UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Efficient Software-based Base Station Design for C-RAN

Mojgan Barahman

Supervisor: Doctor Luís Manuel de Jesus Sousa Correia

Co-Supervisor: Doctor Lúcio Miguel Studer Ferreira

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

2021

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Efficient Software-based Base Station Design for C-RAN

Mojgan Barahman

Supervisor: Doctor Luís Manuel de Jesus Sousa Correia
Co-Supervisor: Doctor Lúcio Miguel Studer Ferreira

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Doctor Isabel Maria Martins Trancoso,
Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:

Doctor Luís Filipe Lourenço Bernardo,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa;

Doctor Luís Manuel de Jesus Sousa Correia,
Instituto Superior Técnico, Universidade de Lisboa;

Doctor Sílvia Ruiz Boqué, Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels, Universitat Politècnica de Catalunya – BarcelonaTech, Espanha;

Doctor António Manuel Raminhos Cordeiro Grilo,
Instituto Superior Técnico, Universidade de Lisboa;

Doctor António José Castelo Branco Rodrigues,
Instituto Superior Técnico, Universidade de Lisboa.

2021

To my beloved parents,
my lovely husband and my little daughter

`

Acknowledgements

This dissertation would not have been possible without the support and encouragement of many people. I am eternally grateful to all of them.

First and foremost, I would like to express my gratitude to my research advisor, Prof. Luis M. Correia, for his patience, invaluable advice, constructive criticism, and for teaching me the value of being thorough. I am particularly grateful for the opportunity to participate in the COST Action CA15104, IRACON meetings, a unique opportunity to meet researchers from all around Europe. I would also like to thank my co-supervisor, Prof. Lúcio S. Ferreira, for his valuable comments in improving this dissertation and for his support during my stay in Lisbon.

Particular appreciation also goes to all current and former members of our Group for Research on Wireless (GROW). Especially, my fellow colleagues and friends Kenan Turbic and Behnam Rouzbehani for sharing their experience and knowledge and Vera Almeida and Ema Caterre, whose help made the work easier.

I am very thankful to my friends and family for motivating me and supporting me during this long journey. I am tremendously grateful to my husband for his support, encouragement, love, and advice when it was needed the most. His support is what had gotten me through when I wanted to give up. I am thankful to my daughter, who joined us when I was writing my dissertation, for giving me unlimited happiness and pleasure and showing patience during my thesis writing.

I also appreciate my parents-in-law for their love and moral support; my dear friend Zahra Mirhoseini, with whom my life experience in Lisbon was certainly much more pleasant; and my siblings whose love and support light up my life forever, no matter where they are.

Finally, my deepest gratitude goes to my parents, who encouraged me to pursue graduate school and provided plenty of inspiration along the way; their unconditional love and support means the world to me.

Abstract

This thesis addresses the problem of computing resource allocation in Cloud Radio Access Networks. A game-based optimization algorithm was developed to distribute the computing resources among BaseBand Units (BBUs) in a BBU-pool whereby resource utilization is maximized. The model allocates computing resources on-demand, based on the instantaneous BBUs requests, using a game-theory bargaining approach; in case the available resources are not enough to fulfil all requests, BBUs are prioritized to ensure the adequate Quality of Service, low-priority ones being always guaranteed a minimum computing resource. The performance of the proposed model is observed over time, concerning resource usage, BBU fulfilment level, fairness and efficiency. Simulations in a group of cells with a mixture of heterogeneous services in tidal traffic conditions show that resources allocated to BBUs are consistent with the priority of ongoing services and in line with real-time demand. Results also confirm that the proposed model manages bottlenecks effectively and shows a higher performance compared with equal and demand proportional resource allocation schemes. There is no wastage in the proposed model during congestions and it fairly distributes 100% of the resources among BBUs in these cases, by shrinking the capacity share of the lower priority BBUs to compensate for the higher priority BBUs' resource shortages. Hence, the proposed model fulfils high prioritized BBUs' demands 13% more compared to the other allocation schemes. Results also show that improving the average fulfilment level from 98% to 100% requires doubling the available resources at the cost of the average resource usage being cut in half.

Keywords

Wireless Communications, Cloud-RAN, Computing Resource Utilization, Resource Allocation Optimization, Fairness.

Resumo

Esta tese aborda o problema de atribuição de recursos de computação em Redes de Acesso rádio em Nuvem. Foi desenvolvido um algoritmo de otimização baseado em teoria de jogos para distribuir os recursos de computação entre as unidades de banda base (BBUs) num agregado de BBUs onde a utilização dos recursos é maximizada. O modelo atribui recursos de computação a pedido, baseado nos pedidos instantâneos das BBUs, usando uma abordagem de negociação baseada em teoria de jogos, quando os recursos disponíveis não são suficientes para satisfazer todos os pedidos, as BBUs são priorizadas, para assegurar a Qualidade de Serviço adequada, garantindo-se sempre um mínimo de recursos de computação para os pedidos de prioridade baixa. O desempenho do modelo proposto foi analisado em termos temporais, relativamente à utilização de recursos, nível de utilização de BBUs, justiça de atribuição e eficiência. Simulações num grupo de células com mistura heterogénea de serviços em condições de tráfego variante no tempo mostram que a atribuição de recursos às BBUs é consistente com a prioridade dos serviços em curso e com os pedidos em tempo-real. Os resultados confirmam também que o modelo proposto gere eficazmente os problemas de estrangulamento e mostram um melhor desempenho comparado com esquemas de atribuição igual ou proporcional de recursos. Não há desperdício no modelo proposto durante congestão e os recursos são distribuídos de maneira justa entre 100% das BBUs nestes casos, através de uma redução da quota da capacidade nas BBUs com prioridade baixa para compensar as faltas das BBUS com prioridade elevada. Assim, o modelo proposto preenche as condições dos pedidos com alta prioridade melhor em 13% comparado com os outros esquemas de atribuição. Os resultados mostram também que o aumento da média do nível de desempenho de 98% para 100% requer a duplicação dos recursos disponíveis, com o custo de a utilização média dos recursos descer para metade.

Palavras–chave

Comunicações sem fios. RANs em Nuvem. Utilização de Recursos de Computação. Otimização de Atribuição de Recursos. Atribuição Justa.

چکیده

این پایان نامه بر روی مسئله تخصیص منابع محاسباتی در شبکه های دسترسی رادیویی ابری تمرکز دارد. یک الگوریتم بهینه سازی مبتنی بر تئوری بازی ارائه شده است که بهره وری از منابع محاسباتی موجود در استخر واحدهای پردازش باند پایه را به حداکثر می رساند. این مدل با استفاده از رویکرد چانه زنی در تئوری بازی طراحی شده است و منابع محاسباتی را بر اساس درخواستهای آنی تخصیص می دهد. در صورتی که منابع موجود برای پاسخگویی به همه درخواستها در یک لحظه کافی نباشد، واحدهای پردازش باند پایه اولویت بندی میشوند تا کیفیت خدمات حفظ شود. در عین حال، مدل پیشنهادی همیشه یک حداقل منبع محاسباتی را برای واحدهای با اولویت پایین تر تضمین میکند تا از ازکارافتادن آنها جلوگیری کند. عملکرد مدل پیشنهادی در گذر زمان مشاهده شده است وکارآیی مدل از نظر استفاده از منابع موجود، سطح تحقق درخواستها و اعتدال نسبی در تخصیص منابع، مورد بررسی قرار گفته است. نتایج بررسی بر روی گروهی از سلولها، که در شرایط نوسان، مخلوطی از خدمات ناهمگن را ارائه میدهند، نشان میدهد که منابع اختصاص یافته با اولویت خدمات در حال انجام و همچنین با مقدار تقاضا متناسب است. علاوه بر این نتایج نشان میدهد که مدل پیشنهادی بحرانهای کمبود منابع را به خوبی مدیریت میکند و نسبت به طرح تخصیص منابع به صورت مساوی و طرح تخصیص منابع فقط مبتنی برتقاضا، عملکرد بالاتری از خود نشان میدهد. دربحرانهای کمبود منابع، در مدل پیشنهادی هدر رفت وجود ندارد و 100٪ منابع منصفانه توزیع میشوند. به نحوی که کمبود واحدهایی که اولویت بالاتری دارند، با کاهش سهم واحدهای کم اولویت تر، جبران میشود. به همین دلیل است که در شبیه سازی انجام شده، تقاضای واحدهای با اولویت بالاتر 13٪ بیشتر از سایر واحدها برآورده شده است. همچنین نتایج نشان می دهند که بهبود سطح تحقق درخواستها از 98٪ به 100٪ نیازمند دو برابر شدن منابع موجود است و این در صورتی است که استفاده از منابع به نصف کاهش می یابد.

کلمات کلیدی

ارتباطات بی سیم ، شبکه های دسترسی رادیویی ابری، بهره وری از منابع محاسباتی، بهینه سازی تخصیص منابع، اعتدال نسبی.

Table of Contents

List of Figures

List of Tables

List of Acronyms

1G	1st Generation
2G	2nd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5GC	5G Core Network
5QI	5G QoS Identifier
AICC	Allocated Computing Capacity
AMF	Access and Mobility management Function
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat reQuest
ASIC	Application-Specific Integrated Circuits
AvCC	Available Computing Capacity
BBU	Baseband Unit
BPSK	Binary Phase Shift Keying
BS	Base Station
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CoMP	Coordinated Multiple Point
CP	Common Processing
CPRI	Common Public Radio Interface
CPU	Central Processing Units
CQI	Channel Quality Indicator
C-RAN	Cloud RAN
CRC	Cyclic Redundancy Check
CSI	Channel State Information
DAS	Demand-aware computing resources Allocation Scheme
DCC	Decentralized Cloud Controller
DFT	Discrete Fourier Transform
DL	DownLink
EAS	Equal computing resources Allocation Scheme

ECOS	Embedded Conic Solver
eNB	evolved NodeB
EPC	Evolved Packet Core
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FFT	Fast Fourier Transform
FPGA	Field-Programmable Gate Array
FR	Frequency Range
GBR	Guaranteed Bit Rate
GNBS	Generalized Nash Bargaining Solution
GPP	General Purpose Processor
GPRS	General Packet Radio Service
GPU	Graphics Processing Unit
GSM	Global System for Mobile Communications
HARQ	Hybrid ARQ
HSS	Home Subscriber Service
HTTP	HyperText Transfer Protocol
ICIC	Inter-Cell Interference Coordination
IFFT	Inverse Fast Fourier Transform
IP	Internet Protocol
IQ	In-phase and Quadrature
IRACON	Inclusive RAdio COmmunication Networks for 5G and beyond
KKT	Karush-Kuhn-Tucker
L1	Layer 1
L2	Layer 2
L3	Layer 3
LTE	Long Term Evolution
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MSS–BBU	Multi–Site/ multi–Standard BaseBand Unit
MTC	Massive machine Type Communications
NAS	Non-Access Stratum
NBS	Nash Bargaining Solution
NFV	Network Functions Virtualization
NG	Next Generation
NG-RAN	Next Generation Radio Access Network
NR	New Radio

OAI	Open Air Interface
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPS	Operations Per Second
OS	Operating System
PC	Personal Computer
PCCC	Parallel Concatenated Convolutional Code
PCRF	Policy and Charging Rules Function
PDB	Packet Delay Budget
PDCP	Packet Data Convergence Protocol
PDF	Probability Density Function
PDN	Packet Data Networks
PDSCH	Physical DL Shared CHannel
PDU	Packet Data Unit
PER	Packet Error Rate
P–GW	Packet Data Network Gateway
PHY	PHYsical
PL	Propagation Loss
PMI	Precoding Matrix Indicator
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QDAS	QoS-Demand-aware computing resources Allocation Scheme
QFI	QoS flow ID
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RCC	Required Computing Capacity
RE	Resource Element
RI	Rank Indicator
RLC	Radio Link Control
RRC	Radio Resource Control
RRH	Remote Radio Head
RS	Reference Signal
RT	Real-time
RTLinux	Real-time Linux
SDAP	Service Data Adaptation Protocol
SCS	SubCarrier Spacing
SC-FDMA	Single Carrier Frequency Division Multiple Access

S-GW	Serving Gateway
SINR	Signal to Interference plus Noise Ratio
SNR	Signal to Noise Ratio
TB	Transport Block
TBS	Transport Block Size
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TM	Transmission Mode
TTI	Time Transmission Intervals
UE	User Equipment
UHD	Ultra-High Definition
UL	UpLink
UMTS	Universal Mobile Telecommunications System
UP	User-specific processing
UPF	User Plane Functions
VoIP	Voice–over–IP
VM	Virtual Machine

List of Symbols

α	Shape parameter of Pareto Distribution
γ	SNR
Δ	Relative deviation
Δf_{BW}	Operating bandwidth
$\overline{\Delta f}_{BWu}$	Average bandwidth of the RBs allocated per user
Δf_{BW}^{ref}	Reference value of channel bandwidth
Δt	Interval between two successive time instants
Δt_c	Coherence time
Δt_{CQIPMI}	CQI/PMI reporting interval
Δt_{RI}	Rank Indicator reporting interval
Δt_{SF}	Sub-frame duration
ΔT^{SIM}	Simulation duration
Δt^{TTI}	TTI duration
ϵ_{solver}	CVX solver's tolerance
η	Computing resource allocation efficiency
$\bar{\eta}$	Average of computing resource allocation efficiency
η_{sch}	Scheduler efficiency
η^{RB_U}	User's RB efficiency
$\eta_{DL\|UL}^{RB_B}$	BBU's RB efficiency in DL/UL
$\vartheta\|\boldsymbol{\xi}$	Lagrange multipliers
λ	Rate parameter of the Exponential Distribution
μ	Mean of a distribution
σ	Standard deviation of a distribution
τ_{AE}	Time period (in speech frames) between VoIP successive active state entries
τ_{SP}	VoIP silence period
τ_{TS}	VoIP talk period
υ	User's mobility speed

υ_{Avg}	Average of the users' mobility speed
a_{pd}	Average power decay
B	Bargaining power
c	Speed of light
C_{BP}	BBU-pool's existing computing capacity
$\mathbf{C}^{Al}$	Allocated computing capacity vector
$\mathbf{C}^{Al^*}$	Optimum allocated computing capacity vector
C^{Al}	BBU's allocated computing capacity
$C^{Al_{DAS}}$	BBU's allocated computing capacity in demand-aware resource allocation scheme
$C^{Al_{EAS}}$	BBU's allocated computing capacity in equal resource allocation scheme
C_{BP}^{Al}	BBU-pool's total allocated computing capacity
C_{BP}^{Av}	BBU-pool's available computing capacity
C^{R}	BBU's required computing capacity
C_{C}^{R}	BBU's required computing capacity for CPs
$C_{DL\|UL}^{R}$	BBU's required computing capacity in DL/UL
$C_{DLc\|ULc}^{R}$	BBU's required computing capacity for all the DL/UL CPs
$C_{DLu\|ULu}^{R}$	BBU's required computing capacity for all the DL/UL UPs
C_{U}^{R}	BBU's required computing capacity for UPs
$\mathbf{C}^{R_{\min}}$	Minimum guaranteed computing capacity vector
$C^{R_{\min}}$	BBU's minimum guaranteed computing capacity
$C^{R_{PEAK}}$	BBU's peak required computing capacity
$C^{R_{usr}}$	User's required computing capacity for a specific processing step
$C_{DL\|UL}^{R_{usr}}$	DL/UL user's total required computing capacity for UPs
C^{ref}	Reference value of a possessing step's required computing capacity
C^{TYP}	Cell type
E	Scaling effect
F	Fairness index
F_{A}	Voice Activity Factor
f_{C}	Carrier frequency
$f_{D,\max}$	Maximum Doppler shift
f^{B}	BBU's fulfilment level
$\overline{f^{B}}$	Average of a BBU's fulfilment level
G_{mud}	Radio capacity gain obtained due to multi-user diversity

G_{pos}	Radio capacity gain obtained due to users positioning in the cell
G_R	Gain of the receiving antenna
G_T	Gain of transmitting antenna
H^{TYP}	Type of RRH
I_{MCS}	Modulation and coding scheme index
I_{TBS}	Transport block size index
k	Scale parameter of Pareto Distribution
$\mathcal{L}$	Lagrange function
L_P	Path loss
m	Modulation order
$m_{\max}$	Maximum modulation order
M_{RI}	Integer coefficient of RI periodicity calculation
m^{ref}	Reference value of modulation order
$\mathbb{N}$	Natural numbers
$\mathcal{N}$	Normal Distribution
N_B	Number of BBUs in the BBU-pool
N_{MIMO}	MIMO order
N_P	Number of the players in the bargaining game
N_{Str}	Number of spatial streams
$N_{\Delta f}^{RB}$	Number of the available RB in a given bandwidth
N_{Al}^{RB}	Number of RBs allocated to a user
N_R^{RB}	Number of user's required RBs
N_{MIMO}^{ref}	Reference value of BS's MIMO order
N_{Str}^{ref}	Reference value of number of spatial streams
N_{CQIPMI}^{SF}	Number of sub-frames in CQI and PMI reporting interval
N^{SRV}	Number of services with different IDs
N^{TS}	Number of time slices in a time interval
N^U	Number of users in a BBU
N_{cell}^u	Number of users in a single cell
N_S^U	Number of users of a specific service in a BBU
p	Processing step
$P_{\tau_{AE}=n}$	Probability that the time period τ_{AE} between successive VoIP active states has duration n speech frames

$P_{\tau_{SP}=n}$	Probability that a silence period τ_{SP} has duration n speech frames
$P_{\tau_{TS}=n}$	Probability that a talk period τ_{TS} has duration n speech frames
P_A	Probability of being in the VoIP active state
P_{AI}	Probability of transitioning from active speech state to the inactive or silent state
P_I	Probability of being in the VoIP inactive state
P_{IA}	Probability of transitioning from inactive speech state to the active state
P_T	Power fed to the antenna
P_R	Power sensitivity at the receiver antenna
P^{chc}	Channel coding processing step
P^{chd}	Channel decoding processing step
P^{che}	Channel estimation processing step
P^{dm}	Demodulation processing step
P^{OFDMA}	OFDMA processing step
P^{md}	Modulation processing step
P^{mdc}	MIMO decoding processing step
P^{mpc}	MIMO precoding processing step
P^{SCFDMA}	SCFDMA processing step
p^{SRV}	3GPP priority level of a service
$p^{SRV}_{\max}$	Maximum value of 3GPP services priority levels
$p^{SRV}_{\min}$	Minimum value of 3GPP services priority levels
Q	Quantization resolution
Q^{ref}	Reference value of quantization resolution
$\mathbb{R}$	Real numbers
r	User's coding ratio
$\mathbf{R}_{\min}$	Minimum guaranteed resources vector
$r_{\min}$	A player's minimum guaranteed resource
$\mathbf{R}^{Al}$	Allocated resources vector
r^{Al}	Allocated resources to a player
$\mathbf{R}^{Al^*}$	Optimum allocated resources vector in the NBS solution
r^{Al^*}	Optimum allocated resources to a player
$\mathbf{R}^{AlG^*}$	Optimum allocated resources to the players in the GNBS solution
R_U^{Arr}	User arrival rate
R^{cell}	Cell radius

r^{ref}	Reference value of coding ratio
s	Service
S^{CP}	Set of CPs
$S^{CP}_{DL\|UL}$	Set of DL/UL CPs
S^{FS}	Feasible solution set
S^{Parm}	Set of effective parameters on the BBUs' required computing capacity
S^{Proc}	Set of the processing steps
s^{U}	Service ID
$S^{U_{DL\|UL}}$	BBU's active DL/UL user set
S^{UP}	Set of UPs
$S^{UP}_{DL\|UL}$	Set of DL/UL UPs
t	Time instant
$\boldsymbol{\mathcal{U}}$	Players' utilities vector in a bargaining game
$\mathcal{U}$	Player's utility function in a bargaining game
U	Resource usage
$\overline{U}$	Average of resource usage
$\mathcal{U}_{BP}$	BBU-pool's utility function
$\mathcal{U}_{LBP}$	Logarithmic transformed form of the BBU-pool's utility function
V^{Pkt}	Packet volume
w^{srv}	service weight
$\overline{w^{srv}}$	Average of service weights in a BBU
X^{aprox}	Approximated values of parameter X
X^{ref}	Reference value of parameter X

List of Software

CVX	CVX turns MATLAB into a modelling language used for solving convex optimization problems.
MATLAB	MATLAB is used for simulation, optimization, and plotting figures.
MS Word 2016	MS Word 2016 is used to edit this thesis and all associated document such as publications.

Chapter 1

Introduction

This chapter provides an overview of the work developed in this dissertation. Section 1.1 presents a brief history of the evolution of mobile communications networks over the past two decades. Section 1.2 describes the motivation and objectives of the research. Section 1.3 addresses the work's novelty, and Section 1.4 provides an overview of the research strategy and lists the publications originated from parts of this work. Finally, Section 1.5 explains the dissertation structure.

1.1 Brief History

Over the past decades, wireless and mobile communications networks have experienced a significant increase in users and data consumption, exploiting diverse applications and services. The growth of communication requests lead to new technologies to fulfil subscriber's requirements. Figure 1.1 briefly presents this evolution; a brief review being presented in what follows.

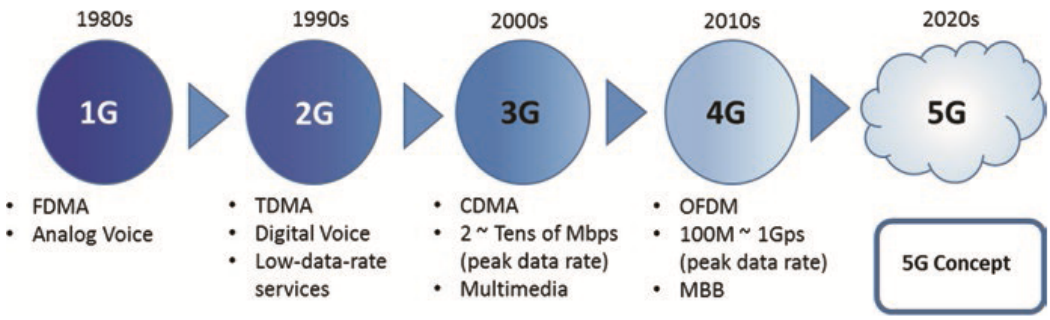


Figure 1.1 – Evolution of cellular standards (based on [LSJY20]).

The first Generation (1G) of mobile communications networks was introduced in the 1980s. It was based on analogue cellular technology and supported voice calls only. By employing Time Division Multiple Access (TDMA) in the 1990s, the second Generation (2G), namely the Global System for Mobile communications (GSM), could support text messaging. Later, the General Packet Radio Service (GPRS) introduced the packet switched domain to network architectures.

Coding Division Multiple Access (CDMA) was employed in the mid-1990 to 2000s to develop the third generation (3G), which continued digital processing. However, the exploitation of more frequency bandwidth and higher symbol rates lead to a much higher peak data rate than 2G. Universal Mobile Telecommunications System (UMTS) started as the joint European and Japanese system for 3G and was standardized by the 3G Partnership Project (3GPP). In UMTS, circuit switching remained besides packet switching.

The fourth Generation (4G) was driven by an increasing demand for capacity, lower-cost data delivery and competition from other technologies. 3GPP developed the Long-Term Evolution (LTE) that exploits Orthogonal Frequency Division Multiple Access (OFDMA) for DownLink (DL) and Single-Carrier FDMA (SC-FDMA) in UpLink (UL). The LTE standard offers more bandwidth and services as well as a significant improvement in capacity. Its design was based on packet switching aiming to supply a high-quality audio/video streaming over end-to-end Internet Protocol (IP) and offers higher peak user throughput to a minimum of 300 Mbps in DL and 75 Mbps in UL.

The fifth Generation (5G) is the next primary mobile telecommunications standard beyond 4G. A data rate greater than 1 Gbps/user is expected in 5G, exploiting reduced cell size, distributed antennas and massive Multiple Input Multiple Output (MIMO) beamforming. Moreover, bandwidth, security and Quality of Service (QoS) are expected to increase while decreasing service costs and delays. The introduction of 5G allows cellular and wireless networks to match data rates and use cases, allowing a higher density

of mobile broadband users and supporting device-to-device, ultra-reliable, and massive machine communications [NGMN15]. Technologies such as virtualization, especially with Cloud Radio Access Network (C-RAN), also play a key role in the evolution of 5G. C-RAN was first introduced by China Mobile Research Institute in 2010: the cloud computing-based architecture for Radio Access Networks (RANs) that utilizes open platforms and real-time virtualization technology for multiplexing and dynamically sharing the Base Stations (BSs)' resources in a datacenter, leading to higher data rates and lower network latencies.

1.2 Thesis Motivation and Objectives

The proliferation of high data rate applications in conjunction with high mobile terminals usage nowadays has triggered a drastic increase in data rate demands [Cerw20], Figure 1.2. Therefore, wireless networks providers must continuously improve their infrastructure to serve data demand accordingly. This presents a difficult challenge, since resource allocation in conventional RANs is inefficient, since it is based on peak-hour traffic requirements. However, since users' demand is time-variant, traffic is not always at the peak level and may be up to 10 times lower in off-peak hours [CCYS14]; thus, a fixed allocation scheme leaves idle resources at various times/areas.

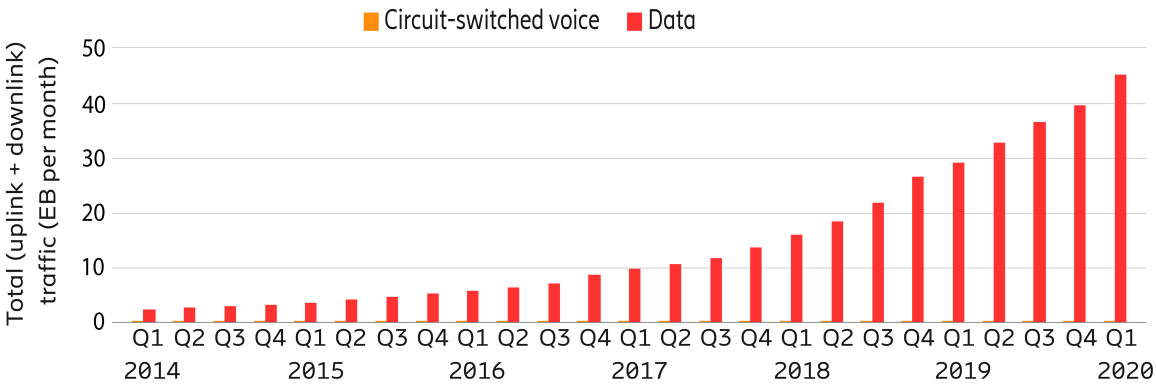


Figure 1.2 – Global total traffic in mobile networks, 2014-2020 (extracted from [Cerw20]).

C-RAN has emerged as a centralized paradigm to provide a solution for higher data rates and capacity demands in a cost-efficient way: Baseband Processing Units (BBUs) of BSs are decoupled from the radio units, known as Remote Radio Heads (RRHs); software-based BBUs are then centralized and consolidated in servers of a data center, known as BBU-pools. C-RAN is a critical enabling technology of 5G [CCYS14], providing higher data rates and lower network latencies by multiplexing the BBU-pool's resources. Resource multiplexing enables over-loaded BBUs to use residual resources left by the underutilized ones; hence, utilization is improved, and fewer resources are required than the sum of stand-alone BBU demands [ARET19], [LZGN16].

Although the consolidation of resources in C-RAN reduces the number of the required resources in the

network, there are still critical challenges for data centers, such as power consumption [DaWF16], [HMDM19] and [WeGP13]: a medium-sized one with 930 m^2 and 288 racks can consume 4 MW in the traffic peak [PMZW10]. Since computing resources, i.e., servers, are the most energy-intensive entities in data centers, it is worthwhile to apply efficient resource management strategies to maximize their utilization and reduce the number of idle ones. An idle server consumes 60% of its peak power usage, although it has no productivity [PMZW10].

However, designing efficient resource management strategies is a complicated process for cloud providers. Due to the variety of network services, user arrival rates and channel conditions, BBU resource demands fluctuates significantly throughout the day. On the one hand, a BBU computing capacity should suffice peak demands; on the other hand, provisioning fixed resources based on peak requirements leads to idle resources for the rest of the day. As a result, an efficient resource management strategy in a BBU-pool should allocate the computing capacity dynamically, following the BBUs' instantaneous demand, while efficiently handling the resources in case of a shortage. Resource shortages are time instants in which the BBU-pool's available resources are less than demand spikes and come into play in two circumstances: when the objective is intentionally to design the pool with minimum computing resources; or, even if there are more computing resources, they cannot be initialized at a rate similar to the one of demand fluctuations (in the scale of milliseconds), due to hardware limitations.

1.3 Novelty and the Main Contributions

In this thesis, a BBU-pool computing resource allocation scheme is proposed within a dynamic traffic demand environment. The proposed model estimates the BBUs' demands and reconfigures BBUs' Allocated Computing Capacity (AICC) accordingly. The main objective is to maximize the utilization of BBU-pool computing resources, which is crucial to guarantee low power consumption in the network.

The novelty of the proposed scheme is the consideration of the limits of the BBU-pool computing resources and the prioritization of BBUs in bottlenecks based on the characteristics of their ongoing services and QoS constraints. Simultaneously, the model guarantees all BBUs with a minimum computing resources to avoid crashing; furthermore, contrary to existing works, the proposed model has a low complexity and provides fairness of resource allocation and system efficiency, which makes it applicable in practical implementations. Considering both QoS and BBU Required Computing Capacities (RCCs) as real-time parameters, i.e., given on the basis of Time Transmission Intervals (TTIs), is essential not only in 4G deployments but also for the upcoming service-oriented 5G and ensures that the BBU-pool is provisioned with an optimum configuration, consistent with BBU demands.

In order to evaluate the model performance, an attempt has been made to emulate a typical day of operation in cellular networks over which the performance of the proposed model is compared against equal and demand-proportional resource allocation schemes, which can be found in the literature as

common allocation approaches. Moreover, the model performance is evaluated in terms of resource usage efficiency and BBU fulfilment level, considering real-time network traffic in a tidal channel condition. Studies highlight how the limit of the Available Computing Capacity (AvCC) is correlated with the BBU demands' fulfillment level. In general terms, more resource availability translates to better fulfilment levels and lower levels of resource usage. However, above certain levels, the provisioning of more resources degrades the average resource usage dramatically while contributing very little (or nothing) to improving the demands' fulfilment level.

1.4 Research Strategy and Impact

The aim of providing an efficient resource allocation strategy in a BBU-pool is to maximize resource utilization. To achieve this goal, resources should be allocated to BBUs based on their real-time demand such that QoS is maintained. Hence, the first step is traffic demand evaluation, and the optimal solution for resource utilization can be found only afterward. In this way, the proposed resource management algorithm comprises two main steps:

1. **RCC estimation**: calculation of instantaneous demand (measured in Operations Per Second [OPS]) of BBUs, according to the real-time network/user parameters.
2. **Computing resource allocation optimization**: obtaining the optimal on-demand computing resource allocation that maximizes both BBU-pool resource utilization and efficiency with respect to the required QoS.

The BBUs' RCC estimation is based on a well-defined model, [MAMM16] and [DeDL15], for the given network and user parameters at a specific time. The results are then fed into the computing resource allocation step in order to find the optimal AICC to BBUs. To this end, computing resource allocation in a BBU-pool is modeled as a game-theory bargaining game. Players, i.e., BBUs, compete for the limited computing resources of the BBU-pool to maximize their processing speed; the Generalized Nash Bargaining Solution (GNBS) with adaptive bargaining powers [Myer91] is applied to find a solution for the bargaining game. The two-fold solution maximizes both the BBU-pool computing resource utilization and the processing speed of the BBUs. In the proposed model, QoS constraints are considered. Additionally, service characteristics are monitored in real-time, which is essential not only in 4G deployments but also for the upcoming service-oriented 5G.

This work was developed within the framework of the COST Action CA15104, Inclusive RAdio COmmunication Networks for 5g and beyond (IRACON) [COST20]. Participating in this project and regularly attending its meetings provided an opportunity to interact with researchers working on similar topics in Europe and the world, with whom the work was discussed and received valuable feedback.

Excerpts from the work presented in this dissertation have already been published in several articles in international journals and conferences and internal reports prepared and presented within IRACON:

- Book Chapters
    - M. Barahman and L.M. Correia, “Cloud Radio Access Network (C-RAN),” in C. Oestges, (Ed.), *Inclusive Radio Communication Networks for 5G and Beyond*, Academic Press, Cambridge, MA, USA, 2020 (in final edition).
- Journals
    - M. Barahman, L.M. Correia and L.S. Ferreira, "A QoS-Demand-Aware Computing Resource Management Scheme in Cloud-RAN," *IEEE Open Journal of the Communications Society*, vol. 1, page 1850-1863, Oct. 2020.
- Conferences
    - M. Barahman, L.M. Correia and L.S. Ferreira, "A Real-time QoS-Demand-Aware Computational Resource Sharing Approach in C-RAN," in *Proc. of EuCNC 2020 - European Conference on Networks and Communications*, Dubrovnik, Croatia, June 2020.
    - M. Barahman, L.M. Correia and L.S. Ferreira, "An Efficient QoS-Aware Computational Resource Allocation Scheme in C-RAN," in *Proc. of WCNC 2020 - IEEE Wireless Communications and Networking Conference*, Seoul, South Korea, Apr. 2020.
    - M. Barahman, LM. Correia and L.S. Ferreira, "A Fair Computational Resource Management Strategy in C-RAN," in *Proc. of CoBCom 2018 - International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications*, Graz, Austria, July 2018.
- Internal Reports
    - M. Barahman, L.M. Correia and L.S. Ferreira, “A Real-time QoS-Demand-Aware Computational Resource Allocation Scheme in C-RAN,” TD (20)13045. Online meeting: COST Action CA15104 (IRACON) meeting, Sep. 2020.
    - M. Barahman, L.M. Correia and L.S. Ferreira, “A Real-time QoS-Demand-Aware Computational Resource Sharing Approach in C-RAN,” TD (20)12027. Louvain-la-Neuve, Belgium: COST Action CA15104 (IRACON) meeting, Jan. 2020.
    - M. Barahman, L.M. Correia and L.S. Ferreira, “A Real-time Computational Resource Usage Optimization in C-RAN,” TD (19)11018. Gdansk, Poland: COST Action CA15104 (IRACON) meeting, Sep. 2019.
    - M. Barahman, L.M. Correia and L.S. Ferreira, “Optimizing Computational Resources Usage in C-RAN,” TD (19)10037. Oulu, Finland: COST Action CA15104 (IRACON) meeting, May 2019.
    - M. Barahman, L.M. Correia and L.S. Ferreira, “A Real-time Computational Resource Management in C-RAN,” TD (18)07064. Cartagena, Spain: COST Action CA15104 (IRACON) meeting, June 2018.
    - M. Barahman, L.M. Correia and L.S. Ferreira, “A Fair Computational Resource Management Strategy in C-RAN,” TD (18) 06040. Nicosia, Cyprus: COST Action CA15104 (IRACON) meeting, Jan. 2018.

1.5 Structure of the Dissertation

This thesis is structured in seven chapters and six annexes. The rest of the document being organized as follows.

Chapter 2 gives an overview of 4G and 5G mobile communications networks within the scope of the thesis. Section 2.1 and Section 2.2 present an overview of the network architectures and radio interfaces for 4G and 5G, respectively. In Section 2.3, QoS is described and Section 2.4 presents a summary of coverage and radio capacity concepts. Section 2.5 is dedicated to the C-RAN concept and virtualization; an overview of virtualization is presented, focusing on BBU-pool virtualization and related approaches. Section 2.6 explains how the concept of game theory is used to solve a resource allocation problem. Finally, Section 2.7 mentions the state of the art related to computing resource management in C-RAN.

Chapter 3 presents the novel model and algorithm for efficient computing resource management in a BBU-pool. Section 3.1 gives a brief description of the chosen C-RAN architecture, strategies used for BBU-pool virtualization and discusses the main network assumptions. Section 3.2 presents an overview of the proposed resource management model. Section 3.3 summarizes the approach for estimating the amount of computing resources that each BBU requires at a given time, and Section 3.4 describes the proposed optimization model for assigning the computing resources across the BBUs in a BBU-pool. Equal and demand proportional resource allocation models are also described in Section 3.5 as two reference resource allocation schemes being compared with the proposed one. The rest of the chapter is dedicated to the definition of evaluation metrics, model implementation, canonical scenario and the simulator/model assessment.

The proposed computing resource management model in Chapter 3 is limited to a single time, the model being improved in Chapter 4 by addressing time-varied traffic and demand in a tidal channel condition. Therefore, Chapter 4 presents an extension to the proposed resource management model and defines a real-time computing resource allocation framework. Section 4.1 gives an overview of the proposed model. Section 4.2 explains a strategy to find a proper time interval between two successive resource allocations. Section 4.3 mentions the metrics that are used in order to evaluate the proposed model in a real-time framework, and the rest of the chapter is dedicated to the simulator implementation, canonical scenario, and simulator assessment.

Chapter 5 analyses the proposed computing resource allocation model's performance in terms of the BBU fulfilment level, resource allocation efficiency, fairness and resource usage. To this end, a reference scenario is characterized first in Section 5.1; BBUs' real-time demands are estimated, and optimal resource allocations are achieved. Accordingly, the evaluation metrics are assessed for one snapshot of the network and for time-varied traffic in Section 5.2 and Section 5.3, respectively.

Chapter 6 compares the performance of the proposed resource allocation model with other resource allocation schemes. Moreover, the effect of the model's input parameters variation on its performance is analyzed. Section 6.1 presents an overview of the chapter. The comparison of the model's performance with equal and demand proportional resource allocations schemes is presented in

Section 6.2. Section 6.3 and Section 6.4 analyze the effect of BBU-pool available computing capacity variation and user arrival rate variations on the model's performance, respectively.

Finally, Chapter 7 concludes this dissertation by summarizing and recalling the presented work's framework and novelty in Section 7.1, the main results in Section 7.2, the key contributions in Section 7.3, and the potential improvements and the directions for future works in Section 7.4.

This dissertation also includes 6 annexes. Annex A presents the convexity proofs. Annex B presents a BBU's RCC variation relative to the variation of the effective parameters. Annex C lists the maximum Doppler shifts associated with user speed in some FDD operating Bands. In Annex D, the services' traffic profiles are described. And finally, Annex E and Annex F include the simulator's assessment results.

Chapter 2

Basic Concepts and State of the Art

This chapter provides a background and fundamental concepts of 4G and 5G networks, C-RAN and virtualization, as they are key topics for the work, as well as radio interfaces, in Section 2.1 and Section 2.2, respectively. Quality of Service, coverage and radio capacity are addressed next in Section 2.3 and Section 2.4. Then, C-RAN architecture and the framework for virtualization are discussed briefly in Section 2.5. Section 2.6 explains how the concept of game theory is used in order to solve a resource allocation problem. The last part of this chapter, Section 2.7, is dedicated to analyzing the state of the art.

2.1 LTE Basic Concepts

In this section, an overview of LTE's network architecture is given, based on [DaPS11], [Ahma13] and [HoTo11].

2.1.1 Network Architecture

LTE's network architecture has evolved from GSM and UMTS. LTE discontinued the circuit switched domain's support, operators being required to transfer their circuit switched services (e.g., voice) to the packet switched domain. The aim is to provide a seamless IP connectivity between User Equipment (UE) and the Packet Data Network (PDN).

LTE uses the concept of radio bearer to transfer data through the network. Each bearer is a flow of IP packets associated with specific QoS parameters related to application requirements. Each user may need several bearers based on the diversity of QoS requirements while using multiple applications, e.g., Voice–over–IP (VoIP) and file transfer, to connect to different PDNs. Logically, LTE network protocols are classified into control plane and user plane: the control plane is responsible for managing the radio access bearers, besides the connection between UE and network, while the user plane is responsible for transporting user traffic. Figure 2.1 illustrates LTE's network architecture.

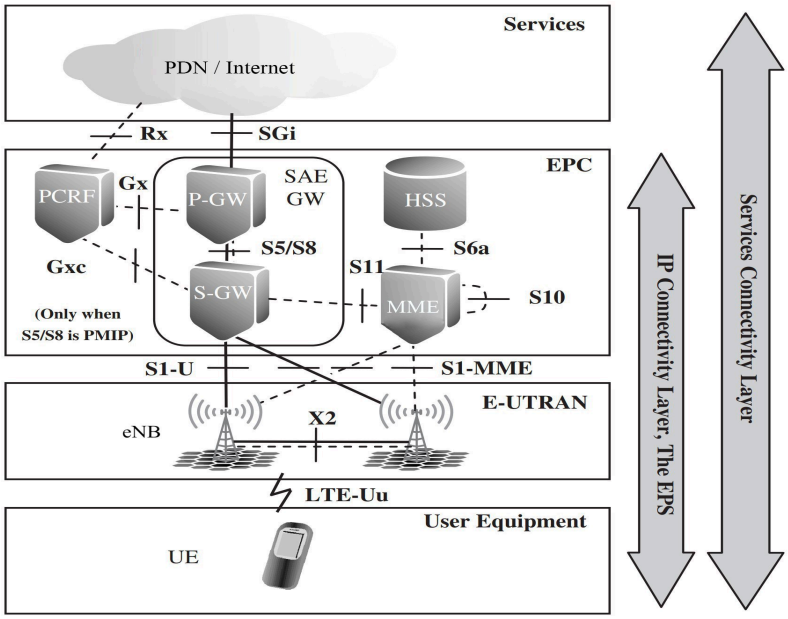


Figure 2.1 – LTE's network architecture (based on [HoTo11]).

The general components of LTE's network are:

- **UE** is the user's device to communicate with the radio network.
- **Evolved UMTS Terrestrial RAN (E–UTRAN),** consists of evolved NodeBs (eNBs). It handles radio communications between the Evolved Packet Core (EPC) and the UE. The eNBs are interconnected with each other using the X2 interface. Other connection interfaces between eNB and network

elements are illustrated in Figure 2.1. The functionalities of eNB are discussed in Section 2.1.3.

- **EPC** is composed of the following elements:
  - **Mobility Management Entity (MME)** is involved in the control plane. Bearer activation and deactivation to the terminals is one of its responsibility. It is also responsible for Non-Access Stratum (NAS) signaling and security, the functionality operating between EPC and terminal.
  - **Serving Gateway (S-GW)** acts as a transporter of user plane IP packets. SGW plays multiple roles (e.g., charging and accounting, information gathering and handovers between eNBs). It is also responsible for interworking with other 3GPP technologies (e.g., GSM).
  - **Home Subscriber Service (HSS)** is the database containing user-related and subscriber-related information. It also supports mobility management functions, call and session setup, user authentication and access authorization.
  - **PDN Gateway (P-GW)** is the IP anchor point acting as the interface between the LTE network and the external IP networks. Its responsibility includes IP address allocating and QoS enforcement for terminals. It is also a mobility anchor for non–3GPP radio access technologies connected to EPC.
  - **Policy and Charging Rules Function (PCRF),** QoS handling and charging is under its responsibility.

