelligent functions. Thus, the traffic models studied and simulated in this work are described below according to [15] and [41].

Full buffer traffic characterizes users who have an infinite amount of data to transmit. In this sense, the implementation allows each user to transmit a single packet of infinite size according to the model in Figure 3.12. It should be noted that there is no latency data for full buffer packets, as these packets are not fully transmitted at the end of the simulation. Furthermore, IoT users follow this type of traffic in the case of this research simulation.

Figure 3.12: Full buffer traffic model. (Adapted from [41]).

HTTP traffic has a bursty profile due to the model of interactions with the World Wide Web (WWW) web page structure. Web pages have a main and several embedded objects, so the number of embedded objects, size of all objects, and reading and parsing time for the main object are essential characteristics of web browsing. Figure 3.13 illustrates the HTTP traffic model.

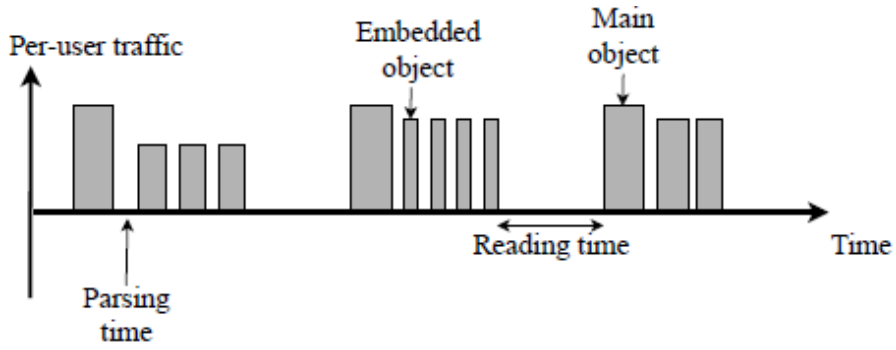


Figure 3.13: HTTP traffic model. (Adapted from [41]).

In turn, the video streaming traffic model is based on video frames that consist of several randomly sized packets arriving at regular time intervals T. The video encoder inserts delay intervals, called inter-arrival times, between the packets in a frame. Thus, video sessions are described by the inter-arrival times between the start of the frame and between the packets of a frame, as well as the number of packets per frame and the size of each packet. The source video rate of 64 kbps is used. Figure 3.14 shows the video streaming traffic model.

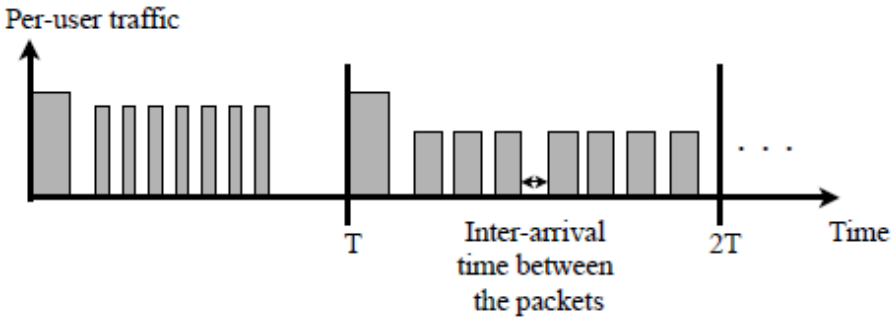


Figure 3.14: Video streaming traffic model. (Adapted from [41]).

The VoIP model uses Adaptive Multi-Rate (AMR) audio coding, an optimized data compression scheme for voice coding. In this case, a data rate of 12.2 kbps was used. Figure 3.15 illustrates the VoIP traffic model. Only one VoIP packet is generated every 20 ms during periods of activity and a Silence Insertion Descriptor (SID) payload is generated every 160 ms during break times.

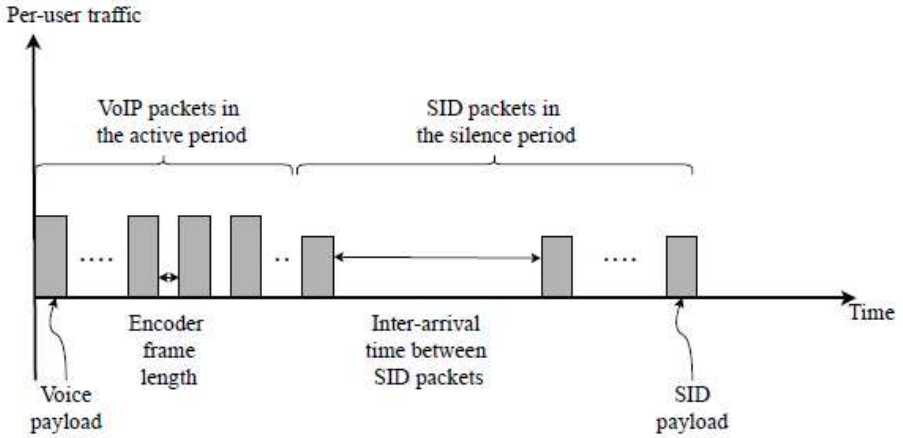


Figure 3.15: VoIP traffic model. (Adapted from [41]).

The gaming traffic model is generated with a uniformly distributed initial time to simulate the random timing relationship of the client traffic packet arrival to the uplink frame boundary. It is also defined based on the inter-arrival time between packets, the size of the packet, and the portion of the UDP header to be inserted in the packet. It should be noted that gaming-type packets are relatively small due to the nature of games. Figure 3.16 shows the gaming traffic model. Finally, the vehicular model is based on [42]. For more details on this subject, we recommend the survey [15].

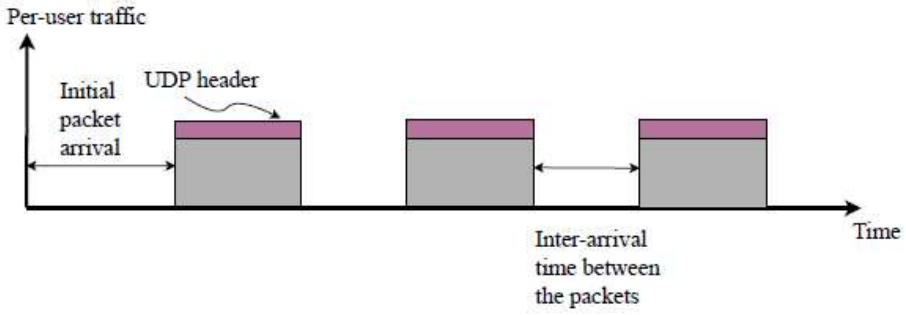


Figure 3.16: Gaming traffic model. (Adapted from [41]).

3.5 Summary

This chapter provides a general characterization of the transmission scheme of 5G systems in both time and frequency, as well as a brief explanation of the main QoS concepts. Next, the basic concepts, advantages, cell types, and configuration models of HetNets were described. Finally, the main traffic models used in this research were highlighted. Next, Chapter 4 studies how gNodeB allocates resources to the users associated with it through scheduling techniques.

"Mathematics is as much an aspect of culture as it is a collection of algorithms."

- Carl Benjamin Boyer

4

5G DOWNLINK PACKET SCHEDULING

RADIO RESOURCES are limited, so managing resources between users is essential. This process is called resource allocation, and one of its application methods involves the implementation of packet scheduling algorithms. This chapter presents a description of the factors that influence scheduler decisions and the metrics analyzed, as well as describing the algorithms implemented.

4.1 Overview of scheduling algorithms

The objectives of Radio Resource Management (RRM) procedures are to manage and execute resource sharing to offer greater QoS performance. Radio resources are defined by frequency, i.e. carrier frequency and bandwidth, and by time, i.e. transmission duration. In this way, the RRM process consists of distributing the limited resources in the frequency spectrum efficiently to the active devices in the network over a period of time, meeting the network's performance requirements [1]. In other words, this procedure involves classifying the priority of the users and mapping the resources to the users by the specific requirements of the system.

In turn, the scheduling process is one of the examples of RRM and considers performance indicators defined for the scenario before the resources are allocated. The literature on RRM differentiates scheduling algorithms between Channel-Independent Scheduling (CIS) and Channel-Dependent Scheduling, also known as dynamic scheduling. CIS-type strategies are also classified as classic scheduling techniques, as they distribute resources to users equally without considering traffic models, channel conditions, and QoS. On the other hand, dynamic techniques allocate resources according to specific information sent from the UE to the gNodeB, for example, the Channel State Information (CSI) Report, the Buffer Status Reports (BSR), and the QoS requirements defined for each traffic model [1].

In general, resource allocation using scheduling algorithms follows this procedure: the k-th RB is allocated to the j-th user if its $m_{j,k}$ metric is the highest, i.e. when it complies with Equation 4.1 [6]. This logic is also illustrated in Figure 4.1.

$$m_{j,k} = max._i \{m_{i,k}\}. \tag{4.1}$$

Hence, considering the types of strategy that define the scheduling algorithms and the logic through classification by metric, we highlight the factors that influence the scheduler's decision and the performance metrics it evaluates. Table 4.1 illustrates the notations used in the following equations.

Table 4.1: Equations symbology for factors and metrics associated to schedulers.

Symbol	Definition
$R_i(t)$	the instantaneous average achieved rate
$R_i(t-1)$	the past average achieved rate
$r_i(t)$	the current achievable transmission rate
t_c	the averaging filter memory
t_{enter}	the time instant at which the packet arrives to the buffer
R_x	received packets
x_i	the data transmitted by the user i
n	the total number of users
T_{rx}	the instant of data reception
T_{tx}	the instant of data transmission
R	the average bit rate
B	Bandwidth
P_{tx}	transmission power

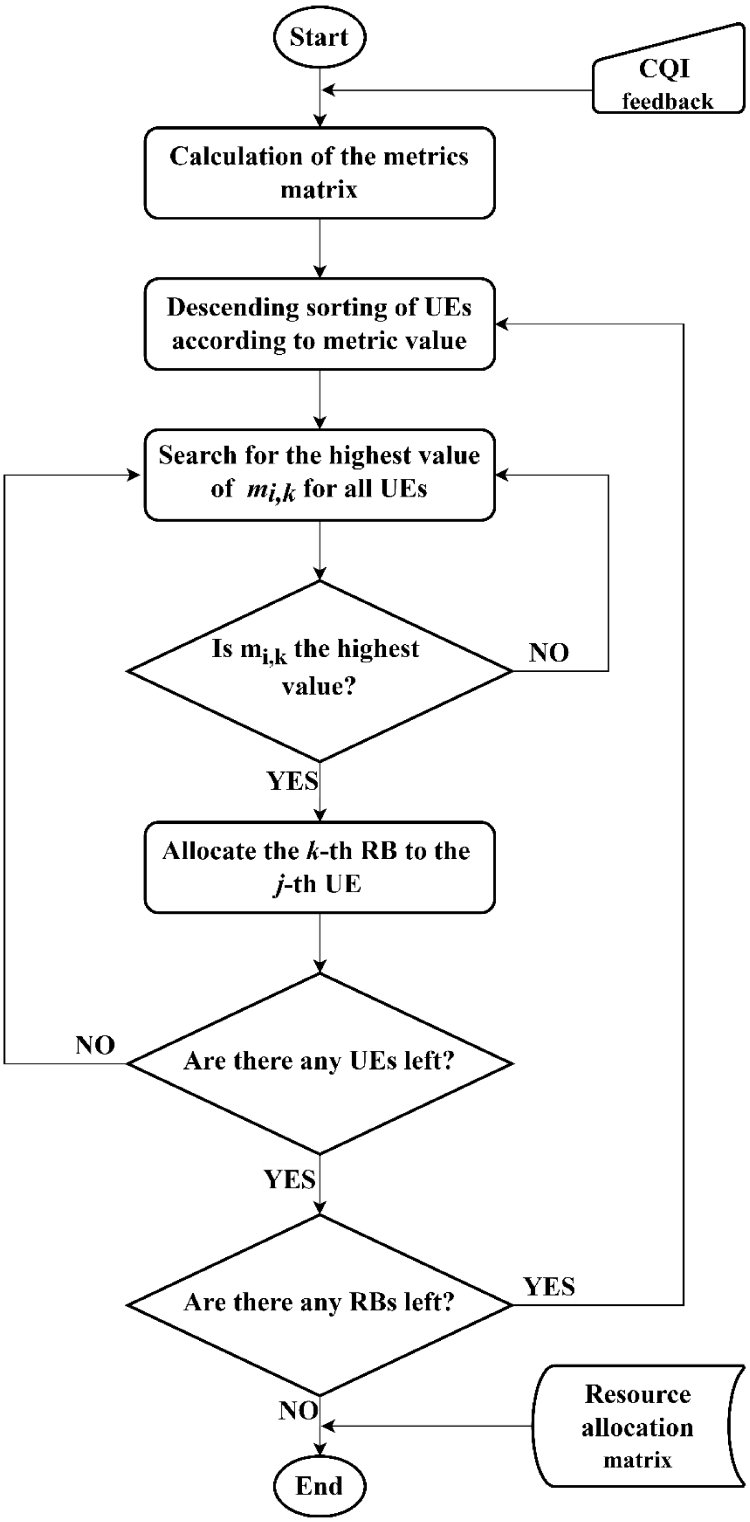


Figure 4.1: Radio resource scheduling algorithm process.

4.1.1 Factors influencing the scheduler's decision

The scheduling algorithms collect various measurements from the reports offered by the UEs to generate metrics that determine when the channel will be available to a given user. Therefore,

this subsection describes the schedulers' inputs and how these parameters direct the algorithms' output.

- **Average Data Rate (Resource allocation history)**

The average data rate of user i in the previous TTI serves as a historical parameter for allocating resources to the user. Thus, the scheduling process can use this data to prioritize users with the lowest past achieved throughput, resulting in a higher rate of fairness in the distribution of resources. In other words, the lower the past achieved throughput, the higher the metric. The instantaneous average achieved rate $R_i(t)$ is updated at each TTI according to the past achieved data rate $R_i(t-1)$ [1]:

$$R_i(t) = \left(1 - \frac{1}{t_c}\right) R_i(t-1) + \frac{1}{t_c} r_i(t). \tag{4.2}$$

The constant t_c indicates the averaging filter memory, where $1/t_c$ is the moving average data rate weight to calculate the average data rate. It should also be noted that the average data rate takes into account the current achievable transmission rate $r_i(t)$ according to the current CQI value for user i.