The functionalities of LTE elements are divided into three layers, each containing multiple sublayers: Layer 1 (L1) is the PHYsical (PHY) layer; Layer 2 (L2) contains Medium Access Control (MAC), Radio Link Control (RLC), and Packet Data Convergence Protocol (PDCP) sublayers; likewise, Layer 3 (L3) is split into Radio Resource Control (RRC) and NAS. The functionality of each of the mentioned layers are explained in detail in Section 2.1.3.

2.1.2 LTE Radio Interface

LTE operates in both Time Division Duplex (TDD) and Frequency Division Duplex (FDD) modes. Currently, 22 paired band and nine unpaired bands have been defined for LTE. Some of the frequency bands are shown in Table 2.1.

LTE uses OFDMA for DL, which is based on OFDM wherein all the bandwidth is divided into subcarriers orthogonal to each other. For an efficient operation of spectrum and a provision of isolation between subcarriers, OFDMA inserts cyclic prefix between subcarriers. Depending on the transmission scenario, the length of cyclic prefix can be normal or expanded, leading to different subcarriers bandwidths, i.e., 7.5 kHz or 15 kHz. As for UL, LTE uses SC-FDMA, which uses orthogonal subcarriers similarly, but while in OFDMA there is a one-to-one mapping between data symbols and subcarriers, SC–FDMA allows a data symbol to be transmitted in parts over multiple subcarriers.

The radio frames, used for signal transmission in LTE, last for 10 ms. Frames are divided into 10 smaller sub-frames. Each sub-frame consists of two 0.5 ms slots, where one slot is further divided into 6 or 7 OFDM symbols in the time domain. The smallest chunk of data transmitted by the LTE eNB is called Resource Block (RB), which consists of all OFDM symbols in a slot in the time domain and 12 or 24 subcarriers, depending on each subcarrier bandwidth being 15 kHz or 7.5 kHz, leading to a 180 kHz

bandwidth in the frequency domain. The smallest amount of data identified in the LTE PHY layer is called a Resource Element (RE) made of one OFDM symbol in the time domain and one subcarrier in the frequency domain. A sample of an RB and RE are shown in Figure 2.2.

Table 2.1 – LTE frequency bands (extracted from [HoTo11]).

Frequency bands for paired bands		
Operating Band	UL [MHz]	DL [MHz]
Band 1	1920-1980	2110-2170
Band 2	1850-1910	1930-1990
Band 3	1710-1785	1805-1880
Band 4	1710-1755	2110-2155
Band 5	824-849	869-894
Band 6	830-840	875-885
Band 7	2500-2570	2620-2690
Band 8	880-915	925-960
Band 9	1750-1785	1845-1880
Band 10	1710-1770	2110-2170
Band 11	1427.9-1452.9	1475.9-1500.9
Band 20	832-862	791-821
Band 21	1447.9-1462.9	1495.9-1510.9
Band 22	3410-3490	3510-3590
Band 23	2000-2020	2180-2200
Band 24	1626.5-1660.5	1525-1559

Frequency bands for unpaired bands	
Operating Band	UL and DL [MHz]
Band 33	1900-1920
Band 34	2010-2025
Band 38	2570-2620
Band 39	1880-1920
Band 40	2300-2400
Band 41	2496-2690

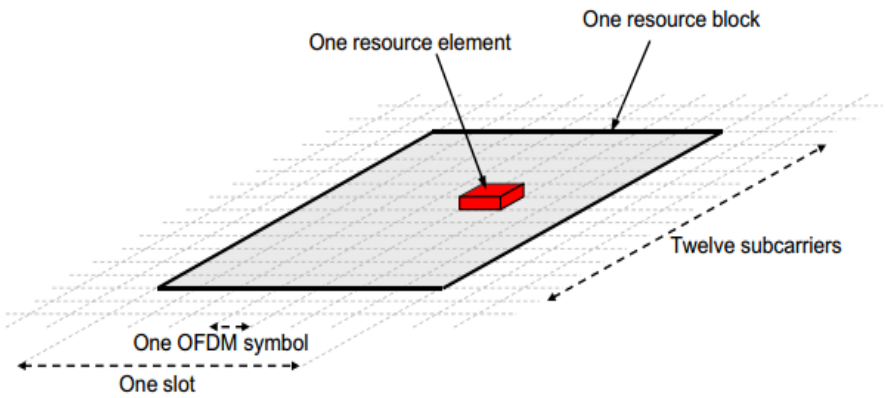


Figure 2.2 – RB in time and frequency domain (extracted from [DaPS11]).

To deliver information faithfully within the network, LTE uses a structure of channels. Data are classified for efficient delivery in the LTE protocol layers, RLC performing this logical categorization in the process named classification or concatenation. For actual transmission, however, data is mapped onto physical channels, each one corresponding to a set of REs. User data in DL is transported in Physical DL Shared Channel (PDSCH), the transmission of data in the PDSCH being made in units known as Transport Blocks (TBs).

LTE was designed to operate in a diverse range of bandwidths from 1.4 MHz to 20 MHz, the number of available RBs in the network depending on the bandwidth. The correspondence between bandwidth and the respective number of RBs is presented in Table 2.2. The transmission bandwidth can be further enlarged by using carrier aggregation, where several Component Carriers are aggregated in order to send/receive data to/from a single user. LTE can aggregate up to 5 component carriers leading to a 100 MHz transmission bandwidth. It should be noticed that in the case of carrier aggregation, the physical layer process applies separately to each component carrier.

Table 2.2 – Number of RBs associated with each LTE channel bandwidth (extracted from [Tols15]).

Channel bandwidth [MHz]	1.4	3.0	5.0	10	15	20
Number of RBs	6	15	25	50	75	100
Peak throughput – DL [Mbps] (4×4)	17.5	44.3	73.4	150.8	220.3	299.6
Peak throughput – UL [Mbps] (1×2)	4.4	11.7	18.3	36.7	55.6	75.4

The number of bits carried by a single RE depends on the Modulation and Coding Scheme (MCS), LTE supporting three modulation schemes: Quadrature Phase Shift Keying (QPSK), 16 Quadrature Amplitude Modulation (16QAM) and 64QAM, corresponding to 2, 4 and 6 bits per modulation symbol. According to the user's Signal to Interference plus Noise Ratio (SINR), an MCS index is assigned to the user that maximizes throughput. The eNB selects an MCS index during data transmission. Measuring SINR, the UE estimates the link quality before transmission and recommends the highest MCS that the UE can decode with a block error rate less than 10% [Ahma13]; the recommended MCS is reported to the eNB as a Channel Quality Indicator (CQI) known for both UE and eNB, and the eNB selects an appropriate MCS index. Figure 2.3 shows the relationship between SINR and data rate for some modulation schemes.

Channel coding is used to enhance communications efficiency and robustness. The encoder on the transmitter side adds redundancy to the data in the form of parity bits, these redundant bits being used on the receiver side to correct a number of channel errors. An encoder receives k bits as an input at a time and produces a codeword of n bits in which the ratio of k/n is the coding ratio, which describes the amount of redundant information used for protecting data; a higher coding ratio decreases channel data error, but the bandwidth efficiency will be decreased.

In early releases, LTE was required to support a peak data rate of 100 Mbps in DL and 50 Mbps in UL, but later 3GPP Release 10 enhanced the capabilities of LTE to a 100 MHz bandwidth (by using five component carriers) in LTE–Advanced, in addition to enhanced MIMO configuration led to a peak data rate of 3 Gbps in DL and 1.5 Gbps in UL. With a higher UL/DL speed, LTE supports various Transmission Modes (TMs) from multiple transmitting antennas, i.e., MIMO. In spatial multiplexing TM, multiple independent and separately encoded data signals called streams of data are transmitted from multiple transmit antennas. The number of streams with unique data indicates the transmitting order. The UE determines the transmitting order compatible with the channel condition. On the other TM, i.e., transmit diversity, the same data is transmitted across the different antennas, at the same time and frequency wherein the transmitting order is equal to one.

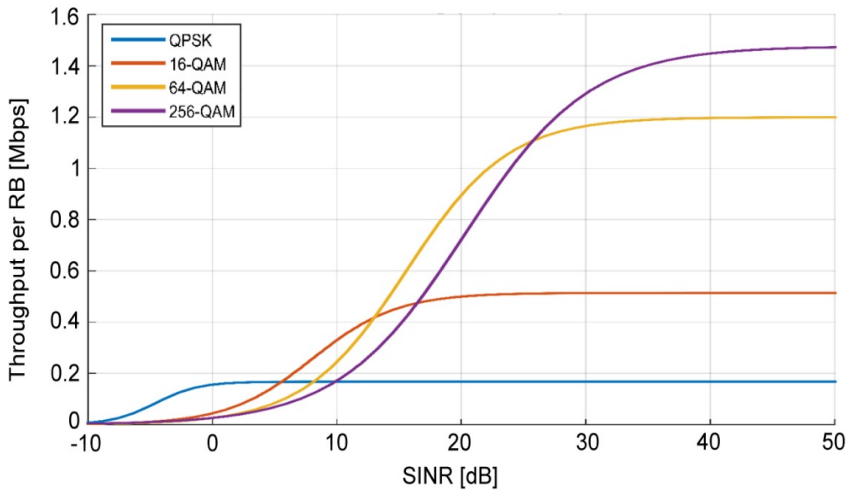


Figure 2.3 – DL throughput, according to SINR (extracted from [Viei18]).

The adjustment of the type of multi-antenna transmission scheme is based on the radio environment. MIMO with a higher order can be used in good channel conditions, i.e., high SINR and low correlation in the antennas. In low SINR, another type of multi-antenna technique should be used, e.g., transmit diversity. To adjust the TM between UE and eNB, the UE requires the eNB system information, i.e., capability and engineering of the eNB cell [3GPP20a]. System information includes most of the essential and frequently used transmission parameters acquired to establish a connection, e.g., MIMO order. The UE applies the system information acquisition procedure upon selecting and upon reselecting a cell, after handover completion, after entering E-UTRAN from another Radio Access Technology (RAT), upon return from out of coverage and upon receiving a notification that the system information has changed. After information acquisition, the UE feedbacks the Channel State Information (CSI) based on the channel condition and reports its preferred TM accordingly. Eventually, the BS designs a TM according to the received feedback and notifies the UE by RRC messages. The CSI report is a composition of one or several pieces of information [Ahma13]:

- **Rank Indicator (RI)**: it is applicable for spatial multiplexing modes, indicating the UE preferred MIMO order under the current channel condition.
- **Precoding Matrix Indicator (PMI)**: it indicates the UE preferred precoding matrix required during the precoding process.
- **CQI**: it represents the highest MCS that, if used, would mean the user plane data transmission using the recommended RI and PMI would be received with a block-error probability of at most 10%.

Reporting can be configured to be aperiodic, being transmitted upon request by the network, or periodic, being delivered with a certain periodicity. In each case, the UE's time and frequency radio resources to report CSI are controlled by the eNB and configured in the higher layer.

Moreover, LTE supports mobility across the cellular network and is optimized for low mobile speed from 0 to 15 km/h; higher mobile speeds between 15 and 120 km/h are also supported with high performance. Mobility across the cellular network is maintained at speeds from 120 to 350 km/h (or even up to 500 km/h depending on the frequency band) [3GPP09]. To facilitate the estimation of coherence time and coherence bandwidth related to the user mobility in the cell and multipath communication channel,

LTE exploits Reference Signals (RSs). The channel frequency response is estimated at the RS locations over the time-frequency grid. Using interpolation techniques makes it possible to estimate the channel at other time-frequency locations. RSs locations in time and frequency should be in such a way to ensure sufficient channel estimation accuracy. The required spacing between RSs in the frequency-domain (coherence frequency) and time-domain (coherence time) depends on the channel's maximum delay spread and maximum Doppler shift, respectively.

2.1.3 Overview of LTE Base Station's Processing

As mentioned before, the LTE elements' functionalities are divided into two planes (user and control planes) and three layers, each containing multiple sublayers (L1 is the PHY layer, L2 contains MAC, RLC and PDCP sublayers, and L3 is split into RRC and NAS sublayers). Figure 2.4 shows the control plane protocol stack.

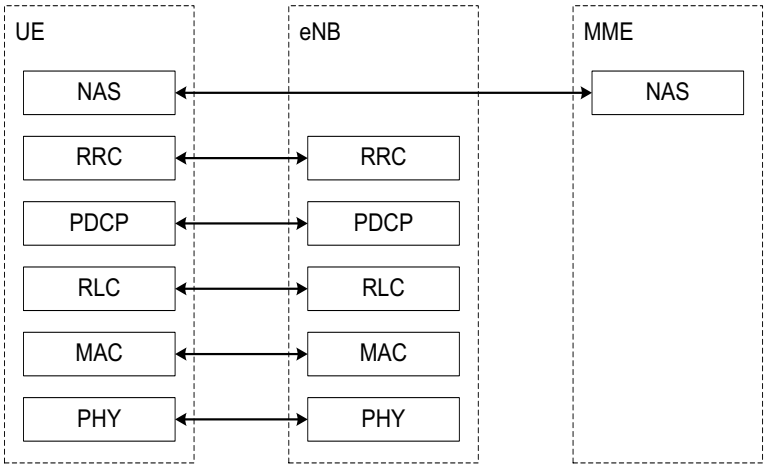


Figure 2.4 – Control plane protocol stack (extracted from [3GPP20b]).

In the following, the functionalities of each layer are explained.

- **NAS** (L3): control protocol, whose primary services and functions include EPS bearer management, Paging origination, authentication and security control.
- **RRC** (L3): it performs functions of Broadcast, Paging, RRC connection management, RB control, mobility functions and UE measurement reporting.
- **PDCP** (L2): it controls RRC messages originating from the control plane and IP packets originating from the user side; for the user plane, PDCP offers ciphering, header compression, reordering and retransmission during handover.
- **RLC** (L2): it comprises Automatic Repeat reQuest (ARQ) functionality and supports data segmentation and concatenation to minimize the protocol overheads independent of the data rate.
- **MAC** (L2): it multiplexes data from different services (radio bearers) onto a MAC Packet Data Unit (PDU), a TB. The MAC layer maintains the negotiated QoS for each radio bearer by instructing the sublayer above, the RLC, about the amount of data transmitted from each radio bearer. Another critical task for the MAC sublayer is scheduling. The scheduler in the eNB controls the assignment

of UL and DL radio resources.

- **PHY** (L1): it implements the functions required for transmitting information across the physical channel. As Figure 2.5 shows, the data arrives at the PHY layer in the form of TBs in each TTI. The functionality of the PHY layer details in the DL side is explained in what follows:
  - **Cyclic Redundancy Check (CRC)** (PHY): error detection is provided on each TB through a CRC. CRC is appended to the TBs received from the MAC layer before being passed to the next step. In the CRC method, a certain number of check bits, often called a checksum, are appended to the message being transmitted. The receiver can determine whether the check bits agree with the data. Furthermore, if the number of bits is more than 6 144 in a block, it is broken into smaller ones. CRC is directly connected to the error correction methods.
  - **Channel coding and rate-matching** (PHY): to enhance wireless communications' efficiency and robustness, channel coding is used. Encoder on the transmitter side adds redundancy to the data in the form of parity bits, which are used to correct a number of channel errors on the receiver side. The channel coding scheme applied to the user plane data is turbo coding, which is a Parallel Concatenated Convolutional Code (PCCC) with two 8-state constituent encoders and one turbo code internal interleaver [3GPP20c]. The coding rate of the turbo encoder is 1/3. If the size of TBs appended with a CRC is larger than the maximum coding block size supported by the turbo coder, the blocks are segmented into smaller code blocks. Channel coding and rate matching are later performed, and the codeblocks are concatenated to create codewords. Turbo coded blocks are individually rate matched. The rate matching block creates an output bitstream with a desired code rate. The resulting rate matched blocks are concatenated to create a single codeword for transmission. The rate matching algorithm is capable of producing any arbitrary rate.
  - **Scrambling** (PHY)**:** by scrambling, the eNB can separate signals coming simultaneously from many different UEs and the UE can separate signals coming simultaneously from many different eNB. Scrambling produces a block of scrambled bits from the input bits. The bits are scrambled with a different scrambling sequence for each codeword coming from the channel coding process.
  - **Baseband Modulation** (PHY): maps the bit values of input to complex modulation symbols with a specified modulation scheme.
  - **Layer mapping and precoding** (PHY): in these steps, input symbols from the modulation phase are mapped to symbols transmitted over multiple transmit antennas. Layer mapping splits data into layers. The number of layers can be up to the MIMO order. There are different layer mapping methods specific to each MIMO mode. After mapping, the layers are precoded, exploiting a precoding matrix. Precoding types depend on radio channel characteristics and MIMO mode. In this stage, the layer matrix is multiplied by a precoding matrix that creates the antenna port subcarrier value for each modulated symbol to be mapped. Results are sent to the next stage to perform resource mapping.
  - **Resource mapping** (PHY): in this phase, the blocks of modulated symbols are mapped onto subcarriers in OFDMA symbols REs. Resource mapping is performed separately for each

antenna port used for transmission. Information transferred over the antenna depends on the selected MIMO mode, e.g., in transmit diversity the same information is transmitted by several antennas.

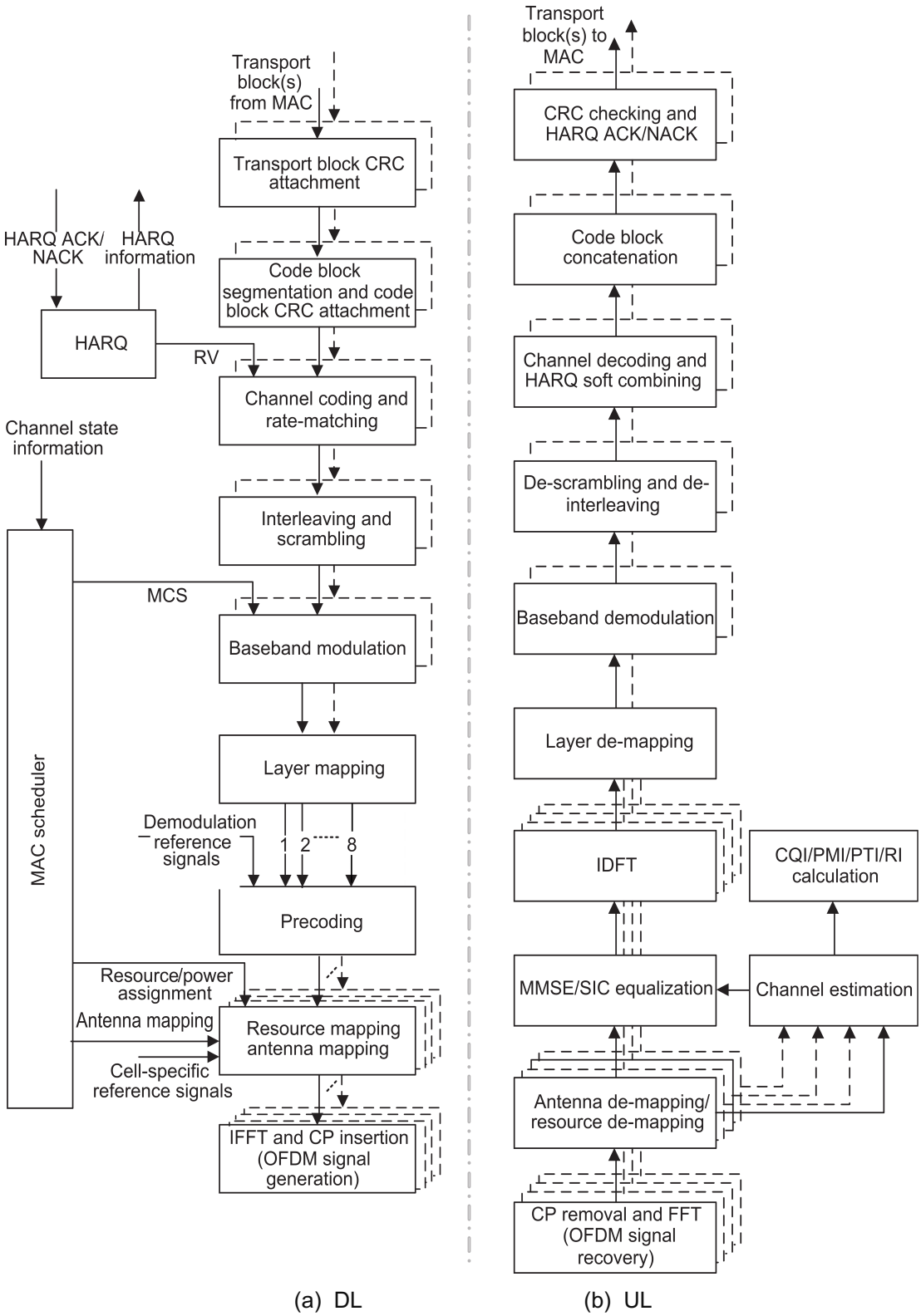


Figure 2.5 – Overview of PHY layer processing in BS (extracted from [Ahma13]).

- **OFDM** (PHY): after assigning modulated symbols to all subcarriers for an antenna port, the symbols are sent to the OFDM modulator. Exploiting Inverse Fast Fourier Transform (IFFT) converts the symbols into the time domain. Then cyclic prefix is inserted, and data is transmitted.

On the UL side, the received symbols are processed in the reverse manner as for the DL one:

- **OFDM Demodulation** (PHY): it demodulates an OFDM input signal. The first cyclic prefix of the OFDM symbol is removed and one Fast Fourier Transform (FFT) operation per received FDMA symbol is performed. The received subcarrier values are recovered.
- **Antenna and resource de-mapping** (PHY): it aims to invert resource mapping operations to extract RS and data.
- **Equalization**: it aims to compensate for channel distortion and restore the original signal. In frequency-domain equalization, the received signal is transformed to the frequency-domain using a Discrete Fourier Transform (DFT) operation. The equalized frequency-domain signal is then transformed to the time-domain using an IDFT operator.
- **Layer de-mapping** (PHY): it aims to invert the operations of layer mapping to separate and detect the received symbols via MIMO antennas.
- **Baseband Demodulation** (PHY): it aims to recover the information content from the modulated carrier wave.
- **Descrambling:** it is the reverse of the scrambling process, returning the unscrambled bit sequence from the received scrambled bit sequence.
- **Decoding CRC check and Hybrid ARQ (HARQ):** In this stage, the aim is to recover data bits and parity bits from descrambled streams. Its functionality is the reverse of the rate matching and channel coding process. A cycle of rate de–matching, turbo decoding, code block concatenation and CRC check functions is required to output data and parity bits. The parity bits are fed back to the HARQ controller block, which controls HARQ transmission for transport channels to generate HARQ control signals, e.g., TB size and retransmission number using the HARQ ACK/NACK feedback from the receiver. Turbo decoder is also comprised of two decoders and one internal interleaver/de-interleaver. A code block is iteratively processed so that the output of one of the decoders is fed into the other one. After each decoder component, CRC is checked. Iterative exchange continues until the maximum number of iterations (specified at the input port), or it will be stopped as soon as CRC check results with success. Moreover, data is reordered by interleaving and de-interleaving blocks in the decoding process.

2.2 5G Basic Concepts

5G evolved from LTE and supports just the packet switch domain. The architecture of 5G is presented in Figure 2.6.

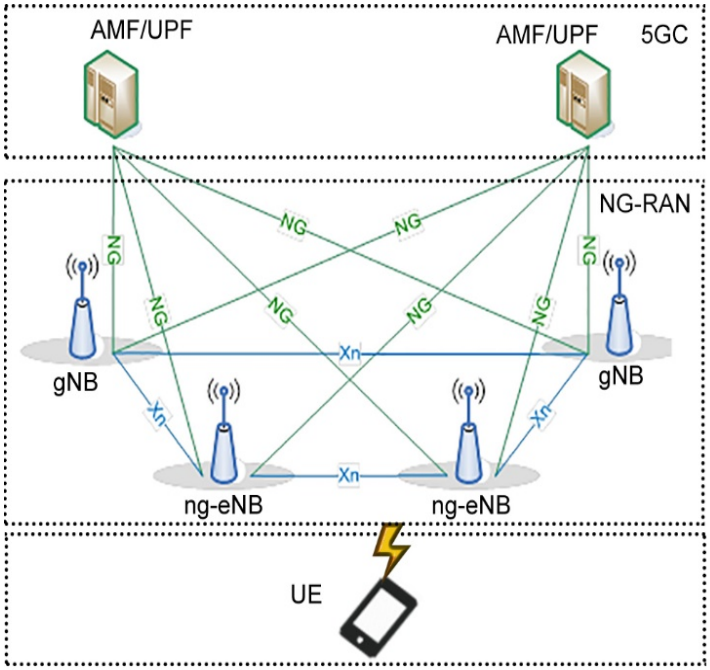


Figure 2.6 – 5G overall architecture (based on [3GPP20d]).

General components of 5G network are:

- **UE** is the user's device to communicate with the radio network.
- **Next Generation Radio Access Network (NG-RAN)** it consists of a set of next generation NodeBs (gNBs) and next generation eNodeB (ng-eNBs), which are connected to the 5G Core Network (5GC) through the Next Generation (NG) interface: gNB and ng-eNB function for radio resource management, i.e., radio bearer control, radio admission control, connection mobility control, dynamic allocation of resources to UEs in both UL and DL (scheduling); IP header compression, encryption and integrity protection of data; gNB and ng-eNB also function for selection of an Access and Mobility management Function (AMF) at UE attachment when no routing to an AMF can be determined from the information provided by the UE; routing of user plane data towards User Plane Functions (UPFs); routing of control plane information towards AMF; connection setup and release; scheduling and transmission of paging messages; scheduling and transmission of system broadcast information; measurement and measurement reporting configuration for mobility and scheduling; transport level packet marking in the UL; session Management; support of network slicing; QoS flow management and mapping to data radio bearers; distribution function for NAS messages; RAN sharing; dual connectivity and tight interworking between New Radio (NR) and EUTRA.
- **5GC** is composed of the following elements:
    - **AMF** is responsible for NAS signaling termination; NAS signaling security; access stratum security control; inter core network node signaling for mobility between 3GPP access networks; idle mode UE reachability (including control and execution of paging retransmission); registration area management; support of intra-system and inter-system mobility; access authentication; access authorization including check of roaming rights; mobility management control (subscription and policies) and support of network slicing.

- **UPF** is the anchor point for Intra-/Inter-RAT mobility (when applicable); external PDU session point of interconnect to data network; packet routing & forwarding; packet inspection and User plane part of Policy rule enforcement; traffic usage reporting; UL classifier to support routing traffic flows to a data network; branching point to support multi-homed PDU session; QoS handling for user plane, e.g., packet filtering, gating, UL/DL rate enforcement; DL packet buffering and DL data notification triggering.

Similar to LTE, 5G network protocols are classified into control and user planes. The NG user plane protocol stack between a gNB node and a UE is illustrated in Figure 2.7. The Service Data Adaptation Protocol (SDAP) sublayer is responsible for mapping between a QoS flow and a data radio bearer and marking QoS flow ID (QFI) in both DL and UL packets. The concept of QoS flow and QFI is explained in more detail in Section 2.3. The rest of the sublayers' functionality is the same as in LTE. A difference is in the physical layer, since, unlike LTE, 5G exploits low density parity check for both DL/UL shared channel coding. For more details, one is referred to [3GPP20d] and [3GPP20e].

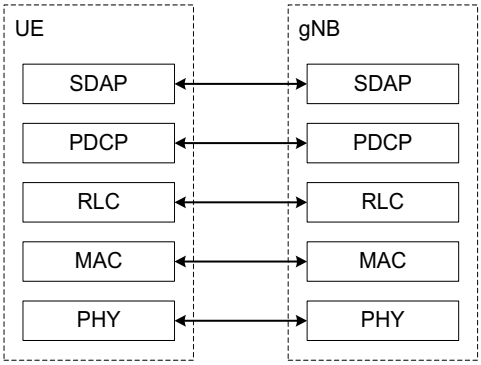


Figure 2.7 – 5G user plane protocol stacks (extracted from [3GPP20d]).

Apart from the bands used for LTE, new ones were added to 5G spanning from 450 MHz to around 6 GHz. Using previously unexploited frequency bands, i.e., millimeter waves, frequency bands above 24 GHz or even 60 GHz are foreseen, enabling 5G with higher data rate and capacity. 5G supports both paired and unpaired bands: in the former, distinct frequency ranges are allocated for UL and DL, and in the latter, a single shared frequency range is allocated to both UL and DL. The Frequency Ranges (FR) in which NR can operate are described in Table 2.3, where FR1 includes all existing and new bands below 6 GHz and FR2 includes new bands in the range of 24 to 53 GHz, [3GPP20f].

5G uses OFDMA for both UL and DL. Unlike LTE, where SC-FDMA is the main and the only data transmission scheme in UL, OFDMA is the main transmission scheme for 5G in UL with the possibility of using SC-FDMA as complementary. Moreover, 5G supports several ranges of SubCarrier Spacing (SCS) as it is designed to support a wide range of deployment scenarios. Therefore, the subcarrier spacing baseline is selected to be 15 kHz, but it can vary in the range of 15 kHz to 120 kHz [3GPP20f]. The number of RBs for each BS channel bandwidth and the SCS is specified in Table 2.4 for FR1 and Table 2.5 for FR2.

Table 2.3 – 5G operating bands in FR (based on [3GPP20f]).

Operating bands in FR1			
operating band	UL [MHz]	DL [MHz]	Duplex Mode
n1	1920– 1980	2110 – 2170	FDD
n2	1850 – 1910	1930 – 1990	TDD
n3	1710 – 1785	1805 – 1880	TDD
n5	824– 849	869 – 894	TDD
n7	2500 – 2570	2620 – 2690	TDD
n8	880 – 915	925 – 960	FDD
n12	699 – 716	729 – 746	FDD
n20	832 – 862	791 – 821	FDD
n25	1850 – 1915	1930 – 1995	FDD
n28	703 – 748	758 – 803	FDD
n34	2010 – 2025	2010 – 2025	TDD
n38	2570 – 2620	2570 – 2620	TDD
n39	1880 – 1920	1880 – 1920	TDD
n40	2300 – 2400	2300 – 2400	TDD
n41	2496 – 2690	2496 – 2690	TDD
n50	1432 – 1517	1432 – 1517	TDD
n51	1427 – 1432	1427 – 1432	TDD
n66	1710 – 1780	2110 – 2200	FDD
n70	1695 – 1710	1995 – 2020	FDD
n71	663 – 698	617 – 652	FDD
n74	1427 – 1470	1475 – 1518	FDD
n75	N/A	1432 – 1517	SDL
n76	N/A	1427 – 1432	SDL
n77	3300 – 4200	3300 – 4200	TDD
n78	3300 – 3800	3300 – 3800	TDD
n79	4400 – 5000	4400 – 5000	TDD
n80	1710 – 1785	N/A	SUL
n81	880 – 915	N/A	SUL
n82	832 – 862	N/A	SUL
n83	703 – 748	N/A	SUL
n84	1920 – 1980	N/A	SUL
n86	1710 – 1780	N/A	SUL

Operating bands in FR2		
operating band	UL and DL [MHz]	Duplex Mode
n257	26500 – 29500	TDD
n258	24250 – 27500	TDD
n260	37000 – 40000	TDD
n261	27500 – 28350	TDD

Table 2.4 – Number of RBs associated with each channel bandwidth for FR1 (based on [3GPP20f].

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

Table 2.5 – Number of RBs associated with each channel bandwidth for FR2 (based on [3GPP20f].

Channel Bandwidth [MHz] / SCS [kHz]	50	100	200	400
60	66	132	264	N. A
120	32	66	132	264

As mentioned in the previous section, the number of bits that can be carried by a single RE depends on the selected MCS. An appropriate MCS is selected, based on the information that the UE sends about its communication channel quality. 5G supports π/2-Binary Phase Shift Keying (π/2-BPSK), BPSK, QPSK, 16QAM, 64QAM and 256QAM [3GPP20g].

Moreover, similar to LTE, 5G exploits channels to deliver information faithfully in the network. Data are classified for efficient delivery in the 5G layers. Since this concept's principles are the same as for LTE, the discussion is not repeated in this subsection.

2.3 Quality of Service

The particular quality a network offers for a service is called QoS [PSAD05]. The number of simultaneous users, bit rate and power level are key factors affecting the quality achieved by users. In order to study QoS, a bearer service with clearly defined characteristics and functionality should be set up from the source to the destination of a service.

Nowadays, people use different service types while using mobile communications systems: one can make a phone call using VoIP while browsing the internet at the same time. Requirements vary depending on the type of service, e.g., VoIP cannot tolerate much delay while a slight delay in web browsing is bearable. Therefore, there are two categories of bearers in LTE: Guaranteed Bit Rate (GBR) bearers guarantee minimum bit rate for their services and are suitable for real-time services, e.g., voice; Non-GBR bearers offer no such guarantees, so they are suitable for non-real-time services, e.g., web browsing. Each bearer has two fundamental parameters: QoS Class Identifier (QCI) and Allocation and Retention Priority (ARP). QCI is a scalar representing predefined values for priority, Packet Delay Budget (PDB), and Packet Error Rate (PER), a bearer being always associated with a QCI, enabling an eNB to decide how to handle a bearer. ARP is used to determine whether to accept or reject a bearer establishment in case of radio congestion. In Table 2.6, standardized QCI characteristics are listed.

In LTE, QoS is divided into four different classes concerning the type of service [3GPP20h]. Table 2.7 summarizes the key features of QoS classes, being defined in what follows.

- **Conversational** contains applications performing real-time conversations between live end-users, e.g., voice over IP and video conferencing. Preserving time relation between information entities of the stream is a characteristic of the applications within this class.

Table 2.6 – Standardized QCI characteristics for LTE (based on [3GPP16a]).

QCI	Resource Type	Priority Level	PDB [ms]	PER	Example Services
1	GBR	2	100	10^{-2}	Conversational Voice
2		4	150	10^{-3}	Conversational Video (Live Streaming)
3		3	50	10^{-3}	Real-time Gaming, V2X messages
4		5	300	10^{-6}	Non-Conversational Video (Buffered Streaming)
65		0.7	75	10^{-2}	Mission Critical user plane Push To Talk voice (e.g., MCPTT)
66		2	100	10^{-2}	Non-Mission-Critical user plane Push to Talk voice
75		2.5	50	10^{-2}	V2X messages
5	Non-GBR	1	100	10^{-6}	IMS Signaling
6		6 (For Multimedia Priority Services subscribers)	300	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, progressive video, etc.)
7		7	100	10^{-3}	Voice, Video (Live Streaming)
8		8 (For premium subscribers)	300	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, progressive video, etc.)
9		9 (for non-privileged subscribers)			
69		0.5	60	10^{-6}	Mission Critical delay sensitive signaling (e.g., MC-PTT signaling)
70		5.5	200	10^{-6}	Mission Critical Data (e.g., example services are the same as QCI 6/8/9)
79		6.5	50	10^{-2}	V2X messages

Table 2.7 – QoS classes (extracted from [Corr14]).

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

- **Streaming** encloses applications that are demanding real-time data flow, e.g., streaming video and audio. This class of applications also preserves time relation between information entities.
- **Interactive** is employed when data is requested from a remote device. Applications, including web-browsing, database retrieval, server access or polling for measurement records and automatic database inquiries fall into this class. Requesting response patterns and preserving payload content are essential characteristics of QoS in this class.

- **Background** is a scheme applied when the application runs in the background, e.g., email, SMS and download of databases. In this class of applications, the destination is not expecting the data within a specific time, but payload content should be preserved.

5G is capable of responding to a diverse range of services and applications. 3GPP categorizes 5G services into five classes [3GPP16b]:

1. enhanced mobile broadband,
2. critical communications,
3. Massive machine Type Communications (MTC),
4. network operation,
5. enhancement of Vehicle-to-everything (eV2X).

A conceptual diagram depicting the service dimensions is presented in Figure 2.8, showing some examples of the services in each class. One should note that the service classes defined in the earlier version of mobile communications, i.e., conversational, streaming, interactive and background, are also foreseen to be supported by 5G.

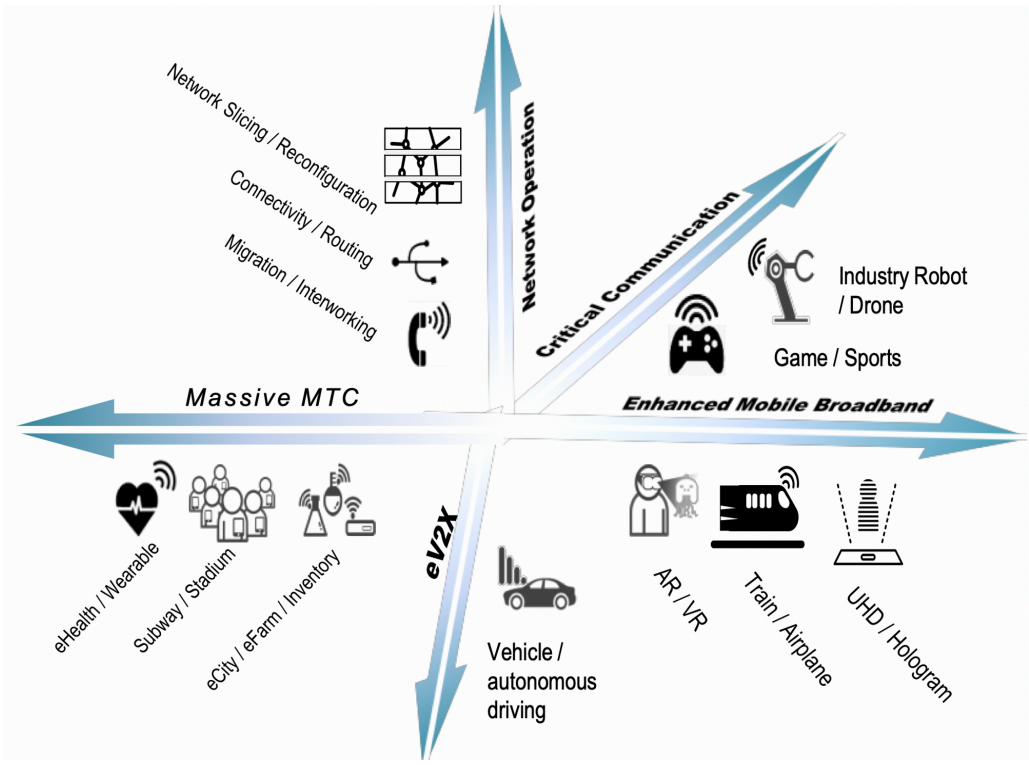


Figure 2.8 – 5G service dimension (extracted from [3GPP16b]).

In order to guarantee QoS, 5G proposes a model presented in Figure 2.9: 5GC establishes one or more PDU session for an individual UE; the process starts by sending a PDU session establishment message from the UE side to the 5GC; once the 5GC receives the request, it sends an inquiry to the gNB for the resources that the PDU session requires; the gNB then asks the UE to establish one or several data radio bearers over which the data is exchanged between UE and gNB. A logical interface between the NG-RAN node and UPF, i.e., NG-U, is also established for data transmission between gNB and UPF.

UPF performs transport level packet transmission on a per QoS Flow basis, making QoS differentiation among the distinct PDU sessions. A PDU session may contain one or multiple QoS flows. Packets of the user plane traffic mapped onto the same QoS flow receive the same traffic forwarding treatment, e.g., scheduling and admission threshold, edge-to-edge between the UE and the UPF. One of the fundamental parameters that 5GC exploits to characterize an individual QoS flow is 5G QoS Identifier (5QI), which is a scalar representing predefined values for resource type, priority level, PDB, PER, averaging window and maximum data burst volume as the performance characteristics that a QoS flow receives. The one-to-one mapping of standardized 5QI values to 5G QoS characteristics are specified in Table 2.8.

Table 2.8 – Standardized 5QI to QoS characteristics mapping (extracted from [3GPP18]).

5QI Value	Resource Type	Default Priority Level	PDB [ms]	PER	Default Maximum Data Burst Volume[B]	Example Services
1	GBR	20	100	10^{-2}	N/A	Conversational Voice
2		40	150	10^{-3}		Conversational Video (Live Streaming)
3		30	50	10^{-3}		Real-time Gaming, V2X messages, Electricity distribution – medium voltage, Process automation - monitoring
4		50	300	10^{-6}		Non-Conversational Video (Buffered Streaming)
65		7	75	10^{-2}		Mission Critical user plane Push To Talk voice (e.g., MCPTT)
66		20	100	10^{-2}		Non-Mission-Critical user plane Push To Talk voice
67		15	100	10^{-3}		Mission Critical Video user plane
5	Non-GBR	10	100	10^{-6}		IMS Signaling
6		60	300	10^{-6}		Video (Buffered Streaming), TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
7		70	100	10^{-3}		Voice, Video (Live Streaming), Interactive Gaming
8		80	300	10^{-6}		Video (Buffered Streaming), TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
9		90				
69		5	60	10^{-6}		Mission Critical delay sensitive signaling (e.g., MC-PTT signaling)
70		55	200	10^{-6}		Mission Critical Data
79		65	50	10^{-2}		V2X messages
80		68	10	10^{-6}		Low Latency eMBB applications Augmented Reality
82	Delay Critical GBR	19	10	10^{-4}	255	Discrete Automation
83		22	10	10^{-4}	1354	Discrete Automation
84		24	30	10^{-5}	1354	Intelligent transport systems
85		21	5	10^{-5}	255	Electricity Distribution- high voltage

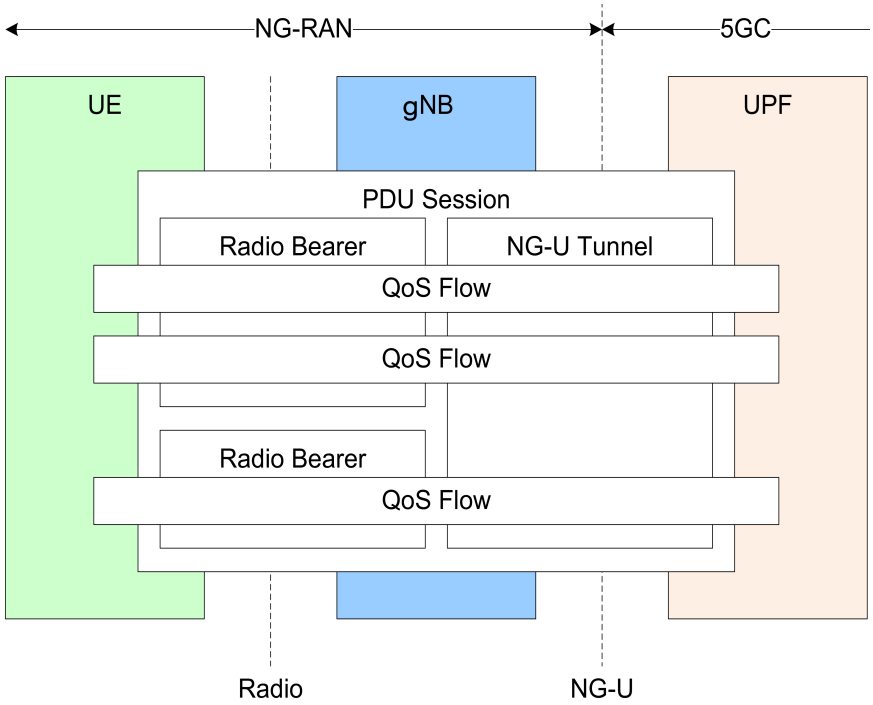


Figure 2.9 – QoS architecture (extracted from [3GPP20d]).

2.4 Coverage and Radio Capacity

Coverage and radio capacity are two major factors limiting cellular network performance. Coverage indicates how strong the transmitted signal should be in order to be recovered by a typical mobile device far from the BS, being more critical in rural areas. To estimate coverage, the Propagation Loss (PL) should be analyzed. To achieve the highest PL, it is required to have the largest value that a transmitter can send and the smallest value at which the receiver can recover the information.

However, these calculations are not enough because the actual PL depends on other factors, i.e., radio propagation antenna and geographical aspects. The effective height of the antenna concerning the ground should be considered as well. An elevated antenna propagates radio waves farther, which leads to a higher coverage; however, since each cell's radio capacity is limited, the extra coverage may not be useful, and it can cause interference in neighboring cells.

The cell coverage radius can be estimated considering the link budget and an appropriate propagation model for the path loss, [Corr20],

$$R_{[\mathrm{km}]}^{cell} = 10^{\frac{P_{T[\mathrm{dBm}]} + G_{T[\mathrm{dBi}]} - P_{R[\mathrm{dBm}]} + G_{R[\mathrm{dBi}]} - L_{P[\mathrm{dB}]}}{10\, a_{pd}}}, \tag{2.1}$$

where:

- P_T: power fed to the antenna,
- G_T: gain of transmitting antenna,

- P_R: power sensitivity at the receiver antenna,
- G_R: gain of the receiving antenna,
- L_P: path loss given by the link budget computation,
- a_{pd}: average power decay.

The above-mentioned factors influence the cell size, cells being categorized as Macro-, Micro-, Pico- and Femto-cells with respect to the size. Table 2.9 shows the typical cell radius.

Table 2.9 – Different cell radius (extracted from [Corr14]).

Cell	Radius [km]
Macro	> 3
Micro	0.1 – 1
Pico	< 0.1
Femto	< 0.05

On the other hand, radio capacity describes the number of devices that a BS can process, limited by the BS data rate, which is crucial in crowded areas, e.g., urban ones. The maximum number of users who can access the network depends on the amount of RBs and data rate allocated to each one according to the modulation and type of service. The total number of users in a single cell can be estimated as [Carr11]

$$N_{cell}^{u} = \frac{\Delta f_{BW[\mathrm{Hz}]}\ \eta_{sch}\ G_{pos}\ G_{mud}}{\overline{\Delta f}_{BWu[\mathrm{Hz}]}}, \tag{2.2}$$

where:

- Δf_{BW}: total bandwidth available,
- G_{pos}: radio capacity gain obtained due to users positioning in the cell,
- G_{mud}: radio capacity gain obtained due to multi-user diversity,
- $\overline{\Delta f}_{BWu}$: average bandwidth of the RBs allocated per user,
- η_{sch}: scheduler efficiency, chosen in [0, 1].

As (2.2) shows, the bandwidth allocated to each cell has a strong effect on capacity. LTE works in a diverse range of bandwidths, which can be enlarged even more with carrier aggregation, therefore, radio capacity is much higher in LTE compared to previous cellular network generations.

5G also supports the concept of carrier aggregation, but it can support the aggregation of up to 16 carrier components with different bandwidths, which leads to a transmission bandwidth of up to 6.4 GHz [DaPS18]. The application of new network topology, i.e., using small cells installed closer to subscribers, Wi-Fi offloading, advanced antenna techniques and MIMO channels, also increase the overall user throughput and radio capacity in 5G.

It is important to note that radio capacity is inversely related to QoS: the higher is QoS, the lower capacity is. Tuning coverage, capacity and QoS is an optimization problem that is addressed in radio network planning.

2.5 C-RAN and Virtualization

The proliferation of high data rate applications in conjunction with high mobile terminals usage nowadays has triggered a drastic increase in data rate demands [Cisc19], therefore, wireless network providers must continuously improve their infrastructure to serve this demand accordingly. The challenge is even more difficult because resource allocation in conventional RANs is inefficient, since it is based on peak-hour traffic requirements, while users' demand is time-varying, hence, traffic not being always at the peak level and being possibly up to 10 times lower in off-peak hours [CCYS14]; thus, a fixed allocation scheme leaves resources idle in various times/areas. C-RAN has emerged as a centralized paradigm to provide a solution for higher data rates and capacity demands in a cost-efficient way. In this section, the C-RAN architecture and its advantages are discussed according to [NGMN13], [CMRI13], [CCYS14] and [HDGK13].

In contrast to the traditional RAN architecture, where radio and baseband processing are integrated inside the BS, C-RAN splits the functions set into two main categories: the radio unit as RRH and signal processing unit as BBU. A C-RAN architecture is illustrated in Figure 2.10: baseband processing units of multiple traditional BSs are separated and aggregated into a central site to form a BBU-pool to which associated RRHs are connected. Multiple BBU-pools are connected via a high-speed optical link at a higher level.

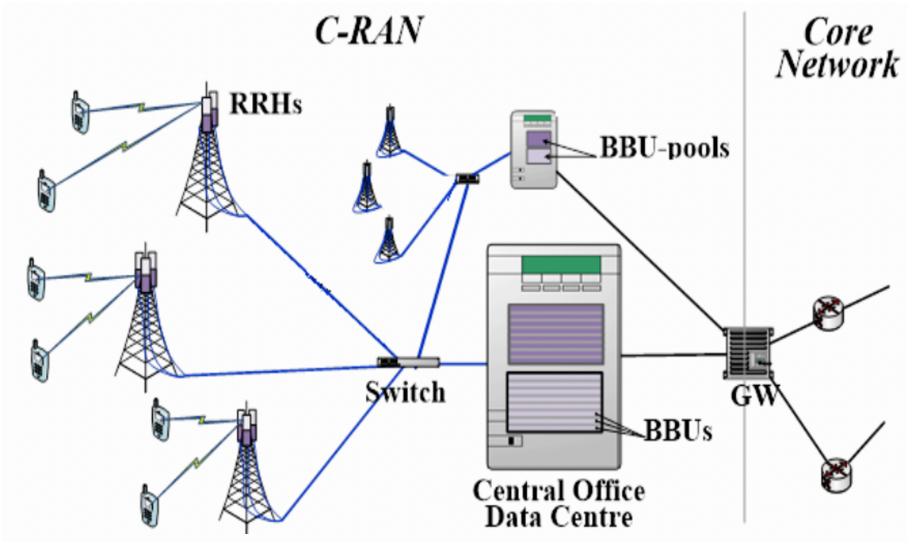


Figure 2.10 – C-RAN architecture (extracted from [FPHG14]).

From a technological viewpoint, BBU implementation can be based on a General-Purpose Processor (GPP), Graphics processing units (GPUs), or traditional BS platform (non–GPP based). As GPP technology has improved in terms of new instructions, processing pipeline, power consumption and powerful cache technology, it is possible to exploit GPP in signal processing and BBU-pools. Furthermore, multiple Central Processing Units (CPUs) can be embedded in one server, each with multiple cores, which increases performance at a high rate and meets all kinds of processing requirements from PHY layer to application layer, control and data plane.

As GPPs (or GPUs) are programmable processors, GPP based systems have advantages in flexibility and configurability. Therefore, it supports easy migration to newer and updated standards. The usage of application-specific hardware, e.g., Field-Programmable Gate Array (FPGA) and Application-Specific Integrated Circuits (ASICs), and software, is another approach for BBU-pool implementation that has been taken by most traditional RANs so far. In any case, to handle the antenna interface function, e.g., CPRI, network switch interface and/or pre-processing, i.e., conjoint processing among users in the area of an RRH, dedicated interfaces are required in a BBU and should be embedded.

Items located in an RRH include radio equipment, antennas, backhaul transmission equipment and other auxiliary equipment, such as power supply equipment, towers and monitoring equipment. A diverse range of network topologies, e.g., star, tree, ring and any combination, can be exploited to connect RRHs by fiber to BBUs. Selecting the type of connection to the BBU depends on fiber accessibility in the geographical region, e.g., in an area with plenty of fiber resources, the star topology is recommended due to the high transmission reliability. The connection between BBUs and RRHs, on the other hand, can be centralized or distributed: in the centralized design, a switch is required to transmit RRHs data to/from BBUs, while in distributed mode, RRH and BBU are connected directly.

Furthermore, the strategy to split the functions set results in full centralized or partial centralized C-RAN architecture: in the former, the BBU consists of L1, L2 and L3 BS functions, while in the latter, the BBU does not include L1 functions, but it integrates all other higher-layer functions. In another words, RRH includes both the PHY and the radio functions [CMRI11]. The two partitioning strategies are illustrated in Figure 2.11. Nevertheless, there are various possibilities on partial centralized C-RAN depending on how splitting PHY functions between RRH and BBU.

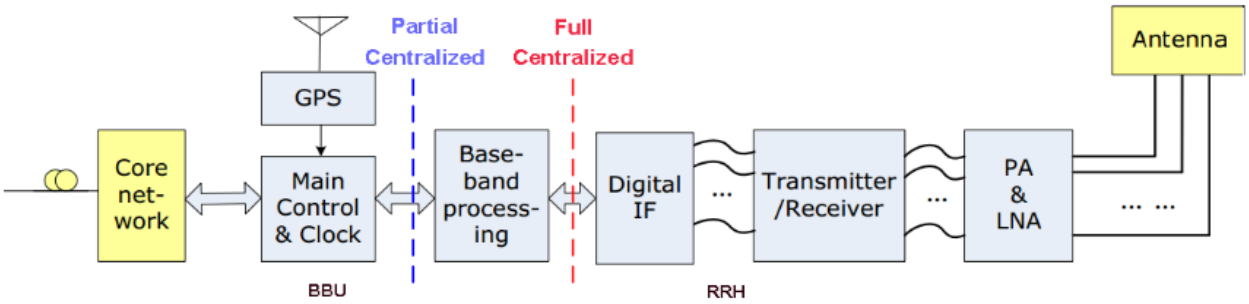


Figure 2.11 – Different separation method for BS functions (extracted from [CMRI11]).

Regardless of different C-RAN splitting point, it has significant benefits, listed as follows [CMRI11]:

- **Adaptability to non–uniform traffic**: During the day, the number of people in a particular area varies, e.g., population is more in residential areas in non-working hours, therefore, for each area, peak traffic loads do not occur at the same hour. However, BSs are tuned to operate correctly even in peak hours, which means a big waste of sources during non–peak hours. Centralizing BSs with different peak-hour traffic, e.g., by mixing residential and business regions in the same BBU-pool, balances resource usage, as the resources of under-loaded BSs at a given time instant can be shared with over-loaded ones. It is then expected to have lower peak resource requirements in a pool than the sum of peak requirements of individual BSs.

- **Energy and cost saving coming from statistical multiplexing gain in BBU-pool**: In total, C-RAN can save 15% capital expenditure and 50% operational expenditure compared to traditional RAN. Providing power to the RRH and BBU and air conditioning spend a considerable amount of energy in a mobile network. As mentioned before, in C-RAN, less BBUs are sufficient to meet network needs, therefore, the electricity cost can be decreased. Besides, in the hours that fewer subscribers are active, it is possible to switch off some BBUs in the pool, which do not impact on overall network coverage. On the other hand, gathering equipment in a central place reduces civil work on remote sites.
- **Increase throughput and decrease delays**: In LTE, radio resources are shared, leading to increased throughput and decreased delays. The idea is to use the same frequencies in all cells, therefore, inter-cell interference is high in these systems, Inter-Cell Interference Coordination (ICIC) being an approach addressing inter-cell interference. With ICIC, in the case of interference, the UE sends feedback to the eNB, and the eNB cooperates with the adjacent cells not to use that specific subcarrier. Coordinated Multi Point (CoMP) is another technique to improve inter-cell interference, where several cells are grouped in a CoMP-set, cooperating with each other to serve one or several UEs depending on their feedbacks. Since in C-RAN several BBU are integrated in one place, it is possible to collect all cells within a CoMP to be served in one BBU-pool, hence, tighter interaction between BSs is achieved. Moreover, ICIC operation can be improved by easier connection between multiple BBUs rather than many cells.
- **Ease in maintenance and network upgrades:** C-RAN can manage the network in peak hours and non–peak hours properly with less human intervention. Besides, hardware upgrade is possible by upgrading just a very few locations in the BBU-pool. With C-RAN, implementing new standards and frequent CPU updates is more comfortable as well. Moreover, exploiting Software Defined Radio and software BSs, makes it achievable to upgrade to new frequencies and new standards through software updates instead of hardware upgrades. On the other hand, to upgrade the system to increase coverage and radio capacity or deploy a new cell, can be done just by adding a new RRH or install a small device to the BBU pool, hence, flexibility is increased.

Although C-RAN provides significant advantages, it raises some challenges as well. To carry the baseband In-phase and Quadrature (IQ) signals, a high bandwidth between the BBU and RRH is required. Furthermore, techniques for BBU cooperation and interconnections, as well as virtualization techniques, should be developed. Moreover, to take advantage of C-RAN benefits, efficient strategies should be applied in order to distribute the resources of the BBU-pool among BBUs. An efficient resource provisioning scheme should minimize both the resource idle times and the BBUs' over-loading.

Virtualization is constructing one or several logical entities on top of an abstracted physical one [CCYS14]. The goal is sharing computing resources with the aid of a set of virtual nodes and links. To this end, several virtual entities coexist in one physical infrastructure. With virtualization, a virtual environment for guest Operating System (OS) and applications is provided. The guest OSs are separated, even though they are running on the same physical machine. The critical point in virtualization is isolating each virtual element from the others. Applying virtualization promotes flexible control and low cost, efficient resource usage.

Virtualization can be done in several scopes in the network, namely computing resources, radio resources and network management application entities. In this thesis, one focus on the BBU-pool computing resource management. Network virtualization separates the BBU-pool's data storage, processing capacity and management control to constructs virtual BSs on top of existing resources, e.g., CPUs, memory and network interface card, Figure 2.12. The virtual BSs run a portion (or full, depending on splitting point) of L1, MAC and upper layers functionality as software applications. Virtualization techniques share a common network environment, programming environment and an IT platform among several BSs.

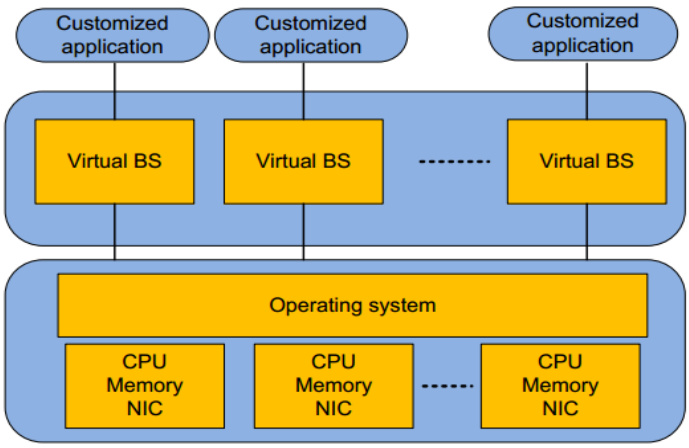


Figure 2.12 – BBU-pool with multiple virtual BSs sharing hardware and systems (extracted from [CCYS14]).

2.6 Resource Allocation and Game Theory

In cloud areas, resource allocation plays an essential role in the performance of the entire system. Since BBU requirements are not uniform across the network, an optimal resource allocation strategy is needed to distribute BBU-pool resources among BBUs fairly. The importance of optimal resource allocation becomes even more significant in the presence of resource shortage, when not all BBUs' demands can be served simultaneously. In these cases, the resource allocator should prioritize BBUs appropriately in order to satisfy QoS constraints. On the other hand, as BBU traffic demand fluctuates over time, the allocator should keep the provided resources as close as possible to the real-time demand to enhance resource usage. As a result, the resource allocation in a C-RAN BBU-pool can be considered as an optimization problem under uncertainty for a dynamic environment.

In order to find an optimal solution for the resource allocation problem in a BBU-pool, the bargaining concept in cooperative game theory can be applied. This concept is applied to competitive situations, where players are strategically competing against one another for the same resources [Myer91]. Computing resource allocation in the BBU-pool can be defined as a bargaining game in which the players, i.e., BBUs, compete for the BBU-pool's AvCC to increase their signal processing speed in order

to satisfy QoS constraints.

In order to solve this problem, the Nash Bargaining Solution (NBS) as a well-known solution to bargaining problems can be applied [Myer91]. NBS fairly splits resources among players by allowing them to bargain with each other. A negotiated outcome is selected from a given set of feasible outcomes. The outcomes are evaluated according to the individual utility functions of the players. The players are bargaining to agree on choosing an element that maximizes their utility. If the players' total resources are less than the available ones, the entire players' requests are satisfied. Otherwise, NBS provides an optimal compromise solution. The players may only obtain the minimum number of resources they expect by joining the game without cooperation, which is called disagreement.

NBS is the unique fair Pareto optimal solution among all feasible ones that maximizes the product of the utility gains over all negotiators, being a useful tool to model interactions among negotiators that guarantees all players acquire the maximum utility with fair concerns [Myer91]. Mathematically, a bargaining problem with N_P players is defined as a pair $\left(S^{FS} \cup \{\mathbf{R}_{\min[N_P\times 1]}\}, \boldsymbol{\mathcal{U}}_{[N_P\times 1]}\right)$ where S^{FS} is convex and a closed set, containing all the feasible solutions for the problem, and $\boldsymbol{\mathcal{U}}$ is the utility functions set, such that

$$\boldsymbol{\mathcal{U}}_{[N_P\times 1]}\left(\boldsymbol{R}^{Al}_{[N_P\times 1]}\right) = \left[\mathcal{U}_1\left(\boldsymbol{R}^{Al}_{[N_P\times 1]}\right), \dots, \mathcal{U}_{N_P}\left(\boldsymbol{R}^{Al}_{[N_P\times 1]}\right)\right]^T, \tag{2.3}$$

where $\mathcal{U}_i: \mathbb{R}^{N_P} \to \mathbb{R}$ is the utility function of player i, and

$$\boldsymbol{R}^{Al}_{[N_P\times 1]} = \left[r_1^{Al}, r_2^{Al}, \dots, r_{N_p}^{Al}\right]^T, \tag{2.4}$$

where r_i^{Al} is the number of resources allocated to player i. The utility function in a game represents the players' preference that is defined when a problem is formulated as a bargaining game. Moreover, $\mathbf{R}_{\min} \in S^{FS}$ is the minimum desired number of resources that should be guaranteed to the players,

$$\mathbf{R}_{\min[N_P\times 1]} = \left[r_{\min 1}, r_{\min 2}, r_{\min 3}, \dots, r_{\min N_P}\right]^{\mathrm{T}}, \tag{2.5}$$

where $r_{\min i}$ is the minimum number of resources that player i expects by joining the game, being also defined in the problem formulation phase.

Given S^{FS}, $\mathbf{R}_{\min}$ and $\boldsymbol{\mathcal{U}}$, the NBS, $\mathbf{R}^{Al^*}_{[N_P\times 1]}$, is achieved by solving the following optimization problem:

$$\mathbf{R}^{Al^*}_{[N_P\times 1]} = \underset{\mathbf{R}^{Al}_{[N_P\times 1]} \in S^{FS}\cup\{\mathbf{R}_{\min[N_P\times 1]}\}}{\operatorname{argmax}} \left(\prod_{i=1}^{N_P} \left[\mathcal{U}_i\left(\mathbf{R}^{Al}_{[N_P\times 1]}\right) - \mathcal{U}_i\left(\mathbf{R}_{\min[N_P\times 1]}\right)\right]\right). \tag{2.6}$$

The aim is to find the optimal solution from the given set of feasible solutions that maximizes all players' utility while fairness among them is satisfied. The NBS is the unique solution for a bargaining game that satisfies Nash axioms as the attributes that any rational solution should meet to come up with fairness and efficiency, being defined as follows [Myer91]:

1. **Strong efficiency:** it asserts that the solution should be feasible, in other words $\mathbf{R}^{Al^*} \in S^{FS}$; the solution should also be Pareto efficient, meaning that none of the players can be made better utility without making at least one player worse utility.
2. **Individual rationality**: it asserts that the solution should be better off than the disagreement point.

3. **Scale covariance**: it asserts that if the same affine transformation is performed on all players' utilities, then the solution will be transformed accordingly.
4. **Independence of irrelevant alternatives:** it asserts that the solution should not be affected if an irrelevant alternative is eliminated, i.e., $\mathbf{R}^{Al^*}$ is a bargaining solution for any subset $(S^{FS})'$ of S^{FS} that contains $\mathbf{R}^{Al^*}$.
5. **Symmetry:** it asserts that in the players' equal situation, the solution should not discriminate among them.

The symmetry axioms mentioned above guarantees the equal priority of players during the bargaining game, meaning that all players involved in the bargaining game are assigned with the same bargaining power. However, if the negotiators in a bargaining game have strategic advantages, an asymmetric solution arises, maximizing the production of weighted utility functions [Binm91], [Myer91], where the weights are positive values reflecting the negotiators' bargaining powers, therefore, the generalized NBS (GNBS) is a unique solution to the defined bargaining problem that satisfies axioms 1 to 4 above. GNBS is characterized as the point $\mathbf{R}^{AlG^*}_{[N_P\times 1]}$, that

$$\mathbf{R}^{AlG^*}_{[N_P\times 1]} = \underset{\mathbf{R}^{Al}_{[N_P\times 1]} \in S^{FS}\cup\{\mathbf{R}_{\min[N_P\times 1]}\}}{\operatorname{argmax}} \left(\prod_{i=1}^{N_p} \left[\mathcal{U}_i\left(\mathbf{R}^{Al}_{[N_P\times 1]}\right) - \mathcal{U}_i\left(\mathbf{R}_{\min[N_P\times 1]}\right)\right]^{B_i}\right), \tag{2.7}$$

where B_i is player i's bargaining power.

2.7 State of the Art

Pooling, cloudification and virtualization have been studied from different viewpoints in the area of telecommunications. This section provides an overview of the state of the art and resource management strategies in C-RAN.

2.7.1 C-RAN Architecture

Several architectures for C-RAN have been proposed, enabling load balancing and computing resource management within the BBU-pool using RAN softwarization and virtualization techniques.

In [NGMN13], suitable scenarios for C-RAN utilization and its major functionalities were studied. An architecture for C-RAN was explained, in which RRH clusters, composed of several RRHs, are connected to associated BBU clusters, each composed of several BBUs, in the BBU-pool. To balance load, data from an arbitrary RRH in the cluster can be switched to any BBU in the associated BBU cluster.

The authors in [HDGK13] go one step further, by introducing a C-RAN architecture with Multi-Site/multi-Standard BaseBand Unit (MSS–BBU) and discussing the prerequisites and challenges for a multi-standard cloud radio BS, where each cluster of several RRHs can exploit a separate pool. At a higher level, pools in various locations are interconnected by high–speed optical links. Furthermore, an element

in every MSS-BBU, called a Decentralized Cloud Controller (DCC), controls the load balancing within a cloud-based BS. It collaborates with other DCCs from adjacent MSS-BBUs. If the available computing resources of an MSS-BBU are insufficient to meet the resource requests, the local DCC asks the computing resources from its remote neighbors by sending the load toward them.

Several software implementations of LTE and 5G eNB have also been developed, e.g., Amari LTE 100 [Amar15] and Open Air Interface (OAI) [OAI14], which enables LTE and 5G RAN functionality as a software implementation over GPPs in virtualized environments. Alyafawi et al. [ASBD15] also studied the processing time deadline to find acceptable execution platforms for C-RAN in virtualized environments. They focused on the LTE FDD PHY layer and infrastructure as a service that handles and manages storage, network and other computing resources among Virtual Machines (VMs). Open Stack is exploited to orchestrate the computing resources on VMs. Furthermore, OAI is deployed as a software BBU running on Real-time Linux (RTLinux) in a GPP host machine. The processing time of two distinct hypervisors, kernel based VM and Linux containers, are compared. They calculated the processing time for sub-frames in UL and DL in the BBU side and multiple MCSs and bandwidths. The results show that, despite all the various configurations, the processing time for de/encoding increases with the MCS index's increase, and a decoding time twice as long as the encoding time. Moreover, it was concluded that VMs with at least 4 GHz CPUs are required to support the LTE-FDD PHY layer with maximum load.

Moreover, in [MCN15], a service-oriented architecture for LTE was introduced. In a way that all the network elements e.g., EPC elements or RAN, are regarded as a service that can be offered as a service instant to an enterprise end-user, i.e., an operator. Service Manager, service orchestrator and cloud controller are three primary functional components of this architecture that manage and control the infrastructure and services. Each service has a service orchestrator and the cloud controller cooperating to manage the configuration coordination of all services instances. The architecture of each service orchestrator and the cloud controller differs according to each service requirement. Once an enterprise end-user request arrives at the service manager, a service orchestrator is created, which communicates its needs and requirements to the cloud controller in order to create, configure, orchestrate and manage the requested service instant, e.g., instances of RAN and EPC. Once a service instant is established and running, the cloud controller delivers an interface to the enterprise end-user. This proposed architecture is compared with the Network Functions Virtualization (NFV) architecture of the European Telecommunications Standards Institute to show its compromise with current industry activities.

The feasibility of exploiting GPUs to build wireless BSs was also studied in [ZCLD15]. The authors developed parallel implementations of the main BS functions to evaluate the GPU utilization as the baseband signal processors. Their result shows that four NVIDIA GTX680 GPUs are required to achieve real-time LTE sub-frame processing.

2.7.2 Computing Demand Estimation and Resource Allocation

Dynamic radio and computing resource prediction and allocation increase the efficiency and the radio and computing capacity of a network. Evaluating the traffic demand is the first step in an efficient

resource allocation; the optimal solution for resource utilization can only be found afterwards. In a cloud networking environment, this topic has received significant attention in recent literature.

[Eart12] and [MAMM16] proposed a model to estimate the amount of RCCs of a BBU. The RCC of a BBU is the amount of computing capacity required so that no computing delay is imposed on the signal processing. By taking some well-defined operating scenarios, the RCC per information bit transmission is estimated by counting the number of mathematical operations that each signal processing step performs. The achieved values are then scaled for any desired scenario, considering the network/user parameters' variations that affect the result, e.g., bandwidth and number of antennas. The estimated RCC values and the scaling rules have been interpreted from various sources either from scientific research or empirical experiments.

After estimating the computing demand of the BBU, the next step is to find the optimal allocation of resources among them. Several resource management approaches have been proposed in the literature, aiming at maximizing C-RAN computing resource utilization.

X. Wang et al. [WTTC16] proposed a model to balance load among the BBUs in a pool. In BBU over-loading, the excess load is migrated to other underutilized active BBUs, enabling the over-loaded BBU to use the extra resources left by the other BBUs in the pool at a specific time instant. Consequently, the load becomes more balanced, leading to improved resource utilization and better energy efficiency. Within this framework, they formulated the C-RAN resource management problem as a linear integer programming challenge. The proposed model re-assigns the processing tasks that cause BBU over-loading to the appropriate underutilized BBUs so that BBU-pool resource utilization is enhanced.

Additionally, load migration enables reducing the number of active BBUs by consolidating the processing task of multiple BBUs in a few ones in off-peak hours, when most of the BBUs in the pool are underutilized. K. Sundaresan et al. [SASR16] suggested a dynamic RRH to BBU mapping framework, which enables a BBU to serve several RRHs at the same time. The goal was to minimize the idle resources by reducing the number of active BBUs when traffic load is low and a single BBU is sufficient, showing a 50% improvement in resource usage compared to the baseline one-to-one RRH to BBU mapping strategy. Similarly, Al-Dulaimi et al. [AlAN19] proposed a model based on graph coloring to switch off low traffic BBUs and divert their processing load to neighboring under-loaded ones in the pool.

The authors in [YoTP18], and [QHSV15] also formulated the BBU-pool resource allocation as a bin packing problem. BBUs are treated as bins with finite computing capabilities and the cell processing tasks as the items that should be packed in the bins so that fewer BBUs are used; they used heuristic algorithms to solve the defined problems.

W. Chien et al. [ChLC19] went beyond the BBU-pool and proposed a resource management model to improve network resource usage by turning off the BBU-pools with low traffic and redirecting their RRHs in the network.

Many works in the literature focus on load migration as a strategy for resource utilization optimization in a BBU-pool. However, this policy imposes additional overheads to the network due to increased data exchanges between the source and the target BBUs [CFHH05]. The migration cost is higher in dense

areas, since handover, CoMP transmission/reception and interference occur more often among small cells [NGMN15]. One approach for reducing the data exchange burden is to serve coordinated RRHs with a single BBU [ZJJL17] and the BBU computing capacity being elastically reconfigured according to its real-time demand.

An adaptive computing capacity strategy is chosen in this thesis in order to optimize the computing resource utilization of the BBU-pool. To the best of our knowledge, to date, only a few works on BBU-pool resource management have considered adaptable computing resources for the BBUs. D. Pompili et al. [PoHT16] proposed a framework for elastic and on-demand computing resource allocation to the BBUs in the pool employing virtualization techniques. The BBU functions are performed on the VMs reposed on top of general-purpose servers. Their model estimates BBU demands, regarding a given pattern, and delivers the BBU-pool computing resources accordingly.

Based on a similar platform, N. Yu et al. [YSDH19] proposed a model to improve the computing resource utilization of a BBU-pool by switching off the low traffic RRHs and their associated BBUs, diverting their processing load to the neighbors in the pool. If required, more resources are allocated to the target BBUs in order to improve their processing capability. The models proposed in [PoHT16] and [YSDH19] improve the computing resource utilization; however, both assume that there are always adequate resources in the pool to meet the peak demands and do not suggest a resource management strategy in the case of a resource shortage.

Chapter 3

Resource Allocation Model

This chapter presents a novel model and algorithm for efficient computing resource management in a BBU-pool. Section 3.1 gives a brief description of the chosen C-RAN architecture, strategies used for BBU-pool virtualization and discusses the main network assumptions. Section 3.2 presents an overview of the proposed resource management model. Section 3.3 summarizes the approach for estimating the amount of computing resources that each BBU requires at a given time instant, and Section 3.4 describes the proposed optimization model for assigning the computing resources across the BBUs in a BBU–pool, accordingly. Equal and demand proportional resource allocation models are also described in Section 3.5 as two reference resource allocation schemes compared with the proposed one. The rest of the chapter is dedicated to the definition of evaluation metrics, model implementation, canonical scenario and the simulator/model assessment.

3.1 Network Architecture and Assumptions

In this thesis, one considers a C-RAN architecture used for both 4G and 5G. The selected architecture is presented in Figure 3.1, where BBUs from multiple BSs are aggregated in a BBU-pool and each BBU is connected to its RRH through a high-speed optical link. The BBU-pools are linked together and connected to the core network via high-speed connections in the upper level [CCYS14]. Some assumptions are taken in order to constitute the network model being explained in what follows.

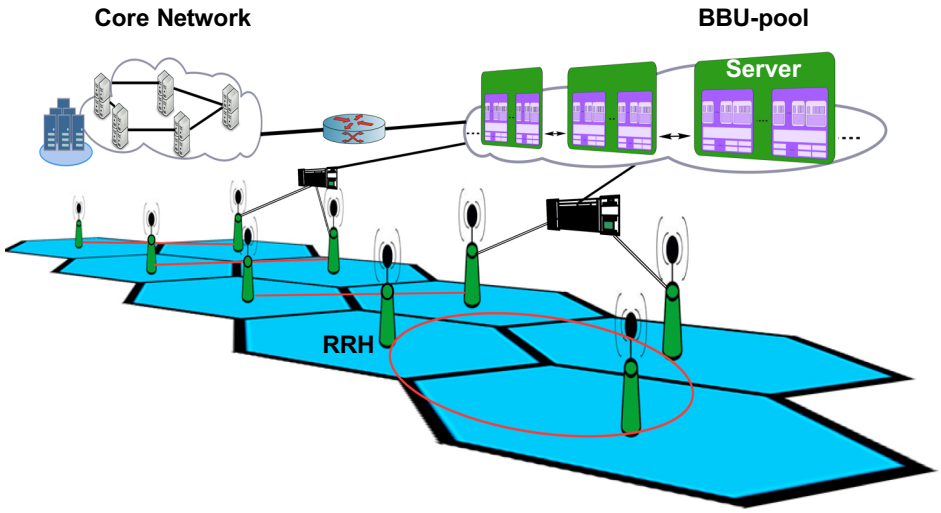


Figure 3.1 – C-RAN architecture.

Although a BBU can transmit/receive a signal to/from several RRHs [CCYS14], for simplicity, it is assumed that each RRH is served by one BBU in the pool and that a BBU serves just one RRH via a high-speed, low latency fiber front-haul with abundant capacity.

Without loss of generality, only user plane data transmission is considered in this thesis. Concentration is on the PHY layer, taking channel de/coding, de/modulation, MIMO de/pre-coding, channel estimation, and OFDMA and SC-FDMA into account as the primary signal processing steps of the BBUs. However, using a model similar to the one presented in Section 3.3, the proposed model can be fitted to the whole protocol stack layers and the control plane data transmission and signaling.

A user is counted active at time instant t_k if it has a packet to be received/transferred at that time. The number of RBs that the user requires at a given time instant is a variable of its packet volume and MCS. The user preferred MCS is derived based on its associated Signal to Noise Ratio (SNR). Since users are considered mobile, they are assigned with a random SNR assumed to be unchanged within the considered coherence time. It is also assumed that in the case of a packet loss, the transmitter resends the same packet under the HARQ process, the retransmitted packet being treated as a newly arrived one. New user arrivals are accepted until there is no more available radio RBs to be allocated. Although a single user can arrive in the network several times and perform several services per day in the real world, it is assumed that a user arrives just once to the network. Any new arrival to the network is counted as a new user.

A BBU-pool is a centralized location, including computing resources of multiple BSs being consolidated in general-purpose servers, which are shared and flexibly allocated to the BBUs based on their real-time demands through virtualization techniques. The architecture of a BBU-pool is illustrated in more detail in Figure 3.2. Each BBU–pool contains several standard IT servers; baseband computing resources being deployed on them. The implementation of baseband functionalities is based on virtualization techniques that allow a single physical machine to act as multiple logical entities using a hypervisor software layer. The logical entities called VMs share the computing, storage and communication resources of the server. Each VM has a real-time guest OS on which an instance of a software based BBU (so-called soft BBU) is implementing so that the BBU functionalities are implemented as applications on the VM.

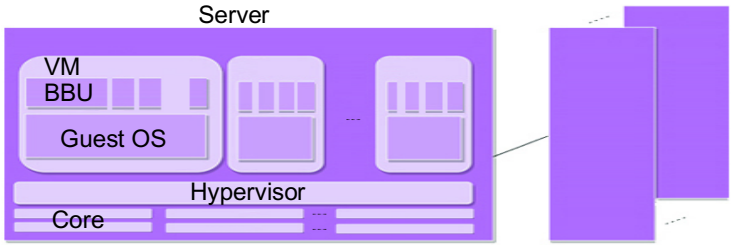


Figure 3.2 – BBU–pool and RAN virtualization.

Although a server's computing resources include input/output, storage, memory, CPU, etc., for simplicity, only CPU (with the same configuration for all servers) is considered the computing resource in the pool. Each BBU-pool owns a specific number of computing resources. In order to share computing resources between VMs, i.e., soft BBUs, processors are time-sliced. Each BBU gets full access to the processors to execute its related processing, and then the next BBU gets them for the second split, and so on. Resource sharing is detailed in Section 3.4.

3.2 Model Overview

An overview of the proposed computing resource management model for a single time instant is presented in Figure 3.3. Different types of inputs are needed to feed the proposed model, which are grouped according to their nature into user parameters and network ones.

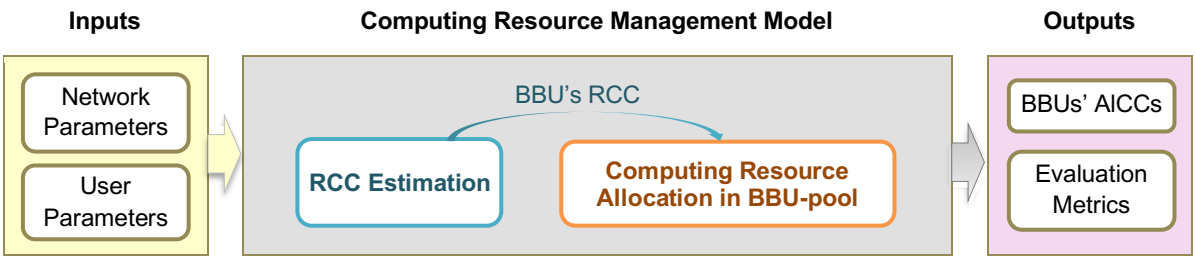


Figure 3.3 – Model overview.

The network inputs are the parameters that provide the characteristics of the network, including:

- Cell-specific info:
  - type of cell, C^{TYP}, i.e., Macro-, Micro-, Pico- and Femto-cell,
  - RRH traffic type, H^{TYP}, i.e., residential or business,
  - operating bandwidth, $\Delta f_{BW[\mathrm{MHz}]}$,
  - quantization resolution, $Q_{[\mathrm{bit}]}$,
  - MIMO order, N_{MIMO};
- BBU-pool specific info:
  - number of the BBUs aggregated in the BBU-pool, N_B,
  - available computing capacity, $C^{Av}_{BP\ t_k[\mathrm{GOPS}]}$.

The user parameters are those that specify the user characteristics that are identified as:

- User-specific info:
  - number of spatial streams, $N_{Str\ u,t_k}$,
  - SNR, $\gamma_{u,t_k[\mathrm{dB}]}$,
  - packet volume, $V^{Pkt}_{u,t_k[\mathrm{B}]}$,
  - service ID, s^U_u;
- User Arrival Rate, $R^{Arr}_{U[\mathrm{user/min}]}$.

The aim of providing an efficient resource allocation strategy in a BBU-pool is to maximize resource utilization. To achieve this goal, resources should be allocated to BBUs based on their real-time demand such that QoS is maintained. Hence, the first step is traffic demand evaluation and the optimal solution for resource utilization can be found only afterwards. In this way, the proposed resource management algorithm comprises two components, Figure 3.3:

1. **RCC estimation**: calculation of instantaneous demand (measured in Operations per Second [OPS]) of BBUs, according to the real-time network/user parameters.
2. **Computing resource allocation in a BBU-pool**: obtaining the optimal on-demand computing resource allocation maximizes both BBU-pool resource utilization and efficiency with respect to the required QoS.

By taking as inputs network and user parameters at a specific time instant, the estimation of the BBUs' RCC is based on a well-defined model [MAMM16] and [DeDL15]. The results are then fed to the computing resource allocation step in order to find the optimal AICC to BBUs. To this end, the BBU-pool computing resource allocation is formulated as a game-theory based bargaining problem, which is solved by the corresponding axiomatic solutions.

The resource management module's output is the BBUs' optimal resource allocations, which maximizes the BBU-pool computing resource utilization, in addition to the evaluation metrics that enable evaluating the proposed computing resource management strategy. Evaluation metrics are explained in detail in Section 3.6, including:

- BBU fulfilment level, f^B_{b,t_k},
- fairness index, F_{t_k},

- efficiency of resource allocation, $\eta_{t_k[\%]}$,
- resource usage, $U_{t_k[\%]}$.

It should be noticed that just one time instant is considered in this chapter. The model for computing resource management in a time-varying network is explained in Chapter 4.

3.3 Required Computing Capacity Estimation

3.3.1 BBU Physical Layer Processing

There are several options for splitting BS functionalities between BBU and RRH. Since the focus in this thesis is on the BS PHY layer, the main BS processing steps that are considered to be performed in a BBU involve channel de/coding and de/modulation, channel estimation, MIMO de/precoding steps, and IFFT/FFT and cyclic prefix insertion (OFDMA/SC-FDMA), being presented in the BS processing set, S^{Proc}, given by

$$S^{Proc} = \{P^{chc}, P^{chd}, P^{md}, P^{dm}, P^{mpc}, P^{mdc}, P^{che}, P^{OFDMA}, P^{SCFDMA}\}, \tag{3.1}$$

where:

- P^{chc}: channel coding,
- P^{chd}: channel decoding,
- P^{md}: modulation,
- P^{dm}: demodulation,
- P^{mpc}: MIMO precoding,
- P^{mdc}: MIMO decoding,
- P^{che}: channel estimation,
- P^{OFDMA}: OFDMA,
- P^{SCFDMA}: SC-FDMA.

In order to exploit parallel processing, the total processing in the BBU should be split into smaller portions, each allocated to a separate processor. Several levels of parallelization can be applied in the BBU processing. As a BS can handle the process of several users in a single sub-frame, a more obvious parallelization is to split the BS processing into any single user's process that each can perform independently and in parallel. The goal is to split the total processing of a BBU into smaller parallelized processing portions. Still, not all BS's processing steps can be classified per user. For example, IFFT/FFT and cyclic prefix insertion (OFDMA/SC-FDMA) cannot be split, as it consists of processing the signal resulting from the combination of all the users' signals. Therefore, the BS processing set, S^{Proc}, is classified into two categories:

- User-specific processing (UP): it includes the signal processing steps that can be split per user and its set, S^{UP}, containing channel de/coding and de/modulation, channel estimation and MIMO

de/precoding, such that

$$S^{UP} = \{P^{chc}, P^{chd}, P^{md}, P^{dm}, P^{mpc}, P^{mdc}, P^{che}\}. \tag{3.2}$$

The UP set for an individual user, depends on the user being in UL or DL. The users' processing in DL includes channel coding, modulation and MIMO precoding,

$$S_{DL}^{UP} = \{P^{chc}, P^{md}, P^{mpc}\}, \tag{3.3}$$

and for UL it includes channel decoding, demodulation, channel estimation and MIMO decoding,

$$S_{UL}^{UP} = \{P^{chd}, P^{dm}, P^{es}, P^{mdc}\}. \tag{3.4}$$

- Common processing (CP): it includes the common signal processing steps among all users for a given RRH and its set, S^{CP}, containing FFT/IFFT and cyclic prefix insertion (OFDMA, SC-FDMA),

$$S^{CP} = \{P^{OFDMA}, P^{SCFDMA}\}. \tag{3.5}$$

Similarly, the CP set for an individual user depends on the user being in UL or DL. The CP in DL includes OFDMA,

$$S_{DL}^{CP} = \{P^{OFDMA}\}, \tag{3.6}$$

and for UL it includes SC-FDMA,

$$S_{UL}^{CP} = \{P^{SCFDMA}\}. \tag{3.7}$$

Each BBU receives multiple UPs from the MAC layer. Figure 3.4 shows the process done in each codeword output from channel coding to the transmission antenna port on the bottom, for both DL and UL. The output of a processing step is fed as input to the next one. A TB passed down from the MAC layer to the PHY one is first channel coded, as depicted in Figure 3.4. In this model, the channel coding scheme is a combination of error detecting, CRC, error-correcting, rate matching and scrambling, as described in Section 2.1.3. The resulting encoded bits, i.e., codewords, are transformed in a corresponding block of modulation symbols, e.g., 64QAM and 256QAM. After that, the MIMO precoding takes place, which includes both MIMO encoding and layer/antenna mapping. Supplying different TM results in different mapping and precoding operation that is specific to it. Then, the resulting signal is mapped onto the time domain by OFDMA modulation, to be transmitted on each antenna port.