- **Head of Line Packet Delay (D_{HOL})**

This parameter indicates how long a packet remains in the buffer before being sent. Thus, for the current time t and for the time instant t_{enter} in which the packet arrives at the gNodeB buffer, the calculation of D_{HOL} is given by [1]:

$$D_{HOL} = t - t_{enter}. \tag{4.3}$$

In this case, there is a status of transmission queues, where the longer the queue, the higher the metric, i.e. the metric is maximized for users who experience long HOL queues, ensuring the delivery of packets with minimum latency.

- **Buffer levels and Queue status**

The buffer level and queue status parameters indicate the data amount the scheduling algorithm must transfer on the uplink and downlink. This ensures the flexible allocation of resources according to the existing flows in the buffer or transmission queue for each active user [6]. Therefore, these two parameters influence the scheduler's decision so that more resources are allocated to users with a greater amount of data to transmit, which reduces latency and improves reliability.

- **Channel Quality Indicators (CQIs)**

The values of the Channel Quality Indicators range from 0 to 15 depending on the integer value encoded in 4 bits. Thus, 0 indicates a very poor channel condition, while 15 is the best possible. In this sense, the reported CQI values ensure that the reported channel conditions are suitable for the performance provided to the user through resource allocation. The AMC module selects the MCS that best suits the reported CQI value, defining the number of bits per resource element [1]. Therefore, the CQI influences throughput: higher CQI values result in higher data rates. It can be seen that CQIs are essential for designing channel-aware schedulers.

- **Quality of Service Requirements**

The Quality of Service Identifier (QCI) value related to each data flow may be used to manage specific policies to meet QoS requirements of various use cases and applications. QoS-aware schedulers are based on QCIs, as these values represent the type of flow, priority level, tolerated packet error loss, and packet budget delay for each application. For example, resource allocation prioritizes users with a higher priority level or lower tolerated packet error loss. In addition, the HARQ procedure of retransmissions takes the CQI values into account to adapt certain services and applications to the expected packet error loss limitations [1].

4.1.2 Performance metrics evaluated

The performance metrics evaluated serve as a classification of schedulers according to the direction of the objectives analyzed through the implementation of the algorithm.

- **Throughput**

Throughput expresses the data transmission rate over the sending time as indicated by Equation 4.4 [43]. Achieving high throughput values is one of the main objectives of mobile communications systems in all their generations.

$$Throughput = \frac{\sum R_x \, Packet\ Size}{Delivery\ Time}. \tag{4.4}$$

- **Block Error Ratio (BLER)**

The Block Error Ratio (BLER) measures the reliability of the communication channel, indicating the rate of erroneous data blocks to the total number of data blocks transmitted. BLER is calculated by dividing the number of erroneous blocks by the total number of blocks transmitted:

$$BLER = \frac{Erroneous\ blocks}{Total\ number\ of\ blocks}. \tag{4.5}$$

- **Fairness Index**

The fairness index metric highlights the network's ability to distribute resources equally among users. A common method for calculating the fairness index is Jain's fairness index expressed by Equation 4.6 [44], where x_i represents the throughput of the i-th user, and n indicates the total number of users.

$$Fairness\ index = \frac{(\sum_{i=1}^{n} x_i)^2}{n\ x\ \sum_{i=1}^{n} {x_i}^2}. \tag{4.6}$$

- **Latency**

Latency measures the time taken for a packet to be received by the user's device. Thus, it measures the perceived delay in successfully receiving packets and can be calculated using Equation 4.7, where T_{rx} and T_{tx} represent, respectively, the instant at which the data is received and the instant of transmission.

$$Latency = T_{rx} - T_{tx}. \tag{4.7}$$

- **Goodput**

Goodput measures the performance of schedulers in terms of their effective data rate. Thus, the difference between goodput and throughput lies in the bits successfully transmitted, since goodput does not consider retransmission packets or header control bits. Equation 4.8 shows the calculation of goodput:

$$Goodput = \frac{Original\ data}{Delivery\ Time}. \tag{4.8}$$

- **Spectral Efficiency**

The spectral efficiency measures the network's effectiveness in managing radio resources and is expressed as the ratio between average data rate and bandwidth (bits per second per Hertz). In Equation 4.9, R indicates the average bit rate, and B represents the bandwidth.

$$Spectral\ Efficiency = \frac{R}{B}. \tag{4.9}$$

- **Energy Efficiency**

Energy efficiency can be defined as the ratio between the average data rate and the transmission power (bits per Joule). It is therefore a parameter used to determine the power

usage by the communication system. In Equation 4.10, R indicates the average bit rate, and P_{tx} represents the transmission power.

$$Energy\ Efficiency = \frac{R}{P_{tx}}. \tag{4.10}$$

4.2 Scheduling Algorithms

Multi-user scheduling is one of the main features of 5G systems, as it allows available resources to be distributed among active users according to the decision factors of each algorithm. The following subsections describe the methodology of the schedulers studied.

4.2.1 Round Robin (RR)

The Round Robin algorithm provides a fair distribution of resources over time for all users so that resource allocation is carried out using a first-in, first-out (FIFO) method. Although this strategy guarantees equality in terms of the time each user occupies the channel, it does not guarantee equality in terms of throughput, as it is a channel-independent scheduling method [6]. Equation 4.11 shows the calculation of the RR metric, ${m_{i,k}}^{RR}$, where t and T_i indicate the current time instant and the last time the user was served.

$${m_{i,k}}^{RR} = t - T_i. \tag{4.11}$$

Figure 4.2 illustrates the summary flowchart of the RR scheduler used in this study, and Algorithm 1 shows the RR logic in pseudocode.

Algorithm 1: Round Robin (RR)

Input: active users
Output: scheduled users
1. Define RB parameters: *currentTime*
2. Reset the resource grid for this slot
3. Get active users
   if there are no active users **then**
      return to Step 1
   else
      proceed to Step 4
   end
4. Allocate the same RBs, which were used in the previous slot, for users that need a retransmission
5. Initialize the RBs over which no user would be scheduled
6. Get number of resources left for scheduling
7. Get active users that does not need retransmissions
8. Schedule last scheduled user from previous slot if they have unscheduled resources left
9. Compute the scheduling order for the next users in queue
10. Set index of the last scheduled user in the queue
11. Save how many resources for the last scheduled user were not scheduled
12. Reset RB grid mask after user allocation is done all users
13. Set information for scheduling in next slot: last user scheduled and unused resources
14. Perform all calculations of the scheduler and schedule users

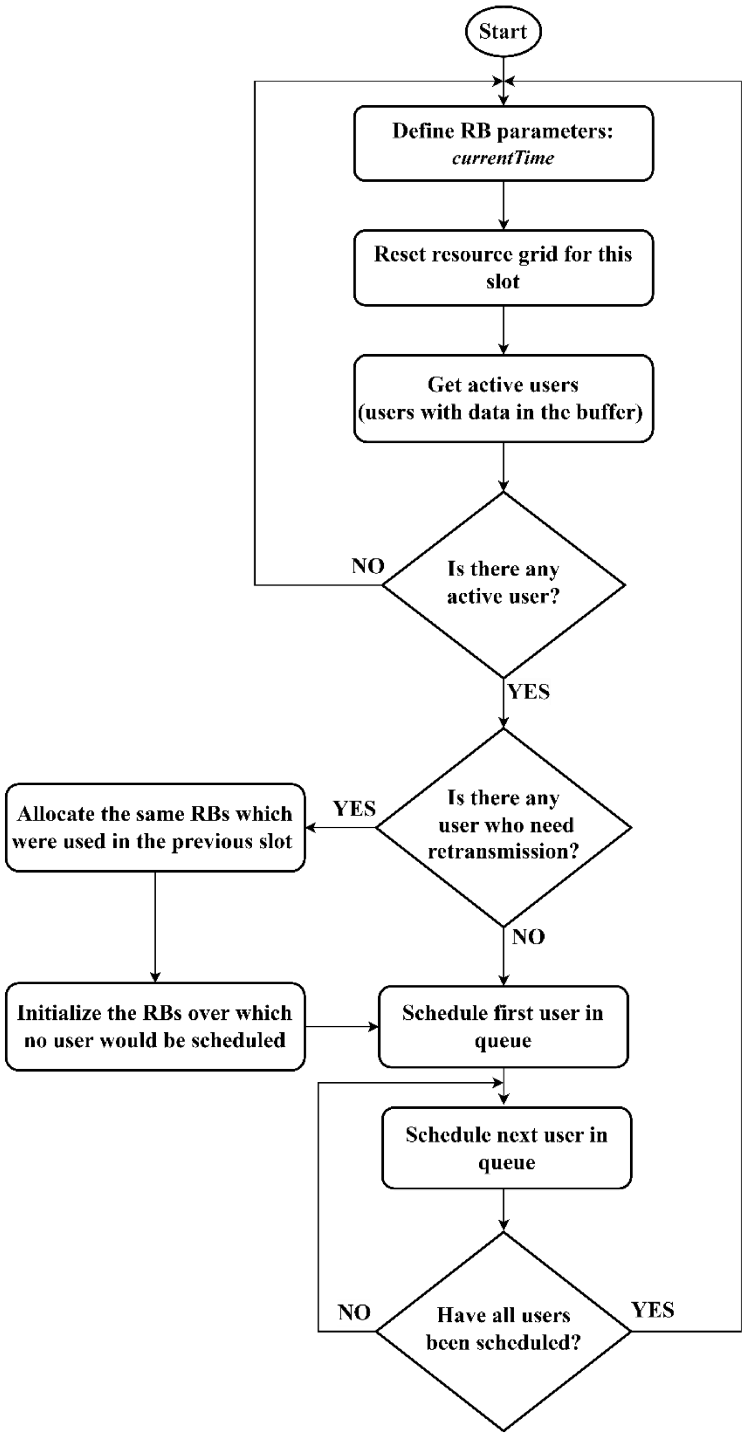


Figure 4.2: RR scheduler summarized flowchart.

4.2.2 Best CQI

The best CQI scheduler allocates the RBs prioritizing the users with the best channel conditions according to Equation 4.12. We highlight ϕ as the CQI value, and i represents the user for which the algorithm's metric is calculated. In this way, the scheduler evaluates the updates between uplink and downlink on the CQIs of each user and prioritizes the active users with the highest CQI value in the resource allocation process [45].

$$m_{i,k}{}^{best\ CQI} = \phi_i(\tau). \tag{4.12}$$

Algorithm 2 shows the logic of the best CQI through pseudocode and Figure 4.3 shows the summarized flowchart of the best CQI scheduler used in this work.

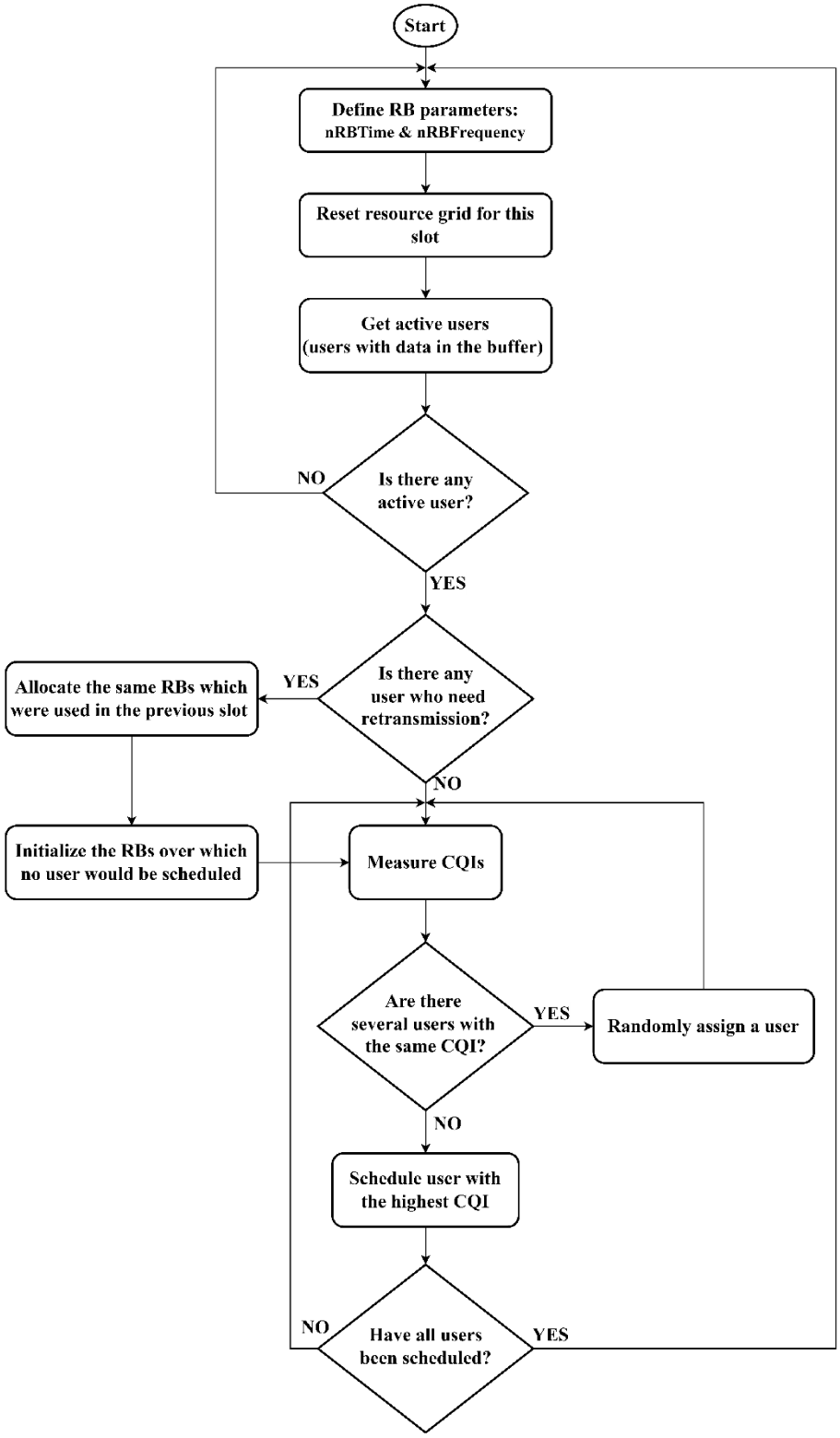


Figure 4.3: best CQI scheduler summarized flowchart.

Algorithm 2: best CQI
Input: active users
Output: scheduled users
1. Get *time* and *frequency* RBs parameters
2. Reset the resource grid for this slot
3. Get active users
   if there are no active users **then**
      return to Step 1
   else
      proceed to Step 4
   end
4. Allocate the same RBs, which were used in the previous slot, for users that need a retransmission
5. Initialize the RBs over which no user would be scheduled
6. Get number of resources left for scheduling
7. Get active users that does not need retransmissions
8. Get parameters of active users
9. Measure CQIs: create a 4-dimensional array of CQIs over frequency, time, and codewords for each active user
10. Update the 3-dimensional arrays of CQIs based on user feedback
11. **if** there are several users with the same CQI **then**
    Find active users with best CQI: find indices and total number of active users with best CQI
    Randomly assign a user
    elseif
    Schedule user with the highest CQI
    end
    Return to Step 9
12. Reset RB grid mask after user allocation is done all users

4.2.3 QoS-Aware Scheduler (QAS)

The study of the implementation of the QoS-Aware scheduling algorithm can be found in [22]. It is based on the weighted sum throughput maximization problem described by Equation 4.13:

$$\underbrace{\arg max}_{\{b_1,\dots,b_{i,c}\}} c + \left(\sum_{i=1}^{I} \zeta_i {t_i}^T b_i\right)$$

subject to:

$$b(n) \in \{0,1\}, \forall n$$

$${b_j}^T b_k = 0, \forall k \neq j$$

$${t_i}^T b_i \geq c\gamma_i, \forall i \in \{non\ full\ buffer\ users\}$$

$$0 \leq c \leq 1$$

$$\sqrt{\mathcal{I}_o I}\|{t_i}^T b_i\|_2 \leq \sum_{i=1}^{I} {t_i}^T b_i\,, \forall i \in \{full\ buffer\ users\}. \tag{4.13}$$

For Equation 4.13, ${t_i}^T$ represents the throughput vector for each user i, and $b_i = \left[b_{1,i}, \dots, b_{n,i}\right]^T$ is the vector of RBs allocated to each user. Also, in $\zeta_i = \alpha^{-\beta i}\sigma^{-\max\{d_{c,i}-d_i\}}$, we have: $\alpha^{-\beta i}$ indicates the reliability parameter, which decreases exponentially according to the base $\alpha = 2$; β_i represents the average BLER over the codewords of user i; the latency priority factor is represented by $\sigma = 1.05$; $d_{c,i} - d_i$ indicates the difference between the characteristic delay constraint of user i, $d_{c,i}$, and the current delay, d_i, of the user. These characteristic delay constraint values are predefined according to the traffic model for RT applications. In this way,

the user's priority depends on how close the current delay of the user's packet is to the delay constraint for that application. It should also be noted that the higher the reliability factor, $\alpha^{-\beta i}$, the higher the user's priority, as it presents highly reliable traffic.

Concerning the constraints, the first indicates that RBs can be treated in binary form, while the second refers to the fact that each RB must be associated with one user at a time. In turn, γ_i indicates the total number of bits in the user i's buffer, so the third constraint ensures that the number of RBs allocated to a user is sufficient for their use. To this end, the variable c is included to ensure feasibility since it proportionally reduces the RBs allocated to all users. Finally, there is a constraint related to the fairness of full buffer users by implementing Jain's fairness index [44] so that $\mathcal{J}_o$ represents the desired fairness index.

This optimization problem is called mixed binary integer programming and is implemented according to Algorithm 3. An open-source MATLAB tool for disciplined convex programming alongside Gurobi Optimizer was used to solve this problem. Next, Figure 4.4 and Algorithm 4 illustrate the summary process of the QAS algorithm.

Algorithm 3: Binary integer optimization problem

1. Starts MATLAB-based modeling system for convex optimization: *cvx_begin*
2. Set variables for binary and feasibility constraint functions
   $variable\ c$
   $variable\ b(numberOfRBs * numberOfActiveUsers)\ binary$
3. Set objective function
   $maximize(tunedThroughput * b + c)$
4. Set constraint functions
   $subject\ to$
   $onesDiagonalMatrix * b = onesVector$
   $fullBufferIndicator = bufferedBits == Inf;$
   $fullBufferUsers = find(fullBufferIndicator);$
   $nonFullBufferUsers = find(\sim fullBufferIndicator);$
   Set a constraint that imposes that the throughput of non full buffer users is sufficient for their needs. Variable c proportionally reduces the assigned
   RBs of non full buffer users to achieve feasibility of the optimization problem
   $0 \leq c \leq 1$
   if nonFullBufferUsers **then**
      $nonFBThroughputMatrix = throughputMatrix;$
      $nonFBThroughputMatrix(fullBufferUsers,:) = 0;$
      $bufferedBits(fullBufferUsers,:) = 0;$
      $nonFBThroughputMatrix * b \geq c.* bufferedBits;$
   end
   Set a constraint that imposes fairness among full buffer users
   if fullBufferUsers **then**
      $fBThroughputMatrix = throughputMatrix;$
      $fBThroughputMatrix(nonFullBufferUsers,,:) = 0;$
      $fBEsimatedThroughput = estimatedThroughput;$
      $fBEstimatedThroughput(nonFullBufferUsers,:) = 0;$
      $\sqrt{J * numberActiveUsers} * |fBThroughputMatrix * b| \leq fullBufferThroughput * b;$
5. Ends MATLAB-based modeling system for convex optimization: *cvx_end*

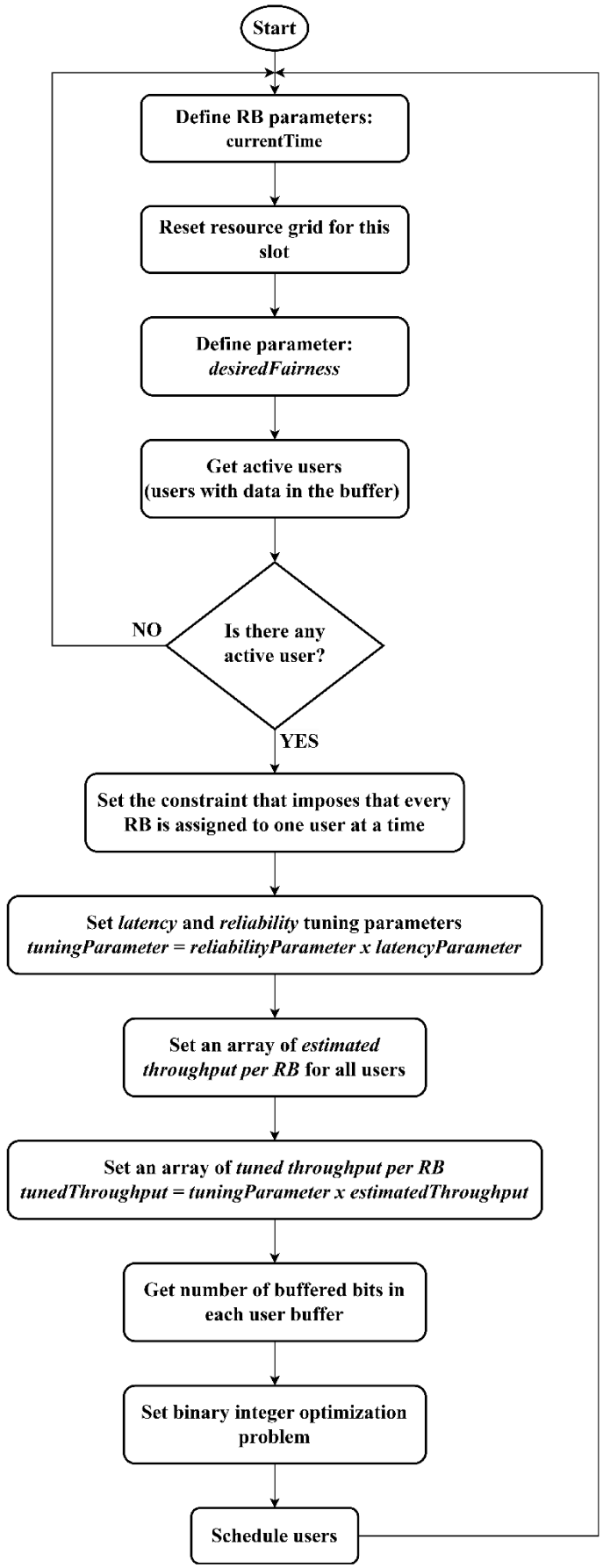


Figure 4.4: QAS scheduler summarized flowchart.

Algorithm 4: QoS-Aware Scheduler (QAS)
Input: active users
Output: scheduled users
1. Define RB parameters: *currentTime*
2. Reset the resource grid for this slot
3. Define parameters: *desiredFairness, reliabilityParameter,* and *latencyParameter*
4. Get active users
   if there are no active users **then**
      return to Step 1
   else
      proceed to Step 5
   end
5. Set parameters: $numberOfActiveUsers$ and $numberOfRBs$
6. Set the constraint that imposes that every RB is assigned to one user at a time
7. Set array of tuning parameter by multiplying latency and reliability tuning parameters for every user:
*tuningParameter = reliabilityParameter*latencyParameter*
8. Set array of estimated throughput for all users
9. Set array of tuned throughput per RB:
*tunedThroughput = tuningParameter*estimatedThroughput*
10. Get number of buffered bits in each user buffer
11. Set binary integer optimization problem as stated in Algorithm 3
12. Schedule users

4.2.4 Weighted QoS-Aware Scheduler (WQAS)

This work proposes a QAS modification to highlight RT traffic users by implementing weights seeking to overcome the challenge of maintaining QoS requirements in scenarios of high network stress. In this way, vehicular, VoIP, gaming, and video streaming users are configured with a scheduling weight of 10, which results in 10 consecutive RBs being allocated to the user if it is scheduled. On the other hand, only one RB is assigned to the NRT users when scheduled. These values were adjusted based on empirical tests. Figure 4.5 illustrates the summarized flowchart for the Weighted QoS Aware Scheduler. Algorithm 5 shows the summarized process of the WQAS algorithm.

To implement the weight rule, the slots are initially distributed according to the Weighted Round Robin (WRR) queuing process. This creates a vector of consecutive numbers of RBs associated with each type of user, differentiating between RT and NRT services. Next, a second scheduling stage begins based on the QoS Aware scheduler, configuring tuning parameters related to reliability and latency, as well as the desired fairness imposed on the system. The parameters for the number of users and RBs available are configured so that each RB is allocated to one user at a time.

Next, the tuning parameter value is calculated based on reliability and latency. Then, it creates an estimated throughput array for all users for subsequent calculation of tuned throughput as the product between tuning parameters and estimated throughput. The information of buffered bits is stored in each user's buffer. The binary integer optimization problem is solved as stated in Algorithm 3 and considering the number of RBs associated with each user by the WRR stage. Finally, all users are scheduled.

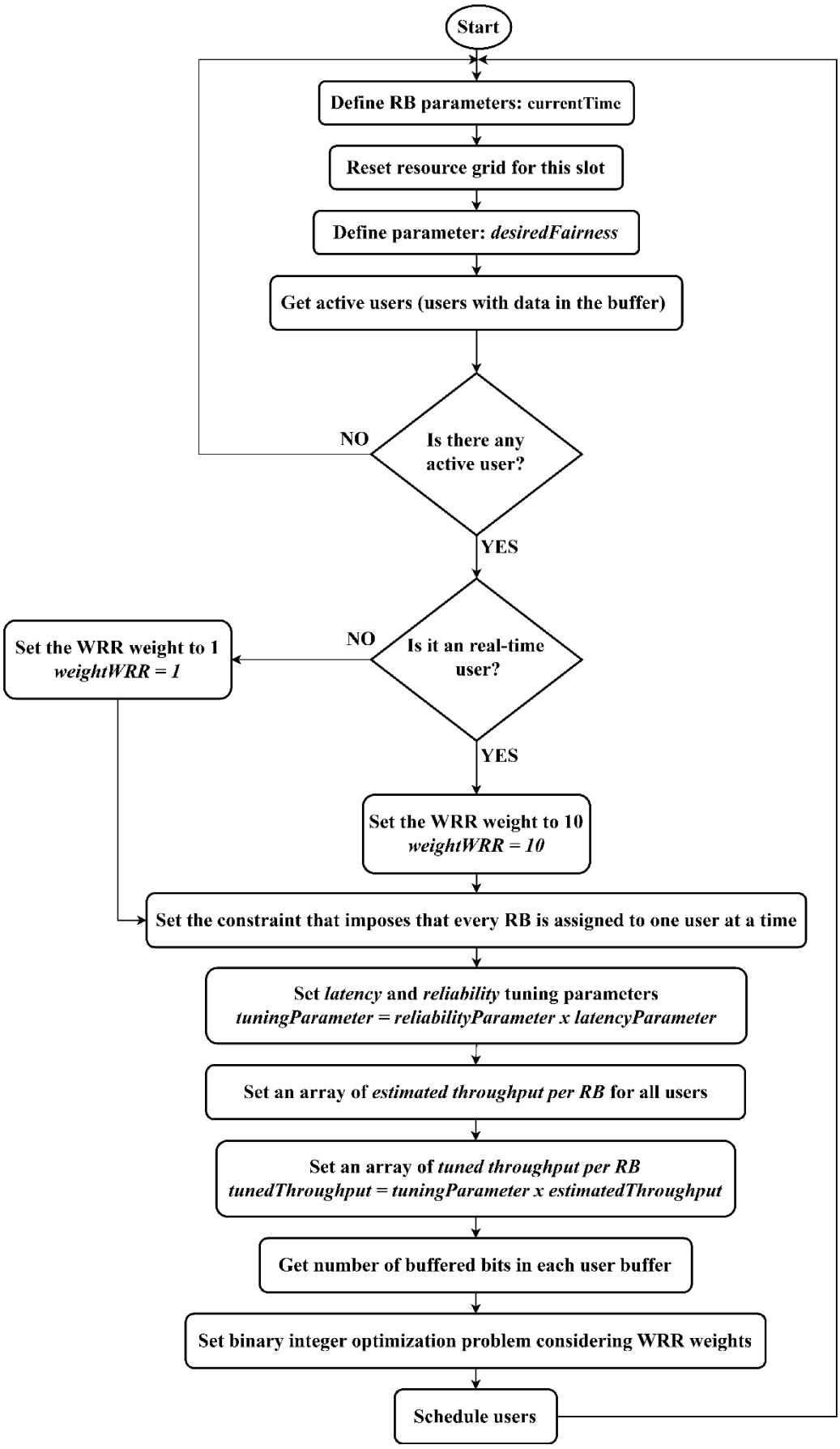


Figure 4.5: WQAS scheduler summarized flowchart.