Data flow in UL has similar reverse steps. After removing the cyclic prefix, received signals first transformed into frequency domain by SC–FDMA demodulation (or OFDMA in 5G). Subsequently, channel estimation and MIMO decoding follows. Channel estimation aims to estimate the channel characteristics; the effect of the channel on the transmitted information is estimated in order to decode received signals correctly. After estimating the channel impulse response, received data blocks are MIMO decoded. The goal is to separate and detect the received symbols via MIMO antennas and reproduce original symbols faithfully. Now, received symbols are demodulated and decoded. These two processing steps invert the operations of modulation and channel coding in the transmitter side. The step tasks are described in more detail in Section 2.1.3.

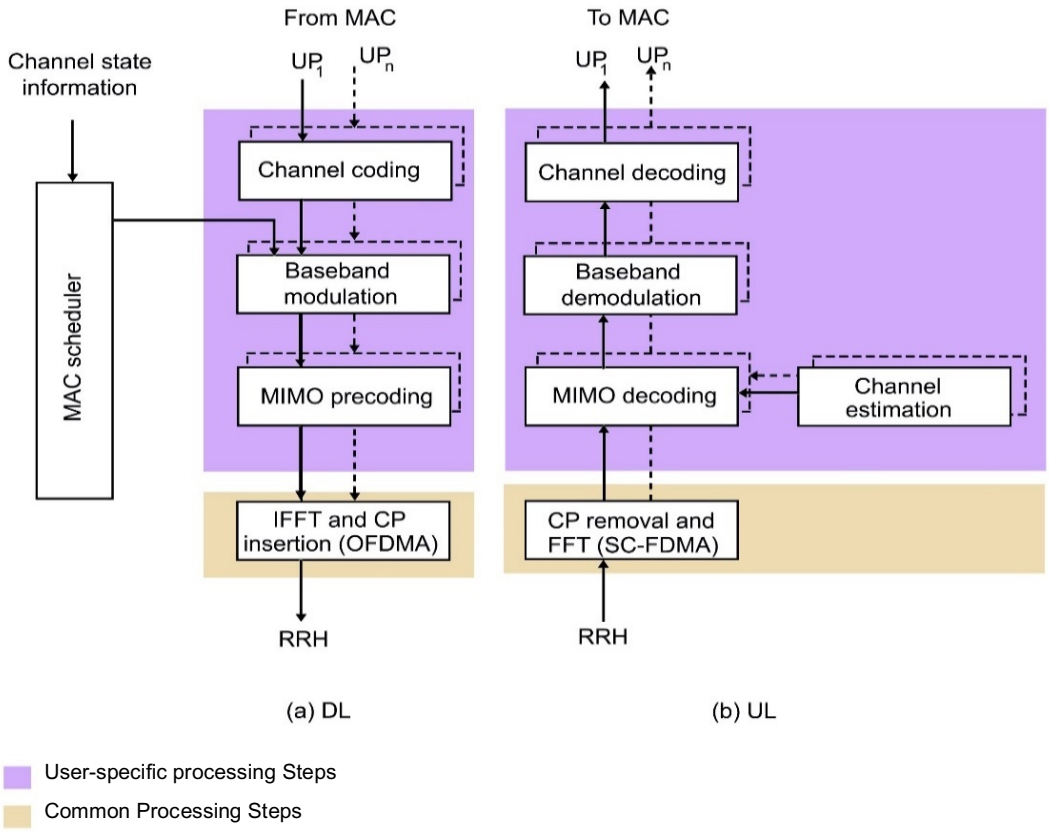


Figure 3.4 – A simplified block diagram for PHY layer process in a BBU.

The next subsection is dedicated to estimating the amount of computing capacity that each BS processing step requires.

3.3.2 BBU Required Computing Capacity

The RCC of a BBU is defined as the minimum computing capacity required to perform its instantaneous signal processing within a TTI. In order to achieve a BBU's RCC at a given time instant t_k, each of the processing step's RCC in the BBU should be calculated. In this thesis, the RCC estimation is based on the model proposed in [MAMM16], being performed by a function of parameters affecting the complexity of signal processing. The effective parameters are listed in the set S^{Parm},

$$S^{Parm} = \{\Delta f_{BW[\mathrm{MHz}]}, N_{MIMO}, Q_{[\mathrm{bit}]}, m_{u[\mathrm{bit/symbol}]}, r_u, N_{Str\ u,t_k}, \eta_u^{RBU}\}, \tag{3.8}$$

where:

- Δf_{BW} : channel bandwidth, e.g., $\Delta f_{BW} \in \{20, 40, 100\}[\mathrm{MHz}]$,
- N_{MIMO}: BS MIMO order, e.g., $N_{MIMO} \in \{1,2,4,8\}$,
- Q: quantization resolution, e.g., $Q \in \{16, 24\}[\mathrm{bit}]$,
- m_u: user u modulation, e.g., $m_u \in \{2,4,6,8,10\}[\mathrm{bit/symbol}]$, corresponding to QPSK, 16QAM, 64QAM, 256QAM and 1024QAM,
- r_u: user u coding ratio, e.g., $r_u \in [1/4, 1]$,
- $N_{Str\ u,t_k}$: user u number of spatial streams, up to the MIMO order,

- $\eta_u^{RB_U}$: user u RB efficiency, $\eta_u^{RB_U} \in [0,1]$.

RB efficiency, η^{RB}, is a parameter that affects the complexity of signal processing: for a single user u at time instant t_k, $\eta_{u,t_k}^{RB_U}$ is the fraction of available RBs in the bandwidth being allocated to the user, hence

$$\eta_{u,t_k}^{RB_U} = \frac{N_{Al\,u,t_k}^{RB}}{N_{\Delta f}^{RB}}, \tag{3.9}$$

where:

- $N_{Al\,u,t_k}^{RB}$: number of allocated RBs to the user u at time instant t_k,
- $N_{\Delta f}^{RB}$: total number of sub-frame's RBs in a given bandwidth, e.g., $N_{\Delta f}^{RB} = 200$, in a 20 MHz bandwidth.

The sum of all active DL/UL users' RB efficiencies in a BBU states the BBU's RB efficiency,

$$\eta_{DL|UL\,b,t_k}^{RB_B} = \sum_{\forall\, u \in S_{b,t_k}^{U_{DL|UL}}} \eta_{u,t_k}^{RB_U}, \tag{3.10}$$

where $S_{b,t_k}^{U_{DL|UL}}$ is the set of all active DL/UL users in BBU b at t_k. The network is fully loaded whenever the total number of allocated RBs is equal to the available ones in the bandwidth.

In order to estimate a BBU's RCC, a reference value is given to each of the effective parameters. Accordingly, an algorithm is selected for every signal processing step. The RCC of a UP/CP step is then acquired by counting the number of arithmetic operations that should be performed per information bit transmission. The reference values assigned to parameters $x \in S^{Parm}$ and the processing step RCCs obtained from them (which are named as the reference RCCs) are listed in Table 3.1 and Table 3.2, respectively.

The reference RCCs can then be scaled to any other desired value of $x \in S^{Parm}$. For each UP step $p \in S^{UP}$ of user u, scaling is given by

$$C_{u,p,t_k[\mathrm{GOPS}]}^{R_{usr}} = C_{p[\mathrm{GOPS}]}^{ref} \left(\frac{\Delta f_{BW[\mathrm{MHz}]}}{\Delta f_{BW[\mathrm{MHz}]}^{ref}}\right)^{E_{\Delta f_{BW},p}} \left(\frac{N_{MIMO}}{N_{MIMO}^{ref}}\right)^{E_{N_{MIMO},p}} \left(\frac{Q_{[\mathrm{bit}]}}{Q_{[\mathrm{bit}]}^{ref}}\right)^{E_{Q,p}} \left(\frac{m_{u,t_k[\mathrm{bit/symbol}]}}{m_{[\mathrm{bit/symbol}]}^{ref}}\right)^{E_{m,p}} \left(\frac{r_{u,t_k}}{r^{ref}}\right)^{E_{r,p}} \left(\frac{N_{Str\,u,t_k}}{N_{Str}^{ref}}\right)^{E_{N_{Str},p}} \left(\eta_{u,t_k}^{RB_U}\right)^{E_{\eta RB,p}}, \tag{3.11}$$

where:

- C_p^{ref}: processing step p's reference RCC, Table 3.2,
- Δf_{BW}^{ref}: channel bandwidth's reference value, where $\Delta f_{BW}^{ref} = 20$ MHz,
- N_{MIMO}^{ref}: MIMO order's reference value, where $N_{MIMO}^{ref} = 1$,
- Q^{ref}: quantization resolution's reference value, where $Q^{ref} \in \{16,24\}[\mathrm{bit}]$,
- m^{ref}: modulation order's reference value, where $m^{ref} = 6$ bit/symbol,
- r^{ref}: coding ratio's reference value, where $r^{ref} = 1$,
- N_{Str}^{ref}: the number of spatial streams' reference value, where $N_{Str}^{ref} = 1$,

- $E_{BW,p}$: bandwidth's scaling exponent for the processing step p's RCC, Table 3.2,
- $E_{MIMO,p}$: MIMO order's scaling exponent for the processing step p's RCC, Table 3.2,
- $E_{Q,p}$: quantization resolution's scaling exponent for the processing step p's RCC, Table 3.2,
- $E_{m,p}$: modulation's scaling exponent for the processing step p's RCC, Table 3.2,
- $E_{r,p}$: coding ratio's scaling exponent for the processing step p's RCC, Table 3.2,
- $E_{Str,p}$:the number of spatial streams' scaling exponent for the processing step p's RCC, Table 3.2,
- $E_{\eta RB,p}$: RB efficiency's scaling exponent for the processing step p's RCC, Table 3.2.

Table 3.1 – Reference values for signal processing's effective parameters (based on [DeDL15]).

Parameter $(x \in S^{Parm})$	$\Delta f_{BW[\mathrm{MHz}]}$	$m_{u[\mathrm{bit/symbol}]}$	r_u	$N_{Str\,u}$	N_{MIMO}	$Q_{[\mathrm{bit}]}$	$\eta_b^{RB_B}$
Reference Value (x^{ref})	20	6	1			16, 24	1

Table 3.2 – Reference RCCs and scaling exponents (based on [DeDL15]).

BS Processing Step $(p \in S^{Proc})$	**Reference RCC** $(C^{ref}_{p[\mathrm{GOPS}]})$	**Effective Parameter** $(x \in S^{Parm})$ **Scaling Exponent** $(E_{x,p})$						
		Δf_{BW}	m_u	r_u	$N_{Str\,u}$	N_{MIMO}	Q	η^{RB}
p^{SCFDMA}	2.7	1.0	-			1.0	1.2	0.5
p^{OFDMA}	1.3							
p^{che}	3.3		0.0		1.0			
p^{mpc}	1.3							1.0
p^{mdc}	2+3.3 N_{MIMO}				0.0	2.0		
p^{dm}	2.7		1.5		1.0	0.0		
p^{md}	1.3							
p^{chd}	8.0		1.0					
p^{chc}	1.3							

Equation (3.11) is applicable for all UP steps in both UL and DL. Once the RCC of any single UP step in DL/UL is achieved, it is possible to calculate the total RCC of a DL/UL user u, $C^{R_{usr}}_{DL|UL\ u,t_k[\mathrm{GOPS}]}$, by summing up the achieved values, therefore,

$$C^{R_{usr}}_{DL|UL\ u,t_k[\mathrm{GOPS}]} = \sum_{p \in S^{UP}_{DL|UL}} C^{R_{usr}}_{u,p,t_k[\mathrm{GOPS}]}. \tag{3.12}$$

Summing up all the active DL/UL users' RCCs gives the BBU's total UP's RCC in DL/UL,

$$C^{R}_{DLu|ULu\ b,t_k[\mathrm{GOPS}]} = \sum_{u \in S^{U_{DL|UL}}_{b,t_k}} C^{R_{usr}}_{DL|UL\ u,t_k[\mathrm{GOPS}]}. \tag{3.13}$$

Finally, the total RCC required for UP in a BBU is achieved by summing all DL and UL ones in the BBU,

$$C^{R}_{U\ b,t_k[\mathrm{GOPS}]} = C^{R}_{DLu\ b,t_k[\mathrm{GOPS}]} + C^{R}_{ULu\ b,t_k[\mathrm{GOPS}]}. \tag{3.14}$$

On the other hand, the CP of a BBU b at time instant t_k for CP in DL/UL is estimated by:

$$C^{R}_{DLc|ULc\ b,t_k[\mathrm{GOPS}]} = \sum_{p\in S^{CP}_{DL|UL}} C^{ref}_{p[\mathrm{GOPS}]} \left(\frac{\Delta f_{BW[\mathrm{MHz}]}}{\Delta f^{ref}_{BW[\mathrm{MHz}]}}\right)^{E_{\Delta f_{BW},p}} \left(\frac{Q_{[\mathrm{bit}]}}{Q^{ref}_{[\mathrm{bit}]}}\right)^{E_{Q,p}} \left(\frac{N_{MIMO}}{N^{ref}_{MIMO}}\right)^{E_{N_{MIMO},p}} \left(\eta^{RB_B}_{DL|UL\ b,t_k}\right)^{E_{\eta RB,p}}. \tag{3.15}$$

Accordingly, the total RCC for CP in BBU b, $C^{R}_{C\ b,t_k[\mathrm{GOPS}]}$, is

$$C^{R}_{C\ b,t_k[\mathrm{GOPS}]} = C^{R}_{DLc\ b,t_k[\mathrm{GOPS}]} + C^{R}_{ULc\ b,t_k[\mathrm{GOPS}]}. \tag{3.16}$$

Finally, summing up all of the UP RCCs and CP RCCs of BBU b at time instant t_k, total RCC of the BBU is achieved,

$$C^{R}_{b,t_k[\mathrm{GOPS}]} = C^{R}_{U\ b,t_k[\mathrm{GOPS}]} + C^{R}_{C\ b,t_k[\mathrm{GOPS}]}. \tag{3.17}$$

In case that several users are active in the BBU, the computing capacity required for the CP should be met, otherwise, none of the user's data can be transferred. Therefore, computing capacity that is required for a BBU's CP is its minimum demand that should be guaranteed,

$$C^{R_{\min}}_{b,t_k[\mathrm{GOPS}]} = C^{R}_{C\ b,t_k[\mathrm{GOPS}]} \quad : \quad b = 1,2,\dots,N_B. \tag{3.18}$$

In case that no user is active in the BBU at a given time instant, $C^{R_{\min}}_{b,t_k}$ is equal to zero. On the contrary, a BBU's RCC is in the peak level if the network is fully loaded, users have the highest MCS, and all users' spatial streams are equal to the MIMO order, based on (3.11). Therefore, a BBU's peak RCC is

$$C^{R_{PEAK}}_{b[\mathrm{GOPS}]} = \sum_{p\in S^{Proc}} C^{ref}_{p} \left(\frac{\Delta f_{BW[\mathrm{MHz}]}}{\Delta f^{ref}_{BW[\mathrm{MHz}]}}\right)^{E_{\Delta f_{BW},p}} \left(\frac{N_{MIMO}}{N^{ref}_{MIMO}}\right)^{E_{N_{MIMO},p}} \left(\frac{Q_{[\mathrm{bit}]}}{Q^{ref}_{[\mathrm{bit}]}}\right)^{E_{Q,p}} \left(\frac{m_{\max[\mathrm{bit/symbol}]}}{m^{ref}_{[\mathrm{bit/symbol}]}}\right)^{E_{m,p}} \left(\frac{N_{MIMO}}{N^{Str}_{ref}}\right)^{E_{N_{Str},p}}. \tag{3.19}$$

where $m_{\max}$ is the order of the highest modulation scheme for wireless technologies, i.e., 1024QAM at this time, as discussed in [3GPP17] for upcoming technologies.

A BBU's RCC at a given time instant can also be classified, per UL and DL connections,

$$C^{R}_{DL|UL\ b,t_k[\mathrm{GOPS}]} = C^{R}_{DLu|ULu\ b,t_k[\mathrm{GOPS}]} + C^{R}_{DLc|ULc\ b,p,t_k[\mathrm{GOPS}]}. \tag{3.20}$$

3.4 QoS-Demand-Aware Computing Resources Allocation

3.4.1 Overview

As mentioned in Section 2.5, C-RAN integrates the BBUs of multiple BSs in a BBU-pool and increases the network resources' utility by multiplexing BBU resources in the pool. Resources multiplexing enables over-loaded BBUs to use residual resources left by underutilized ones. Hence, utilization is improved, and fewer resources are required than the sum of stand-alone BBU demands. To take advantage of C-RAN benefits, efficient strategies should be applied in order to distribute the resources of the BBU-pool among BBUs. An efficient resource provisioning scheme should minimize the resource idle times and the BBUs' over-loading.

However, designing efficient resource management strategies is a complicated process for cloud providers. Due to the variety of network services, user arrival rates and channel conditions, BBU demand fluctuates significantly throughout the day. On the one hand, a BBU computing capacity should suffice peak demands; on the other hand, provisioning fixed resources based on peak requirements leads to idle resources for the rest of the day.

As a result, an efficient resource management strategy in a BBU-pool should allocate the computing capacity dynamically, following the BBUs' instantaneous demand, while efficiently handling the resources in the case of a shortage. Resource shortages are time instants in which the BBU-pool's available resources are less than demand spikes and come into play in two circumstances: when the objective is intentionally to design the pool with minimum computing resources; or, even if there are more computing resources, they cannot be initialized at a rate similar to the one of demand fluctuations (in the scale of milliseconds), due to hardware limitations.

In this section, a QoS-Demand-aware computing resources Allocation Scheme (QDAS) is proposed to solve the resource allocation problem. The proposed model is built on the concept of NBS in cooperative game theory. The BBU-pool computing resources allocation in a single time instant is modeled as a cooperative bargaining game. The players, i.e., BBUs, are trying to reach an agreement that gives a mutual advantage. During the bargaining game, each BBU is assigned a utility function and a bargaining power: the utility function of a BBU expresses the portion of the BBU's demand that is served, and the achievement of its bargaining power is based on the priority level of the active services in the BBU. Each BBU tries to request more computing resources to maximize its utility to speed up its processing time. Meanwhile, the BBU claims a minimum number of computing resources at a given time instant. GNBS is applied in order to find optimal computing resource allocation that maximizes the BBU utilities while the utility of the entire BBU-pool is considered. As mentioned in Section 2.6, GNBS is a very effective tool to model interactions among negotiators that guarantees all players to acquire the maximum utility with fair concerns.

An abstract view of the proposed model for computing resource management is presented in Figure 3.5. The resource allocation module receives average weight of BBUs' active services, RCCs, minimum guaranteed RCCs and the BBU-pool's AvCC as inputs. It calculates BBUs' bargaining powers in the

first step according to their estimated RCCs and the priority of their active services. Together with the BBU-pool's AvCC and BBUs' minimum guaranteed requirements, the results are fed to the next step in order to find the optimal computing resources allocation.

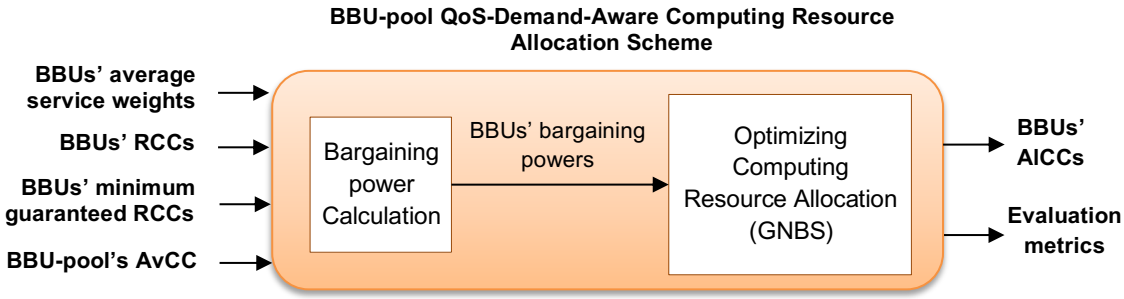


Figure 3.5 – Abstract view of QoS-demand-aware computing resource allocation scheme.

In what follows, modelling the computing resources allocation problem as a bargaining problem and applying GNBS to find the defined problem's solution is described in detail.

3.4.2 Definition of Utility Functions

The problem of finding an efficient resource allocation in the BBU-pool is comparable with a bargaining game in cooperative game-theory [Myer91]. BBUs are counted as players negotiating over a limited number of computing resources of the BBU-pool to increase their processing capacities, while taking resource utilization maximization as a mutual benefit. The outcome is an agreement on selecting one resource allocation strategy, i.e., a feasible solution from many possible choices. A resource allocation strategy at a certain time instant t_k is given by vector $\mathbf{C}^{Al}_{t_k\,[N_B\times 1]}$,

$$\mathbf{C}^{Al}_{t_k} = [C^{Al}_{1,t_k[\mathrm{GOPS}]}, C^{Al}_{2,t_k[\mathrm{GOPS}]}, \dots, C^{Al}_{N_B,t_k[\mathrm{GOPS}]}]^{\mathrm{T}}, \tag{3.21}$$

where:

- C^{Al}_{b,t_k}: BBU b AICC at time instant t_k,
- N_B: number of BBUs in the pool.

Each BBU evaluates its preference over a selected strategy by its utility function individually. BBU b's utility is defined by a function $\mathcal{U}_{b,t_k}: \mathbb{R}^{N_B} \to \mathbb{R}$ reflecting the portion of the BBU's request that is satisfied,

$$\mathcal{U}_{b,t_k}(\mathbf{C}^{Al}_{t_k}) = \frac{C^{Al}_{b,t_k[\mathrm{GOPS}]}}{C^{R}_{b,t_k[\mathrm{GOPS}]}}. \tag{3.22}$$

If the total computing demand is less than the available resources in the BBU-pool, then all BBUs' demands are satisfied; otherwise, a compromise solution is selected in which the minimum guaranteed RCC of BBUs, $\mathbf{C}^{R_{\min}}_{t_k\,[N_B\times 1]}$, are served,

$$\mathbf{C}^{R_{\min}}_{t_k} = [C^{R_{\min}}_{1,t_k[\mathrm{GOPS}]}, C^{R_{\min}}_{2,t_k[\mathrm{GOPS}]}, \dots, C^{R_{\min}}_{N_B,t_k[\mathrm{GOPS}]}]^{\mathrm{T}}. \tag{3.23}$$

During the bargaining, BBUs attempt to get more computing resources to increase their utility. However,

three limitations are imposed:

1) The total AlCC in a feasible solution should not exceed the BBU-pool's AvCC,

$$C_{BP\ t_k[\mathrm{GOPS}]}^{Al} \leq C_{BP\ t_k[\mathrm{GOPS}]}^{Av}, \tag{3.24}$$

where $C_{BP\ t_k}^{Av}$ is the BBU-pool AvCC at t_k, and

$$C_{BP\ t_k[\mathrm{GOPS}]}^{Al} = \sum_{b=1}^{N_B} C_{b,t_k[\mathrm{GOPS}]}^{Al} \quad : \quad b = 1,2,\dots,N_B. \tag{3.25}$$

2) The resource allocator must provide the minimum guaranteed RCC of an individual BBU,

$$C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}} < C_{b,t_k[\mathrm{GOPS}]}^{Al} \quad : \quad b = 1,2,\dots,N_B. \tag{3.26}$$

3) Each BBU may not ask for more capacity than its RCC at a specific time,

$$C_{b,t_k[\mathrm{GOPS}]}^{Al} \leq C_{b,t_k[\mathrm{GOPS}]}^{R} \quad : \quad b = 1,2,\dots,N_B \tag{3.27}$$

As a result of these critical constraints, the feasible solution set is bounded as

$$S_{t_k}^{FS} = \left\{ \mathbf{C}_{t_k}^{Al} \mid \sum_{b=1}^{N_B} C_{b,t_k}^{Al} \leq C_{BP\ t_k}^{Av},\ C_{b,t_k}^{R_{\min}} < C_{b,t_k}^{Al} \leq C_{b,t_k}^{R} \right\}. \tag{3.28}$$

$S_{t_k}^{FS}$ is a subset of $\mathbb{R}^{N_B}$ that contains all the feasible solutions of the bargaining problem. The goal is to find the optimal solution that maximizes the BBUs utility functions while satisfying fairness among them. As mentioned in Section 2.6, NBS is suitable for solving the problem and guarantees to find the optimal solution if the BBUs utility functions as well as the defined solution set are convex and closed [Binm91]. $S_{t_k}^{FS}$ is a convex set, because the line segment between any desired pair of points in the set lies entirely within the set and the convexity conditions is hold also for $\mathcal{U}_{b,t_k}$ (a prof is given in Annex A). Since both $\mathcal{U}_{b,t_k}$ and $S_{t_k}^{FS}$ are convex, the pair $(S_{t_k}^{FS} \cup \{\mathbf{C}_{t_k}^{R_{\min}}\}, \boldsymbol{\mathcal{U}}_{t_k}(\mathbf{C}_{t_k}^{Al}))$ defines the bargaining problem for the computing resource allocation in a BBU-pool [Myer91],

$$\boldsymbol{\mathcal{U}}_{t_k}(\mathbf{C}_{t_k}^{Al}) = \left[\mathcal{U}_{1,t_k}(\mathbf{C}_{t_k}^{Al}), \mathcal{U}_{2,t_k}(\mathbf{C}_{t_k}^{Al}), \dots, \mathcal{U}_{N_B,t_k}(\mathbf{C}_{t_k}^{Al})\right]^{\mathrm{T}}. \tag{3.29}$$

Equation (3.29) shows a vector function in which each component function, $\mathcal{U}_{b,t_k}(\mathbf{C}_{t_k}^{Al})$, presents the utility function of BBU b.

3.4.3 Bargaining Power

In order to increase the user's satisfaction level, an efficient resource allocator should be able to support QoS constraints. To fulfill QoS requirements and to improve efficiency and fairness, an individual BBU is assigned bargaining powers, which are composed of the BBUs' RCCs and their active services' average weight, being used as the power factor for the BBUs that reflect their priority while allocating resources. Services' weights result from the normalization of Priority Level of services defined in 3GPP, Table 2.6 and Table 2.8, in the range of [1, 100]. The rationale behind it is that the Priority Level is a

characteristic by which 3GPP specifies QoS requirements and determines the packet forwarding treatment, Section 2.3. The weight of service s is

$$w_s^{srv} = 1 + \frac{99\,(P_{\max}^{srv} - P_s^{srv})}{(P_{\max}^{srv} - P_{\min}^{srv})}, \tag{3.30}$$

where:

- P_s^{srv}: the Priority Level of service s, given by 3GPP, Table 2.6 and Table 2.8,
- $P_{\min|\max}^{srv}$: the minimum/maximum of 3GPP service Priority Levels,
- 99 is used as a normalization factor, being the difference between the maximum and the minimum parameters' values.

Accordingly, the average weight of ongoing services in a BBU is denoted as

$$\overline{w_{b,t_k}^{srv}} = \frac{\sum_{s=1}^{N^{srv}} N_{S\,b,s,t_k}^{U}\, w_s^{srv}}{N_{b,t_k}^{U}}, \tag{3.31}$$

where:

- $N_{S\,b,s,t_k}^{U}$: number of users of service s in BBU b at t_k,
- N_{b,t_k}^{U}: total number of users in BBU b at t_k.

Finally, the combination of BBUs' RCCs and the services' average weight defines the BBU bargaining powers as

$$B_{b,t_k} = \frac{\overline{w_{b,t_k}^{srv}}\left(C_{b,t_k[\mathrm{GOPS}]}^{R} - C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}\right)}{\sum_{l=1}^{N_B}\left(\overline{w_{l,t_k}^{srv}}\left(C_{l,t_k[\mathrm{GOPS}]}^{R} - C_{l,t_k[\mathrm{GOPS}]}^{R_{\min}}\right)\right)}. \tag{3.32}$$

A BBU bargaining power is a positive value within [0, 1], so that in a time instant one has

$$\sum_{b=1}^{N_B} B_{b,t_k} = 1. \tag{3.33}$$

Equation (3.32) implies that once the minimum guaranteed RCC is allocated to BBUs, i.e., $C_{b,t_k}^{R_{\min}}$, the rest of the resources are distributed such that QoS is maintained, hence, services with a higher priority should be allocated with more resources. In this context, in the next sections, maintaining QoS is equivalent to BBU prioritization based on service weights.

3.4.4 Generalized Nash Bargaining Solution

By modelling the BBU-pool computing resource allocation as a bargaining game, the GNBS can be used as the unique fair Pareto optimal solution among all feasible ones existing in $S_{t_k}^{FS}$. GNBS satisfies Nash axioms as the attributes that any rational solution should meet to come up with fairness and efficiency, and is achieved by maximizing the product of the BBU utility functions weighted by the BBU bargaining powers [Myer91], [Binm91]. By defining $\mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right)$ as the utility function of the BBU-pool,

$$\mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right) = \prod_{b=1}^{N_B}\left(\mathcal{U}_{b,t_k}\left(\mathbf{C}_{t_k}^{Al}\right) - \mathcal{U}_{b,t_k}\left(\mathbf{C}_{t_k}^{R_{\min}}\right)\right)^{B_{b,t_k}} = \prod_{b=1}^{N_B}\left(\frac{C_{b,t_k[\mathrm{GOPS}]}^{Al} - C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}}{C_{b,t_k[\mathrm{GOPS}]}^{R}}\right)^{B_{b,t_k}}, \tag{3.34}$$

GNBS provides a unique solution $\mathbf{C}_{t_k}^{Al^*}$ for the defined bargaining game by solving the following optimization problem:

$$\mathbf{C}_{t_k}^{Al^*} = \underset{\forall \boldsymbol{C}_{t_k}^{\boldsymbol{Al}} \in S_{t_k}^{FS} \cup \left\{\mathbf{C}_{t_k}^{R_{\min}}\right\}}{\operatorname{argmax}} \left(\mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right)\right). \tag{3.35}$$

For clarity, the process of computing resource allocation in a BBU-pool is shown in Figure 3.6. Given the BBUs' RCCs, minimum guaranteed RCCs, the average weight of ongoing services and the BBU-pool's AvCC as inputs, all BBU bargaining powers are calculated in the first step, line 2. In the case that the total resource demand is less than or equal to the AvCC, all BBUs are allocated with the computing resources fulfilling their demands, line 5; otherwise, GNBS is achieved as an optimal compromise solution by solving (3.35), line 7.

Input: $\mathbf{C}_{t_k}^{R}, \mathbf{C}_{t_k}^{R_{\min}}, \overline{\boldsymbol{w}_{t_k}^{srv}}, C_{BP\ t_k}^{Av}$

Output: $\mathbf{C}_{t_k}^{Al^*}$

1: **For** b = 1 **to** N_B **do**

2: $\quad B_{b,t_k} \leftarrow \left(\overline{w_{b,t_k}^{srv}}\left(C_{b,t_k}^{R} - C_{b,t_k}^{R_{\min}}\right)\right) / \sum_{l=1}^{N_B}\left(\overline{w_{l,t_k}^{srv}}\left(C_{l,t_k}^{R} - C_{l,t_k}^{R_{\min}}\right)\right)$

3: **end for**

4: **If** $\sum_{b=1}^{N_B} C_{b,t_k}^{R} \leq C_{BP\ t_k}^{Av}$

5: $\quad \mathbf{C}_{t_k}^{Al^*} \leftarrow \mathbf{C}_{t_k}^{R}$

6: **else**

7: $\quad \mathbf{C}_{t_k}^{Al^*} \leftarrow \underset{\forall \mathbf{C}_{t_k}^{Al} \in S_{t_k}^{FS} \cup \left\{\mathbf{C}_{t_k}^{R_{\min}}\right\}}{\operatorname{argmax}} \left(\mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right)\right)$

8: **end if**

9: **return** $\mathbf{C}_{t_k}^{Al^*}$

Figure 3.6 – Algorithm for QoS-demand-aware computing resource allocation at time instant t_k.

In order to solve (3.35), first the following optimization problem is put forward:

$$\underset{\boldsymbol{C}_{t_k}^{\boldsymbol{Al}}}{\text{maximize}} \quad \mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right), \tag{3.36a}$$

subject to

$$\sum_{b=1}^{N_B} C_{b,t_k}^{Al} \leq C_{BP\ t_k}^{Av}, \tag{3.36b}$$

$$C_{b,t_k}^{R_{\min}} < C_{b,t_k}^{Al} \quad : \ b = 1,2,\ldots,N_B, \tag{3.36c}$$

$$C_{b,t_k}^{Al} \leq C_{b,t_k}^{R} \quad : \ b = 1,2,\ldots,N_B, \tag{3.36d}$$

where (3.36b) to (3.36d) take the constraints given in $S_{t_k}^{FS}$ into account. The objective function is then transformed into the logarithmic form in order to facilitate problem solving; due to the monotonic behavior of the logarithm function, the logarithm of $\mathcal{U}_{BP}(\mathbf{C}_{t_k}^{Al})$ does not change the result [BoVa04]; moreover, since C_{b,t_k}^{R} is a constant value for BBU b at a given time instant t_k, it has no effect on the final optimization result, hence, it can be eliminated. Therefore, the objective function can be rewritten as:

$$\mathcal{U}_{LBP}(\mathbf{C}_{\mathrm{t_k}}^{Al}) = \sum_{b=1}^{N_B} B_{b,t_k} \log\left(C_{b,t_k}^{Al} - C_{b,t_k}^{R_{\min}}\right) \tag{3.37}$$

$\mathcal{U}_{LBP}(\mathbf{C}_{\mathrm{t_k}}^{Al})$ tends to -∞ when C_{b,t_k}^{Al} approaches $C_{b,t_k}^{R_{\min}}$, hence, constraint (3.36c) is automatically satisfied, and it can be relaxed, the optimization problem being rewritten as:

$$\underset{\mathbf{C}_{t_k}^{Al}}{\text{maximize}} \quad \mathcal{U}_{LBP}(\mathbf{C}_{t_k}^{Al}), \tag{3.38a}$$

subject to

$$\sum_{b=1}^{N_B} C_{b,t_k}^{Al} \leq C_{BP\ t_k}^{Av}, \tag{3.38b}$$

$$C_{b,t_k}^{Al} \leq C_{b,t_k}^{R} \quad : \quad b = 1,2,\dots,N_B. \tag{3.38c}$$

Equation (3.38) is convex, since all constraints are linear inequalities and the objective function is the sum of the concave functions [BoVa04], therefore, it has a unique optimal solution. One can find a detailed discussion on solving the problem in [KhLL14] with linear time complexity in the order of $O(N_B)$.

3.5 Equal and Demand-Proportional Resource Allocation

In order to evaluate the performance of the proposed model, other resource allocation schemes found in the literature were also implemented, hence, enabling a comparison, the equal and demand-proportional allocation approaches having been taken, [FMPS20], [KoSo19] and [KeMT98]. These resource allocation schemes do not provide any optimization, still they serve as a good comparison:

- **Equal resource Allocation Scheme (EAS)**: An overview of the equal computing resource allocation scheme is illustrated in Figure 3.7. EAS is a simple method that, given the AvCC of a BBU-pool, equally distributes computing resources among BBUs, regardless of the BBUs' demands and active services' priority,

$$C_{b,t_k[\mathrm{GOPS}]}^{Al_{EAS}} = \frac{C_{BP\ t_k[\mathrm{GOPS}]}^{Av}}{N_B}. \tag{3.39}$$

Figure 3.7 – Overview of equal computing resource allocation scheme.

- **Demand-proportional resource Allocation Scheme (DAS)**: An overview of demand-proportional computing resource allocation scheme is illustrated in Figure 3.8. Given the BBU-pool's AvCC and BBUs' RCCs and minimum guaranteed RCCs, DAS ensures minimum guaranteed resources to each BBU and distributes the remaining resources among them proportionally to their user processing demands

$$C_{b,t_k[\mathrm{GOPS}]}^{Al_{DAS}} = \left(C_{BP\ t_k[\mathrm{GOPS}]}^{Av} - \sum_{l=1}^{N_B} C_{l,t_k[\mathrm{GOPS}]}^{R_{\min}}\right) \frac{C_{U\ b,t_k[\mathrm{GOPS}]}^{R}}{\sum_{l=1}^{N_B} C_{U\ l,t_k[\mathrm{GOPS}]}^{R}} + C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}. \tag{3.40}$$

DAS is more complex than EAS, since BBUs' demands should be achieved before resource provisioning.

Figure 3.8 – Overview of demand-proportional computing resource allocation scheme.

The mentioned computing resource allocation schemes are compared with more details in Section 6.2.

3.6 Evaluation Metrics

As explained before, resource management aims to enhance resource utilization while maintaining QoS, for which resource allocation should uphold the priority of ongoing services. Different metrics are defined in this section in order to assess the performance of the proposed model. The metrics are explained in detail in what follows:

- **BBU fulfilment level:** a value within [0, 1] measuring the fraction of BBU b's UP RCC that is satisfied without any processing delay,

$$f_{b,t_k}^{B} = \frac{\left(\min\{C_{b,t_k[\mathrm{GOPS}]}^{Al}, C_{b,t_k[\mathrm{GOPS}]}^{R}\}\right) - C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}}{C_{U\ b,t_k[\mathrm{GOPS}]}^{R}}, \tag{3.41}$$

higher values of f_{b,t_k}^{B} indicate that a larger portion of the BBU's UP demands is met.

- **Fairness index:** a parameter that compares the fairness of the proposed resource allocation scheme with Jain's fairness indicator [JaCH84]. The fairness index is applied in the resource shortages when total RCC exceeds the amount of BBU-pool's AvCC at a time instant,

$$F_{t_k} = \frac{\left(\sum_{b=1}^{N_B} \frac{f_{b,t_k}^{B}}{w_{b,t_k}^{srv}}\right)^2}{N_B \sum_{l=1}^{N_B} \left(\frac{f_{l,t_k}^{B}}{w_{l,t_k}^{srv}}\right)^2}. \tag{3.42}$$

The range of F_{t_k} is within $[1/N_B, 1]$, which defines the closeness of the BBUs' fulfilment level to the average weight of the BBUs' active services. The higher the value of F_{t_k}, the higher the fairness of the resource allocation is.

- **Resource usage:** a value within [0, 100] % indicating the proportion of BBU-pool's AICC used for signal processing, to the BBU-pool's existing computing capacity,

$$U_{t_k[\%]} = \frac{\sum_{b=1}^{N_B} \min\{C_{b,t_k[\mathrm{GOPS}]}^{Al}, C_{b,t_k[\mathrm{GOPS}]}^{R}\}}{C_{BP\ [\mathrm{GOPS}]}} 100, \tag{3.43}$$

where C_{BP} is the BBU-pool's existing resources; higher values of U_{t_k} indicate a lower resource wastage, i.e., a larger portion of the existing computing capacity is used.

- **Dynamic resource allocation efficiency**: achieved by comparing the total amount of AICC that the model suggests to a BBU-pool in a specific time instant, with the traditional approaches that allocate a static amount of computing capacity to the BBU-pool according to the BBUs' RCCs in peak hours. Hence, the efficiency of dynamic resource allocation can be quantified by

$$\eta_{t_k[\%]} = \left(1 - \frac{C_{BP\ t_k[\mathrm{GOPS}]}^{Al}}{\sum_{b=1}^{N_B} C_{b[\mathrm{GOPS}]}^{R_{PEAK}}}\right) 100 \tag{3.44}$$

η_{t_k} is a value within [0, 100] %, with the higher values indicating a more efficient resource allocation among the BBUs in the BBU-pool.

3.7 Model Implementation

3.7.1 Implementation Overview

The model implementation is explained in detail in this section. Figure 3.9 presents the flowchart of the simulation process. Given the input parameters for a single time instant t_k, the BBUs' RCCs are estimated in the first step. As mentioned in Section 3.3, the number of RBs that a user requires, $N_{u,t_k}^{RB_U}$, the users' modulation, m_{u,t_k}, and coding ratio, r_{u,t_k}, are three key parameters in the BBUs' RCC estimation process. Based on the user's SNR, γ_{u,t_k}, an MCS index is proposed that maximizes the user's throughput. The user's packet volume, V_{u,t_k}^{Pkt}, and MCS are two critical parameters in the estimation of the number of its required RBs. The process of mapping SNR to MCS and extracting the number of required RBs for data transfer is explained in detail in the next subsection.

Once $N_{u,t_k}^{RB_U}$, m_{u,t_k}, and r_{u,t_k} are achieved, the BBU's RCCs is estimated as explained in Section 3.3. In the process of RCC estimation, the RCC of both CP and UP are achieved. When all users' RCCs are estimated, the achieved values and the other required parameters are fed to the next step to calculate

the BBUs' bargaining powers according to Section 3.4.3. Just after the bargaining powers calculation, the computing resource allocation in the BBU-pool is achieved. In case the resource allocation scheme in the BBU-pool is the QDAS, the optimum computing resource allocation is acquired by solving (3.38). In this step, (3.38) is solved using CVX (a modelling system for constructing and solving disciplined convex programs, developed by Stanford University [CVX20]), The outputs of the optimization step are the optimum computing resources allocation and values of performance metrics defined in Section 3.6.

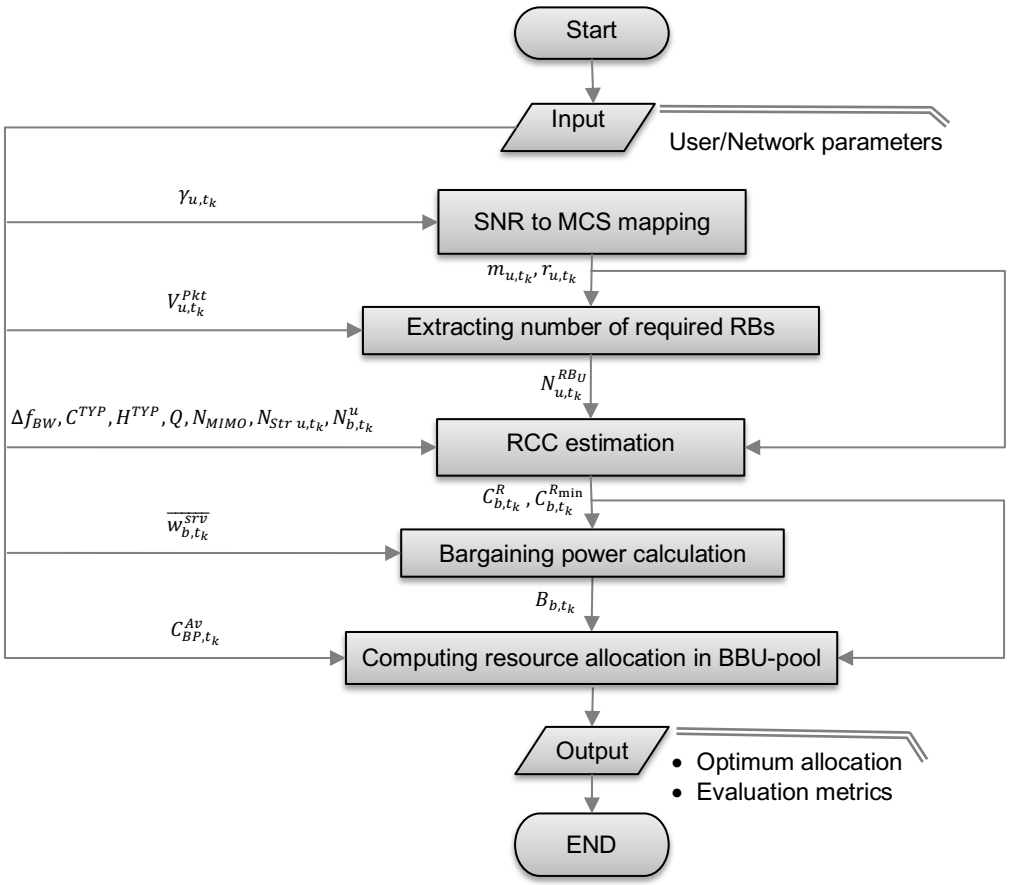


Figure 3.9 – Flowchart of the model implementation.

3.7.2 Extracting Number of Required Resource Blocks

The number of RBs that a user requires at a given time instant, $N_{R\ u,t_k}^{RB}$, depends on the user's MCS and the packet size that is going to be transferred. The simulation process for extracting the number of user u's required RBs for a packet transfer is summarized in Figure 3.10, supported on Table 3.3 and Table 3.4.

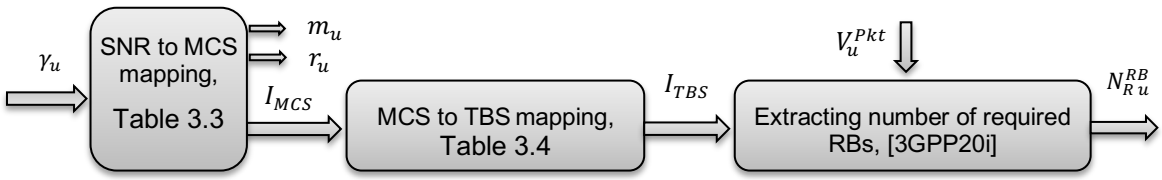


Figure 3.10 – The procedure for extracting number of required RBs.

Table 3.3 – Modulation, coding ratio and required SNR for LTE and 5G (extracted from [3GPP17]).

	SNR [dB]	m_u	r_u	Efficiency	I_{MCS}
0	$\gamma < 0$	No transmission			
1	$0 < \gamma < 1.8$	QPSK	0.08	0.15	0
2	$1.8 \leq \gamma < 4.5$		0.44	0.88	1
3	$4.5 \leq \gamma < 8.2$	16QAM	0.37	1.48	3
4	$8.2 \leq \gamma < 10.1$		0.60	2.41	5
5	$10.1 \leq \gamma < 11.9$	64QAM	0.46	2.73	7
6	$11.9 \leq \gamma < 13.8$		0.55	3.32	9
7	$13.8 \leq \gamma < 15.6$		0.65	3.90	11
8	$15.6 \leq \gamma < 17.5$		0.75	4.52	13
9	$17.5 \leq \gamma < 19.5$		0.85	5.12	15
10	$19.5 \leq \gamma < 21.1$	256QAM	0.69	5.55	17
11	$21.1 \leq \gamma < 23.2$		0.78	6.23	19
12	$23.2 \leq \gamma < 25$		0.86	6.91	21
13	$25 \leq \gamma < 27.8$		0.93	7.41	22
14	$27.8 \leq \gamma < 30$	1024QAM	0.81	8.12	24
15	$30 \leq \gamma$		0.89	8.87	26

Table 3.4 – TB size for LTE and 5G (extracted from [3GPP17]).

I_{MCS}	m_u	I_{TBS}
0	QPSK	0
1		4
2		8
3	16QAM	11
4		13
5		14
6		15
7	64QAM	16
8		17
9		18
10		19
11		20
12		21
13		22
14		23
15		24
16	256QAM	25
17		27
18		28
19		29
20		30
21		31
22		32
23		33
24	1024QAM	35
25		36
26		37
27	Reserved	
28		
29		
30		
31		

According to a user's SNR at time instant t_k, an MCS index, I_{MCS}, is proposed to maximize the user's throughput. Although the association between measured SNR and MCS index selected by eNB/gNB are vendor specific, for the sake of simplicity, however, Table 3.3 is used in this thesis, in order to map measured SNR onto MCS and to map the recommended MCS onto the I_{MCS}. Given I_{MCS}, the Transport Block Size (TBS) index, I_{TBS} is extracted from a given lookup table, i.e., Table 3.4. Finally, I_{TBS} and the packet volume are used in order to retrieve the number of required RBs for packet transmission, $N_{R\,u}^{RB}$. Mapping I_{TBS} onto the number of required RBs is according to a lookup table proposed by 3GPP [3GPP20i], the process being the same for both UL and DL.

3.8 Canonical Scenario

A simple urban scenario is defined in this section to evaluate the proposed computing resource management model's performance and show the functional correctness of the implemented algorithm. The test environment is composed of two residential micro-cells with continuous coverage. Each RRH is connected to a BBU inside the BBU-pool. It is assumed that both BBUs are clustered in the same BBU-pool. The BSs are assumed with a 4×4 MIMO order and work with 20 MHz channel bandwidth.

The number of users is assumed to vary within [1, 100]. Users are assumed outdoors, distributed uniformly over the whole area, all with two spatial streams and 19 dB SNR. For simplicity, it is assumed that the users are in UL and run just one service at a time. The user and network parameters are summarized in Table 3.5.

Table 3.5 – Input parameters in canonical scenario.

Parameter \ BBU Index		1 (BBU-FT)	2 (BBU-Voi)
User	# spatial streams $(N_{Str\,u,t_k})$	2	
	SNR $(\gamma_{u,t_k[\mathrm{dB}]})$	19	
	User Arrival Rate $(R_{U[\mathrm{user/ms}]}^{Arr})$	[1, 100]	
	Service (s_u^U)	File transfer	VoIP
	Packet volume $(V_{u,t_k[\mathrm{B}]}^{Pkt})$	40	
Network	Cell type (C^{TYP})	Micro	
	RRH traffic type (H^{TYP})	Residential	
	Channel bandwidth, $(\Delta f_{BW[\mathrm{MHz}]})$	20	
	Quantization resolution $(Q_{[\mathrm{bit}]})$	24	
	MIMO order, (N_{MIMO})	4×4	
	# BBUs in the BBU-pool, (N_B)	2	
	BBU-pool's AvCC $(C_{BP\,t_k[\mathrm{GOPS}]}^{Av})$	300	

Two kinds of services are assumed, i.e., file transfer and VoIP. In order to evaluate the behavior of the model in relation to the priorities of active services, the running services are considered different in the BBUs, so that all users in the first BBU use file transfer (BBU-FT) while in the second one they use VoIP (BBU-Voi), leading to the active services' average weights of 36 and 83, respectively, (3.31). The packet size is considered the same, i.e., 40 B, for all of them, for which two RBs are required for packet transfer, Section 3.7.2; it is assumed that user's RB demand is totally met. The considered scenario leads to 387 GOPS of each BBUs' peak demand, $C_b^{R_{PEAK}}$.

On the side of the BBU-pool, it is assumed that 300 GOPS is the BBU-pool's existing computing capacity. For simplicity, it is assumed that 100% of the resources are available for the BBUs.

3.9 Model and Simulator Assessment

3.9.1 Assessment Overview

In this section, both the simulator's functional correctness and the proposed model's performance are assessed. In order to show the functional correctness of the implemented algorithm, the outputs of the canonical scenario defined in Section 3.8 are subjected to a set of empirical tests listed in Table 3.6, among which resource allocation's reliability and CVX assessment are presented in Section 3.9.2. The compatibility of the BBUs' RCCs with the increasing values of the effective parameters are also presented in Annex B.

Table 3.6 – List of empirical tests that were made to validate the model performance.

Test	Description
1	Validating the model assumptions (Check if the input parameters correspond to the structural assumptions made about the system in the defined scenario).
2	Check the compatibility of the BBUs' RCCs with the increasing values of the effective parameters, i.e., $\Delta f_{BW[\mathrm{MHz}]}, N_{MIMO}, Q_{[\mathrm{bit}]}, m_{u,t_k[\mathrm{bit/symbol}]}, r_{u,t_k}, N_{Str\ u,t_k}, \eta_{u,t_k}^{RB_U}$.
3	Check the BBUs' RCC variation with the number of the users.
4	Check the result of the resource allocation if the total demand is more, equal, or less than the BBU-pool's AvCC.
5	Comparing results of the input parameters' variations to see if it is compatible with the expectations: 1. variable service types with constant demand, 2. increasing demand with constant service types.
6	Check if the output parameters are equal to the manually handcrafted test results.
7	Check if the implemented result of the optimization problem is the same as the optimal point achieved by the algebraic approach (CVX Assessment).
8	Verification of the correct plot of all outputs.

The proposed model's evaluation is also done by comparing its performance against EAS and DAS in Section 3.9.3. To this end, the resource allocation phase is repeated for each allocation scheme separately; the evaluation metrics are being assessed afterwards. The performance of the proposed model is further assessed in Section 3.9.4, by analyzing the effect of the BBUs' demands variation on the allocation results.

3.9.2 Reliability of Computing Resource Allocation and CVX Assessment

The functional correctness of the implemented algorithm and the proposed computing resource allocation's reliability is assessed by comparing the implemented results with the optimal point achieved by an algebraic approach and visualizing the results. To this end, the BBUs' RCCs are estimated as the first step of the proposed resource management model, Figure 3.3. The achieved results together with the active services' average weights and the BBU-pool's AvCC are fed to the computing resource allocation module as inputs, Figure 3.5. In this step, the BBUs' bargaining powers are calculated first, and the optimal resource allocation is found afterwards.

Considering 95 active users (selected randomly) results in 300 GOPS computing capacity demands in each BBU, 15 GOPS is for their CP and the minimum capacity that should be guaranteed, (3.17) and (3.16). The BBUs' bargaining powers are achieved in the next step resulting in 0.3 and 0.7 for the BBU-FT and the BBU-Voi, respectively, (3.32). The difference in BBUs' bargaining powers stems from the difference in the weight of their active services, i.e., 36 and 83 for BBU-FT and the BBU-Voi, respectively. The results are summarized in Table 3.7.

Table 3.7 – The proposed computing resource management's results.

Inputs							Outputs	
$\overline{w_{b,t_k}^{srv}}$		$C_{b,t_k[\mathrm{GOPS}]}^{R}$		$C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}$		$C_{BP\ t_k}^{Av}$	$C_{b,t_k[\mathrm{GOPS}]}^{Al}$	
BBU-FT	BBU-Voi	BBU-FT	BBU-Voi	BBU-FT	BBU-Voi		BBU-FT	BBU-Voi
36	83	300		15		300	96	204

After calculating the BBUs' bargaining powers and RCCs, the last step is to find the optimal resource allocation strategy that maximizes the BBU-pool's resource utilization. The optimal solution is obtained by solving (3.38) using CVX. The results are listed in Table 3.7 (the other outputs are discussed in Section 3.9.3). The accuracy of the implemented results is assessed by comparing them with the Karush-Kuhn-Tucker (KKT) approach [BoVa04]. In the first step, associated Lagrange function is made,

$$\mathcal{L}\left(\mathbf{C}_{t_k[N_B\times 1]}^{Al}, \boldsymbol{\xi}_{[N_B\times 1]}, \vartheta\right) = \sum_{b=1}^{N_B} B_{b,t_k} \log\left(C_{1,t_k}^{Al} - C_{b,t_k}^{R_{\min}}\right) - \vartheta\left(C_{BP\ t_k}^{Al} - C_{BP\ t_k}^{Av}\right) - \sum_{b=1}^{N_B} \xi_b\left(C_{b,t_k}^{Al} - C_{b,t_k}^{R}\right), \quad (3.45)$$

where $\boldsymbol{\xi}_{[N_B\times 1]}$ and ϑ are Lagrange multipliers. As the given optimization problem is convex, the point $\left(\mathbf{C}_{t_k[N_B\times 1]}^{Al^*}, \boldsymbol{\xi}_{[N_B\times 1]}, \vartheta\right)$ is optimal if it satisfies all the KKT conditions:

$$\frac{\partial L(\mathbf{C}_{t_k}^{Al}, \boldsymbol{\xi}, \vartheta)}{\partial C_{b,t_k}^{Al^*}} = \frac{B_{b,t_k}}{C_{b,t_k}^{Al} - C_{b,t_k}^{R_{\min}}} - \xi_b - \vartheta = 0 \qquad : \; b = 1,2, \tag{3.46}$$

$$C_{BP\,t_k}^{Al} - C_{BP\,t_k}^{Av} = 0, \tag{3.47}$$

$$C_{b,t_k}^{Al} - C_{b,t_k}^{R} \leq 0, \tag{3.48}$$

$$\xi_b \left(C_{b,t_k[\mathrm{GOPS}]}^{Al} - C_{b,t_k[\mathrm{GOPS}]}^{R}\right) = 0, \tag{3.49}$$

$$\xi_b \geq 0. \tag{3.50}$$

Considering (3.46) to (3.50), the only situation that does not lead to a contradiction is the assumption that both ξ_1 and ξ_2 are equal to zero, (3.46) which leads to

$$\frac{B_{b,t_k}}{C_{b,t_k}^{Al^*} - C_{b,t_k}^{R_{\min}}} = \vartheta \qquad : \; b = 1,2. \tag{3.51}$$

As ϑ is a constant value, it is concluded that

$$\frac{B_{1,t_k}}{C_{1,t_k}^{Al^*} - C_{1,t_k}^{R_{\min}}} = \frac{B_{2,t_k}}{C_{2,t_k}^{Al^*} - C_{2,t_k}^{R_{\min}}} \Longrightarrow C_{1,t_k}^{Al^*} = \frac{B_{1,t_k} C_{2,t_k}^{Al^*} - B_{1,t_k} C_{2,t_k}^{R_{\min}} + B_{2,t_k} C_{1,t_k}^{R_{\min}}}{B_{2,t_k}}, \tag{3.52}$$

which means that the first BBU's RCC can be expressed by the second BBU's RCC. Following (3.47) and the information given in Table 3.7, in the case AvCC is 300 GOPS, the BBUs' AlCC are achieved as 94 GOPS and 206 GOPS for BBU1 and BBU2, respectively. As all of the KKT conditions are satisfied, it is concluded that the achieved result is the optimum point that is the same as the implemented result presented in Table 3.7. The results are also presented in Figure 3.11, illustrating that the results of the model implementation are coherent with those from the algebraic approach.

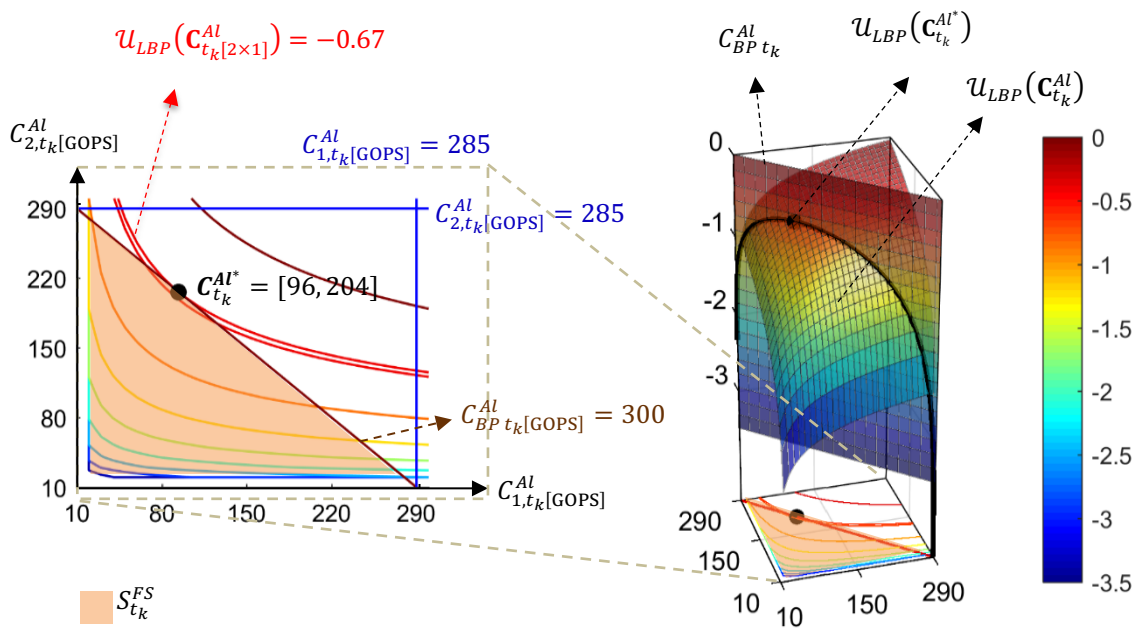


Figure 3.11 – Visualization of resource allocation results.

3.9.3 Comparison among Different Allocation Schemes

The proposed model's evaluation is assessed in this section by comparing its performance against the other resource allocation schemes mentioned in Section 3.5, i.e., EAS and DAS. To this end, the canonical scenario, Section 3.8, is assumed in the condition that 35 users taking the file transferring service are active in the first BBU (BBU-FT) and that 95 users are active in the second BBU (BBU-Voi); resulting in the BBUs' RCCs of 120 GOPS and 300 GOPS, the minimum guaranteed RCC of 9 GOPS and 15 GOPS and the bargaining power of 0.14 and 0.86 for BBU-FT and BBU-Voi, respectively.

Table 3.8 lists the performance achieved by each of the computing resource allocation schemes. EAS allocates resources equally among BBUs, regardless of service priorities or BBU demands. Although EAS is a fast resource allocation scheme without too much complexity, it wastes resources. While BBU-Voi is encountered with resource shortage, BBU-FT's AlCC is more than its RCC. This is the reason that the resource usage, $U_{t_k[\%]}$, is not 100%. Moreover, EAS leads to the smallest fairness index since BBUs' fulfilment levels are not proportional to their service weights. BBU-Voi has less fulfilment level even though its active services have the highest weight. The main reason is that EAS takes neither QoS nor the BBUs' demands into account while distributing resources among them.

Table 3.8 – Comparison among different resource allocation schemes.

<table>
<tr><th colspan="3" rowspan="2">Parameter</th><th colspan="3">Computing Resource Allocation Schemes</th></tr>
<tr><th>QDAS</th><th>DAS</th><th>EAS</th></tr>
<tr><td rowspan="8">Input</td><td rowspan="2">$C^{R}_{b,t_k[\mathrm{GOPS}]}$</td><td>BBU-FT</td><td colspan="2">120</td><td rowspan="2">-</td></tr>
<tr><td>BBU-Voi</td><td colspan="2">300</td></tr>
<tr><td rowspan="2">$C^{R_{\min}}_{b,t_k[\mathrm{GOPS}]}$</td><td>BBU-FT</td><td colspan="3">9</td></tr>
<tr><td>BBU-Voi</td><td colspan="3">15</td></tr>
<tr><td rowspan="2">$\overline{w^{SRV}_{b,t_k}}$</td><td>BBU-FT</td><td>36</td><td colspan="2">-</td></tr>
<tr><td>BBU-Voi</td><td>83</td><td colspan="2">-</td></tr>
<tr><td colspan="2">$C^{Av}_{BP\ t_k[\mathrm{GOPS}]}$</td><td colspan="3">300</td></tr>
<tr><td rowspan="8">Output</td><td rowspan="2">$C^{Al}_{b,t_k[\mathrm{GOPS}]}$</td><td>BBU-FT</td><td>48</td><td>86</td><td rowspan="2">150</td></tr>
<tr><td>BBU-Voi</td><td>252</td><td>214</td></tr>
<tr><td rowspan="2">f^{B}_{b,t_k}</td><td>BBU-FT</td><td>0.37</td><td rowspan="2">0.70</td><td>1</td></tr>
<tr><td>BBU-Voi</td><td>0.85</td><td>0.47</td></tr>
<tr><td colspan="2">F_{t_k}</td><td>1</td><td>0.87</td><td>0.70</td></tr>
<tr><td colspan="2">$\eta_{t_k[\%]}$</td><td colspan="3">61</td></tr>
<tr><td colspan="2">$U_{t_k[\%]}$</td><td colspan="2">100</td><td>90</td></tr>
</table>

DAS takes the real-time demand of BBUs into account. It allocates the minimum guaranteed resources, $C^{R_{\min}}_{b,t_k}$, to each BBU and distributes the remaining resources proportionally to their UP demands, $C^{R}_{U\ b,t_k}$, which is the reason why BBU-Voi receives more resources than BBU-FT. Since the BBUs' AlCCs are

proportional to their RCCs in DAS, the BBUs' fulfilment levels are equal. The fairness index is still lower than QDAS because DAS does not consider QoS.

Besides BBUs' demands, QDAS takes QoS, hence, service priorities, into account while distributing resources. QDAS shrinks the capacity share of the lower priority BBUs in the bottlenecks to compensate for the higher priority BBU resource shortage. This is why it allocates more resources to BBU-Voi with higher average service weights compared with DAS. Since it takes QoS into account, BBU fulfilment levels are proportional to the average service weights. Therefore, the best fairness index among all three provisioning schemes is achieved. Moreover, no waste of resources occurs and it fully uses the available resources since there is a shortage and QDAS bounds the BBU's AlCC to their demands.

It should be noticed that the resource allocation efficiency is the same for all three provisioning schemes as 100% of the BBU-pool's AvCC are allocated to BBUs in all of them.

3.9.4 Performance Evaluation of the Proposed Scheme

As mentioned before, the proposed computing resource allocation scheme receives average weights of BBUs' active services, RCCs, minimum guaranteed RCCs and BBU-pool's AvCC as inputs. It calculates the optimal computing resources allocation and its outputs are the BBUs' optimal AlCCs and the evaluation metrics, Figure 3.5. The proposed model's performance is evaluated in this section, by analyzing the effect of BBUs' RCC variation as one of the inputs. Considering the canonical scenario defined in Section 3.8, the resource allocation phase is repeated with the BBUs' RCCs varying in [9, 320] GOPS (equivalent to increasing the number of users from 1 to 100), while both BBUs' bargaining powers remain as 0.3 and 0.7, respectively. The results are explained in what follows:

- **Allocated Computing Capacity**: An essential factor that should be considered is the compatibility of the AlCC with instantaneous demand in the BBU-pool. In case the BBU-pool's total RCC is equal or greater than its AvCC, 100% of the computing resources should be used. Conversely, if demand is less than AvCC, the available resources should be used to the required extent and the remaining resources should be idle or shut down. Figure 3.12 represents the capacity allocated to each BBU.

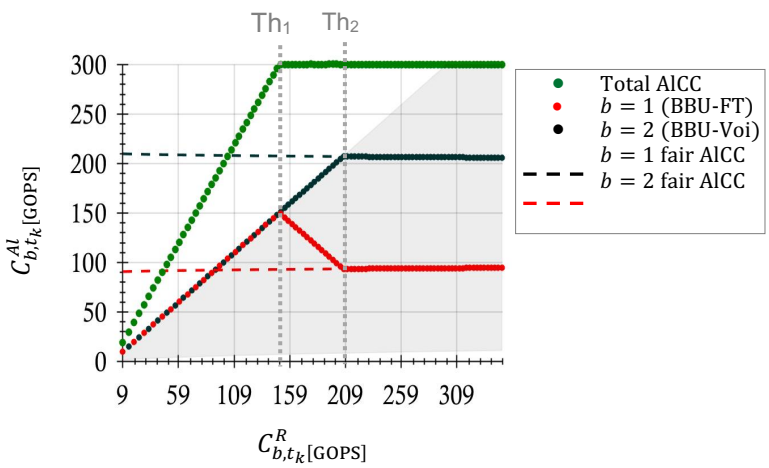


Figure 3.12 – Sensitivity of the BBUs' AlCCs to their RCC in QDAS.

The lower bound of the shaded area in the figure shows the minimum amount of capacity that should be guaranteed to any single BBU and the upper bound represents BBUs' demands. As the figure shows, while total demand is less than BBU-pools' AvCC (300 GOPS), both BBUs' requirements are met and BBUs' AlCC does not exceed their RCC. The resources are allocated as needed; the rest remaining idle. For RCCs larger than threshold Th1, both BBUs' demands cannot be met at the same time due to resource shortage, thus, BBU-FT's AlCC that has a lower bargaining power drops, while BBU-Voi's AlCC is still equal to its RCC. Beyond Th2, BBU-Voi's RCC cannot be met entirely since the resource allocation should be fair and BBUs' AlCC should be proportional to their active services' average weight in a fair allocation, (3.42). Fair allocations are depicted in Figure 3.12 by dash lines. Since there is a shortage, we know that all the resources are used. Moreover, the BBUs' RCCs are equal, and also their minimum guaranteed RCCs are the same. Therefore, in this case, the resource allocation is fair when

$$C^{Al}_{b,t_k[\mathrm{GOPS}]} = \left(300 - 2C^{R_{\min}}_{b,t_k[\mathrm{GOPS}]}\right) B_{b,t_k} + C^{R_{\min}}_{b,t_k[\mathrm{GOPS}]} \quad : \quad b = 1,2. \tag{3.53}$$

- **Fulfilment Level:** The BBUs fulfilment levels are illustrated in Figure 3.13. For RCCs larger than Th_1, there are not enough resources to meet both BBUs requirements. Therefore, the fulfilment level of the BBU-FT with less bargaining power starts to drop since allocated resources are less than its demands. At the same time, BBU-Voi's demand (with higher bargaining power) are 100% fulfilled. Beyond Th_2, BBU-Voi's fulfilment level also starts to drop due to resource shortage as well as to the fact that the BBUs allocated resources should be fair and proportional to the BBUs' bargaining powers. The results confirm that the proposed resource allocation scheme considers the service weights and the priority of the BBUs while distributing resources among the BBUs.

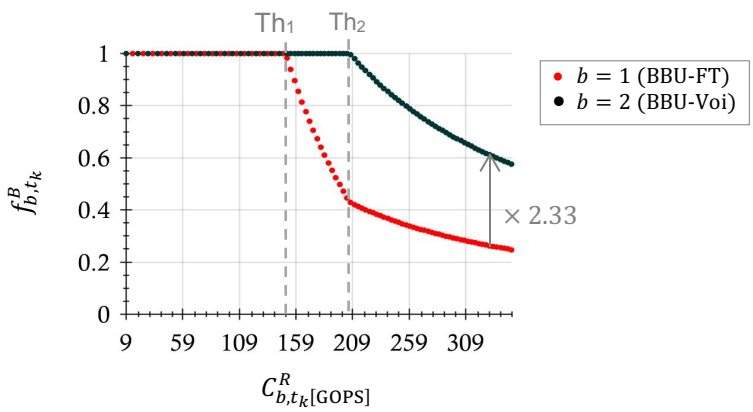


Figure 3.13 – Effect of the BBUs' bargaining powers on the fulfilment level in QDAS.

- **Fairness Index:** Jain's fairness index is another metric that is evaluated for the same case study. The fairness index is valid if the total demand exceeds the BBU-pool's AvCC at a time instant, i.e., beyond Th_1 in Figure 3.12. The allocation is defined to be fair if the fulfilment levels maintain the same proportion of the average weights of active services. The fair allocations are depicted with dashed lines in Figure 3.12, and Figure 3.14 presents the correspondence fairness index.

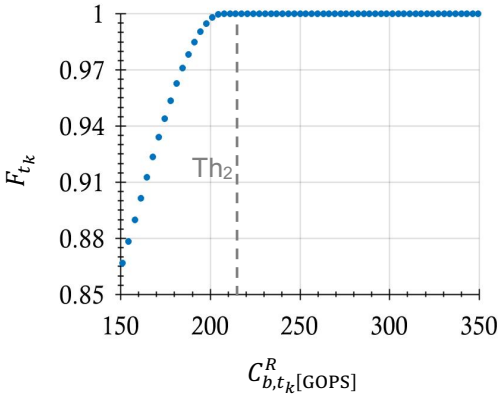


Figure 3.14 – Fairness index in QDAS.

As Figure 3.14 shows, the allocation's fairness is low in the beginning. The reason in that the computing capacity proportional to the service weights is more than the RCC of the BBU with higher priority services, i.e., BBU-Voi. The resource allocation strategy bounds the BBU-Voi's AICC to its RCC, the remaining capacity being allocated to BBU-FT with lower service priority and as a result, BBU-Voi's AICC is less than its fair allocation; on the contrary, BBU-FT receives more. This will end up to a small value for the defined fairness index, as the fairness condition does not hold. By incrementing BBU-Voi's demand, fewer resources remain available for BBU-FT, and the fairness index increases. Beyond Th_1, the fairness condition holds and as Figure 3.13 presents, the fulfilment levels maintain the same proportion of the average weights of active services, i.e., 2.33. The fairness index results confirm that the resource allocator takes not only the priority of services but also instantaneous requirement of BBUs into account while distributing resources among them.

- **Efficiency of Dynamic Resource Allocation**: The efficiency of the dynamic resource allocation is achieved by comparing the total AICC that the proposed model suggests in a BBU-pool with traditional approaches that a static amount of computing capacity is allocated to the BBUs based on their peak hour RCC. Considering BBU's peak RCC, i.e., 387 GOPS per BBU in the canonical scenario, Figure 3.15 shows the proposed resource allocation's efficiency.

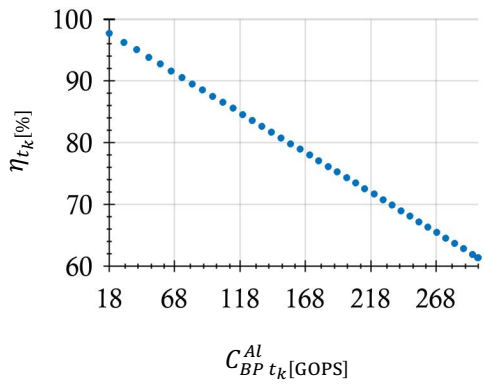


Figure 3.15 – Efficiency of resource allocation in QDAS.

In the beginning, the allocation efficiency is high, the reason being that BBUs' demands are small. The proposed resource allocation model bounds the BBUs' AICCs to their RCC due to its allocation

strategy. As a result, the total allocated resources are much less than the sum of the BBUs' peak requirement that is equal to the resources statically allocated to the BBUs in the traditional approaches; hence, efficiency is high. By increment of BBUs' resource demands, more resources are allocated, hence, efficiency decreases. In the worst case, the efficiency is almost 61% as all the BBU-pools' available resources are allocated, almost 40% of the sum of BBUs' peak demands.

- **AvCC Usage:** Figure 3.16 illustrates the resource usage as the proportion of BBU-pool's total AICC to its AvCC. In the beginning when BBUs' demands are small, the resource usage is low since the allocator bound the BBUs' AICCs to their RCCs. By increasing RCCs, the resources usage also increases until Th_1 when the total demand surpasses the AvCC. The resource allocator distributes the entire available resources among the BBUs and bounds the total AICC to the BBUs' available resources. The results confirm that there is no wastage in the proposed resource allocation scheme and the resources are allocated to the BBUs as needed. On the other hand, in case of resource shortages 100% of the available resources are distributed among the BBUs.

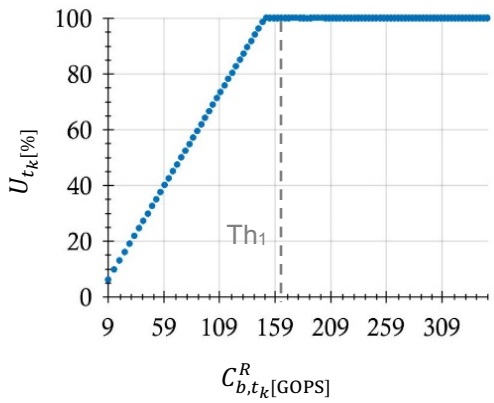


Figure 3.16 – AvCC usage in QDAS.

Chapter 4

Real-time Computing Resource Allocation Framework

This chapter presents an extension to the resource management model proposed previously and defines a real-time computing resource allocation framework considering time-varying traffic and demands in a tidal channel condition. Section 4.1 gives an overview of the proposed model, and Section 4.2 explains a strategy to find a proper time interval between two successive resource allocations. Section 4.3 mentions the metrics used to evaluate the proposed model in a real-time framework. The rest of the chapter is dedicated to the simulator implementation, canonical scenario, and simulator assessment.

4.1 Model Overview

As mentioned before, C-RAN provides both higher data rates and lower network latencies by consolidating and multiplexing BBU resources in a data center called BBU-pool. Although the consolidation of resources in C-RAN reduces the number of the required resources in the network, there are still critical challenges for data centers, such as power consumption [DaWF16], [HMDM19] and [WeGP13]: a medium-sized one with 930 m^2 and 288 racks can consume 4 MW in the traffic peak [PMZW10]. Since computing resources, i.e., servers, are the most energy-intensive entities in data centers, it is worthwhile to apply efficient resource management strategies to maximize their utilization and reduce the number of idle ones. An idle server consumes 60% of its peak power usage, although it has no productivity [PMZW10].

However, designing efficient resource management strategies is a complicated process for cloud providers. Due to the variety of network services, user arrival rates and channel conditions, BBUs' resource demands fluctuate significantly throughout the day. On the one hand, a BBU computing capacity should suffice peak demands; on the other hand, provisioning fixed resources based on peak requirements leads to idle resources in the rest of the day. As a result, an efficient resource management strategy in a BBU-pool should allocate the computing capacity dynamically, in accordance with the BBUs' instantaneous demand, while efficiently handling resources in the case of a shortage. Resource shortages are time instants in which the BBU-pool's available resources are less than demand spikes and come into play in two circumstances: when the objective is intentionally to design the pool with minimum computing resources; or, even if there are more computing resources, they cannot be initialized at a rate similar to the one of demand fluctuations (in the scale of milliseconds), due to hardware limitations.

This chapter presents an extension to the proposed resource management model presented in Chapter 3 (which is limited to a single time) and defines a real-time computing resource allocation framework. In this context, the previously computing resource allocation model is calculated repeatedly over time, and the BBU-pool computing resources are allocated dynamically, based on the BBU's instantaneous demand. The goal is to prevent BBU's over/under-loaded situations before they occur by dynamically influencing the BBUs' RCCs in advance of the actual computing resource allocation. An overview of the proposed computing resource management model is presented in Figure 4.1.

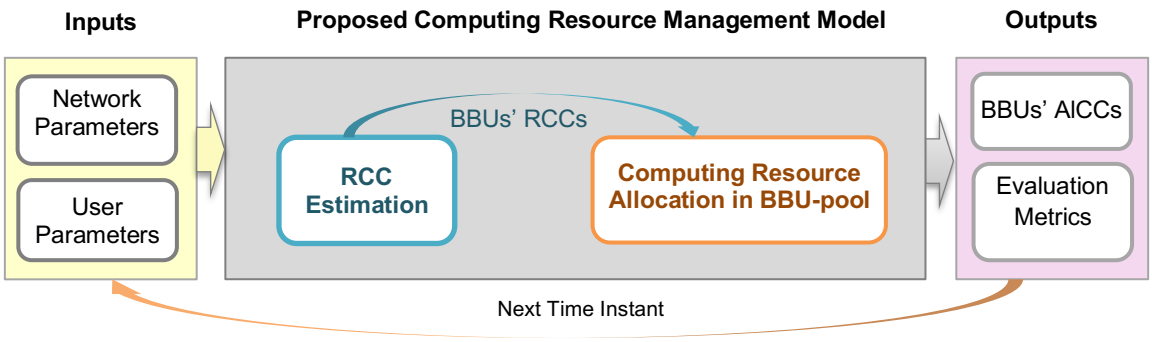


Figure 4.1 – Global model overview.

A BBU's RCC depends on the user/network parameters that are fluctuating within a TTI. Therefore, a defined optimization problem should be solved once with new input parameters, in a TTI, in order to define a real-time computing resource allocation framework. The proposed optimization problem should be small enough to be solved extremely fast, coordinated with a TTI. There are several works in the literature on proposing models to solve real-time convex optimization problems in the order of milliseconds or microseconds [PaEI10]. For more reliability, however, a time framework is defined in this thesis to evaluate an appropriate time interval between two successive problem solving, larger than one TTI, in which the load fluctuation is minimal. To this end, besides the input parameters that are mentioned in Section 3.2, the following parameters are also required:

- User specific info:
  - mobility speed, $\upsilon_{Avg[\mathrm{km/h}]}$,
- Cell specific info:
  - carrier frequency, $f_{C[\mathrm{MHz}]}$.

Section 4.2 explains a strategy to find a proper time interval between two successive resource allocations.

Figure 4.2 presents the process of the computing resource allocation module over time in more detail. Taking as inputs network and user parameters at a specific time instant, the BBUs' RCCs are estimated in the first step. The results are then fed to the computing resource allocation module in order to find the optimal AICC to BBUs. In the next time instant, the resource management process is re-instantiated over new input parameters.

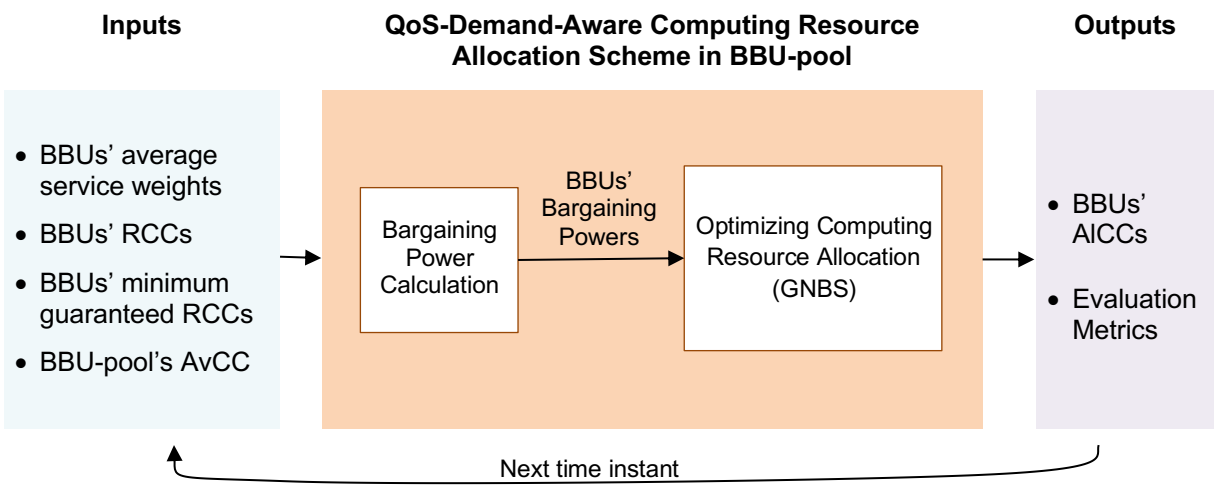


Figure 4.2 – Overview of the QoS-demand-aware computing resource allocation scheme over time.

Besides the BBUs' AICCs, the following performance metrics are calculated per a time interval as the outputs in order to evaluate the model:

- BBU's average fulfilment level, $\overline{f^B}$,
- dynamic resource allocation's average efficiency, $\overline{\eta_{[\%]}}$,
- resource usage, $\overline{U_{[\%]}}$.

4.2 Time Framework

4.2.1 Coherence Time

In order to allocate computing resources adaptively so that BBUs' resource demands are met, their instantaneous RCCs should be known. A BBU's RCC is composed of all the BS processing steps' RCC in the BBU, which are time-variant. The network/user parameters affecting each BS processing step's RCC are listed in S^{Parm}, (3.8). As network parameters are constant, the variation of the BS processing's RCC depends on the user parameters variation, i.e., MCS, the number of streams and the number of allocated RBs. As mentioned in Section 3.7.2, the number of RBs allocated to a user depends on both the user's MCS and the service's bit rate. The bit rate fluctuates per TTI, but the variation of MCS and the number of streams depends on the coherence time and CSI reporting periodicity. Therefore, in order to evaluate an appropriate time interval between two successive RCC evaluations, coherence time and RI reporting periodicity should be considered besides the TTI.

The coherence time, Δt_C, is the expected time duration over which the channel's response is essentially invariant. The value of coherence time depends on the users' mobility speed and maximum Doppler shift [Ahma13]. A popular rule in order to calculate Δt_C is defined as [Rapp96]:

$$\Delta t_{C[\mathrm{s}]} = \sqrt{\frac{9}{(16\pi\, f_{D,\max[\mathrm{Hz}]}^2)}}, \tag{4.1}$$

where $f_{D,\max}$ is maximum Doppler shift,

$$f_{D,\max\,[\mathrm{Hz}]} = f_{C[\mathrm{Hz}]} \frac{\upsilon_{u[\mathrm{m/s}]}}{c_{[\mathrm{m/s}]}}, \tag{4.2}$$

where:

- f_C: carrier frequency,
- υ_u: user's speed u,
- c: light speed, where $c_{[\mathrm{m/s}]} = 299{,}792{,}458$.

In this thesis, $\Delta t_{C[\mathrm{s}]}$ is not studied on a specific user speed. The speed range that LTE supports (to date, the range of 0 to 500 km/h) is split into multiple intervals and, for each, an average speed value is considered. The users can thus be classified as:

- Very-low-speed, e.g., pedestrian, 5 km/h,
- Low-speed, e.g., cyclist, vehicular urban, 50 km/h,
- Mid-speed, e.g., vehicular sub-urban, 90 km/h,
- High-speed, e.g., vehicular rural, 120 km/h,
- Very-high-speed, e.g., high-speed train, 500 km/h.

Accordingly, the values of $\Delta t_{C[\mathrm{s}]}$ and $f_{D,\max\,[\mathrm{Hz}]}$ are calculated in accordance with (4.1) and (4.2) for each speed class and supported operating bands. The UE speeds and related maximum Doppler shifts are listed in Annex C for the supported carrier frequencies. The result of minimum and maximum coherence time corresponding to each speed class are summarized in Table 4.1.

Table 4.1 – User speeds and associated coherence time.

Speed [km/h]	Maximum Doppler Shift [Hz]		Coherence Time [ms]	
	Min	Max	Min	Max
5	3.66	16.68	25.38	115.64
50	36.65	166.78	2.54	11.55
90	65.96	300.21	1.41	6.42
120	87.95	400.28	1.06	4.81
500	366.46	1 667.82	0.25	1.16

As mentioned in Section 2.1.2, gaining a higher UL/DL speed, LTE supports various TMs. TM and the number of users' streams depends on transmitter/receiver capability and CSI reporting. CSI reporting is the UE feedbacks, reporting its preferred TM, based on the channel condition. The eNB configures CSI reporting format for each UE in RRC signaling. Besides the TM, the CSI is a critical parameter on the MCS selection.