Algorithm 5: Summarized Weighted QoS Aware Scheduler
Input: active users
Output: scheduled users
1. Define RB parameters: $currentTime$
2. Reset the resource grid for this slot
3. Define fairness parameter: $desiredFairness$
4. Get active users
5. **WRR level**
   if $user == RT\ user$ **then**
   $weightWRR = 10$
   else
   $weightWRR = 1$
   end if
6. Set the constraint that imposes that every RB is assigned to one user at a time
7. Set tuning parameter by multiplying latency and reliability tuning parameters:
$$tuningP = reliabilityParam.x\ latencyParam.$$
8.Set array of tuned throughput per RB:
$tunedThroughput = tuningP\ x\ estimatedThroughput$
9. Set binary integer optimization problem
10. Schedule users

4.2.5 Channel and QoS-Aware Scheduler (CQAS)

In order to evaluate the impact of a strategy that encompasses channel quality and QoS requirements, this work proposes the implementation of a Channel and QoS-Aware Scheduler (CQAS). To this end, a third tuning parameter is configured for the QAS algorithm, the CQI tuning parameter. Hence, Equation 4.14 derives from Equation 4.13 and shows the deployment of the term Φ_i that refers to the normalization of the CQI values, ϕ_i , to the interval $I = [1.1,1.15]$.

$$\zeta_i = \alpha^{-\beta i}\sigma^{-\max\{d_{c,i}-d_i\}}\Phi_i. \tag{4.14}$$

Figure 4.6 illustrates the summary flowchart for the proposed CQAS algorithm, while Algorithm 6 represents the algorithm's summary process. The algorithm starts by initializing the parameters related to the RBs and redefining the resource grid for the slot. The desired fairness, reliability, and latency parameters are then defined. The active users are then obtained, and an array of users is created to associate the CQI value with each user. Originally, the CQI value, ϕ_i, ranges from 1 to 15, so to match the reliability, $\alpha = 2$, and the latency priority fator, $\sigma = 1.05$, the CQI values are adjusted to the interval $I = [1.1,1.15]$ becoming Φ_i.

Next, the tuning parameter is configured through the product of latency, reliability, and CQI parameters. This creates an array of estimated throughput per RB for all users and, finally, the tuned throughput per RB is calculated by multiplying the tuning parameter by the estimated throughput. The number of bits in each user's buffer is stored and then the step in Algorithm 3 corresponding to the integer binary optimization problem is carried out. Finally, it schedules the users.

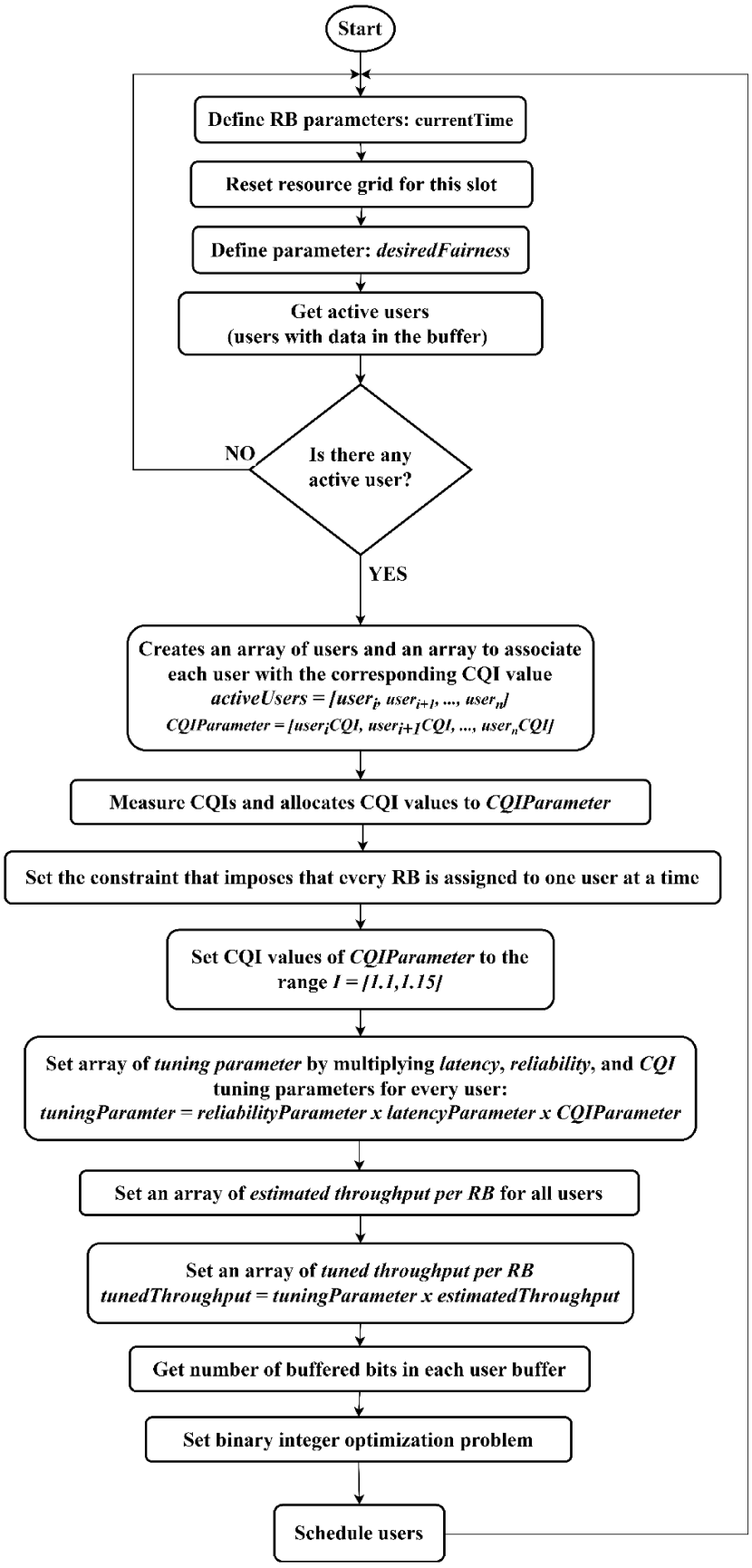


Figure 4.6: CQAS scheduler summarized flowchart.

Algorithm 5: Summarized Channel and QoS-Aware Scheduler

Input: active users
Output: scheduled users
1. Define RB parameters: *currentTime*
2. Reset the resource grid for this slot
3. Define parameters: *desiredFairness, reliabilityParameter,* and *latencyParameter*
4. Get active users
   if there are no active users **then**
      return to Step 1
   else
      proceed to Step 5
   end
5. Creates an array of users and an array to associate each user with the corresponding CQI value
activeUsers = [user$_i$, user$_{i+1}$, ..., user$_n$]
CQIParameter = [user$_i$CQI, user$_{i+1}$CQI, ..., user$_n$CQI]
6. Measure CQIs and allocates CQI values to *CQIParameter*
7. Set the constraint that imposes that every RB is assigned to one user at a time
8. Set CQI values of *CQIParameter* to the range *I = [1.1,1.15]*
9. Set array of tuning parameter by multiplying latency, reliability, and CQI tuning parameters for every user:
*tuningParameter = reliabilityParameter*latencyParameter*CQIParameter.*
10. Set array of estimated throughput for all users
11. Set array of tuned throughput per RB:
*tunedThroughput = tuningParameter*estimatedThroughput*
12. Set binary integer optimization problem as stated in Algorithm 3
13. Schedule users

4.3 Summary

This chapter studied the principles of scheduling algorithms, highlighting both the factors that influence the scheduler's decision and the performance metrics to be analyzed. Thus, the equations and logic through flowcharts and algorithms for the schedulers studied in this work were highlighted, namely RR, the best CQI, QAS, and the implemented algorithms WQAS and CQAS. In the next chapter, the algorithms are compared in terms of verifying their capabilities through a performance evaluation.

"Technology has a banana's shelf life."

- Scott McNealy

5

PERFORMANCE EVALUATION

PERFORMANCE EVALUATION is the stage of verifying the capabilities of the studied algorithms. This chapter shows the configuration of the parameters of the studied scenario and analyzes the algorithm proposals described in Chapter 4. The performance metrics are outlined, and the results are presented and discussed.

5.1 Simulation Environment

The Vienna 5G System Level Simulator MATLAB-based [46] is used to implement the WQAS and CQAS algorithms and compare them to RR, best CQI, and QAS already implemented in the simulator. The proposed scenario is a HetNet with multiple traffic models that include RT and NRT applications for indoor and outdoor users. The traffic models are full buffer (including IoT users), HTTP, video, VoIP, gaming, and vehicular. The number of users for each traffic model was varied by 50, 100, 150, and 200, i.e. the total number of users was 350, 700, 1050, or 1400 to assess network stress. In this way, the results presented were obtained through the average of ten simulation rounds for each scheduler and number of users to achieve confidence about possible variations. Despite this, the variations were minimal, so the confidence intervals have been omitted from the figures. Figure 5.1 shows the scenario of 350 total users and Table 5.1 brings together the main simulation parameters.

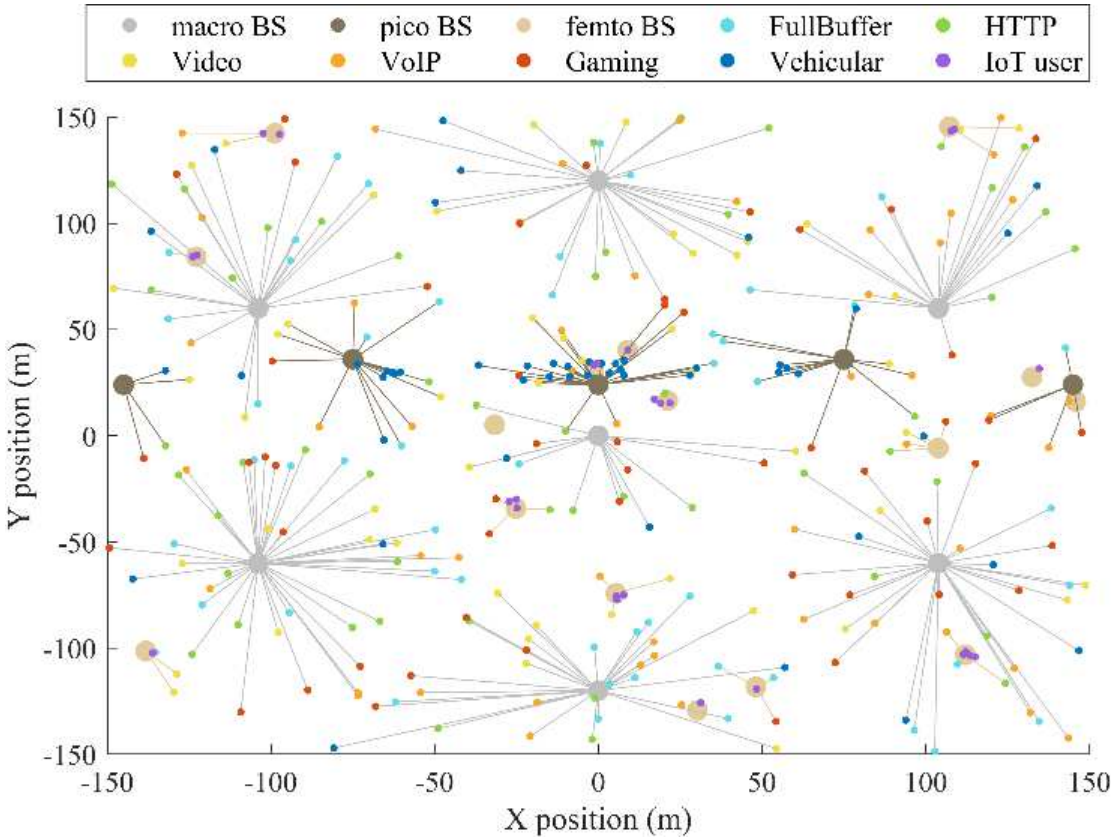


Figure 5.1: Simulation scenario for the HetNet with multiple traffic models for 350 total users.

The macro BSs are arranged in a hexagonal grid structure, while the pico BSs are positioned along the streets next to the vehicular users, and the femto BSs occupy the central position of the user clusters, mainly serving the IoT users. The other users are distributed according to a 2D Poisson distribution function [46].

The study on channel models for frequencies from 0.5 to 100 GHz in 5G networks [47] is the main reference for the system-level simulations used to develop the urban, indoor, and street canyon path loss models for macrocells and femtocells. In turn, picocells follow the free-space path loss model presented in [48].

Regarding the channel models used for pedestrian and vehicular users, the Technical Specification Group Radio Access Network brings together the specifications for High-Speed Downlink Packet Access (HSDPA): UE Radio Transmission and Reception [49]. On the other hand, IoT users follow the Rayleigh channel model.

Considering the association of users with different types of base stations, femtocells are responsible for allocating resources to IoT users. Vehicular users are mostly assigned to picocells but are also served by macrocells. It should be noted that the movement model used for vehicular users is called Random Direction [46], as a random direction is adopted in the first slot so that the user moves in that direction at a constant speed. Finally, it is important to note

that indoor/outdoor and Line-Of-Sight (LOS)/Non Line-Of-Sight (NLOS) decisions are made according to the type of user. In addition, Power Delay Profile (PDP) channel models are configured for pedestrian and vehicular users, and an AWGN channel is used for users in clusters around femtocells [46]. The delay constraints imposed on RT users are 20 ms, 40 ms, 60 ms, and 100 ms for the vehicular, VoIP, gaming, and video models, respectively [22, 46].

Table 5.1: Main simulation parameters.

Parameters	**Values/Meaning**
Simulation duration	2000 time slots
Time slot duration	1 ms
Carrier center frequency	2 GHz
Bandwidth	10 MHz
Schedulers	RR, best CQI, QAS, WQAS, and CQAS
Traffic models	RT: vehicular, VoIP, gaming, and video.
	NRT: HTTP, and full buffer (IoT users are also configured as full buffer)
Delay constraints	Vehicular: 20 ms
	VoIP: 40 ms
	Gaming: 60 ms
	Video: 100 ms
Number of total users	350, 700, 1050 or 1400
Number of BSs/Transmit power	7 macro BSs/46 dBm
	5 pico BSs/43 dBm
	16 femto BSs/30 dBm
Path loss	Macro BSs: UrbanMacro5G [47]
	Pico BSs: free space [48]
	Femto BSs: indoor or Street Canyon (outdoor) [47]
Channel model	IoT users: Rayleigh
	Other users: vehicular or pedestrian as in [49]

5.2 Simulation Results

5.2.1 Throughput

The average throughput per traffic model and scheduler for 350 total users is illustrated in Figure 5.2, while the variation of the average throughput as a function of number of users for full buffer, IoT, HTTP, video, gaming, VoIP, and vehicular is shown in Figure 5.3 to Figure 5.9. Table 5.2 shows the general variation of the average throughput highlighting the proposed schedulers in the most stressed network scenario (1400 total users).

Figure 5.2 separates NRT and RT traffic models into two sets. Considering NRT services, the proposed WQAS algorithm shows worse results than QAS for all three models, which is to be expected given that weights are applied to RT applications so that more RBs are allocated to these types of users. WQAS obtained the worst results for full buffer and IoT users. In turn, CQAS outperforms QAS for all traffic models, except HTTP, due to the CQI factor

being incorporated into its structure as one of its tuning parameters. It is remarkable how the CQI tuning parameter implementation in CQAS circumvents the limitation of QAS for full buffer and IoT users.

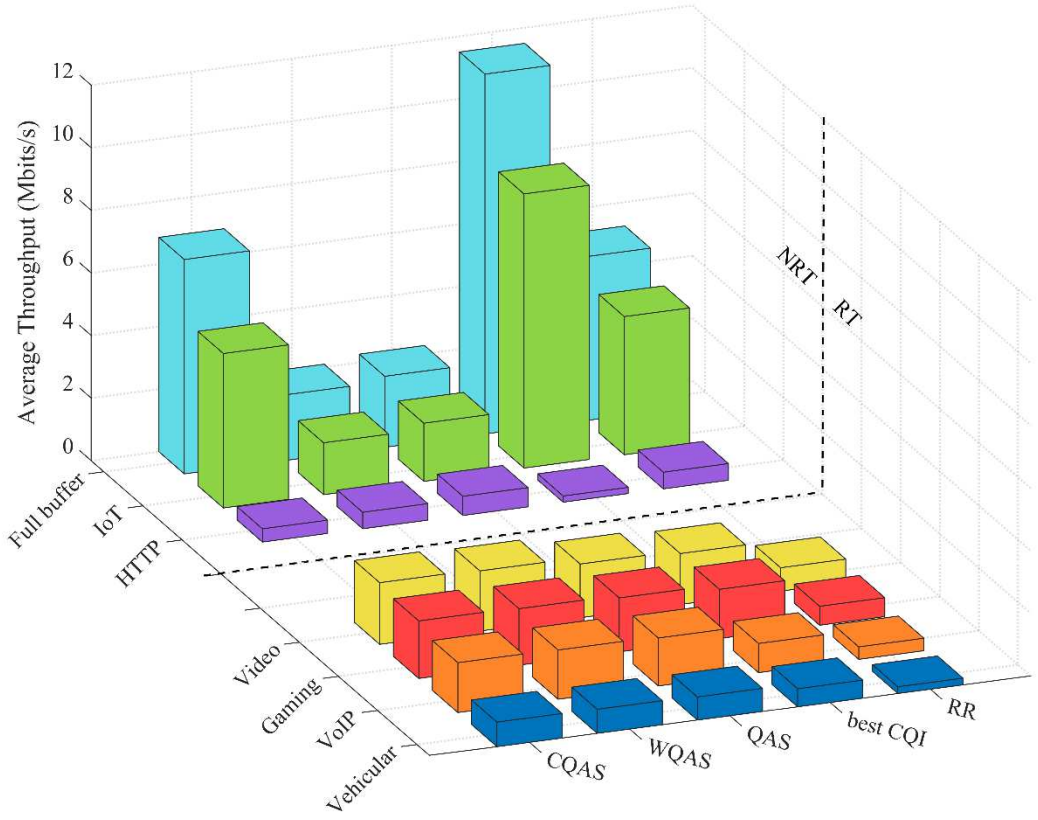


Figure 5.2: Average Throughput per traffic model and scheduler for 350 users.

Figure 5.3 to Figure 5.9 show that there is a tendency for the average throughput to decrease as the number of users varies. Furthermore, Figure 5.3 and Figure 5.4 illustrate how the best CQI algorithm performs better for full buffer and IoT users, but worst for HTTP traffic, with values below 0.2 Mbits/s for all user variations. In addition, Figure 5.3 shows how CQAS outperforms QAS, achieving gains of 202.2%, 204.1%, 205.6%, and 208.7% for the 350 to 1400 user scenarios.

Similarly, Figure 5.4 shows the limitation of QAS for IoT traffic, given that even the RR scheduler obtained better results. The contribution of CQAS in terms of throughput for IoT users compared to QAS was 165.6%, 163.4%, 166.7%, and 171.2% for 50, 100, 150, and 200 users, respectively. Also, according to Figure 5.4, CQAS is superior to RR by 11.2%, inferior to best CQI by 46.6%, and superior to WQAS by 215.3% for 200 users.

Finally, Figure 5.5 shows the limitation of CQAS in terms of HTTP traffic, as it only has higher throughput values than the best CQI algorithm. This shows one of the downsides of

implementing the CQI tuning parameter: CQAS is worse than QAS by 30.9%, 30.6%, 34%, and 31.6% for 50, 100, 150, and 200 users, respectively. In addition, WQAS showed similar results to RR, reaching values close to 0.5 Mbits/s and showing a drop in overall performance for NRT services when compared to QAS.

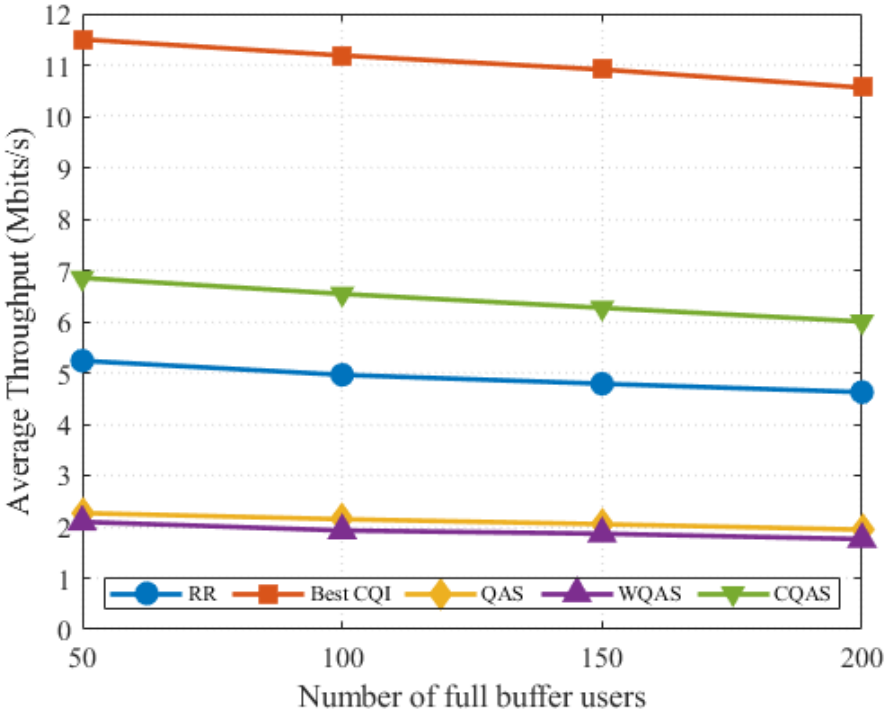


Figure 5.3: Full buffer users' average throughput as a function of the number of users.

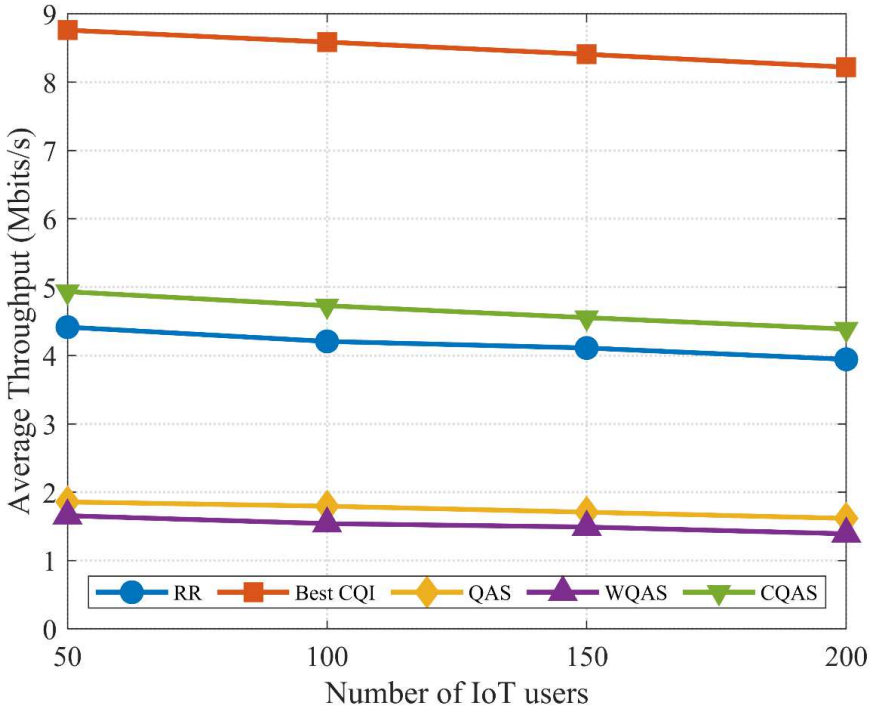


Figure 5.4: IoT users' average throughput as a function of the number of users.

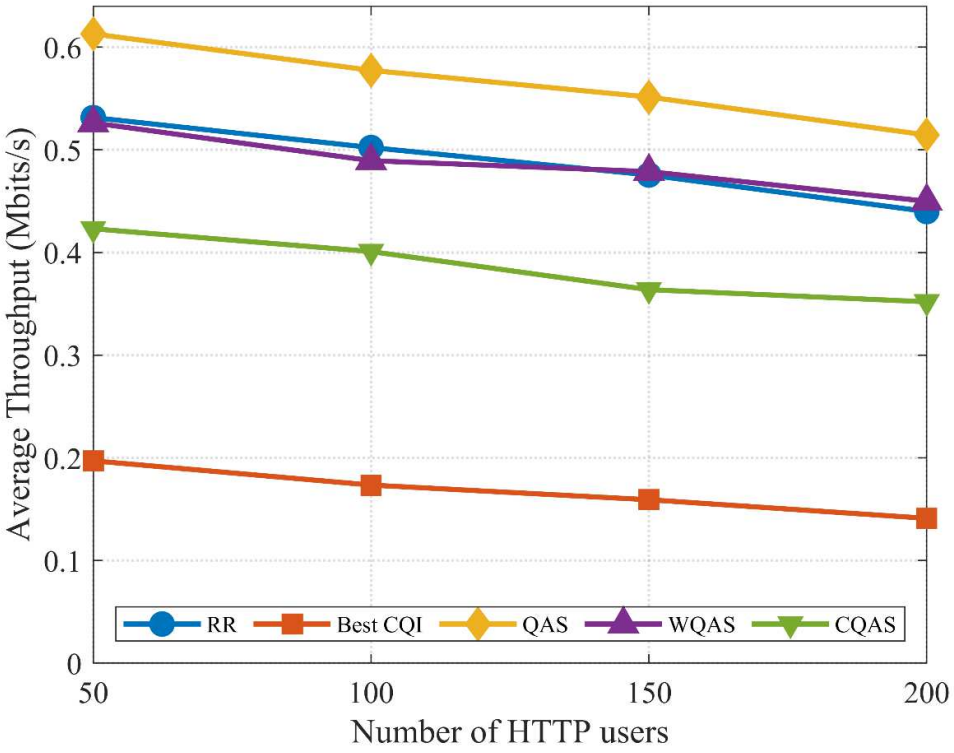


Figure 5.5: HTTP users' average throughput as a function of the number of users.

Figure 5.6 illustrates the variation in the average throughput of video users as a function of the number of users. CQAS shows gains over QAS of 14.8%, 14.7%, 14.5%, and 16.3% for 50, 100, 150, and 200 users, respectively. For the same interval, WQAS outperforms QAS by 12.3%, 11.3%, 14.7%, and 16.2%. This shows how WQAS outperforms CQAS for 150 users and reaches close values in the other scenarios. In general, for RT applications, the schedulers can be ranked from best to worst in the following sequence: CQAS, WQAS, QAS, best CQI, and RR. The CQAS and WQAS algorithms performed almost identically for RT applications, so CQAS was chosen due to its better overall performance, including NRT applications.