For the sake of simplicity, in this thesis, it is assumed that CSI reporting for all users is equally configured to be sent periodically according to the entire bandwidth quality, i.e., wideband reporting. The CSI report is a composition of RI, PMI, and CQI. Based on [3GPP20i], for wideband periodic, the CQI/PMI reporting interval can be configured as {2, 5, 10, 16, 20, 32, 40, 64, 80, 128, 160} sub-frames, therefore, the reporting period, $\Delta t_{CQIPMI[\mathrm{ms}]}$, is equal to the number of sub-frames times the sub-frame duration,

$$\Delta t_{CQIPMI[\mathrm{ms}]} = N_{CQIPMI}^{SF}\, \Delta t_{SF[\mathrm{ms}]}, \tag{4.3}$$

where:

- N_{CQIPMI}^{SF}: number of sub-frames in selected CQI/PMI reporting interval, where $N_{CQIPMI}^{SF} \in \{2, 5, 10, 16, 20, 32, 40, 64, 80, 128, 160\}$ (Configurable in higher layer signaling),
- Δt_{SF}: a sub-frame duration.

In case RI reporting is configured as well, the reporting interval of the RI, $\Delta t_{RI[\mathrm{ms}]}$, is an integer coefficient of $\Delta t_{CQIPMI[\mathrm{ms}]}$, [3GPP20i],

$$\Delta t_{RI[\mathrm{ms}]} = M_{RI} \cdot \Delta t_{CQIPMI[\mathrm{ms}]}, \tag{4.4}$$

where $M_{RI} \in \{1, 2, 4, 8, 16, 32\}$ and is configurable in higher layer signaling. Equation (4.4) indicates that the minimum periodicity of RI reporting is the duration of two sub-frames. Therefore, the transmitting mode is consistent at least for the duration of two sub-frames.

4.2.2 Time Slicing

In order to estimate a BBUs' RCC, the proposed model considers a single snapshot in time, i.e., time instant t_k, and uses the parameter measurements taken at that time instant. The interval between two successive time instants in which a BBU's demand is estimated is denoted by Δt,

$$\Delta t_{[\mathrm{ms}]} = t_{k+1[\mathrm{ms}]} - t_{k[\mathrm{ms}]}. \tag{4.5}$$

The value of Δt is considered constant and proportional to the average of the time that the variation of the BBUs' RCC is minimum. As mentioned previously, the RCC fluctuation depends on the coherence time, Δt_C, RI reporting periodicity, Δt_{RI}, and sub-frame duration, Δt_{SF}. An appropriate TTI is assumed to be equal to a sub-frame duration. In order to allocate computing resources adaptively, so that the BBUs' resource demands are met, Δt should be the minimum value among Δt_C, Δt_{RI} and Δt_{SF}. However, to decrease the excess burden while keeping efficient allocation, the proposed model assumes Δt to be a small integer coefficient of the minimum value between Δt_C and Δt_{RI},

$$\Delta t_{[\mathrm{ms}]} = n\left(\min\{\Delta t_{c[\mathrm{ms}]}, \Delta t_{RI[\mathrm{ms}]}\}\right) \quad \forall n \in \mathbb{N}. \tag{4.6}$$

Twenty-four hours in a day are considered as $\lfloor 24_{[\mathrm{h}]}/\Delta t_{[\mathrm{h}]} \rfloor$ equal and successive time slots (the last time slot may have a lower duration depending on Δt). Moreover, small subintervals, Δt^{TTI}, are set up within larger intervals, each with one TTI duration. The BBU RCCs are evaluated at t_k and t_{k+1}, and, accordingly, RCC values are assumed the same for all time instants in between. The principle of time slicing can be best described using the illustration in Figure 4.3. RCC evaluations are provided at the beginning of an interval, i.e., at t_k, and accordingly, RCC values are estimated for all the time instants $t_{k,n}$ in between. RCC estimation is described in detail in Section 3.3.2.

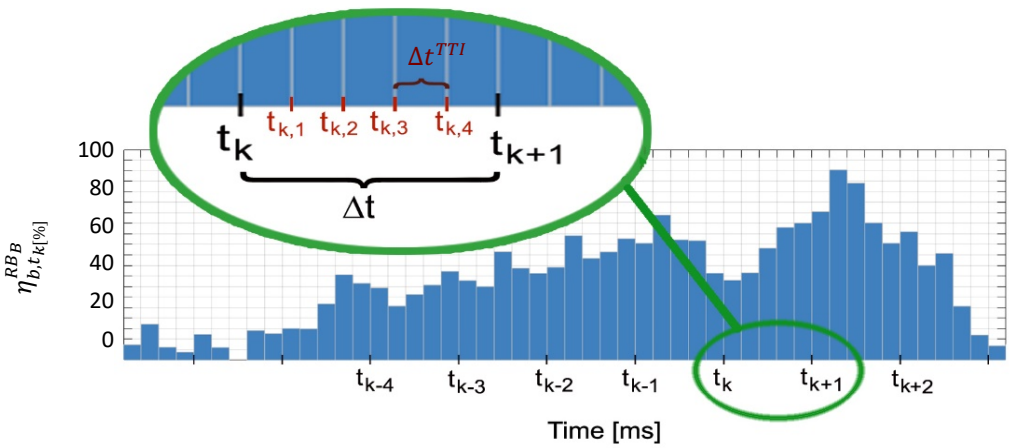


Figure 4.3 – Time slicing, Δt =5 ms, Δt^{TTI} =1 ms.

4.3 Evaluation Metrics

Different metrics are defined in order to evaluate the proposed model's performance in a real-time framework. The metrics are explained in detail in what follows:

- **Average of a BBU's fulfilment level** is the average of a BBU's fulfilment levels for all time slices

in a given duration,

$$\overline{f_b^B} = \frac{\sum_{k=0}^{N^{TS}} f_{b,t_k}^B}{N^{TS}}, \tag{4.7}$$

where N^{TS} is the number of time slices in a simulation interval,

$$N^{TS} = \frac{\Delta T_{[\mathrm{ms}]}^{SIM}}{\Delta t_{[\mathrm{ms}]}}, \tag{4.8}$$

and $\Delta T_{[\mathrm{ms}]}^{SIM}$ is the simulation interval.

- **Average of dynamic resource allocation efficiency** is the average of resource allocation efficiency within a simulation interval,

$$\overline{\eta_{[\%]}} = \frac{\sum_{k=0}^{N^{TS}} \eta_{t_k[\%]}}{N^{TS}}. \tag{4.9}$$

- **Average resource usage** is the average of resource usage within a simulation interval,

$$\overline{U_{[\%]}} = \frac{\sum_{k=0}^{N^{TS}} U_{t_k[\%]}}{N^{TS}}. \tag{4.10}$$

4.4 Simulator Implementation

4.4.1 Overview

This subsection aims at presenting the details of the proposed models' implementation in a real-time framework, namely, traffic and end-users generation, and algorithms being used. Figure 4.4 illustrates the major functional elements of the simulation together with the way they interact with during operation. As the figure shows, the simulation is composed of three main components. In the first module, the network traffic is generated for a selected time interval: end-users are generated, and random SNR are assigned to them; the generation of users' packets is based on stochastic distributions dedicated to each service. After that, the number of RBs required for data transmission is extracted according to the users' SNR and their associated packet size. The results, together with the input network/user parameters are fed to the RCC estimation module. The RCC of the BBUs are calculated as explained in Section 3.3. The results acquired in the first two modules together with the input parameters are fed to the resource allocation module, in which the bargaining powers are calculated first based on Section 3.4.3. Accordingly, the optimal computing resource allocation is acquired exploiting CVX.

All the archived results are used in order to evaluate the performance of the proposed model using the metrics defined in Section 4.3. Traffic simulation and the implementation flowcharts are explained in

more detail in the following sections.

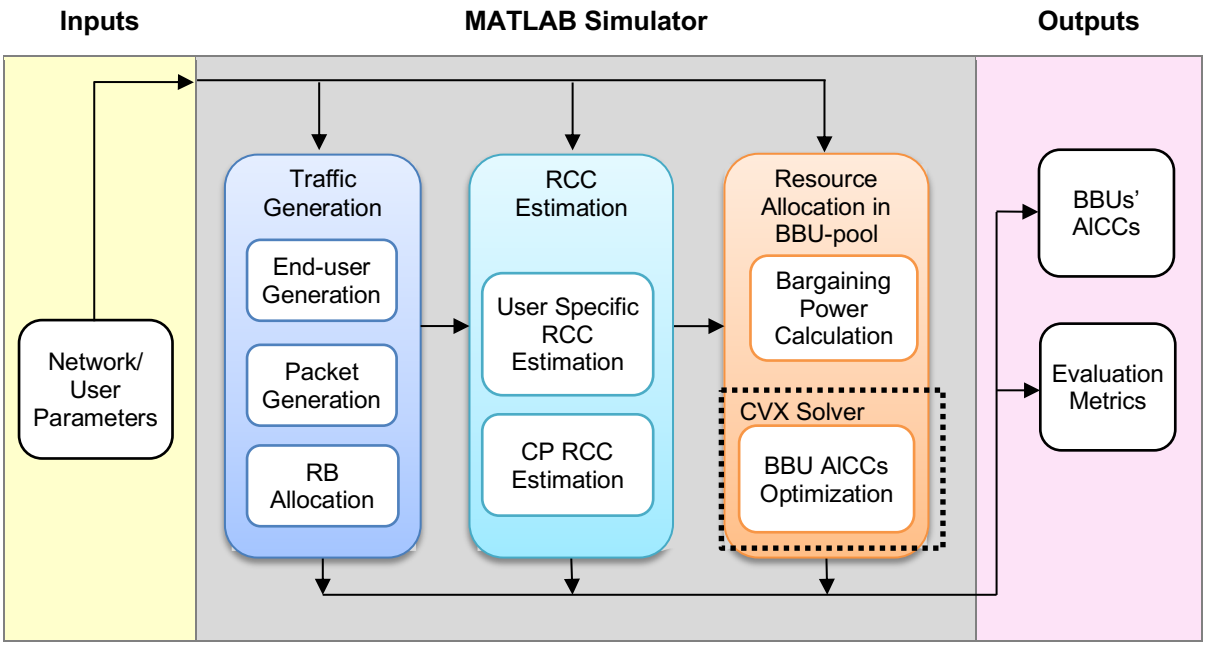


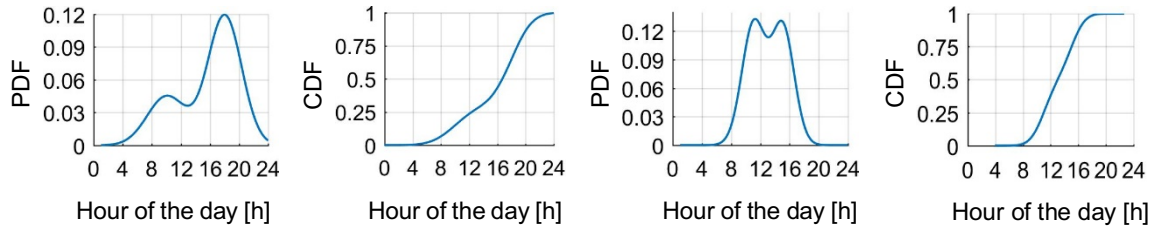
Figure 4.4 – Simulator overview.

4.4.2 Traffic Generation

Network behavior is simulated within a time interval in order to evaluate the proposed computing resource management model. The generation of traffic is based on the generic model defined in [HaGB05]. End-user generation, packet generation and radio resource allocation are three primary considerations while generating the network traffic described in what follows:

- **End-user generation:** end-users are generated at the beginning of the simulation. A user's process starts once the user arrives at the network. Users' arrival rate is given at the beginning of the simulation following a mixture of two Truncated Normal Distributions for both residential and business areas: for the former, the mean values are 10 AM and 6 PM, standard deviations are 160 min and 140 min, and mixing proportions are 30% and 70% for the first and second distribution, respectively; in the latter, mean values are 11 AM and 3 PM for the first and second distribution respectively, both with the standard deviation of 95 min and 50% of mixing proportion. Figure 4.5 illustrates the Probability Density Function (PDF) and the Cumulative Distribution Function (CDF) of users' arrival rate in both areas, user peak hours being taken from [PLLL11], and traffic outside peak hours is selected in such a way that it gradually increases until the peak and then decreases.

  The simulator initially considers users who enter the network during the simulation period. It also targets the users who have entered before and are still active at the beginning of the simulation. Users are considered mobile, therefore, an individual user is assigned with a random SNR that is variable per TTI, based on a Uniform Distribution in [1, 35] dB. Still, the user SNR is assumed to be unchanged within the coherence time, which is considered equal for all users connected to a BS and is calculated based on the BS frequency band and the average speed of the users.



(a) Residential (b) Residential (c) Business (d) Business

Figure 4.5 – PDF and CDF of the user arrival rate in residential/business areas.

- **Packet generation:** running services are determined at the beginning of the simulation based on a defined traffic mixture. According to the service type, its duration is taken as listed in Table 4.2. Traffic is generated in three levels (session, activity and packet) as explained in what follows, further details being presented Annex D:

Table 4.2 – Service characteristics.

Service	Service Parameter		Distribution	Mean	Standard Deviation
VoIP	Packet Inter-Arrival Time		Deterministic	20ms	-
	Duration		Poisson	120s	11s
Video	Frame Packets Inter-Arrival Time		Pareto	6.1ms	3.6ms
	Duration		Poisson	300s	17.3s
	Packet Volume		Pareto	1.3MB	257B
Web Browsing	Packet Inter-Arrival Time	Reading Time	Exponential	30s	
		Parsing Time		130ms	
	Duration		Poisson	420s	20.5s
	Packet Volume	Main Object Size	Lognormal	11MB	25.3MB
		Embedded Object Size		8.2MB	47.3MB
		Number of Embedded Object per Page	Pareto	7.6	10.4
File Transferring	File Size		Lognormal	2MB	700B
E-Mail				1.3MB	380B

  - **Session level:** A session begins when the user starts an application until being disconnected from the network, so once the user arrives at the network, his/her session starts. The session duration depends on the service that the user is performing: for VoIP, video streaming/calling and web browsing, the session's duration is characterized by the Poisson Distribution; for file transfer and email, the session's duration depends on the data volume that is transferred and on the amount of the available RBs in each sub-frame, the Lognormal Distribution being used for the data volume. Web browsing packets' size depends on the main object size, embedded object size and number of embedded objects per web page. Video packet's size relies on the

size and number of packets of a frame and frame rate, while for VoIP it depends on the user's active/inactive state duration.

- **Application activity level:** The application activity level determines the density of information in a service. In this layer, a session is decomposed into the state of being active or inactive. For example, a web browsing service is in the active state when a web page is downloading, but after that, the session is in an inactive state while the user reads the downloaded page; for VoIP, while the user is talking, the session is in the active state, but otherwise, the state is inactive. The duration of being in an active or inactive state depends on the service profile.
- **Packet level:** The packet level is the basic one of traffic generation, deciding how service packets are generated and transferred. Packets' sizes and inter-arrivals follow a specific distribution based on the service profile.

Figure 4.6 presents the relation between levels: during a session the user sends data (activity level), which is then put into packets with intervals between any pair of packets (packet level).

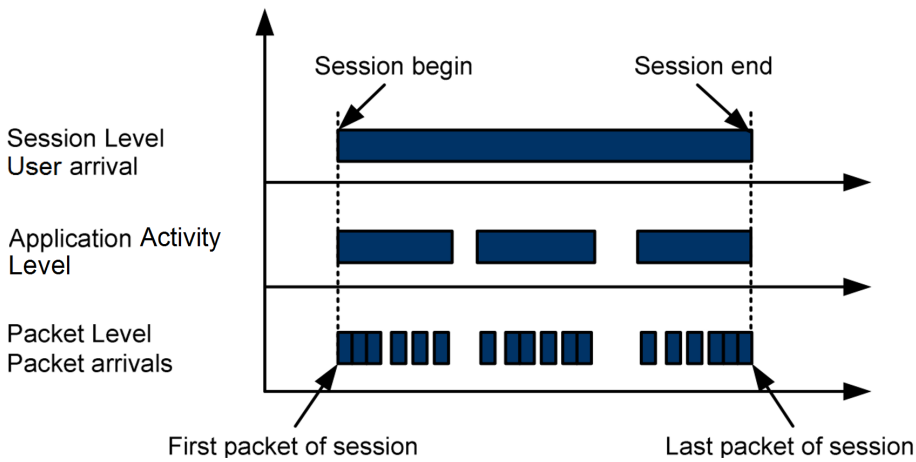


Figure 4.6 – Generic Traffic Source Model (extracted from [HaGB05]).

- **RB allocation**: As mentioned in Section 3.3.2, the number of RBs allocated to a user at a given time instant is an important parameter that is essential for the estimation of a BBU RCC. A simple strategy is used in two steps in order to distribute the available radio resources among active users: the amount of RBs that an active user requires at an individual time instant, $N_{R\,u,t_k}^{RB}$, is calculated as explained in Section 3.7.2; then, the available RBs in the given bandwidth are distributed among users according to their requirement and service priority level.

  Radio resources are limited, and total demand may exceed the available RBs at a given time instant, therefore, the allocation strategy considers the priority of services: the available RBs are allocated to the services with a higher priority level with a guaranteed bit rate, i.e., VoIP and video streaming. The allocation is based on the service data rate, and in case there are not enough RBs in a sub-frame to meet the high prioritized service requirement, the RBs of the subsequent sub-frame will be allocated to them. Once higher priority service requirements are met, the remaining RBs in each sub-frame are evenly distributed among the other services as needed. Likewise, the following sub-frames compensate for the shortcomings.

4.4.3 Overview of CVX Solver

Since the optimization problem defined in this thesis is solved in CVX, this section provides an overview of CVX performance.

CVX is a modelling system for constructing and solving disciplined convex problems. CVX supports several standard problem types, including linear, quadratic, second-order cone and semidefinite ones. This solver is implemented in MATLAB, effectively turning MATLAB into an optimization modelling language. Model specifications are constructed using common MATLAB operations and functions, and standard MATLAB code can be freely mixed with these specifications. This combination makes it simple to perform the calculations needed to form optimization problems or process the results obtained from their solution.

Within a CVX specification, optimization variables have no numerical value; instead, they are special MATLAB objects. This enables MATLAB to distinguish between ordinary commands and CVX objective functions and constraints. CVX reads a problem specification and builds an internal representation of the optimization problem. If it encounters a violation of disciplined convex programming (such as an invalid use of a composition rule or an invalid constraint), an error message is generated. MATLAB converts the CVX specification to a canonical form and calls the underlying core solver to solve it.

If the optimization is successful, the optimization variables declared in the CVX specification are converted from objects to ordinary MATLAB numerical values that can be used in any further MATLAB calculations. CVX also assigns a few other related MATLAB variables, e.g.: one gives the status of the problem, i.e., whether an optimal solution was found, or the problem was determined to be infeasible or unbounded; another gives the optimal value of the problem. Dual variables can also be assigned.

Numerical results of CVX are computed within a predefined precision or tolerance. CVX considers three different tolerance levels $\epsilon_{solver} \leq \epsilon_{standard} \leq \epsilon_{reduced}$ when solving a model:

- The solver tolerance ϵ_{solver} is the level requested of the solver, and the solver will stop as soon as it achieves this level, or until no further progress is possible.
- The standard tolerance $\epsilon_{standard}$ is the level at which CVX considers the model solved to full precision.
- The reduced tolerance $\epsilon_{reduced}$ is the level at which CVX considers the model "inaccurately" solved, returning a status with the Inaccurate prefix; if this tolerance cannot be achieved, CVX returns Failed status, and the values of the variables should not be considered reliable.

The CVX default values of $[\epsilon_{solver}, \epsilon_{standard}, \epsilon_{reduced}]$ are set to $\left[\epsilon^{1/2},\ \epsilon^{1/2},\ \epsilon^{1/4}\right]$, where $\epsilon = 2.22 \times 10^{-16}$ is the machine precision. These tolerance levels were chosen in this thesis, since they are sufficient for most of the applications, including the computing resource allocation in BBU-pool.

It is also noted that CVX supports several solvers, each with different capabilities, the Embedded Conic Solver (ECOS) having been selected in this thesis, which is one of the solvers that relies on a successive approximation method that supports geometric problems and models using functions from the exponential and logarithm families.

4.4.4 Implementation Flowcharts

As mentioned before, the simulation is composed of three main modules: traffic generation, RCC estimation and resource allocation. Each module's process is explained in detail in what follows.

Given the simulation interval and simulation starting time, the BBU-pool's instantaneous load is produced by generating any single BBU's load and RCC in the pool, Figure 4.7. A BBU's load generation is re-called as a subfunction, i.e., the red block, its process being explained in detail in Figure 4.8.

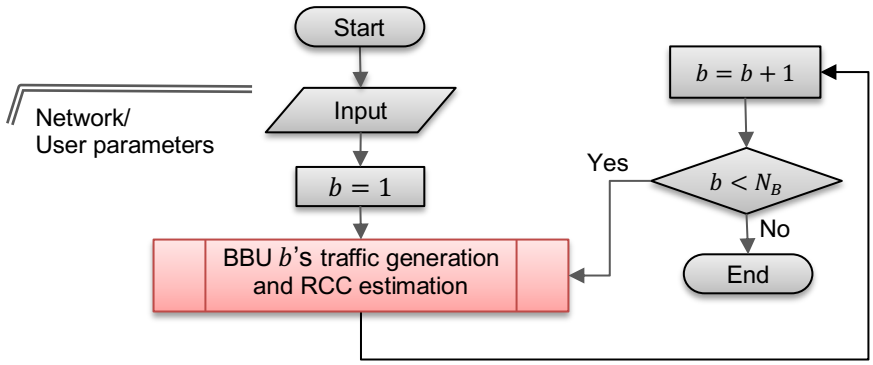


Figure 4.7 – The algorithm of traffic simulation.

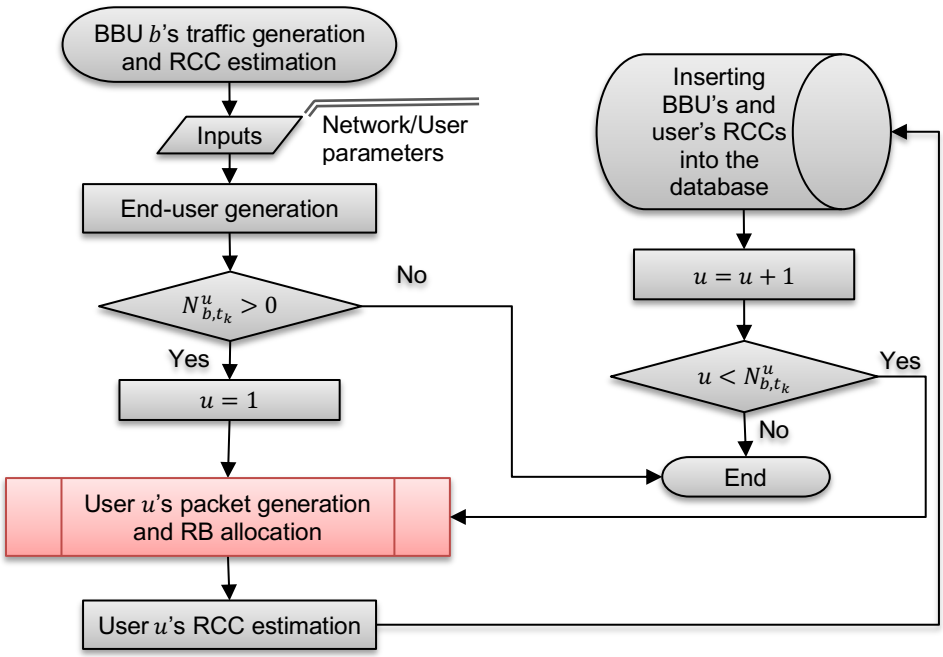


Figure 4.8 – The algorithm of calculating a BBU's traffic generation and RCC estimation.

Given network and user parameters as input, e.g., cell-specific info, traffic mixture, user arrival rate and users' average speed, end-users are generated in the first step for 24 hours. In this step, users' parameters, i.e., SNR, service type, service duration and arrival time, are initialized according to their statistical distributions, Section 4.4.2. After that, load is generated in a loop for all the users who are active within the simulation interval, i.e., both the users whose arrival time is within the simulation interval

and those whose arrival time is before the simulation starts but their associated session has not ended yet. In case there is no user activity within the simulation interval, zero will be returned as the BBU load and RCC. The load generation process includes both services' packet generation and RB allocation, which is called as a subfunction, the red block, being explained in Figure 4.9.

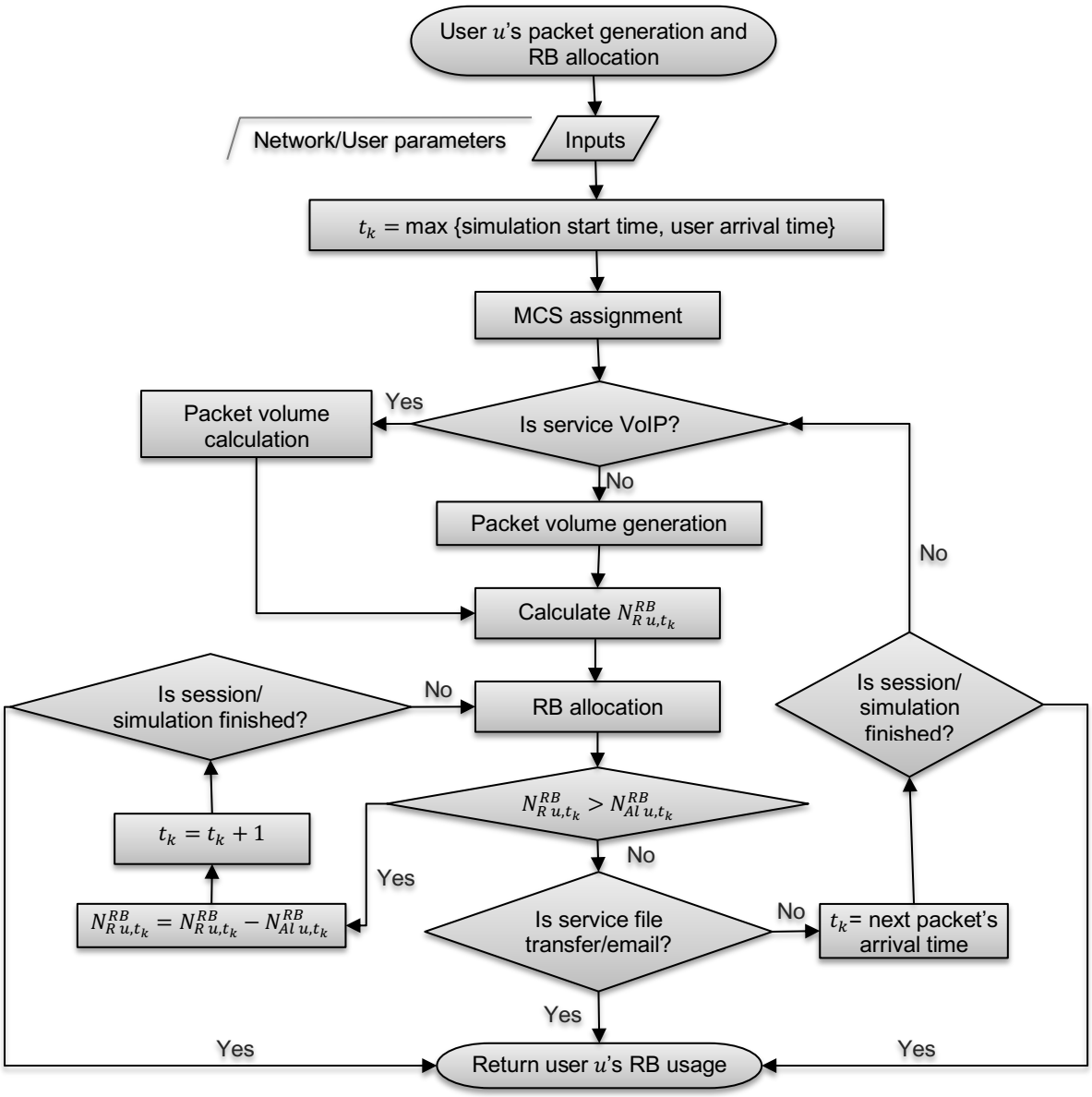


Figure 4.9 – The algorithm of calculating a user's RB usage.

As the network simulation's granularity is defined in a 1 ms scale (LTE TTI duration), the users' load generation is per millisecond. Summing up the entire users' load, the BBU's total load is achieved and the same for RCCs. All the results are stored then in the database. A users' load depends on the service being performed. The load generation process follows the traffic model defined in Annex D, which is particular for each service. As Figure 4.9 shows, given the user SNR, service type, service duration and arrival time as inputs, fed by the previous module, an MCS is assigned to the user in the first step. The MCS assignment is based on the user's SNR, as explained in Section 3.7.2. The users' load is generated

then according to the service type.

The load generation granularity for VoIP service is 20 ms, since VoIP transfers its associated packet once every 20 ms in LTE (VoLTE); however, the VoIP packet size depends on the ratio of time the user is in an active (talking) state, Annex D. The available RBs are allocated to users in accordance with their services' priorities. In case that the total RB demand is higher than the available ones in a given time instant, the capacity share of low-priority services is degraded, and their packets are delayed to compensate for the resource shortage of high-priority services, Section 4.4.2. Delayed packets are processed in the upcoming under-loaded sub-frames. For the other services, the packet size is generated randomly based on their given statistical distributions, Annex D. According to the packet size, the number of resource blocks required to transfer the packets, $N_{R\,u,t_k}^{RB}$, is calculated in the next step. However, the number of the available RBs bounds the maximum amount of RBs allocated to a user at a given time instant. Given the number of user's allocated RBs, $N_{Al\,u,t_k}^{RB}$, and the user/network parameters, the user RCC is calculated in the next step for the current time instant, Section 3.3.2. In case the service is file transfer or email, the traffic generation process is finished after the RB allocation. The user u's instantaneous load and RCC are then returned to the BBU b's traffic generation module as a result. Instead, a packet inter-arrival time is generated for the other services, and the whole process repeats until the simulation or the user session is completed. Once all the BBUs' load and accordingly, their RCCs are calculated, results are inserted to a database to be fetched for the resource allocation process, Figure 4.8.

Figure 4.10 shows the resource allocation algorithm.

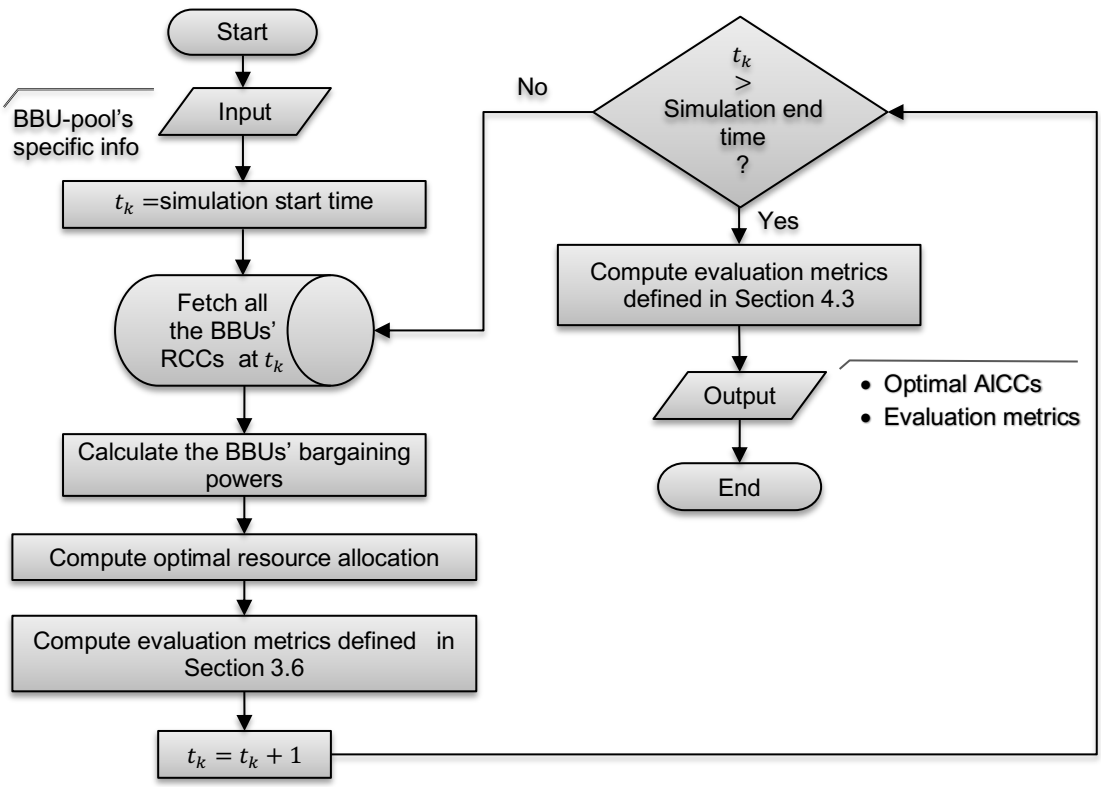


Figure 4.10 – The proposed computing resource allocation algorithm and the model evaluation.

Given the BBU-pool's specific info as inputs, BBUs' RCCs are fetched from the database in the first step. Accordingly, the BBUs' bargaining powers are calculated based on Section 3.4.3 in the next step. Results are then fed to the resource allocation block. In this phase, the optimization problem defined in Section 3.4, (3.38), is solved in order to find the optimum computing resource allocation for an individual time instant. The solution is achieved by exploiting CVX in MATLAB. Achieved results are then used for the metric evaluations, Section 3.5. The whole process repeats for the next time instant and the time dependent evaluation metrics are calculated, Section 4.3.

4.5 Canonical Scenario

The canonical scenario defined in Section 3.8 is extended here to define a time-varying network for the model assessment. The input parameters are summarized in Table 4.3. The number of active users, N_{b,t_k}^{u}, depend on the user arrival rate and the services' durations. The user arrival rate follows a mixture of Truncated Normal Distributions with the peak hours as 11 AM and 6 PM, the simulation being for 25 minutes from 6 PM onwards. The average user arrival rate is 155 and 110 users per minute for DL and UL, respectively. Considering the user arrival rate, traffic mixture, average service durations and packets inter arrival rate, it is expected to have on average 133 and 125 active users per second in DL and UL, respectively (for more information on these parameters, one is referred to Section 4.4.2).

Table 4.3 – Input parameters in canonical scenario.

Parameter \ BBU Index (b)		1	2
User	# spatial streams $(N_{Str\ u,t_k})$	2	
	SNR $(\gamma_{u,t_k[\mathrm{dB}]})$	19	
	User Arrival Rate $(R_{U[\mathrm{user/min}]}^{Arr})$	DL: 155, UL:110	
	Average mobility speed $(\upsilon_{Avg[\mathrm{km/h}]})$	30	
	Service (s_u^U)	{VoIP, Video, Web, file transfer, Email}	
	Packet volume $(V_{u,t_k[\mathrm{B}]}^{Pkt})$	Table 4.2	
Network	Cell Type (C^{TYP})	Micro	
	RRH traffic type (H^{TYP})	Residential	
	Channel Bandwidth $(\Delta f_{BW[\mathrm{MHz}]})$	20	
	Carrier Frequency $(f_{c[\mathrm{MHz}]})$	DL: 2110, UL:1920	
	quantization resolution $(Q_{[\mathrm{bit}]})$	24	
	MIMO Order (N_{MIMO})	4×4	
	# BBUs in the BBU-pool, (N_B)	2	
	BBU-pool AvCC $(C_{BP\ t_k[\mathrm{GOPS}]}^{Av})$	300	

The traffic mixture and service penetration are presented in Table 4.4. VoIP, video streaming, web browsing, file transfer and email services are considered in DL, while just VoIP, file transfer and email

are assumed for UL; this is due to the fact that video streaming and web browsing can be considered as file transfer in UL. The traffic mixture is considered the same for both BBUs. The service penetration shows the percentage of users per 24 hours performing that specific service, however, services' volume shares are the percentage of RB usage from the total one. Table 4.4 clearly shows that VoIP is the most requested service. Although video is the lowest requested service, it holds the highest volume share as the video data rate is the highest among all other services, Annex D. The considered user arrival rate and services lead to an average of 110 and 155 users in UL and DL, respectively. All users are assumed with 19 dB SNR for simplicity.

Table 4.4 – BBUs' service penetrations and traffic volume shares in canonical scenario.

Service	UL	DL	Service Weight	Service Penetration [%]		Volume Share [%]	
				UL	DL	UL	DL
VoIP	✓	✓	83	68	40	10	2
Video Streaming		✓	48	-	2	-	57
Web Browsing		✓	36	-	27	-	11
File Transfer	✓	✓		20	24	65	25
E-mail	✓	✓		12	7	25	5

4.6 Simulator Assessment

The simulator assessment is made in four steps:

1. analyzing the transitory interval at the beginning of a simulation,
2. acquiring the runtime of the simulator,
3. evaluating the traffic generation and comparing the generated samples frequency with their associated density functions,
4. studying the number of simulations that are required in order to have reliable results,

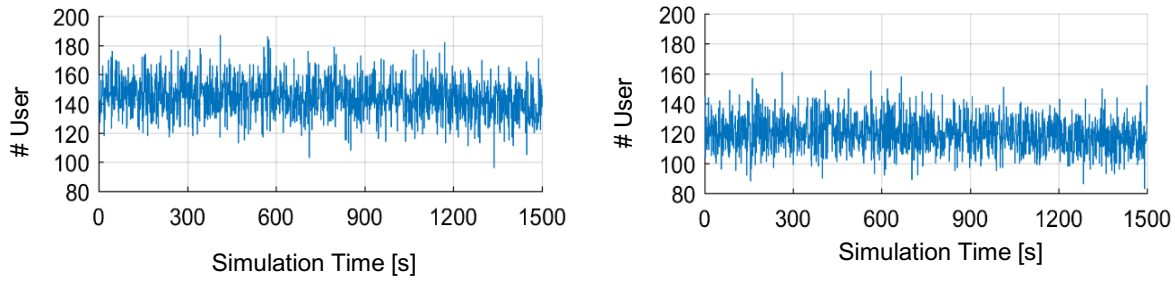
The assessment results are explained in detail in what follows.

4.6.1 Analysis of the Simulator's Transitory Interval

The simulator's transitory interval is evaluated in this section, based on the canonical scenario previously defined. The analyzed parameters are: BBUs' RB efficiency, RCC, AICC and the evaluation metrics defined in Section 4.3, i.e., average of BBUs' fulfilment levels, resource allocations' efficiency and resource usage. Although two BBUs are considered in the BBU-pool, the results are presented for only one of them, since the simulator's behavior is identical for both.

As mentioned earlier, users' arrival time and related service durations are generated randomly at the beginning of the simulation. The simulator initially considers users who enter the network during the simulation period. It also targets the users who have entered before and are still active when the

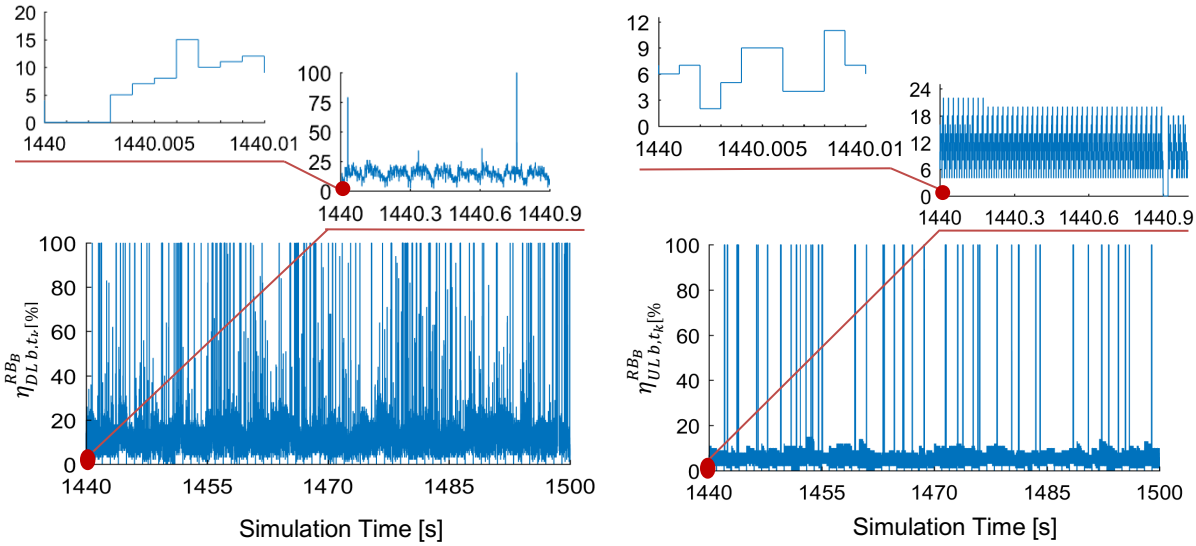
simulation starts. Therefore, irrelevant to the simulation's time, the network behavior is maintained in just a few milliseconds after the simulation starts. The number of active users per second for one of the BBUs is presented in Figure 4.11, where 138 and 123 average users per second are active in DL and UL in the BBU, which is coherent with the expected number mentioned in Section 4.5.



(a) DL (b) UL

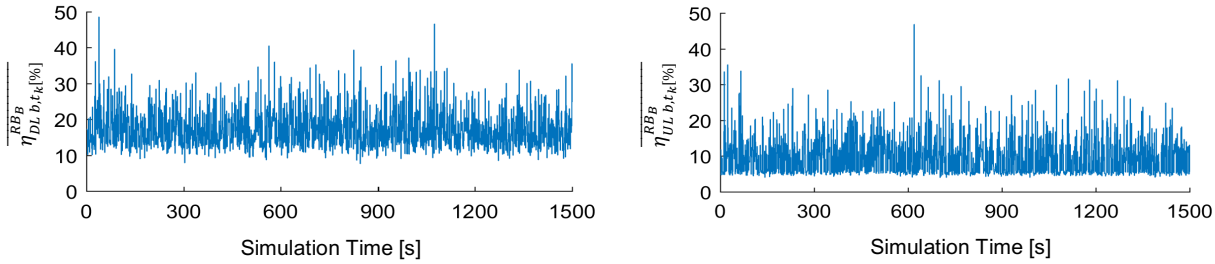
Figure 4.11 – Number of active users per second.

Figure 4.12 and Figure 4.13 present the RB efficiency per millisecond and per second, respectively, for the same BBU, confirming that it has a similar behavior in relation to the average value, from the beginning and over time.



(a) DL (b) UL

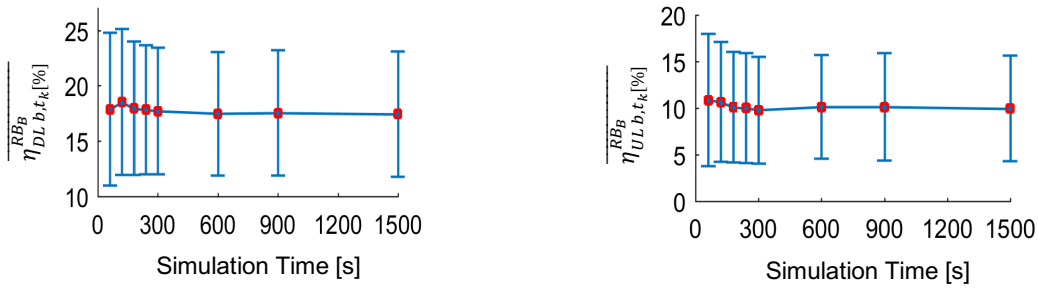
Figure 4.12 – RB efficiency of a BBU per millisecond.