Figure 5.7 shows how the throughput performance for gaming users is similar to video traffic. In this sense, CQAS shows gains of 6.6%, 7.5%, 6.8%, and 9.1% over QAS for 50, 100, 150, and 200 users, respectively. About VoIP traffic, Figure 5.8 shows how the QoS-aware algorithms stand out compared to the best CQI and RR. WQAS shows gains of 2.7%, 3.4%, 7.9%, and 3.3% over QAS for 50, 100, 150 and 200 users, respectively. In turn, CQAS shows gains of 4%, 4.1%, 6.1%, and 2% compared to QAS for the same conditions. Figure 5.9 illustrates how CQAS stands out for vehicular traffic when considering 50 and 100 users, achieving gains of 6% and 5.4% compared to QAS. For 150 and 200 users, the values achieved by WQAS and CQAS are similar.

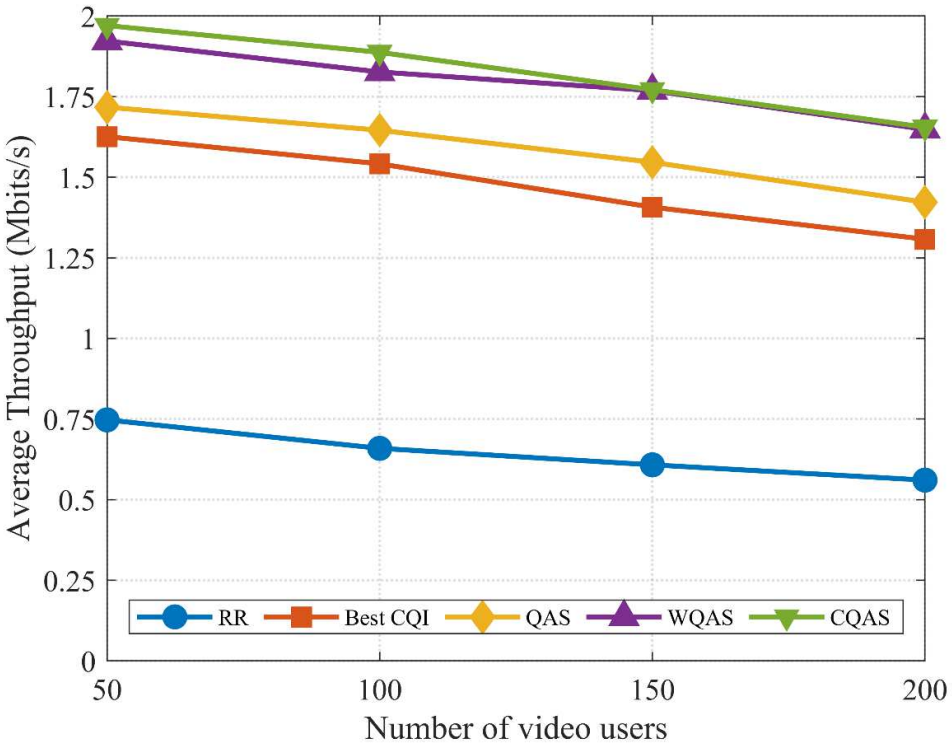


Figure 5.6: Video users' average throughput as a function of the number of users.

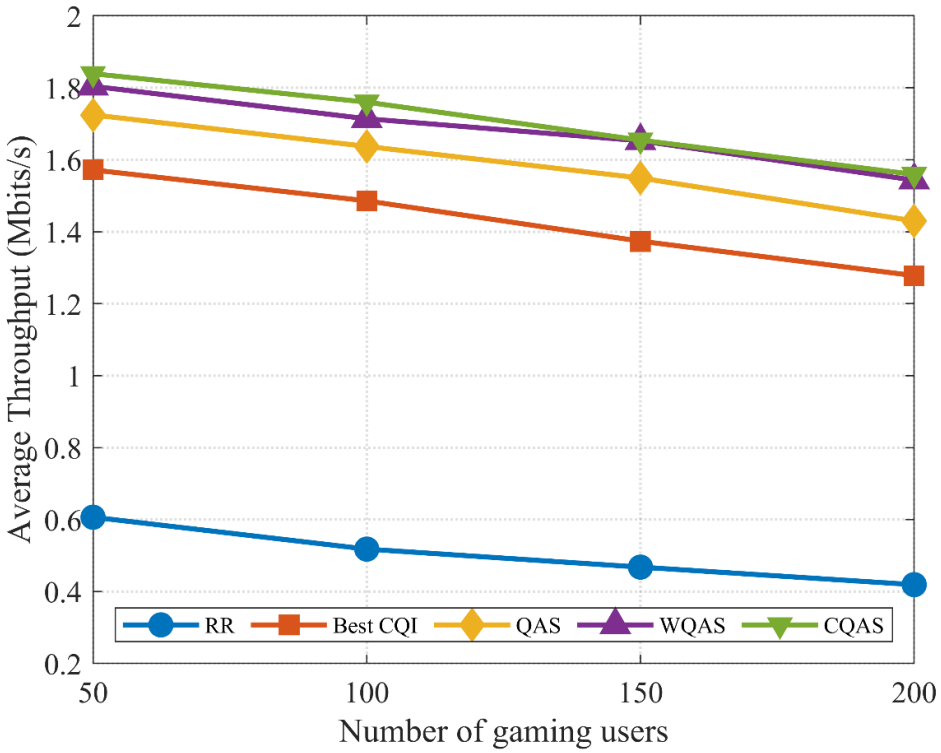


Figure 5.7: Gaming users' average throughput as a function of the number of users.

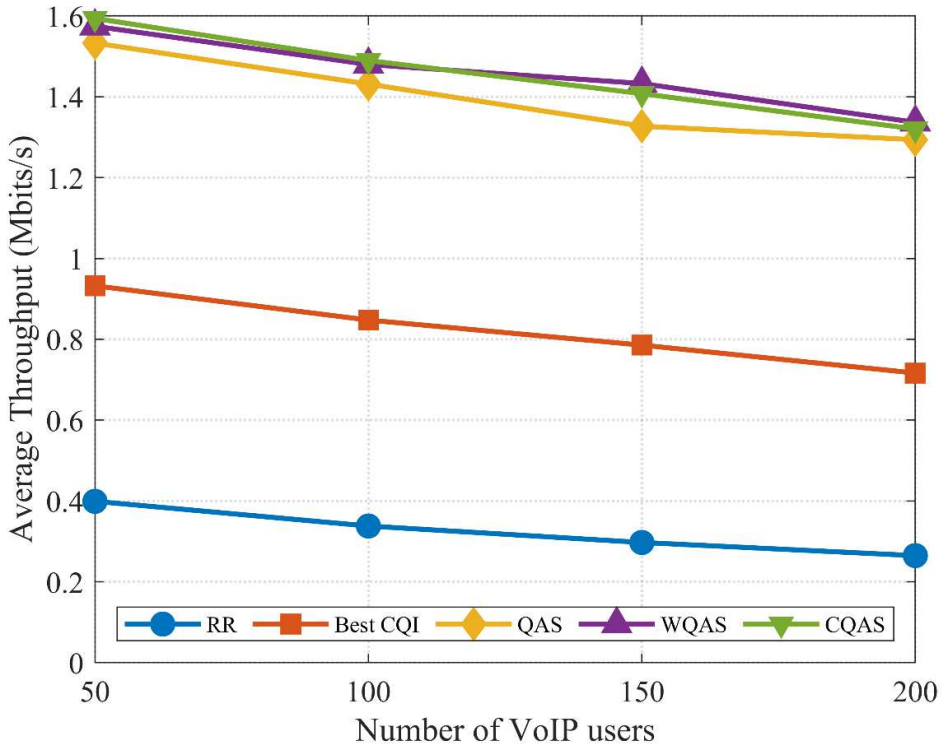


Figure 5.8: VoIP users' average throughput as a function of the number of users.

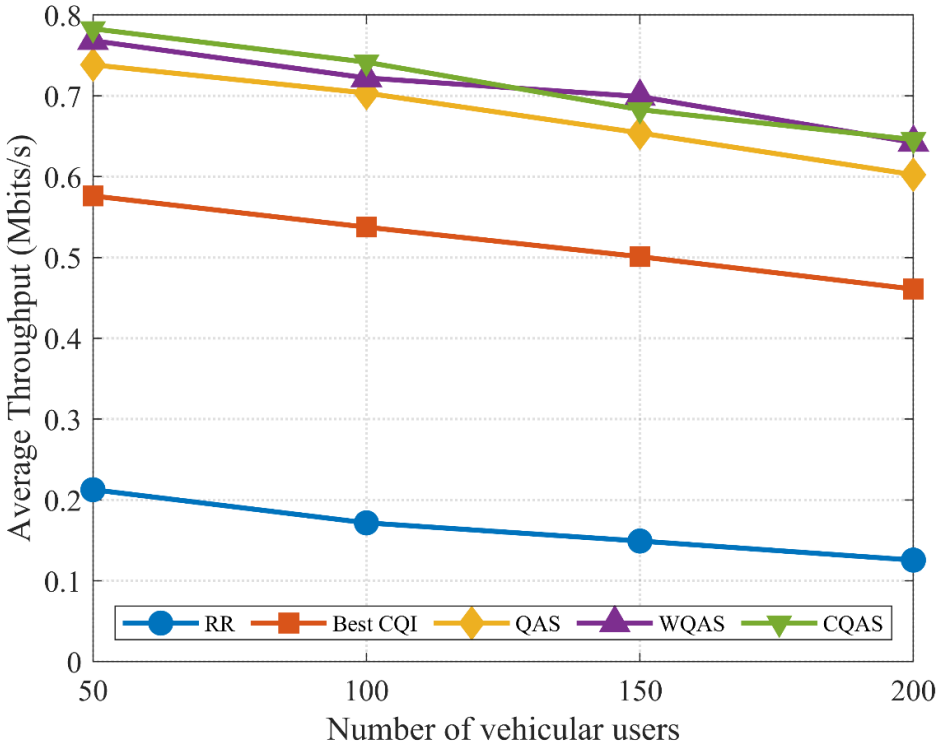


Figure 5.9: Vehicular users' average throughput as a function of the number of users.

Finally, Table 5.2 shows the overall variation in average throughput, highlighting the proposed schedulers. It is noticeable that WQAS performed almost identically to CQAS for RT traffic and outperformed QAS except for NRT models, which is expected due to the reservation

of RBs for RT users. In turn, CQAS performed better than QAS for all traffic models except HTTP, which suggests this is an impact of implementing the CQI tuning parameter.

Table 5.2: Average throughput ratio between the proposed schedulers and the other schedulers for RT services in the most stressed network scenario (1400 total users).

	RR	**best CQI**	**QAS**	**WQAS**	**CQAS**
			Full buffer		
WQAS	0.38	0.17	0.91	1.00	0.29
CQAS	1.30	0.57	3.09	3.41	1.00
			IoT		
WQAS	0.35	0.17	0.86	1.00	0.32
CQAS	1.11	0.53	2.71	3.15	1.00
			HTTP		
WQAS	1.02	3.19	0.87	1.00	1.28
CQAS	0.80	2.50	0.68	0.78	1.00
			Video		
WQAS	2.94	1.26	1.16	1.00	1.00
CQAS	2.95	1.27	1.16	1.00	1.00
			Gaming		
WQAS	3.68	1.21	1.09	1.00	0.99
CQAS	3.72	1.22	1.09	1.01	1.00
			VoIP		
WQAS	5.05	1.87	1.03	1.00	1.01
CQAS	4.99	1.84	1.02	0.99	1.00
			Vehicular		
WQAS	5.11	1.39	1.07	1.00	0.99
CQAS	5.14	1.40	1.07	1.01	1.00

5.2.2 Reliability

As explained in Equation 4.13 of Chapter 4, lower average BLER values indicate greater reliability and, therefore, a higher priority for allocating resources to the user. As such, the CQAS resource allocation procedure is suggested to be effective, given that it has intermediate BLER values between the RR and best CQI algorithms. On the other hand, it has slightly higher values than the QAS and WQAS schedulers. The following order of increasing average BLER values can be observed: RR, QAS, WQAS, CQAS, and best CQI. In addition, reliability is a determining factor in throughput performance, so Table 5.3 shows an overview of the average BLER about the variation in the number of total users.

The RR algorithm allocates resources equally to all users, while the best CQI evaluates channel conditions in such a way that some transmissions may fail and, consequently, reliability is reduced. Concerning WQAS, the average BLER values obtained were similar to those observed when QAS was used, showing that the implementation of weights did not significantly affect BLER. In turn, CQAS shows a slight increase in BLER compared to QAS and WQAS

due to the CQI parameter implementation. Thus, the CQAS, WQAS, and QAS algorithms have better average throughput combinations while maintaining interesting average BLER values.

Table 5.3: Average BLER.

	RR	**best CQI**	**QAS**	**WQAS**	**CQAS**
		350 users			
Full buffer	0.15	0.33	0.18	0.18	0.22
IoT	0.15	0.32	0.19	0.20	0.22
HTTP	0.10	0.25	0.12	0.13	0.20
Video	0.19	0.33	0.25	0.24	0.27
Gaming	0.16	0.34	0.21	0.22	0.25
VoIP	0.18	0.35	0.24	0.26	0.27
Vehicular	0.13	0.30	0.17	0.19	0.19
		700 users			
Full buffer	0.17	0.34	0.19	0.20	0.23
IoT	0.15	0.33	0.21	0.21	0.23
HTTP	0.10	0.27	0.13	0.15	0.20
Video	0.20	0.35	0.27	0.26	0.29
Gaming	0.17	0.35	0.23	0.23	0.27
VoIP	0.19	0.37	0.25	0.27	0.28
Vehicular	0.15	0.31	0.18	0.20	0.21
		1050 users			
Full buffer	0.17	0.35	0.20	0.22	0.25
IoT	0.16	0.35	0.22	0.23	0.24
HTTP	0.11	0.28	0.14	0.16	0.21
Video	0.21	0.35	0.29	0.28	0.30
Gaming	0.18	0.36	0.24	0.25	0.28
VoIP	0.19	0.37	0.27	0.29	0.30
Vehicular	0.15	0.30	0.19	0.21	0.22
		1400 users			
Full buffer	0.18	0.36	0.22	0.23	0.27
IoT	0.17	0.36	0.24	0.25	0.25
HTTP	0.11	0.29	0.15	0.16	0.23
Video	0.22	0.36	0.29	0.29	0.31
Gaming	0.18	0.37	0.25	0.27	0.30
VoIP	0.20	0.38	0.27	0.29	0.31
Vehicular	0.16	0.31	0.20	0.23	0.23

Full buffer traffic, including IoT users, have higher average throughput values as they have lower average BLER as shown in Table 5.3. It should also be noted that HTTP users have fewer active users per interval and, therefore, more users have a BLER value of zero, reducing the value of the average BLER. Taking RT applications into account, Table 5.3 shows that video users have slightly better reliability than VoIP users since they produce larger amounts of larger packets, which results in higher average throughput. Furthermore, gaming users have similar average BLER and average throughput to video users due to the higher data flow characteristic of this traffic model. Finally, it stands out that vehicular users tend to have lower

BLER values than the other RT traffic models due to the vehicular users' association with picocells, which makes them less bottlenecked.

5.2.3 Fairness Index

The fairness index is analyzed to determine whether users receive resources in a balanced way. Thus, Jain's fairness index method [44] calculates fairness from the vector storing the data rate values. Figure 5.10 illustrates the variation of the fairness index as a function of the total number of users.

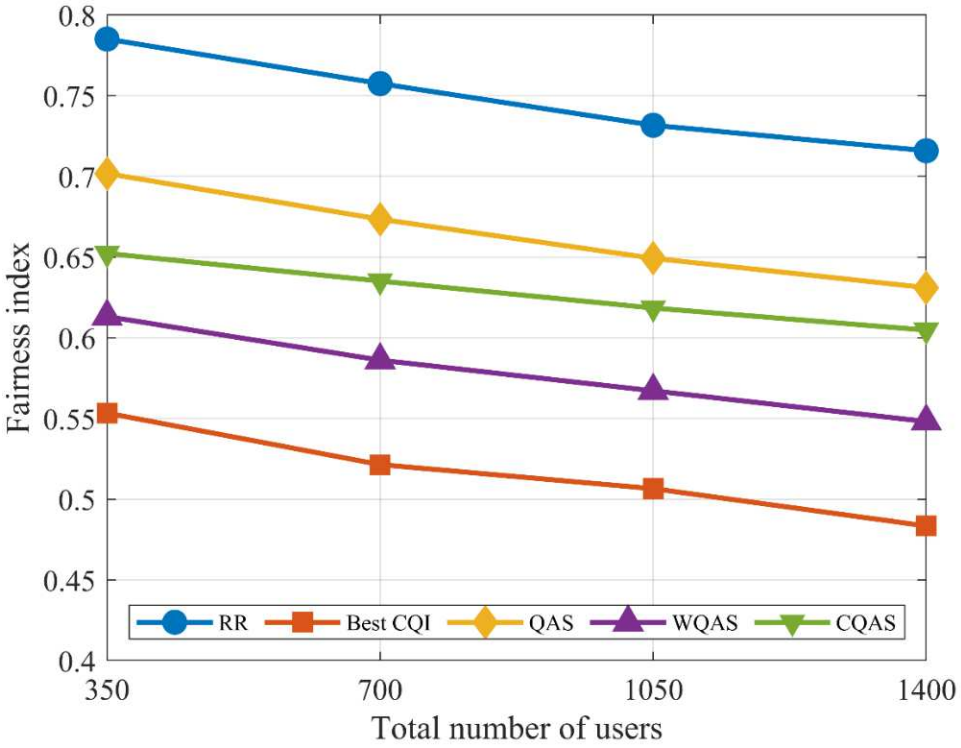


Figure 5.10: Fairness index as a function of the total number of users.

The fairness index decreases as the total number of users increases. It is also worth noting that the highest values are attributed to the RR scheduler as expected since this algorithm allocates resources to all users according to a first-come, first-served procedure. On the other hand, the best CQI scheduler makes its decisions based on the channel conditions reported by the UEs, so users with poor channel conditions tend not to benefit from the resource allocation process. In turn, the algorithms with the desired fairness index configured as a tuning parameter, QAS, WQAS, and CQAS, obtained fairness index values intermediate between best CQI and RR.

It should be noted that among these last three algorithms, QAS has the highest fairness index. WQAS has the lowest value, which suggests that prioritizing users of RT applications

reduces the fairness index by allocating more RBs in sequence to such services. Finally, the CQI tuning parameter implementation for CQAS generated the cost of fairness reduction while producing overall gains in throughput.

5.2.4 Latency

Regarding latency, the closer the latency value experienced by the user is to the delay constraint imposed on the traffic model, the higher the user's priority when allocating resources to the QAS, WQAS, and CQAS algorithms, which are designed considering the operation of the network in the face of mixed RT applications. In this sense, the Delay Constraints (DCs) for vehicular, VoIP, gaming, and video users are 20, 40, 60, and 100 ms [22, 46]. Table 5.4 shows the average percentage ($\overline{\%}$) of users under the average maximum latency values ($\overline{max.lat.}$) for all studied scenarios. Figure 5.11 and Figure 5.12 illustrate the latency Empirical Cumulative Distribution Function (ECDF) per real-time traffic model for the two new proposed algorithms, WQAS and CQAS, considering 1400 total users.

The expected performance that follows the latency sequence according to the delay constraints (vehicular, VoIP, gaming, and video) was achieved only by the QAS, WQAS, and CQAS algorithms. Figure 5.11 shows that WQAS enables 100% of users to reach values lower than those imposed by the DCs for all RT traffic models in the most stressed network scenario. The following values stand out: 100% of vehicular users are under 17 ms, 100% of VoIP users are under 37 ms, 100% of gaming users are under 53 ms, and 100% of video users are under 94 ms. Observing Table 5.4, the results suggest that WQAS performs better in terms of latency when compared to QAS: it has a lower average latency than QAS by 3 ms for vehicular users, 1 ms for VoIP users, 3 ms for gaming users, and 2 ms for video users. Therefore, the results indicate that WQAS has a higher tendency to maintain DCs for RT applications in scenarios of high network stress for more situations than QAS.

Furthermore, QAS shows 100% of RT users below the values imposed as delay constraints. On the other hand, considering CQAS, Figure 5.12 shows that 99% of users achieved latency values lower than the delay requirements for all RT traffic models, and the sequence of DCs was followed. Considering the RR, all the DCs were exceeded, for example, 90% of the vehicular users are under a maximum of 41 ms in the 1400 total users' scenario. In turn, the best CQI scheduler did not achieve good results for RT applications. Considering the same scenario of 1400 total users, 69% of vehicular users are under 2080 ms, 80% of video users are under 1909 ms, 77% of VoIP users are under 1936 ms, and 75% of gaming users are under 2039 ms.

Table 5.4: Average % of users under average maximum latency values.

	RR		Best CQI		QAS		WQAS		CQAS	
350 total users										
	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$
Vehicular	97	28	77	1907	100	14	100	12	100	16
VoIP	100	13	90	1753	100	27	100	26	100	30
Gaming	100	22	88	1756	100	37	100	34	100	40
Video	100	25	86	1845	100	81	100	78	100	85
HTTP	94	1731	60	1446	96	815	94	988	80	1256
700 total users										
	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$
Vehicular	96	31	74	1957	100	15	100	14	100	18
VoIP	100	17	87	1779	100	30	100	30	100	33
Gaming	100	26	85	1756	100	43	100	41	100	47
Video	100	30	82	1893	100	85	100	83	100	89
HTTP	92	1836	56	1522	93	858	91	1084	75	1340
1050 total users										
	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$
Vehicular	92	34	72	1999	100	18	100	16	99	21
VoIP	100	18	84	1854	100	34	100	34	100	36
Gaming	100	27	82	1864	100	47	100	46	100	52
Video	100	31	79	1969	100	90	100	89	100	95
HTTP	89	1913	53	1600	91	906	89	1158	70	1458
1400 total users										
	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$	$\overline{\%}$	$\overline{max.lat}$
Vehicular	90	41	69	2080	100	20	100	17	99	22
VoIP	100	22	80	1909	100	38	100	37	99	41
Gaming	100	32	77	1936	100	56	100	53	99	61
Video	100	35	75	2039	100	96	100	94	99	101
HTTP	86	1996	50	1967	88	932	85	1241	66	1536

Thus, the results suggest that WQAS has the lowest latency values among the algorithms that met the requirements related to delay. However, WQAS and QAS algorithms penalize NRT traffic to the detriment of RT, while the RR and best CQI schedulers are unable to handle RT applications. Finally, CQAS is more balanced between its commitment to serving applications with QoS requirements and providing more transmissions, and consequently, shows higher average throughput for RT and NRT applications.

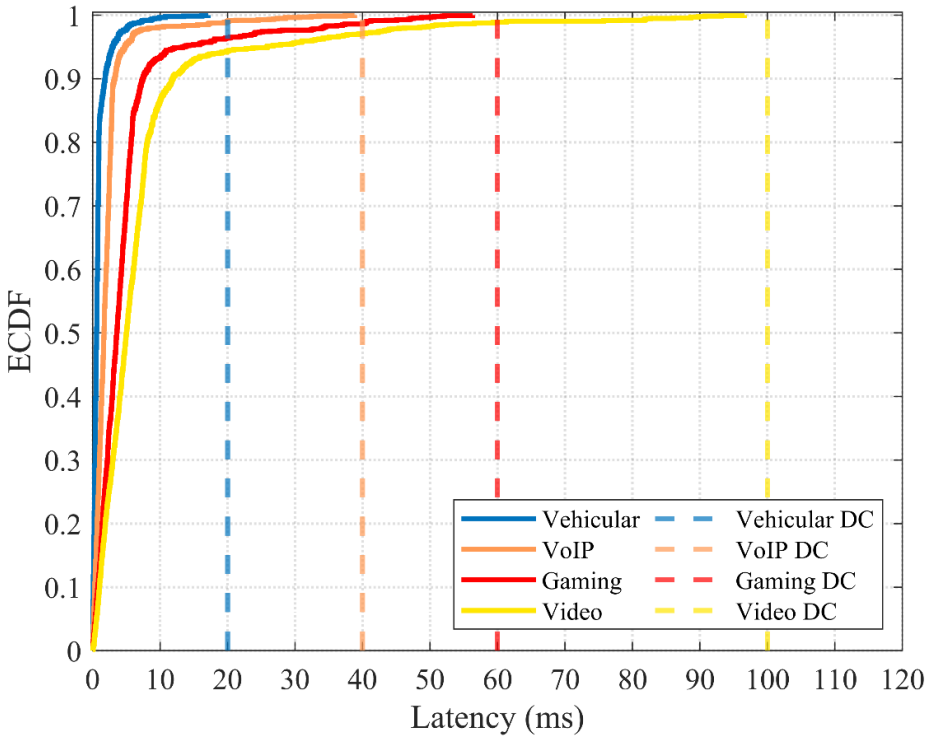


Figure 5.11: Latency ECDF per real-time traffic model for WQAS and 1400 total users.

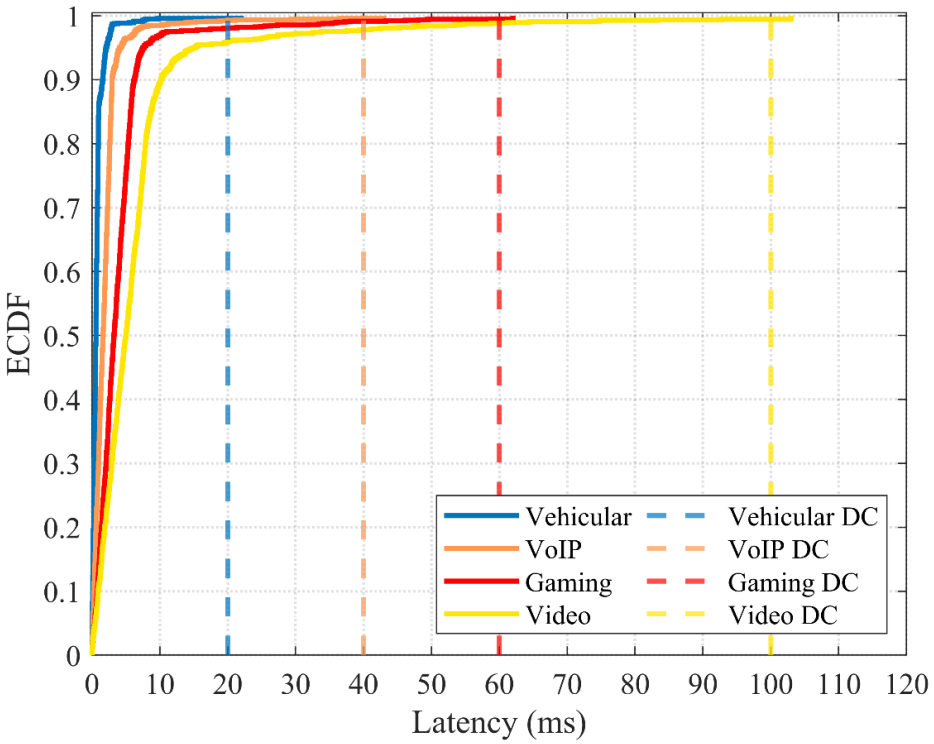


Figure 5.12: Latency ECDF per real-time traffic model for CQAS and 1400 total users.

5.2.5 Simulation Time

Finally, as the optimization problem of QAS, WQAS, and CQAS involve integer programming, they fall into the nonlinear programming (NLP) problems category so make heavy use of

computational resources intensively. Hence, it is important to compare the simulation time data for each algorithm, given the variation in the number of users. Table 5.5 shows the average simulation time in minutes considering the 10 runs for each scenario.