(a) DL (b) UL

Figure 4.13 – Average RB efficiency of a BBU per second.

On the other hand, Figure 4.14 shows the average RB efficiency per second of the BBU in various simulations with different durations, where one can see that it changes slightly for the simulations with less than 200 s, but it remains almost constant for simulations with more than 200 s.



(a) DL (b) UL

Figure 4.14 – Average RB efficiency for different simulation intervals.

The figures related to the other considered parameters are presented in Annex E. It can be observed that the network behavior is maintained just after some milliseconds, and that all parameters have similar behavior in relation to the average value, at the beginning and over time. The simulator interval is analyzed based on the relative deviation percentage given by

$$\Delta_{[\%]} = \frac{\left|X^{aprox} - X^{ref}\right|}{X^{ref}} \tag{4.11}$$

where:

- X^{aprox}: value of parameter X,
- X^{ref}: reference value of parameter X.

The relative deviation for each of n millisecond simulations is achieved by comparing X^{aprox} with the average of all values collected for the total set of simulations as X^{ref}. The results are presented in detail in Annex E, where it can be observed that the relative deviation is less than or of the order of 0.03% for the simulations more than 200 s. Therefore, 200 s can be considered as the simulator transitory interval, and this initial time interval was not considered for algorithms assessment.

4.6.2 Runtime of the Simulator

In this subsection, the runtime of the simulator is studied. All simulations were performed on a desktop Personal Computer (PC) with a two-core Intel® Core™ i3-4150 3.50 GHz processor and 8 GB of memory. The implementation contains 3 000 lines of MATLAB code, of which about half are comments. While evaluating the simulator's speed, the MATLAB priority is high on the windows operating system.

As mentioned before, the simulation contains three main steps: traffic generation, RCC estimation and resource allocation. The simulator's runtime is not equal for all services' traffic generation and depends on the traffic mixture, due to the variety of packets arrival rates and volumes, e.g., more packets are generated for video and its volumes are larger than in other services, hence, the runtime is longer.

On the other hand, in the resource allocation step, the simulator exploits CVX as a modelling system for

constructing and solving convex problems. Exploiting ECOS as the solver in CVX and tuning the solver's solution tolerance to $\epsilon^{1/4}$ allow converging to the optimal solution faster while an acceptable precision is achieved. The computing resource allocation is executed once in a millisecond in accordance with the granularity of the network simulation. In general, the simulator takes 36 ms on average to find the optimal solution, (3.38), once.

Table 4.5 lists the simulators' average runtime for simulating one network minute in the canonical scenario defined in Section 4.5. For comparison, Table 4.5 also shows the simulator's runtime if the number of active users increases three times, being seen that the simulator takes 21 minutes on average to simulate one network minute traffic for an individual BBU, but by tripling the number of users, the simulator takes five times longer.

Table 4.5 – Average duration of the simulator's runtime for one network minute simulation per BBU.

		Average duration of the simulator's runtime [min]		
Average number of active users per minute (UL+DL)	Traffic mixture	Traffic generation and RCC estimation phase	Resource allocation phase	Total
800	Table 4.4	21	36	57
2400		110		146

Figure 4.15 presents the simulator's runtime related to the simulation interval for the same scenario, showing that runtime increases almost linearly with the increment of the simulation interval, the reason being that after the one-minute simulation, results are saved in the hard disc and memory is released.

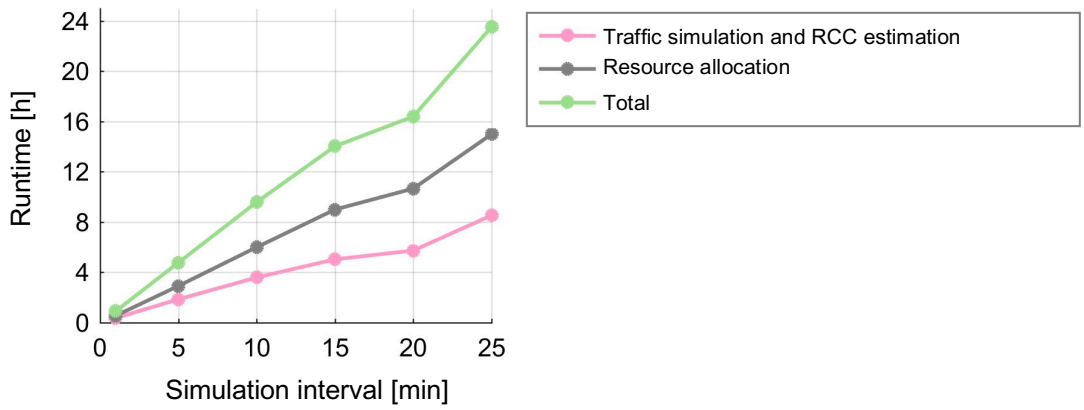


Figure 4.15 – Average duration to simulate one network minute per BBU in the canonical scenario.

The results confirm that it would be impractical to capture the long-time network behavior due to the long simulation runtime. However, since the simulator's transitory interval is 200 s, Section 4.6.1, and the most extended service takes on average seven minutes in defined scenarios, one can say that a ten-minutes simulation interval is enough to evaluate the model performance with desired accuracy.

4.6.3 Traffic Simulation

As described in Section 4.4, traffic is generated considering session, activity and packet levels. Figure 4.16 shows the traffic that a user generates per millisecond in the canonical scenario, Section 4.5, together with the RB efficiency for five kinds of services, i.e., VoIP, video streaming, web browsing, file transfer and email. The traffic behavior depends on the service's traffic model described in Annex C. The VoIP RB usage is in every 20 ms, and for the video service 14 packets arrive at the network every 100 ms, each with a randomly generated inter-arrival time. On the other hand, once the packet of web browsing, file transfer or email services arrive, the entire available bandwidth is allocated to it until the whole packet is transferred.

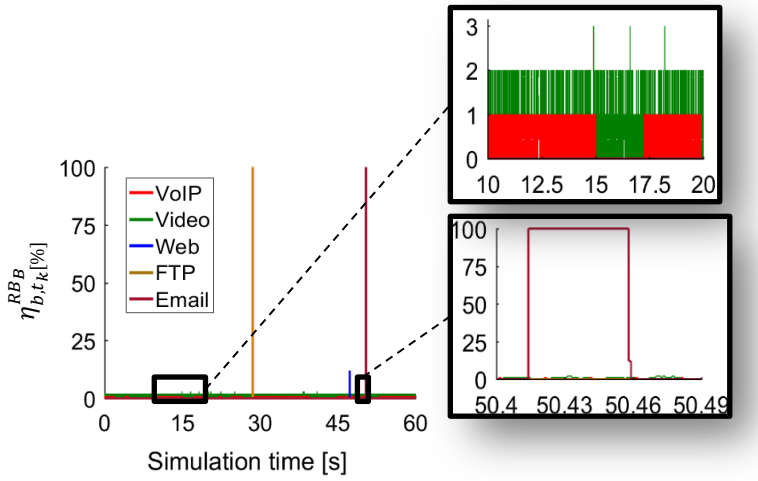


Figure 4.16 – Generated traffic for a single user.

The simulator's behavior is further studied by analyzing the generated samples' frequency. To this end, 35 simulations were observed, each with a 10-minute duration; and the PDF of any set of the generated samples was compared with the given PDF, in Annex D. It is noticed that a 10-minutes simulation is large enough to evaluate the simulator's behavior, as the network RB efficiency is stable after 200 s.

Figure 4.17 shows the given PDF for the file size in the file transfer service as an example. As explained in Annex D, a Truncated Lognormal Distribution is used to generate the samples with the mean and standard deviation of 1.996 and 0.7, respectively. The frequency and size of 40 random generated variables are presented in Figure 4.18. The mean value of the generated variables and their standard deviation are 1.989 and 0.69, respectively.

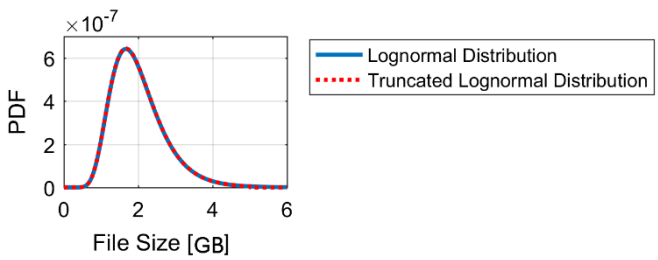
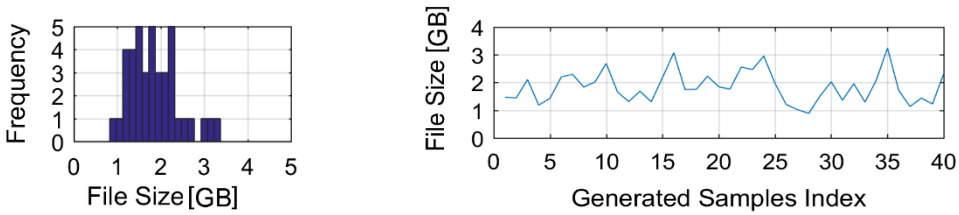


Figure 4.17 – PDF of the file size in file transfer.



(a) Samples frequency (b) Samples size

Figure 4.18 – Generated samples for the file size in file transfer service.

The mean and standard deviation of generated samples are compared with the theoretic ones in the given PDF to check the sample's validity. To this end, the means and the standard deviations of the samples in 35 simulations were obtained. The average of the obtained values was compared with the theoretic mean and standard deviation of the associated PDFs. The validation of the samples is done in two steps. In the first step, a comparison between the samples' mean/standard deviation and the theoretic ones is made by using a one-sample t-test; the comparison mentioned above is valid since the number of generated samples is large enough, i.e., more than 20, and the samples are independent and continuous. The null hypothesis in the test assumes there is no difference between the theoretic mean/standard deviation and the average of 35 simulations mean/standard deviation. The purpose of the one-sample t-test is to check if the null hypothesis is rejected.

Using MATLAB and applying one-sample t-test, the null hypothesis is rejected for none of the sample sets at the significance level of 5%, therefore, the average mean/standard deviation of the generated samples is statistically indistinguishable with the given PDFs mean/standard deviation. Results are listed in Table 4.6 and Table 4.7 for the file transfer service.

Table 4.6 – The average mean of generated samples in comparison with the theoretic mean.

Theoretic Mean	Avg. of 35 Simulations Mean	Null Hypothesis	Result (at significant level of 5%)
1.996	1.989	Avg. of 35 simulations' mean is equal to theoretic mean	Null hypothesis cannot be rejected

Table 4.7 – The average standard deviation of generated samples related to the theoretic one.

Theoretic Standard Deviation	Avg. of 35 Simulations Standard Deviation	Null Hypothesis	Result (at significant level of 5%)
0.7	0.69	Avg. of 35 simulations' standard deviation is equal to theoretic standard deviation	Null hypothesis cannot be rejected

In the next step, the generated samples are analyzed based on the relative deviation percentage (4.11). The randomly generated samples' standard deviation and mean, as the approximated values, X^{aprox}, are compared with the theoretic standard deviation and mean of the given PDF, as reference values, X^{ref}. The results are presented in detail in Annex F. The average of the relative deviation is less than

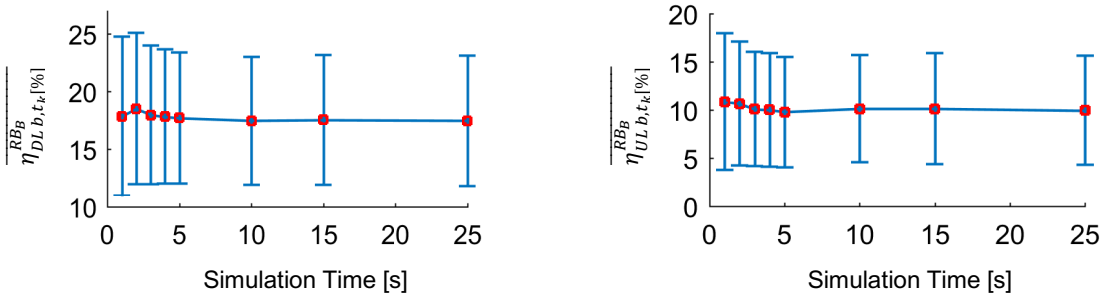
0.2% for all parameters, which is considered an acceptable accuracy for the generated samples, hence, one can say that all simulator's generated samples follow their associated PDF as listed in Annex D.

4.6.4 Sensitivity Analysis on the Number of Simulations

In this subsection, the sensitivity of results is analyzed relative to the number of simulations. To this end, 25 simulations have been performed, each with a 3-minute duration after the initial transitory interval of 200 s, output parameters being taken every millisecond. For each run, the generation of random values is done according to a different seed, affecting the values of the following input variables:

- Number of active users, N^U,
- Packet volume, V^{Pkt},
- User's SNR, γ.

The other input parameters of Table 4.3 are fixed in each simulation. Section E.2 contains the output parameters depicted as a function of the number of simulations. It can be observed that most of the average values are almost constant, independently of the number of simulations. The average of BBU's RB efficiency and fulfilment level are presented in Figure 4.19 and Figure 4.20, as examples.



(c) DL (d) UL

Figure 4.19 – Average of RB efficiency for different number of simulations.

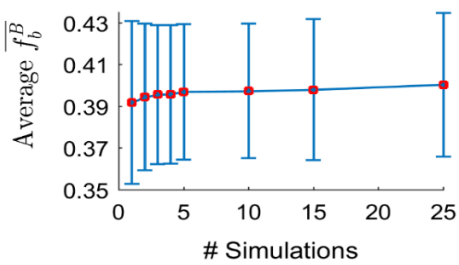


Figure 4.20 – Average of the BBU fulfillment level for different number of simulations.

In order to quantify the observations, the deviation percentage relative to the average of all values obtained for the simulations is computed from (4.11), where X^{aprox} is considered as the value of parameter X for n simulations, results being listed in Annex E. It can be observed that, for the analyzed parameters, the relative deviation of the average values is less than or equal to 0.1%. In conclusion, one can say that one simulation is enough to obtain values with the desired accuracy.

Chapter 5

Reference Scenario and Corresponding Results

This chapter aims to analyze the proposed computing resource allocation model's performance in terms of the BBU fulfilment level, resource allocation efficiency, fairness and resource usage. To this end, a reference scenario is characterized first in Section 5.1. BBUs' real-time demands are estimated, and optimal resource allocations are achieved. Accordingly, the evaluation metrics are assessed for one snapshot of the network and for time-varying traffic in Section 5.2 and Section 5.3, respectively.

5.1 Reference scenario

In order to evaluate the performance of the proposed computing resource management model, a reference scenario is defined in this section. An overall view of the scenario in depicted in Figure 5.1.

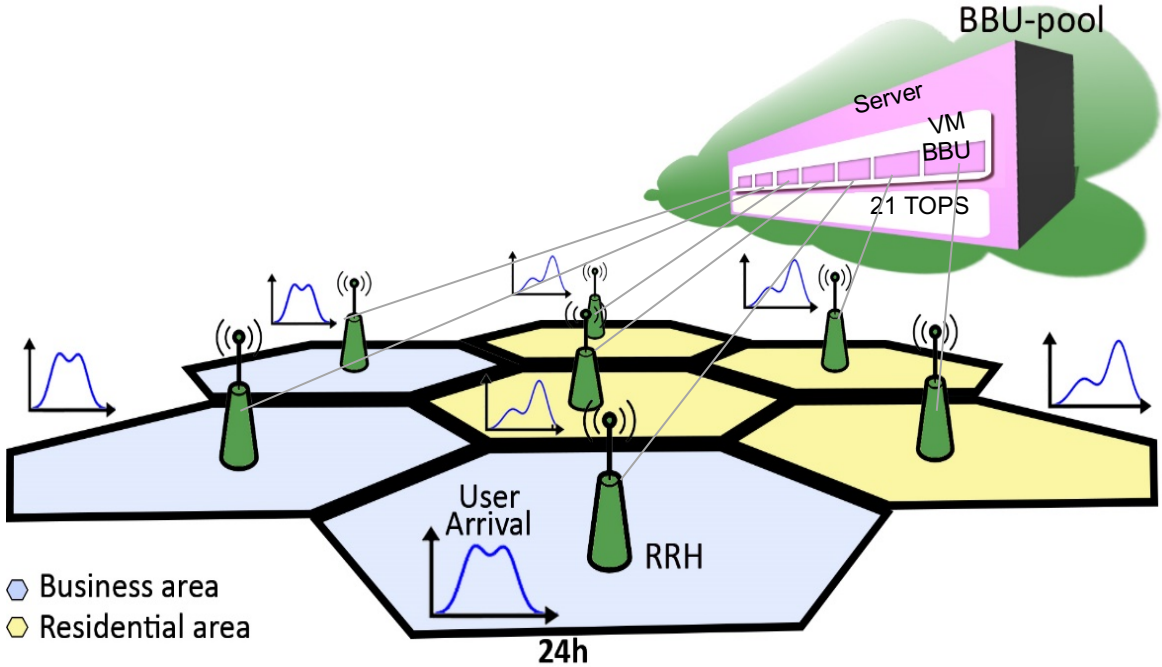


Figure 5.1 – Reference scenario.

Although a BBU-pool can host a diverse number of BBUs, [WRBL14], for clarity, however, the operating scenario includes 4 micro-cell RRHs in a residential area and other 3 in a business one. The RRHs are driven by 7 instances of BBUs, co-located in a single BBU-pool, where each BBU instance in the pool is associated with a single RRH. All BSs are configured with channel bandwidths of 100 MHz, 24-bit quantization resolution and support for 8×8 MIMO.

On the user side, terminals are assumed to have 8 spatial streams, enabling to have the optimum MIMO utilization. Users are outdoor with the average speed of 30 km/h, being distributed uniformly over the whole area. The user's SNR is represented by a random variable taken uniformly in [1, 35] dB at each time instant; accordingly, the modulation and coding ratio for a user is extracted from [3GPP17]. One should note that 1024 QAM is assumed to be the highest modulation offered by the network, leading to a BBU peak RCC of 12 TOPS for the proposed scenario, based on (3.19). To generate traffic demand, an attempt has been done to emulate a typical day of operation in cellular networks, the user arrival rate following the distributions presented in Section 4.4.2; however, due to hardware limitations, this resulted in a too long simulation time, hence, only 10 minutes of network time was simulated, starting at 6 PM (one of the peaks), with a time granularity of 1 ms. The average arrival rate is 140 and 883 users per minute in UL and 195 and 1243 users in DL for business and residential areas, respectively. The aforementioned network/user parameters that are required for BBUs' RCC estimations based on (3.11) and (3.15) are summarized in Table 5.1.

Table 5.1 – Input parameters in reference scenario.

Parameter \ RRH Traffic Type (H^{TYP})	Business	Residential
# Spatial Streams ($N_{Str\ u,t_k}$)	8	
SNR ($\gamma_{u,t_k[\mathrm{dB}]}$)	Uniform [1,35]	
User Arrival Rate ($R^{Arr}_{U[\mathrm{user/min}]}$)	UL:140, DL:195	UL:883, DL:1243
Average Mobility Speed ($\upsilon_{Avg[\mathrm{km/h}]}$)	30	
Service (s^U_u)	{VoIP, video streaming, video calling, web browsing, file transfer, email}	
Packet Volume ($V^{Pkt}_{[\mathrm{B}]}$)	Table 5.3	
Cell Type (C^{TYP})	Micro	
Channel Bandwidth ($\Delta f_{BW[\mathrm{MHz}]}$)	100	
Quantization Resolution ($Q_{[\mathrm{bit}]}$)	24	
MIMO Order (N_{MIMO})	8×8	
# BBUs in the BBU-Pool (N_B)	7	
BBU-Pool AvCC ($C^{Av}_{BP\ t_k[\mathrm{TOPS}]}$)	17.5	

The simulation includes a combination of heterogeneous services, i.e., VoIP, video calling/streaming, file transfer, email and web browsing. These 5 types of services were chosen according to the estimation that, until 2025, more than 90% of mobile traffic will be composed of the proposed service mix [Cerw20] (social networking and software down- and upload are considered as file transfer). Table 5.2 lists the achieved service weights based on (3.30) together with the link, i.e., UL or DL: VoIP and Video calling are simultaneous in both UP and DL; the other services can also be performed in both links but not simultaneously, but since video streaming and web browsing are usually in DL, they are considered as file transfer in UL.

Table 5.2 – Service Weights.

Service	DL	UL	Service Weight
VoIP	✓	✓	83
Video Streaming	✓	-	48
Video Calling	✓	✓	59
Web Browsing	✓	-	36
File Transfer	✓	✓	
E-mail	✓	✓	

For simplicity, it is assumed that users request only one type of service at a time. Service durations are randomly generated for VoIP, video calling/streaming and web browsing, based on a Poisson

Distribution with the mean values of 120 s, 300 s and 420 s, respectively. File transfer and email service durations, however, rely on the user file size and network data rate. Moreover, traffic generation is done at the packet level, where packet size and flow are characterized by stochastic models defined exclusively for each service [NGMN08]. According to the user modulation and coding ratio, the number of RBs that are required to transfer the generated packet is extracted from [3GPP17]. The service characteristics, e.g., the service duration and the packet volume, are summarized in Table 5.3 (for more details one is referred to Annex D).

Table 5.3 – Service characteristics.

Service	Service Parameter		Distribution	Mean	Standard Deviation
VoIP	Packet Inter-Arrival Time		Deterministic	20ms	-
	Duration		Poisson	120s	11s
Video	Frame Packets Inter-Arrival Time		Pareto	6.1ms	3.6ms
	Duration		Poisson	300s	17.3s
	Packet Volume		Pareto	1.3MB	257B
Web Browsing	Packet Inter-Arrival Time	Reading Time	Exponential	30s	
		Parsing Time		130ms	
	Duration		Poisson	420s	20.5s
	Packet Volume	Main Object Size	Lognormal	11MB	25.3MB
		Embedded Object Size		8.2MB	47.3MB
		Number of Embedded Object per Page	Pareto	7.6	10.4
File Transferring	File Size		Lognormal	2MB	700B
E-Mail				1.3MB	380B

The service penetration per cell, i.e., the percentage of the users per 24 hours that are using a specific service is summarized in Table 5.4, profiles with a dominance of VoIP (V) or File transfer (F) and Mixed without dominance (M) being used. The service penetration in Table 5.4 was designed so that each BBU has a different type of service as the highest ratio of running service, so that one can analyze model performance in allocating resources based on service priorities. The BBUs might serve RRHs being in a Residential (R) or Business (B) areas. BBU names in Table 5.4 denote both the area location and service dominance:

- RV: Residential area with VoIP dominance,
- BV: Business area with VoIP dominance,
- RF: Residential area with File transfer dominance,
- BF: Business area with File transfer dominance,
- RM1/RM2: Residential area without service dominance (Mix) (area location and traffic mixture are considered the same for these BBUs, the goal being to analyze the model behavior for BBUs with equal conditions),
- BM: Business area without service dominance (Mix).

Table 5.4 – Service penetration of the BBUs in reference scenario.

Service ID	Service Penetration [%] / BBU Index ($\boldsymbol{b}$)					
	VoIP dominance (RV, BV)		File transfer dominance (RF, BF)		Without service dominance (RM1, RM2, BM)	
	DL	UL	DL	UL	DL	UL
VoIP	60	71	2	3	15	18
Video Calling	1	1	1	1	1	1
Video Streaming	1	-	1	-	1	-
File Transfer	22	26	80	94	67	79
E-mail	2	2	2	2	2	2
Web Browsing	14	-	14	-	14	-

The traffic volume shares per BBU being generated by given service penetration are summarized in Table 5.5. The volume shares of services are the percentage of the RB request from the total available ones. Given the number of the packets per user per second, the packet volumes in MB (being generated randomly) and the user SNR (also being generated randomly), the number of the required RBs are specified. Traffic simulation and implementations have already been explained in detail in Section 4.4.4.

Table 5.5 – Traffic volume share of the BBUs in reference scenario.

Service ID	Traffic Volume Share [%] / BBU Index ($\boldsymbol{b}$)					
	RM1, RM2, BM		RV, BV		RF, BF	
	DL	UL	DL	UL	DL	UL
VoIP	0.5	1	3	6	0.1	0.1
Video Streaming	21	-	32	-	19	-
Video Calling	21	25	32	47	19	22
Web Browsing	4	-	8	-	4	-
File Transfer	52.5	73	26	46	56.9	76.9
E-mail	1	1	1	1	1	1

The deployed computing capacity of the BBU-pool is assumed to be 21 TOPS, based on its average RCC in the peak hours for the defined scenario, from which 83% (17.5 TOPS) is assumed to be the maximum resources that all BBUs are allowed to utilize for signal processing, in order to avoid data-center saturation, and the rest remaining for the other functionalities of BBU-pool.

In what follows, the BBUs' RCCs are estimated, and the BBUs' bargaining powers are calculated as the first step of the proposed resource management model. Afterwards, the optimal resource allocation is found and the performance metrics, defined in Section 3.6 and Section 4.3, are evaluated, accordingly. To have a closer look at the model's performance, one time instant (selected arbitrarily) is taken at first, and then the performance of the model is evaluated over time for the 10 minutes of the simulated network traffic.

5.2 Time Instant Analysis

In order to evaluate the proposed model's performance in more detail, a single snapshot of the network is taken in this section and the evaluation metrics are assessed accordingly. Traffic demands are evaluated as the first step of the proposed resource management model, and the optimal solution for resource allocation are achieved afterwards, the model's performance being assessed, accordingly.

Results in the values of RCC, the average weight of active services, minimum guaranteed computing capacity and bargaining power per BBU are listed in Table 5.6. The presented RCCs show the computing demands of active users at the taken snapshot, wherein BBUs are sorted by demand, i.e., from the lowest to the highest one. Since the simulation is at the packet level, a user is counted active in a given time instant if s/he is transmitting a packet at that time. However, packet transmission is not continuous (the packet inter-arrival time is variable, depending on the type of the service, Annex D), hence, a user may not transmit a packet at a specific time instant and being counted inactive while s/he still has an active session, Figure 4.6. This is the reason why the BBU demands presented in Table 5.6 do not follow exactly the traffic pattern listed in Table 5.5 (e.g., estimated RCCs for BBUs RM1, RM2 are not the same, even though both of them have the same user arrival rate and traffic volume share).

Table 5.6 – BBUs' RCC, average weight of active services, minimum guaranteed AICC and BP at t_k.

BBU Index ($\boldsymbol{b}$)	BM	BF	BV	RF	RM2	RM1	RV
$C^{R}_{b,t_k[\mathrm{TOPS}]}$	0.06	1.17	1.33	3.15	3.54	5.45	10.3
$C^{R_{\min}}_{b,t_k[\mathrm{GOPS}]}$	20.5	95.37	51.54	141.82	97.11	159.97	147.23
$\overline{w^{SRV}_{b,t_k}}$	55.00	45.00	71.20	56.00	69.37	64.74	76.10
$B_{b,t_k[\%]}$	0.13	2.91	5.47	10.14	14.35	20.57	46.44

Table 5.6 presents a higher RCC for residential BBUs compared to business ones, since the chosen scenario has a residential peak traffic demand leading to a higher number of active users, hence, a higher RCC for serving BBUs. The table also shows that the BBUs RV and BV have higher average of service weight, since the majority of their services are VoIP (the most top service priority). On the opposite, the BBU BF, with no VoIP, has the lowest average weight. Moreover, since BBU bargaining powers are combinations of both BBU demands and average weights of active services, one can see that BBU RV with both the highest weight and RCC has also the highest bargaining powers among all.

Once BBU demands are estimated, results are fed to the next step in order to find the optimal resources allocation. Figure 5.2 shows BBU AICCs in comparison with their RCCs. Although none of the BBU demands can be fully met due to the resource shortage, the minimum guaranteed computing capacity i.e., $C^{R_{\min}}_{b,t_k}$, are provided for all BBUs. Once the allocator assigns $C^{R_{\min}}_{b,t_k}$ to BBUs, it distributes the rest of the available resources among the BBUs with respect to the priority of each one, i.e., BBU bargaining powers. BBUs with higher bargaining powers are allocated with more resources while none of their AICCs exceed their demands.

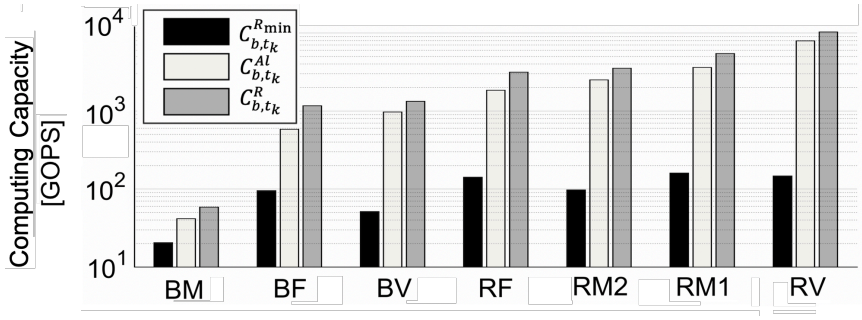


Figure 5.2 – BBU AlCCs vs. RCCs.

Comparing BBUs' fulfilment levels, Figure 5.3, also confirms that the resource allocator takes priority of active services into account while distributing resource among BBUs. Regardless of the BBUs' requirements, all their demands are fulfilled proportionally to the average weight of their active services. BBUs RV and BV's demands are fulfilled more than the other BBUs, since their active services have the highest average weight and hence the highest priority. It is also evident that although the demand of BBU BV is much less than the BBU RF ones, its fulfilment level is much higher for the same reason. Results confirm that the resource allocation is 100% fair, fairness being defined as the closeness of fulfilment level of BBUs to the weight of their ongoing services, (3.42).

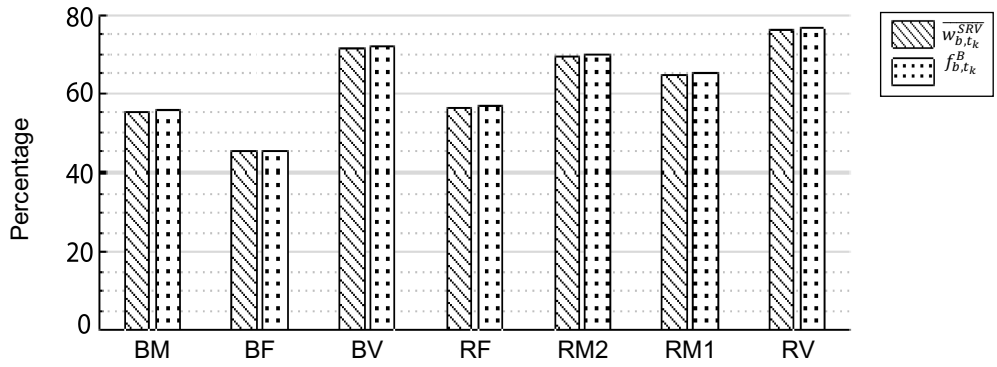


Figure 5.3 – BBU fulfilment levels vs. average of the service weights.

Moreover, since total demand, i.e., 21 TOPS, at the taken time instant is higher than the BBU-pool AvCC, i.e., 17.5 TOPS, the allocator assigns the available resources to BBUs entirely, leading to 83% usage of existing resources (17% being preserved for signaling overhead and saturation prevention). Full use of available resources also leads to the allocation efficiency being 82% higher than the traditional approaches, which assign resources to the BSs statically based on their peak demands. BBU-pool evaluation metrics are summarized in Table 5.7.

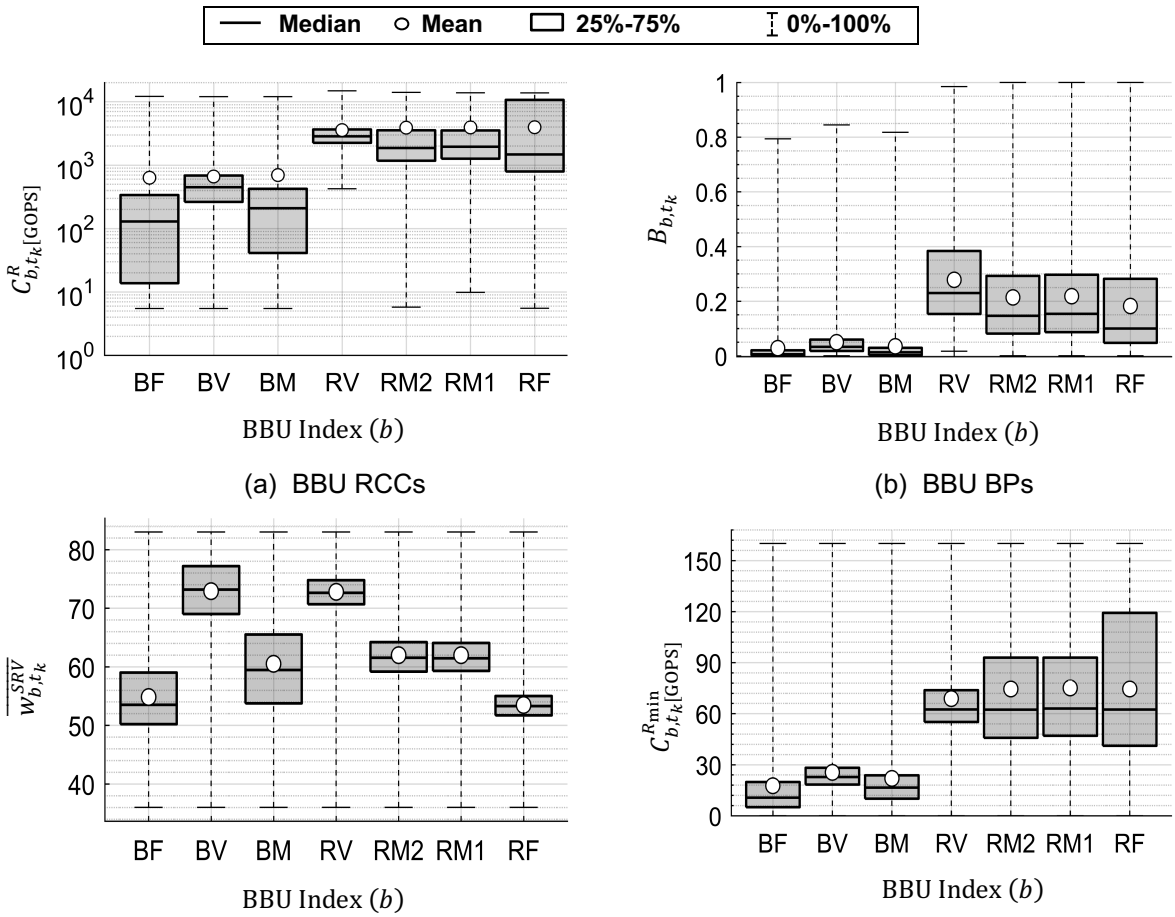
Table 5.7 – Evaluation metrics at t_k.

$F_{t_k[\%]}$	$U_{t_k[\%]}$	$\eta_{t_k[\%]}$
100	83	82

5.3 Time Dependence Analysis

This section addresses the performance of the proposed resources allocation model over time-varying traffic and demand. BBUs' RCC are estimated for each time instant separately and optimal resources allocations are achieved. Accordingly, the performance metrics defined in Section 4.3 are assessed.

Following the model described in Chapter 4, the average of BBUs' RCCs, bargaining powers, ongoing service weights and minimum guaranteed RCCs are computed for 10-minutes of simulated network traffic. Simulation results are shown in Figure 5.4, where BBUs are sorted by demand, i.e., from the lowest to the highest one.



(a) BBU RCCs

(b) BBU BPs

(c) BBU service weights

(d) BBU minimum guaranteed RCCs

Figure 5.4 – Average of BBU RCCs, bargaining powers, service weights, minimum guaranteed RCCs and number of active users within simulation interval.

Figure 5.4(a) presents a higher RCC for residential BBUs compared to business ones, since the chosen scenario has a residential peak traffic demand leading to a higher number of active users, hence, a higher RCC for serving BBUs. For the same reason, the minimum guaranteed RCC of BBUs follows a similar pattern, Figure 5.4(d); the values are relatively smaller, since the minimum guaranteed RCC accounts only for the CP processing steps' demands.

Regardless of the RRH type, simulation results show a reasonably equal average service weights for BBUs with the same traffic mixture, Figure 5.4(c). In addition, since VoIP has the highest service weight, BBUs RV and BV, with the highest proportion of VoIP, have the highest average among all other BBUs. Despite the same average of service weights, BBU RV has a bargaining power higher than the BBU BV one, Figure 5.4(b), which is due to the fact that it is a function of both service weights and RCC, thus, an unequal RCC may lead to different BBU bargaining powers.

Given BBUs' RCCs, minimum guaranteed RCCs and the average weight of ongoing services as inputs, the allocator calculates BBUs' bargaining powers and distributes the BBU-pool AvCC among BBUs proportional to their bargaining powers so that resource utilization is maximized. Figure 5.5 shows the total AICC of the BBU-pool in comparison with its RCC per millisecond, the total amount of the allocated resources never exceeding the available ones (17.5 TOPS), since 100% of the available resources are used in bottlenecks. In contrast, when the total demand is less than the BBU-pool's AvCC, resource allocation is bounded by the amount that is required, the rest of resources being remained available.

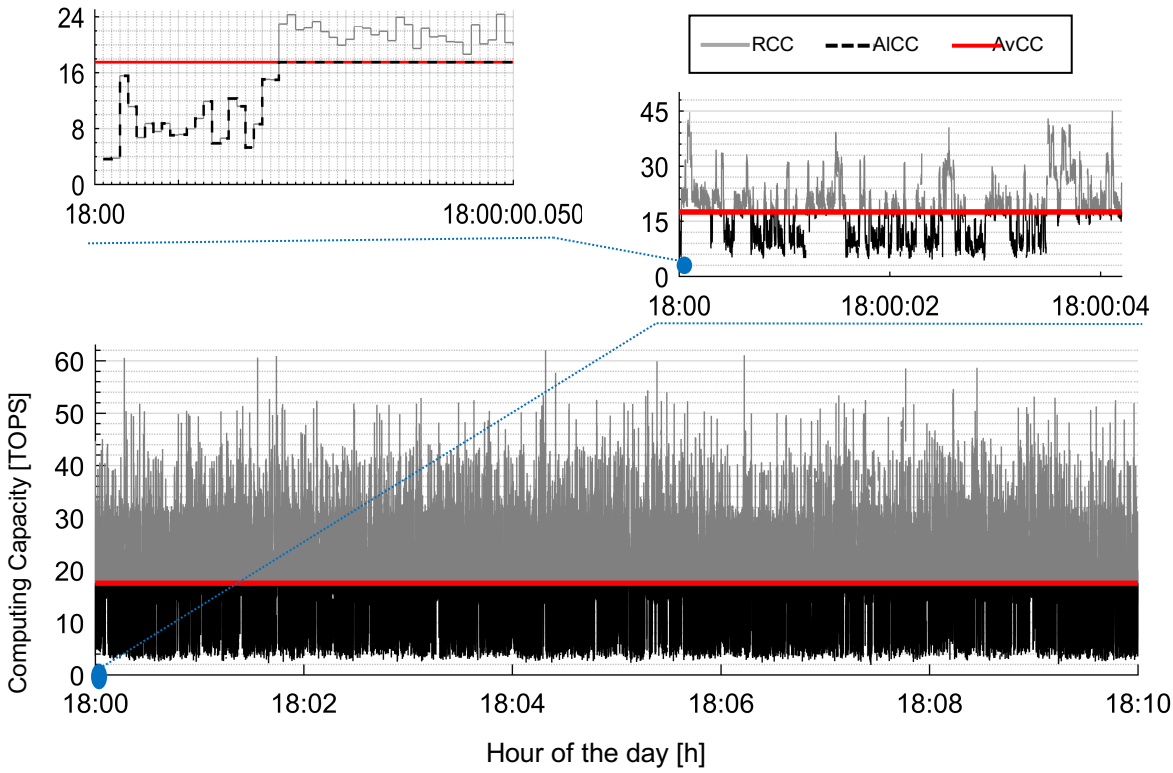


Figure 5.5 – BBU-pool total AICCs vs. its total RCCs per millisecond.

Figure 5.6 compares BBUs' AICCs with their RCCs, over the 10 minutes simulated network traffic. The allocator assigns the minimum guaranteed resources, $C_{b,t_k}^{R_{\min}}$, to each BBU and distributes the remaining resources proportionally to their user processing requirements, $C_{U\ b,t_k}^{R}$, therefore, the allocation follows a pattern similar to BBUs' demands. BBUs in the residential area, with higher demands, receive more resources than in business ones; moreover, BBU RF and BF receive the highest and the lowest resources, since they have the highest and the lowest demand among all BBUs, respectively.

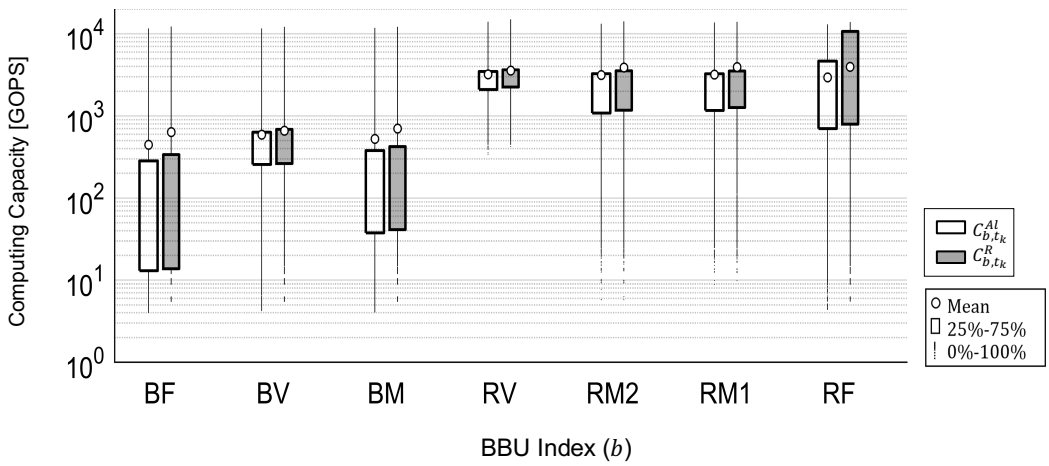


Figure 5.6 – BBU AlCCs vs. RCCs.

The difference between BBUs' RCC and their AlCC in Figure 5.6, stems from two facts: due to the dynamicity of the network, BBUs' demands fluctuate over time, hence, there are time instants that the BBU-pool's total demand exceeds the available resources, and BBUs' AlCCs are less than their RCCs due to the resource shortage; since the allocator takes QoS, hence, service priorities, into account, while distributing resources among the BBUs, it assigns more (less) resources to BBUs with higher (lower) average service weights.

The consideration of priority of active services is more apparent by comparing the fulfilment level of BBUs in Figure 5.7. BBUs with higher service priorities have higher fulfilment levels. One can see the effect of service priority by comparing BBUs RF and BV: although the RCC of RF is much higher than BV's, Figure 5.4(a), its fulfilment level is smaller than BV, since it has a lower average of ongoing service weight, Figure 5.4(c). Moreover, BBUs RV and BV have the highest fulfilment level among all, as their services have the highest weights on average.