Table 5.5: Average simulation time in minutes.

Total users	RR	best CQI	QAS	WQAS	CQAS
350	19	28	48	55	62
700	35	52	91	103	116
1050	66	98	175	195	221
1400	126	193	343	395	445

CQAS is the scheduler with the longest total simulation duration. It shows an increase in total simulation time of 225%, 131%, 30%, and 13% compared to the RR, best CQI, QAS, and WQAS algorithms for the most stressed network scenario. Given the same conditions, WQAS shows a variation of 214%, 105%, 15%, and -11% when compared to RR, best CQI, QAS, and CQAS, respectively. Also noteworthy is the average value of 445 minutes to simulate a 1400-user scenario for CQAS. Because of this, it is interesting to study the application of Reinforcement Learning (RL) methods to reduce the total simulation time in scenarios of greater network stress.

5.3 Summary

This chapter is dedicated to evaluating the performance of the proposed algorithms. The Vienna 5G System Level Simulator MATLAB-based was chosen as the simulation environment to conduct the evaluation. The results show that the WQAS and CQAS algorithms can achieve good performance, respecting the QoS requirements of RT applications, and excelling in the metrics analyzed.

It turns out that the RR and best CQI algorithms had unsatisfactory results for RT traffic, while the QAS and WQAS schedulers had the best results for RT traffic but were detrimental to NRT traffic. Lastly, the CQAS scheduler met the requirements of the RT models and was less harmful to NRT traffic, so it proved to be the most balanced of the algorithms. It also highlights the importance of a Channel and QoS-aware algorithm.

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.”

- Marie Curie

6

CONCLUSIONS

5G NETWORKS are an important mobile communications technology. Demand for this technology is increasing as new applications and services emerge that seek to integrate more users and devices into the network infrastructure.

In this sense, the heterogeneity of 5G systems stands out, both in terms of network elements and the variety of types of users. As such, resource allocation is a critical factor in network configuration and operation, especially scheduling algorithm techniques that are responsible for configuring the efficient use of resources.

Considering the wide range of applications of modern communications systems, the resource allocation methods development that consider QoS requirements for each traffic model is an influential factor during resource allocation. As RT and NRT applications stand out in a heterogeneous environment, the study of channel and QoS-aware solutions stands out.

Despite the existence of several proposals for schedulers, there is a lack of work aimed at allocating resources in heterogeneous networks configured with multiple traffic models in 5G systems that consider both channel conditions and QoS requirements. Therefore, the research described in this dissertation develops two new packet scheduling algorithms, called Weighted QoS Aware Scheduler (WQAS) and Channel and QoS Aware Scheduler (CQAS).

The proposed algorithms were implemented and evaluated in a simulation environment. The MATLAB-based Vienna 5G System Level Simulator was selected as the environment for the evaluation. Thus, the schedulers RR, best CQI, and QAS were chosen for comparison and verification of the relevance of the proposed algorithms. These algorithms were selected because of their literature relevance and, in the case of QAS because it is the basis for creating WQAS and CQAS.

The evaluation was carried out on a network based on macrocells, picocells, and femtocells, with users varying according to the full buffer (including IoT users), HTTP, vehicular, VoIP, gaming, and video traffic models. The comparison was made considering the metrics of average throughput, reliability, fairness index, latency, and computational time.

The results show that both WQAS and CQAS met the QoS requirements of RT applications in all the scenarios evaluated. The WQAS algorithm stood out in terms of throughput gains for RT applications, while CQAS showed better overall performance in terms of throughput, also providing gains for NRT services. Therefore, performance results suggest that with CQAS the network serves more users and applications in the simulated scenarios as required by 5G systems.

For future research, one of the points to be studied and improved about CQAS is the simulation time, so the reinforcement learning methodology is a strong candidate for solving the optimization problem and improving the results of this research. Finally, there is interest in applying the techniques of the WQAS and CQAS algorithms with input parameters adjusted according to network traffic to further enhance the RT and NRT application's performance.

” Everything has been thought of before,
but the problem is to think of it again.”

- Johann Wolfgang von Goethe

References

[1] A. Mamane, M. Fattah, M. El Ghazi, M. El Bekkali, Y. Balboul, and S. Mazer, "Scheduling algorithms for 5G networks and beyond: Classification and survey," *IEEe Access,* vol. 10, pp. 51643-51661, 2022. doi: 10.1109/ACCESS.2022.3174579.

[2] E. Hossain and M. Hasan, "5G cellular: key enabling technologies and research challenges," *IEEE Instrumentation & Measurement Magazine,* vol. 18, no. 3, pp. 11-21, 2015. doi: 10.1109/MIM.2015.7108393.

[3] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2020.

[4] A. Osseiran, J. F. Monserrat, and P. Marsch, *5G mobile and wireless communications technology*. Cambridge University Press, 2016.

[5] N. Bhushan *et al.*, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine,* vol. 52, no. 2, pp. 82-89, 2014. doi: 10.1109/MCOM.2014.6736747.

[6] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE communications surveys & tutorials,* vol. 15, no. 2, pp. 678-700, 2012. doi: 10.1109/SURV.2012.060912.00100.

[7] A. H. Ali and M. Nazir, "Radio resource management with QoS guarantees for LTE-A systems: a review focused on employing the multi-objective optimization techniques," *Telecommunication Systems,* vol. 67, no. 2, pp. 349-365, 2018. doi: 10.1007/s11235-017-0342-z.

[8] M. M. Nasralla, N. Khan, and M. G. Martini, "Content-aware downlink scheduling for LTE wireless systems: A survey and performance comparison of key approaches," *Computer Communications,* vol. 130, pp. 78-100, 2018. doi: 10.1016/j.comcom.2018.08.009.

[9] N. Sharma and K. Kumar, "Resource allocation trends for ultra dense networks in 5G and beyond networks: A classification and comprehensive survey," *Physical Communication,* vol. 48, p. 101415, 2021. doi: 10.1016/j.phycom.2021.101415.

[10] M. E. Haque, F. Tariq, M. R. Khandaker, K.-K. Wong, and Y. Zhang, "A survey of scheduling in 5g urllc and outlook for emerging 6g systems," *IEEE access,* 2023. doi: 10.1109/ACCESS.2023.3264592.

[11] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Communications Surveys & Tutorials,* vol. 23, no. 2, pp. 668-695, 2021. doi: 10.1109/COMST.2021.3059896.

[12] S. Manap, K. Dimyati, M. N. Hindia, M. S. A. Talip, and R. Tafazolli, "Survey of radio resource management in 5G heterogeneous networks," *IEEE access,* vol. 8, pp. 131202-131223, 2020. doi: 10.1109/ACCESS.2020.3002252.

[13] B. Agarwal, M. A. Togou, M. Marco, and G.-M. Muntean, "A comprehensive survey on radio resource management in 5G HetNets: Current solutions, future trends and open issues," *IEEE Communications Surveys & Tutorials,* vol. 24, no. 4, pp. 2495-2534, 2022. doi: 10.1109/COMST.2022.3207967.

[14] R. Kumar, D. Sinwar, and V. Singh, "QoS aware resource allocation for coexistence mechanisms between eMBB and URLLC: Issues, challenges, and future directions in 5G," *Computer Communications,* 2023. doi: 10.1016/j.comcom.2023.10.024.

[15] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Communications Surveys & Tutorials,* vol. 22, no. 2, pp. 905-929, 2020. doi: 10.1109/COMST.2020.2971781.

[16] C. F. Müller, G. Galaviz, Á. G. Andrade, I. Kaiser, and W. Fengler, "Evaluation of scheduling algorithms for 5g mobile systems," *Computer Science and Engineering—Theory and Applications,* pp. 213-233, 2018. doi: 10.1007/978-3-319-74060-7_12.

[17] B. S. Monikandan, A. Sivasubramanian, S. Babu, G. Prasanna Venkatesan, and C. Arunachalaperumal, "Channel aware optimized proportional fair scheduler for LTE

downlink," *Peer-to-Peer Networking and Applications,* vol. 13, pp. 2135-2144, 2020. doi: 10.1007/s12083-019-00826-z.

[18] A. Abdulazeez, M. M. Yahaya, I. B. Yabo, A. Bello, M. M. Umar, and A. Mohammed, "Prioritized quality of service-aware downlink scheduling algorithm for LTE network," in *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, 2022: IEEE, pp. 1-5. doi: 10.1109/NIGERCON54645.2022.9803123.

[19] M. I. Saglam and M. Kartal, "5G enhanced mobile broadband downlink scheduler," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, 2019: IEEE, pp. 687-692. doi: 10.23919/ELECO47770.2019.8990378.

[20] R. Joda *et al.*, "QoS-aware joint component carrier selection and resource allocation for carrier aggregation in 5G," in *ICC 2021-IEEE International Conference on Communications*, 2021: IEEE, pp. 1-6. doi: 10.1109/ICC42927.2021.9500923.

[21] D. H. Taha, H. Haci, and A. Serener, "Novel Channel/QoS Aware Downlink Scheduler for Next-Generation Cellular Networks," *Electronics,* vol. 11, no. 18, p. 2895, 2022. doi: 10.3390/electronics11182895.

[22] A. Shiyahin, S. Schwarz, and M. Rupp, "Quality of Service Aware Scheduling in Mixed Traffic Wireless Networks," in *2022 IEEE 27th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2022: IEEE, pp. 159-165. doi: 10.1109/CAMAD55695.2022.9966904.

[23] M. K. Müller *et al.*, "Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator," *EURASIP Journal on Wireless Communications and Networking,* vol. 2018, pp. 1-17, 2018. doi: 10.1186/s13638-018-1238-7.

[24] O. T. Eluwole, N. Udoh, M. Ojo, C. Okoro, and A. J. Akinyoade, "From 1G to 5G, what next?," *IAENG International Journal of Computer Science,* vol. 45, no. 3, 2018.

[25] J. H. Schiller, *Mobile communications*. Pearson education, 2003.

[26] T. S. Rappaport, *Wireless communications: principles and practice*. Cambridge University Press, 2024. doi: 10.1017/9781009489843.

[27] ITU-R, "Recommendation M.687-2 International Mobile Telecommunications-2000 (IMT-200)," 1997.

[28] C. Cox, *An introduction to LTE: LTE, LTE-advanced, SAE and 4G mobile communications*. John Wiley & Sons, 2012. doi: 10.1002/9781119942825.

[29] G. Liu, Y. Huang, Z. Chen, L. Liu, Q. Wang, and N. Li, "5G deployment: Standalone vs. non-standalone from the operator perspective," *IEEE Communications Magazine,* vol. 58, no. 11, pp. 83-89, 2020. doi: 10.1109/MCOM.001.2000230.

[30] E. M. Report, "Ericsson Mobility Report November 2024," 2024.

[31] M. Series, "IMT Vision–Framework and overall objectives of the future development of IMT for 2020 and beyond," *Recommendation ITU,* vol. 2083, no. 0, pp. 1-21, 2015.

[32] J. T. Penttinen, *5G explained: security and deployment of advanced mobile communications*. John Wiley & Sons, 2019. doi: 10.1002/9781119275695.

[33] S. Rommer, P. Hedman, M. Olsson, L. Frid, S. Sultana, and C. Mulligan, *5G Core Networks: Powering Digitalization*. Academic Press, 2019.

[34] 3GPP, "5G; NR; NR and NG-RAN Overall description; Stage 2," 3rd-Generation Partnership Project (3GPP), 2022.

[35] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Communications Magazine,* vol. 56, no. 3, pp. 210-217, 2018. doi: 10.1109/MCOM.2017.1700517.

[36] 3GPP, "Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN (Release 12)," in "TR 36.932 (V12.1.0)," 2013.

[37] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in fourth generation heterogeneous networks," *IEEE Communications Magazine,* vol. 44, no. 10, pp. 96-103, 2006. doi: 10.1109/MCOM.2006.1710420.

[38] A. Ghosh *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE communications magazine,* vol. 50, no. 6, pp. 54-64, 2012. doi: 10.1109/MCOM.2012.6211486.

[39] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications magazine,* vol. 46, no. 9, pp. 59-67, 2008. doi: 10.1109/MCOM.2008.4623708.

[40] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine,* vol. 53, no. 9, pp. 74-81, 2015. doi: 10.1109/MCOM.2015.7263349.

[41] T. W. I. o. Telecommunications. "The Vienna 5G System Level Simulator: User Mannual." https://owncloud.tuwien.ac.at/index.php/s/izVNiuNNwnw1VS8 (accessed July 02, 2024).

[42] 3GPP, "Service requirements for V2X services," 3rd-Generation Partnership Project (3GPP), 2010.

[43] S. M. Al-Shehri, P. Loskot, T. Numanoglu, and M. Mert, "Common metrics for analyzing, developing and managing telecommunication networks," *arXiv preprint arXiv:1707.03290,* 2017. doi: 10.48550/arXiv.1707.03290.

[44] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measurement of fairness and discrimination for resource allocation in shared computer system," *Eastern Research Laboratory, Digital Equipment Corporation: Hudson, MA, USA,* vol. 2, 1984. doi: 10.48550/arXiv.cs/9809099.

[45] P. Thienthong, N. Teerasuttakorn, K. Nuanyai, and S. Chantaraskul, "Comparative study of scheduling algorithms in lte hetnets with almost blank subframe," *Engineering Journal,* vol. 25, no. 8, pp. 39-50, 2021. doi: 10.4186/ej.2021.25.8.39.

[46] M. K. Müller *et al.*, "Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator," *EURASIP Journal on Wireless Communications and Networking,* vol. 2018, no. 1, pp. 1-17, 2018. doi: 10.1186/s13638-018-1238-7.

[47] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," *3rd Generation Partnership Project (3GPP),* 2017.

[48] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012.

[49] 3GPP, "High Speed Downlink Packet Access (HSDPA); User Equipment (UE) radio transmission and reception (FDD)," *3rd Generation Partnership Project (3GPP)* 2002.