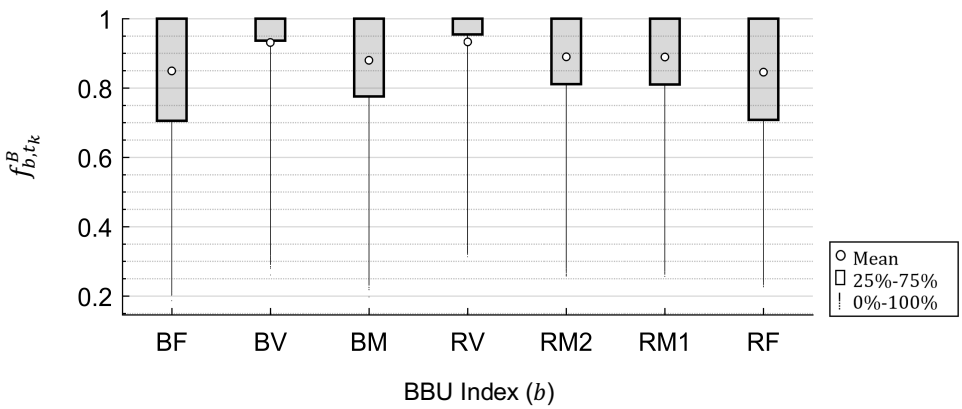


Figure 5.7 – BBU AlCCs vs. RCCs.

The values for BBUs' fulfilment levels are the result of using 66% of the existing average resources, Table 5.8. Resource usage can vary in the range of [0, 83] % in general, depending on the available resources being fully used or not (17% of the existing resources are preserved for signaling overhead

and saturation prevention). Because of the load fluctuations, the BBU-pool's total demand may be less, equal, or more than the BBU-pool's AvCC at a given time instant. The available resources are entirely used if they are less or equal than the total demand, otherwise, they remain idle since the allocator bounds the BBUs' assigned resources to their demand, due to the allocation strategy.

Table 5.8 – Evaluation metrics at t_k.

	Mean [%]	Standard Deviation [%]	Minimum [%]	First Quartile [%]	Median [%]	Third Quartile [%]	Maximum [%]
U_{t_k}	66	21	2	45	82	83	
η_{t_k}	84	5	80			89	99

The efficiency of the proposed computing resource management model is also presented in Table 5.8, showing that it is 84% more efficient than the fixed resource provisioning based on the peak traffic demands. The minimum efficiency never drops below 80%, which stems from the fact that the peak amount that the proposed resource management model allocates to the BBUs, i.e., 17.5 TOPS, is 80% less than the fixed amount that the traditional approaches assign to them, i.e., 12 TOPS.

The results presented in this chapter confirm that the proposed model can efficiently manage the available resources of the BBU-pool in congestions. In these cases, 100% of resources are used and resources provided to the BBUs are consistent with their real-time demands and proportional to the priority of ongoing services, meaning that the model considers QoS while distributing resources among BBUs. Results also confirm that resource provisioning is 100% fair, fairness indicating closeness of the proportion of BBU AICCs to the average priority level of their ongoing services. In the next chapter, the effect of the model's input parameters variation on its performance is assessed.

Chapter 6

Scenarios and Analysis of Results

This chapter compares the performance of the proposed resource allocation model with other resource allocation schemes. Moreover, the effect of the model's input parameters variation on its performance is analyzed. Section 6.1 presents an overview of the chapter. The comparison of the model's performance with equal and demand proportional resource allocations schemes is presented in Section 6.2. Section 6.3 and Section 6.4 analyze the effect of BBU-pool available computing capacity variation and user arrival rate variations on the model's performance, respectively.

6.1 Overview

The proposed resource allocation scheme's performance is compared in this section with two other resource provisioning strategies, namely EAS and DAS (defined in Section 3.5). To this end, the reference scenario in Section 5.1 is considered and the resource allocation phase is repeated for each allocation scheme separately; the evaluation metrics are being assessed afterwards.

The effect of the input parameters on the proposed model's performance is also assessed in this chapter. To this end, a series of new scenarios are considered over the reference scenario (defined in Section 5.1) by varying a set of relevant input parameters of the computing resource allocation module (Figure 4.2) as follows:

- **BBU-pool's AvCC variation**: As one of the inputs of the computing resource management module, the effect of the BBU-pool's computing capacity on the proposed model's performance is assessed by varying its AvCC within [0.4, 83] TOPS, which is equivalent to the BBU-pool's existing resources being varied within [0.5, 99] TOPS (17% of the resources are preserved to prevent saturation).
- **BBUs' RCCs variation**: In order to assess the effect of BBU demands on the proposed model's efficiency, different hours of the day (which leads to different rates of user arrivals) are considered, Table 6.1. Due to the user arrival rate variation during the day, BBUs' RCCs fluctuates, which leads to the variation of RB efficiency, η^{RB}, in on/off-peak hours; this is an input parameter to the RCC estimation module, Figure 4.1, which is significant on the complexity of a BBU signal processing, (3.11) and (3.15), therefore, its fluctuation leads to RCC variations. Since the ultimate goal is to assess the effect of BBUs' RCCs variation on the proposed resource allocation model, there is no difference on which input parameter on the RCC estimation is selected to be changed.

By variation of a BBU's RCC, its minimum guaranteed RCC, which is another input parameter to the resource allocation module, is also fluctuating. Moreover, variations of BBUs' service weights are not evaluated separately, since different service mixtures are considered for the BBUs in the reference scenario, enabling the assessment of the effect of service weights on a BBU's allocated resources. Table 6.1 summarizes the scenario road map.

The outputs (evaluation metrics) considered for model assessment are the ones defined in Section 3.6, for one time instant, namely:

- BBU fulfilment level,
- fairness index,
- efficiency of dynamic resource allocation,
- resource usage,

and in Section 4.3, for time-varying demands, namely:

- average BBU fulfilment level,
- average efficiency of dynamic resource allocation,
- average resource usage.

Table 6.1 – Simulation and the input parameters value that are considered to be changed.

Input Parameter		Effecting	Reference Scenario	Values
$C^{Av}_{BP\ t_k[\mathrm{TOPS}]}$ ($C_{BP[\mathrm{TOPS}]}$)		BBU-pool's Capacity	17.5 (21)	[0.4, 83] ([0.5,100])
$R^{Arr}_{U[\frac{\mathrm{user}}{\mathrm{min}}]}$	Residential areas	BBUs' RCCs	{DL:1243, UL: 883} (06:00PM)	{DL: 15, UL: 11 } (03:00AM) {DL: 368, UL: 260 } (08:00AM) {DL: 455, UL: 323 } (11:00AM) {DL: 373, UL: 268 } (01:00PM) {DL: 668, UL: 475 } (03:00PM) {DL: 113, UL: 80 } (11:00PM)
	Business areas		{DL:195, UL: 140} (06:00PM)	{DL: 0.25, UL: 0.13} (03:00AM) {DL: 258, UL: 183 } (08:00AM) {DL:1350, UL: 958 } (11:00AM) {DL:1180, UL: 838 } (01:00PM) {DL:1375, UL: 978 } (03:00PM) {DL: 0.25, UL: 0.13} (11:00PM)

The rest of the chapter is organized as follows. Section 6.2, compares the proposed QoS demand aware computing resource management scheme with the equal and demand proportional allocation schemes. The effect of BBU-pool's AvCC variation is assessed for both a given time instant and time-varying traffic demands in Section 6.3, and Section 6.4 analyses the impact of demand, i.e., user arrival rate variations.

6.2 Comparison among Different Allocation Schemes

This subsection compares the performance of the proposed model, QDAS, with the two other reference ones, EAS and DAS. To this end, the resource allocation phase is repeated for each of the allocation schemes separately over 10 minutes of simulated network traffic and the evaluation is done accordingly. The maximum resources that all BBUs are allowed to utilize in all experiments is 83%, in order to avoid data-center saturation. The results are narrowed down only to the bottlenecks. The goal is to evaluate the proposed model's performance compared to the other allocation schemes in congestions when total demand in the BBU-pool exceeds the available capacity.

The comparison is done considering all the evaluation metrics defined in Section 4.3 except the allocation efficiency. In bottlenecks, all mentioned schemes allocate 100% of available resources to BBUs due to their allocation strategies. As a result, efficiency does not vary remaining at 80%, which stems from the fact that the peak amount that the proposed resource management schemes allocate to BBUs, i.e., 17.5 TOPS, is 80% of the fixed amount that the traditional approaches assign, i.e., 12 TOPS.

Figure 6.1 compares BBUs' AICCs in the three allocation schemes. EAS allocates resources equally among BBUs, regardless of service priorities or BBU demands. Although EAS is a fast resource allocation scheme without too much complexity, it leads to a waste of resources if the BBU demand is less than its share. In such cases, some allocated resources are unused while a neighboring BBU may experience shortage. In contrast, DAS takes real-time demand of BBUs into account, allocating the minimum guaranteed resources, $C_{b,t_k}^{R_{\min}}$, to each BBU and distributing the remaining resources proportionally to their user processing requirements, $C_{U\,b,t_k}^{R}$; as a result, no BBU encounters a resource shortage in this scheme, while its neighboring BBUs are underutilized. For the same reason, as presented in Figure 6.1, BBUs in the residential area, with higher demands, receive more resources than in business ones; moreover, BBUs RF and BF receive the highest and the lowest number of resources, since they have the highest and the lowest demand among BBUs, respectively, Figure 5.4(a).

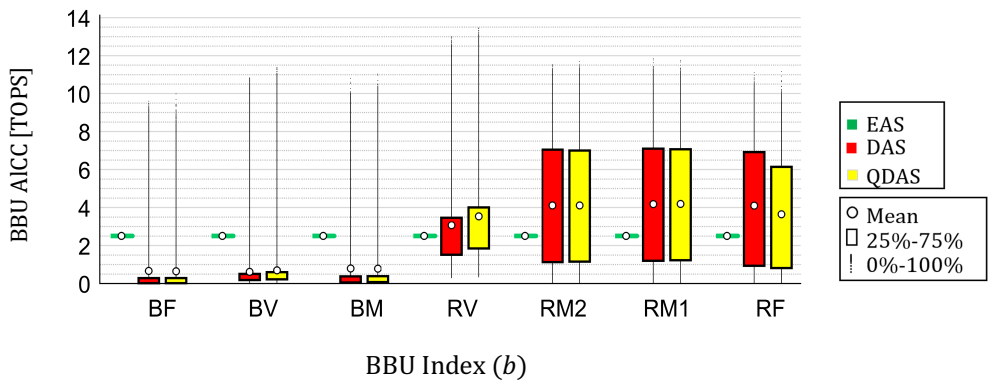


Figure 6.1 – AICC in different allocation schemes.

Similar to DAS, QDAS takes BBU demands into account, hence, resource allocation follows a similar pattern. However, the difference between these two approaches stems from the fact that, in addition to BBUs' demands, QDAS takes QoS, hence, service priorities, into account, thus, QDAS allocates more (less) resources to BBUs with higher (lower) average service weights, compared with DAS. The effect of service priority is apparent when comparing the AICC of BBUs RV and RF in DAS with the one from QDAS: Figure 6.1 shows that DAS allocates on average 3 TOPS to RV while QDAS increases its AICC to 3.5 TOPS, which is 16% more; in contrast, QDAS decreases the resources allocated to RF by 12% (from 4.1 TOPS to 3.6 TOPS) compared to DAS, since its services are not as critical as the ones in RV.

The overall resource usage is presented in Figure 6.2. It can vary in the range of [0, 83] %, depending on the available resources being fully used or not. Due to the dynamicity of the network, BBUs' demands fluctuate over time, hence, the BBU-pool's total demand may be less, equal, or more than the available resources at a given time instant. In the event that the total demand surpasses the available resources and none of the BBU's allocated resources exceed its demand, the available resources are fully utilized, hence, there is no wastage. In contrast, wastage may happen in two circumstances: when the available resources exceed the sum of all BBUs' demands, irrespective of the allocation policy; or, when the available resources are less or equal than the total demand, but a poor allocation policy distributes more resources to one (or more) BBUs than their demand.

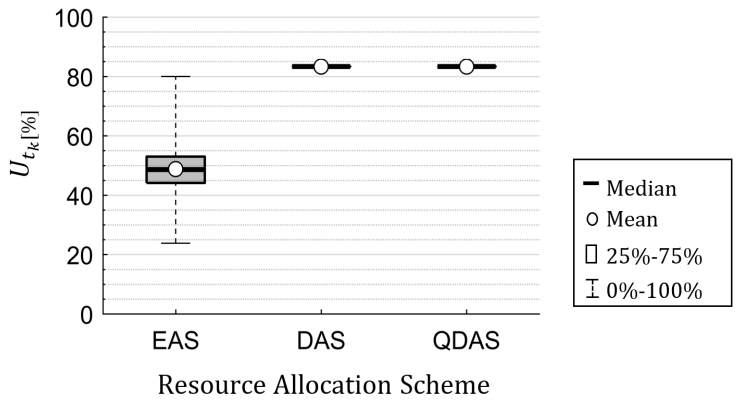


Figure 6.2 – Resource usage in different resource allocation schemes.

The low resource usage in EAS, Figure 6.2, is an example of resource wastage in the second circumstances, since it distributes resources evenly, regardless of the BBUs' demands, resulting that business BBUs are underutilized while residential ones are over-loaded. On the other hand, DAS and QDAS take BBUs' demands into account, thus, resources are fully utilized in both, since none of allocated resources exceed their demands.

Figure 6.3 illustrates that DAS fulfils all BBU's demands equally, irrespective of the priority of ongoing services, therefore, the resource allocation is not fair in the case of a shortage, because BBUs running critical services, i.e., services with lower delay budget and higher priorities, Table 2.8, require more resources to keep up with QoS. In contrast, QDAS supports QoS, so, BBUs with higher service priorities have higher fulfilment levels. One can see the effect of service priority by comparing BBUs RF and BV: although the RCC of RF is much higher than BV's, Figure 5.4(a), its fulfilment level is smaller than BV since it has a lower average of ongoing service weight, Figure 5.4(c). Moreover, BBUs RV and BV have the highest fulfilment level among all, as their services have the highest weights on average. By comparing with DAS, it is also apparent that QDAS shows a higher performance and increases the fulfilment level of BBUs RV and BV, by 13%, for the same reason.

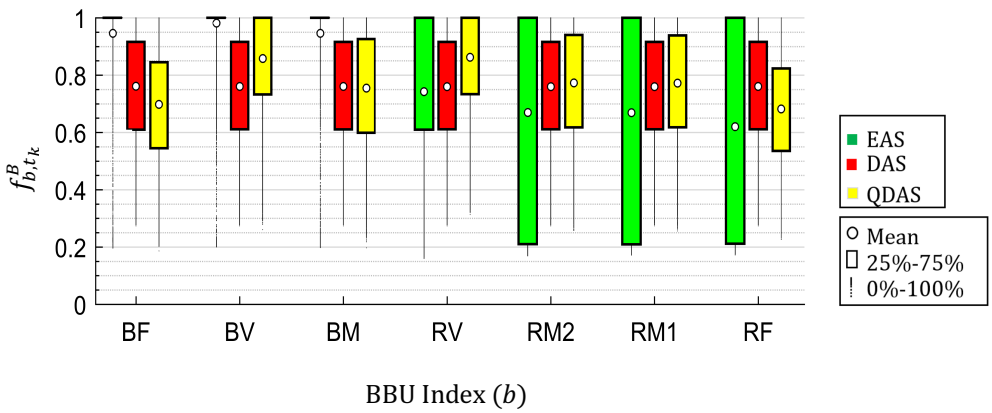


Figure 6.3 – BBU fulfilment levels in different allocation schemes.

Figure 6.3 also shows that EAS fulfils more BBU demands in business areas than in the residential ones, given the uniform resource allocation. This is an example of resource wastage, because for BBUs in business areas the demand is often less than the allocated resources, while, at the same time, BBUs in residential areas run into resource shortage, the outcome being a high (low) fulfilment level for the BBUs in the business (residential) areas.

As shown by results, the proposed model manages bottlenecks effectively and shows a higher performance compared with EAS and DAS. Unlike EAS, there is no wastage in QDAS during the congestions and it uses the available resources entirely in these cases. QDAS shrinks the capacity share of the lower priority BBUs in the bottlenecks to compensate for the higher priority BBU resource shortage. This is why the high prioritized BBUs' demands are fulfilled 13% more in QDAS than DAS, confirming that it considers the QoS while distributing resources among the BBUs.

6.3 Analysis of Available Computing Capacity Variation

The analysis of the effect of the BBU-pool AvCC on the proposed model performance is presented in this section. A single time instant is considered in the first subsection and performance metrics are evaluated accordingly. The model performance is assessed then in a real-time analytics platform in the following subsection.

6.3.1 Time Instant Analysis

In this subsection, the model performance is evaluated for a single snapshot. The results in the values of RCC, average weight of active services, minimum guaranteed computing capacity and bargaining powers for BBUs in the selected time instant are listed in Table 5.6. Figure 6.4 shows AICC in a BBU when BBU-pool existing resources, C_{BP} increases within [0.5, 30] TOPS. Since 17% of resources are preserved for signaling overhead and saturation prevention, this amount is equal to the AvCC being increased within [0.4, 26] TOPS.

When C_{BP} equals 0.5 TOPS, only the minimum guaranteed computing capacity is allocated to the BBUs due to the resource shortage. None of the BBU demands can be fully met before Th_1 since the sum of the RCCs is higher than AvCC. Once the minimum guaranteed requirements are allocated, the rest of the resources are distributed among BBUs with respect to the priority of each BBU, i.e., BBU BP. The effect of the BP is apparent when BBU index BV is compared to BBU BF. BBU BF receives more resources in the beginning because its minimum guaranteed requirement is higher than BBU BV. However, when AvCC increases, the AICC of BBU BV exceeds the AICC of BBU BF due to the fact that BBU BV has higher BP than BBU BF, hence, a higher priority in resource distribution in the pool. Figure 6.4 also depicts that the BBU minimum requirements are always guaranteed and that the BBU AICC never exceeds the RCC.

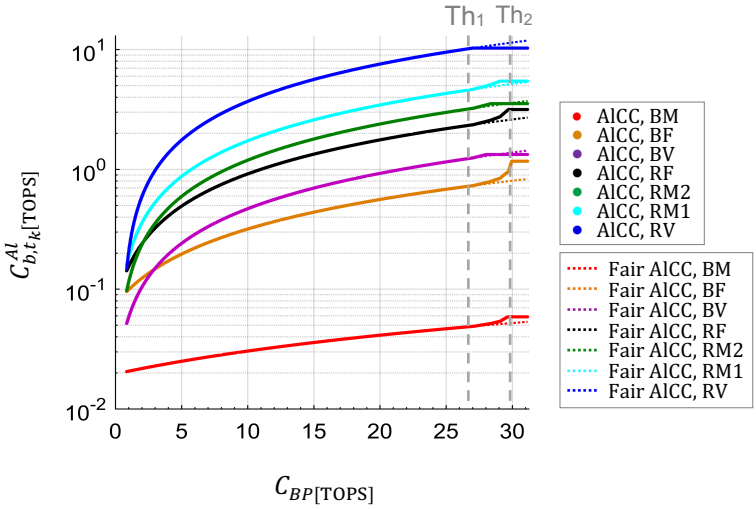


Figure 6.4 – BBU AlCCs.

The BBUs' fulfilment level is presented in Figure 6.5. By increasing AvCC, the BBU fulfilment level is also improved, proportionally to the average weight of the active services before threshold Th1. By comparing BBUs BF and RV, it is confirmed that fulfilment levels have the same proportion of the average weights of active services, i.e., 1.69 up to Th_1. Between Th_1 and Th_2, however, the fulfilment level of BBU BF grows faster, the reason being that the demand of BBUs with higher priority have already been met before Th_1; since the allocated resources to the BBU cannot exceed the demand, with the increase of AvCC, the remaining resources become available to the lower prioritized BBUs. It is also seen in Figure 6.5 that BBU RV is the first to receive 100% of its demand with the increase of AvCC since it has the highest average service weight among other BBUs in the pool; on the contrary, BBU BF is the last one that is fulfilled, since its active services have the lowest average weight.

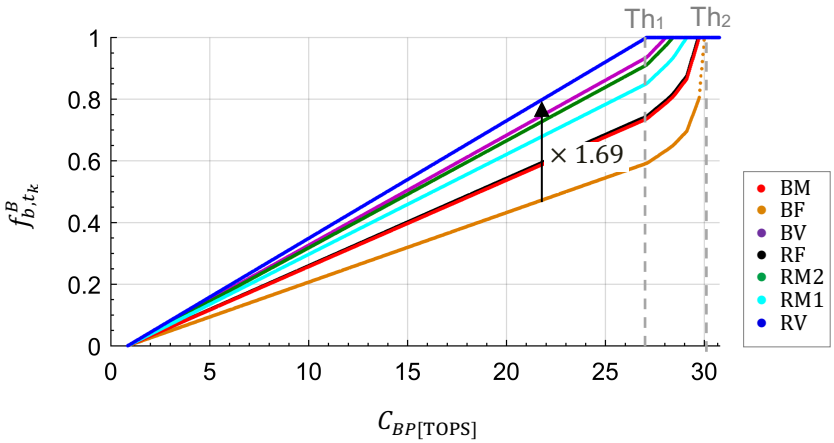


Figure 6.5 – BBU fulfilment levels.

The efficiency of the proposed resource allocation scheme is presented in Figure 6.6. With the increase of AvCC, the efficiency decreases, as more resources are used. Although AvCC is still increasing beyond Th_2, the resource usage does not increase anymore. The reason is that the resource-allocating scheme stops allocating more resources to the BBUs once their demand is fully met, hence, efficiency does not fall below 83%.

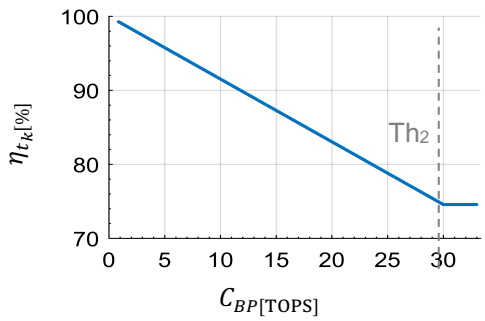


Figure 6.6 – Resource allocation efficiency.

Jain's fairness indicator, the last evaluation metric, is presented in Figure 6.7. The allocation is defined to be fair if the fulfilment levels maintain the same proportion of the average weights of active services. The fairness condition holds before Th_1, but beyond Th_1, however, the fairness indicator decreases due to the fact that the computing capacity proportional to the service weights is more than the RCC for the BBUs with high priority services. The resource allocation strategy bounds the AICC in BBUs to their RCC, so that the remaining capacity is distributed among those with lower service priority. As a result, the AICC of the BBUs with high priority services is less than the amount that is proportional to the average weight of their active services; on the other hand, the BBUs with lower service priority receive more than the average of their active services ratio. This ends with the decrease of the defined fairness index, as the fairness condition does not hold. The reduction of defined fairness index confirms that the resource allocator takes not only the priority of services but also the instantaneous requirement of the BBUs into account while distributing resources among them.

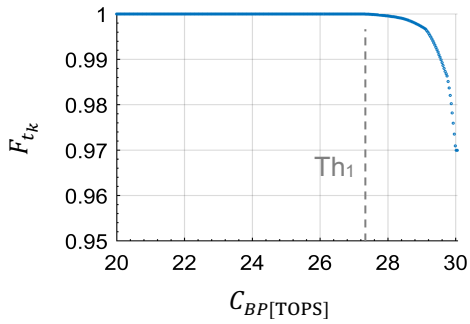


Figure 6.7 – Jain's fairness index.

6.3.2 Time Dependence Analysis

In order to analyze the impact of computing capacity on BBU fulfilment levels, resource usage and efficiency in a real-time platform, 10 minutes of the network is simulated with a time granularity of 1 ms. The resource allocation phase is repeated with the computing capacity of BBU-pool taken in [0.5, 100] TOPS. Each experiment runs for the 10-minutes simulated network traffic and model performance is assessed. It should also be noted that the maximum resources that all BBUs are allowed to utilize in each experiment is 83%, in order to avoid data-center saturation.

Figure 6.8 shows the capacity share of BBUs when the BBU-pool computing capacity, C_{BP}, , increases

from 0.5 to 100 TOPS. It is apparent that BBUs with higher bargaining powers, i.e., higher priorities, are allocated with more resources in the presence of a resource shortage, i.e., before Th_1. One can see the effect of the bargaining power by comparing BBUs RF and RM1. Although their RCCs have similar mean values, Figure 5.4(a), RM1 is allocated with more computing capacity before Th_1, since it has a higher bargaining power, hence, higher priority, while the computing resources are being allocated to BBUs. The resource allocator shrinks the capacity share of the lower priority BBUs in order to compensate for the higher priority BBU resource shortage. Beyond Th_2, 100% of BBU requests are served since the available resources are more than the overall demand.

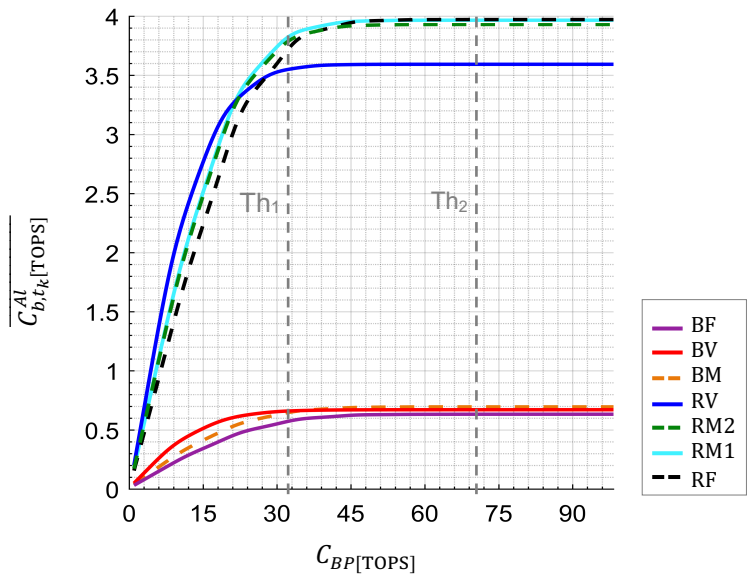


Figure 6.8 – Average of the BBU AlCCs.

The impact of BBU-pool capacity variations on the fulfillment level of BBUs in the pool is presented in Figure 6.9.

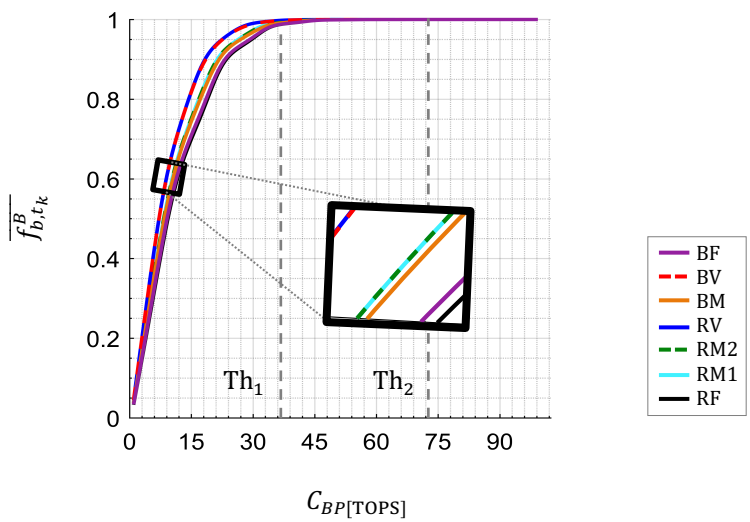


Figure 6.9 – Average of the BBU fulfilment levels.

Regardless of the demand, BBUs with higher service priorities account for higher fulfillment levels in the presence of a resource shortage. Moreover, the fulfilment levels for BBUs with a similar average of service weights are equal, since the proposed resource allocator keeps BBU AICCs proportional to the weight of their ongoing services. One can see the effect of the service weights by comparing BBUs RV and BV: although the RCC of RV is much higher than BV's, Figure 5.4(a), both are fulfilled reasonably equal. These BBUs also have the highest fulfilment level among all the others, since they have the highest average of the service weights.

The reader should also note that although increasing C_{BP} improves the average fulfillment level, as shown in Figure 6.9, correlation is not linear. For instance, when C_{BP} is doubled from 36 to 72 TOPS, the average fulfilment level is improved by only 2%, from 98% to near 100%. This becomes more important when the same boost in C_{BP} from 36 to 72 TOPS incurs a near 20% drop in average resource usage, as depicted in Figure 6.10. This behavior indicates that cloud providers should carefully consider the trade-off between BBU fulfilment levels and resource usage. An idea to decrease resource wastage, in this case, can be to reduce the available computing capacity in the BBU-pool, while degrading the capacity share of the delay-tolerant services in the BBU, to compensate for real-time services resource shortage.

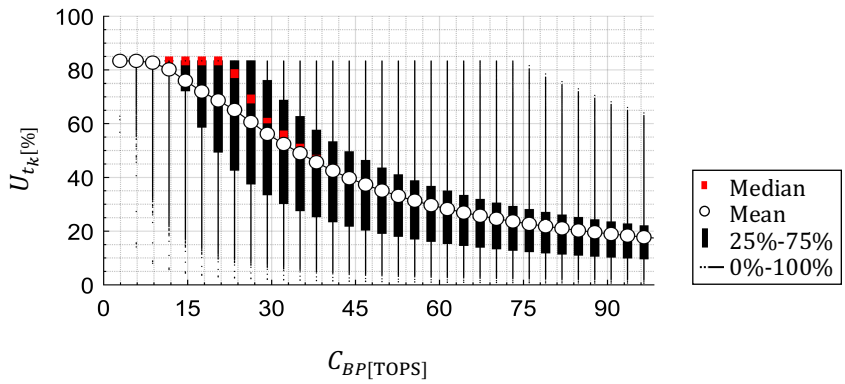


Figure 6.10 – Resource usage.

Figure 6.10 also shows that due to the severe resource shortage in the beginning, when C_{BP} is small, the available resources of the BBU-pool are almost entirely allocated among BBUs (17% of resources are reserved to prevent the datacenter from saturation). However, by increasing C_{BP}, the resource usage degrades: due to the dynamicity of the network, BBUs' demands fluctuate over time, leading to situations where, in some time instants, the total demand is less than the available resources, in these cases resources are not fully utilized, since the allocator bounds the BBU AICCs to their real-time demands. When C_{BP} rises, more resources remain unused, and hence, the resource usage drops.

The efficiency of the proposed resource allocation model is another metric that is calculated based on (4.9) and presented in Figure 6.11. The average efficiency of the pool declines when C_{BP} increases, the decline being faster in the beginning when C_{BP} increases from 0.5 to 22 TOPS: in this range, there is a resource shortage, so the available resources are instantly allocated, the direct outcome being the decline in efficiency as more resources become available in the beginning. Once the requirements of BBUs are fully met, and there is no more shortage, the allocator stops assigning more resources to BBUs (due to the

allocation strategy). Resources that become available afterwards, remain un-allocated, and efficiency drops slower beyond 22 TOPS. However, the average efficiency never drops below 83%, which is the total demand divided by the sum of separate peak demand of BBUs.

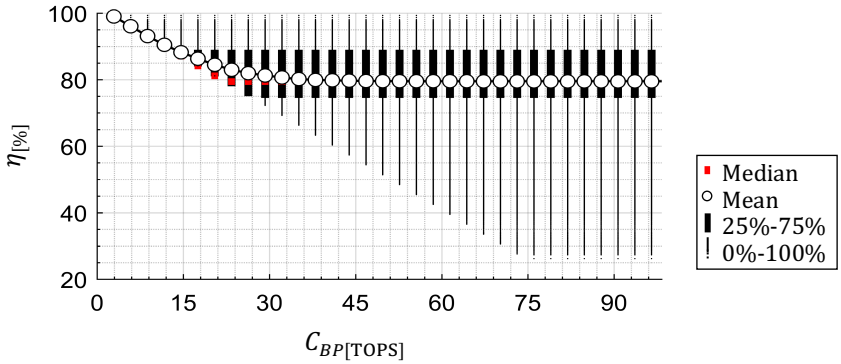
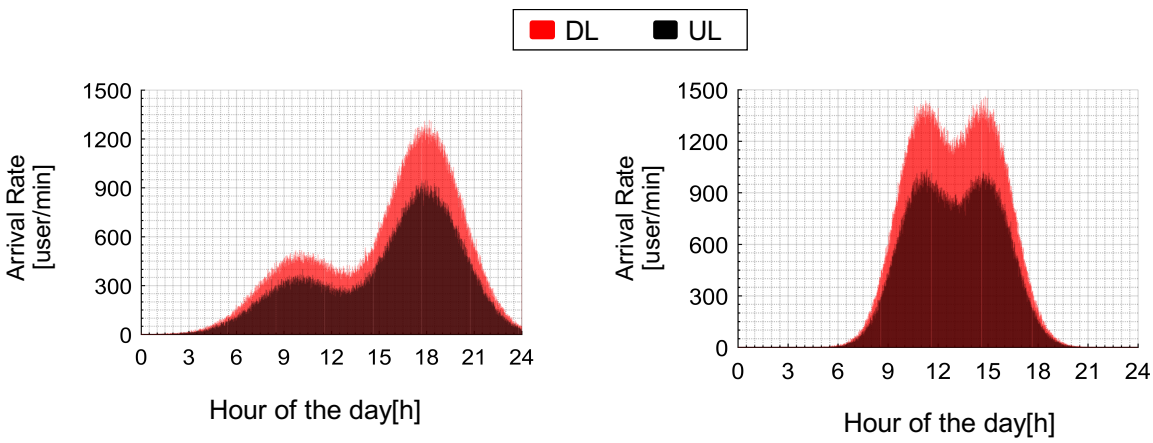


Figure 6.11 – Resource allocation efficiency.

6.4 Analysis of the Effect of User Arrival Rate Variation

In this section, the effect of the user arrival rate on the model's performance is evaluated by monitoring the network behavior during the day. Network traffic is simulated at 7 different hours, i.e., 03h00, 08h00, 11h00, 13h00, 15h00, 18h00 and 23h00, each lasting for a 10-minute interval. The selected hours include both peak hours and off-peak ones, the model's performance being evaluated for each one separately. The users' arrival rate is generated randomly following the pattern that was described in Section 4.4.2. The generated values are depicted in Figure 6.12 as the number of users per minute, showing the number of active users rising in the peak hours, i.e., 10h00 and 15h00 for residential areas and 11h00 and 18h00 for business ones.



a. Residential

b. Business

Figure 6.12 – User arrival rate.

Following the model described in Section 3.3, the average amount of the BBU RCCs, minimum

guaranteed RCCs, services weights and bargaining powers are calculated in the first step, Figure 6.13. As Figure 6.13(a) and Figure 6.13(b) present, the RCC rate is in line with the user arrival rate, Figure 6.12. The BBU computing capacity demand decreases in the off-peak hours, since less users are active, i.e., at 03h00 and 23h00. Figure 6.13(c) shows the BBU average service weights. BBUs with the same traffic mixture, Table 5.4, have almost the same average of the service weights. BBU RV and BV with the highest proportion of VoIP have the highest service weights since VoIP has the highest priority among all of the other services. Moreover, a BBU average weight varies slightly throughout the day as the traffic mixture is considered fixed. However, a BBU bargaining power fluctuates from one extreme to the other during the day, Figure 6.13 (d), the reason being that a BBU bargaining power is a variable of not only ongoing service weights but also the BBU demand, i.e., RCC. Since BBU RCCs fluctuate during the day, the bargaining powers vary proportionally.

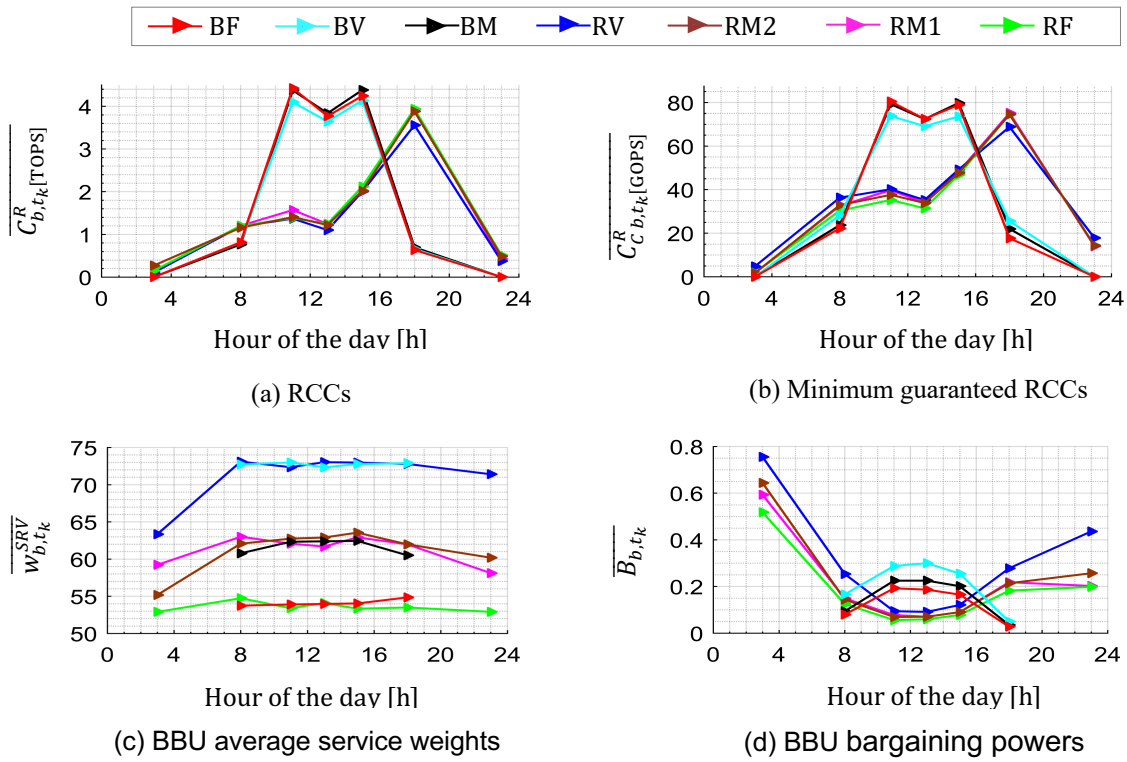
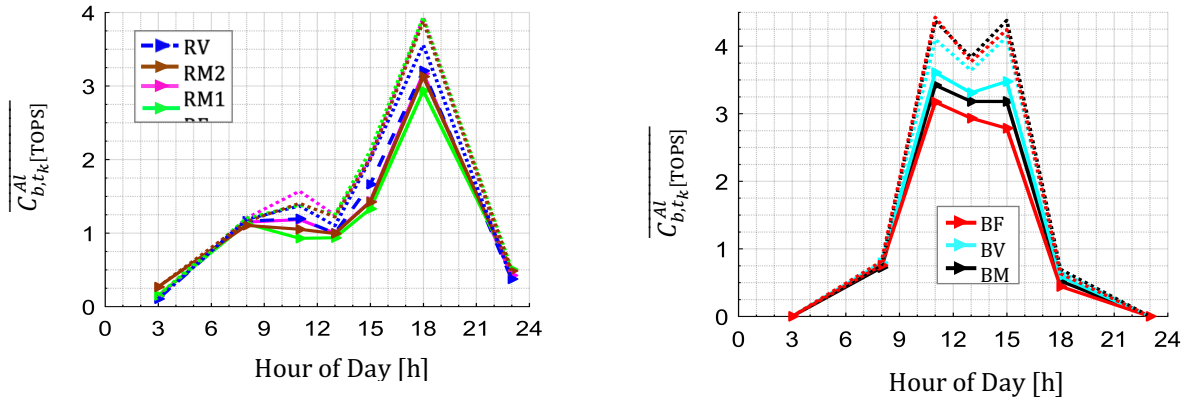


Figure 6.13 – Average of the RCCs, minimum guaranteed RCCs, service weights and bargaining powers within simulation intervals.

Figure 6.14 presents the effect of load variation on the capacity share of BBUs; the dashed lines represent the BBU RCCs. A BBU AICC changes during the day corresponding to its RCC fluctuation, meanwhile, it never exceeds the BBU RCC. As long as the available resources are large enough, all BBU demands are served. In the presence of resource shortage, however, BBUs are prioritized according to their BPs; in this case, the resource allocator decreases the capacity share of the lower priority BBUs, i.e., with lower BPs, in order to compensate for the higher priority BBU resource shortage.

Figure 6.15 shows the effect of traffic load on BBU fulfilment levels, which decrease with load

increments, since a lower proportion of demand is served due to resource limitation. In the conditions that none of the BBU demands can be served entirely, fulfilment levels are in proportion to service weights. So, fulfilment level of BBUs with the same traffic mixture, Table 5.4, are almost the same.



(a) Residential area (b) Business area

Figure 6.14 – Average of the BBU AlCCs during the day.

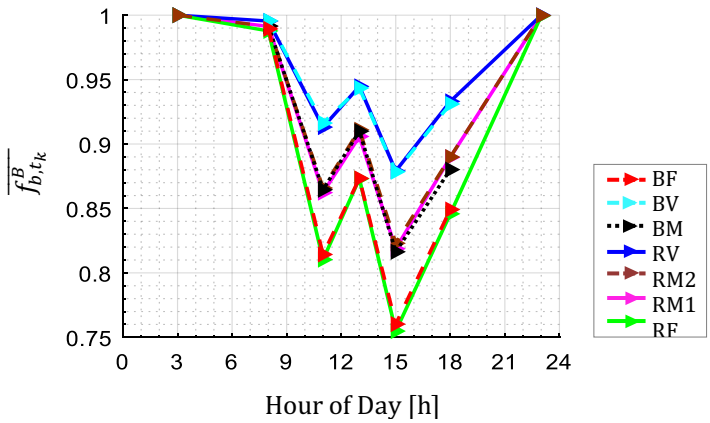


Figure 6.15 – Average of the BBU fulfilment levels.

One can see the effect of service weights on BBU fulfilment levels by comparing BBUs RV, BV and BF at 15h00. Regardless of BBU demands, the fulfilment levels of BBU RV and BV are almost the same, since their traffic mixture is similar, resulting in the same average of service weights. BBUs RV and BV fulfilment levels are also higher than in the other BBUs, since they are processing services with the highest priority levels on average. It is also apparent that the fulfilment level of BBU RV is higher than the BF one, although its demand is much less than BF. The results confirm that the resource allocator always takes the priority of ongoing services into account while distributing resources among BBUs.

The scatterplot of percentage of the BBU-pool resource usage is illustrated in Figure 6.16, the dotted line showing the BBU-pool average resource demand in terms of percentage of the existing resources. The resource usage decreases in the traffic off-peak hours, since the overall demand is lower, and the allocator terminates assigning more resources to the BBUs once their demand is entirely met. By

contrast, usage is increased in the peak traffic hours in line with the demand increment, however, it never exceeds 83%, since this is considered as the peak portion of the existing resources that is allowed to be used in order to prevent data-center saturation.

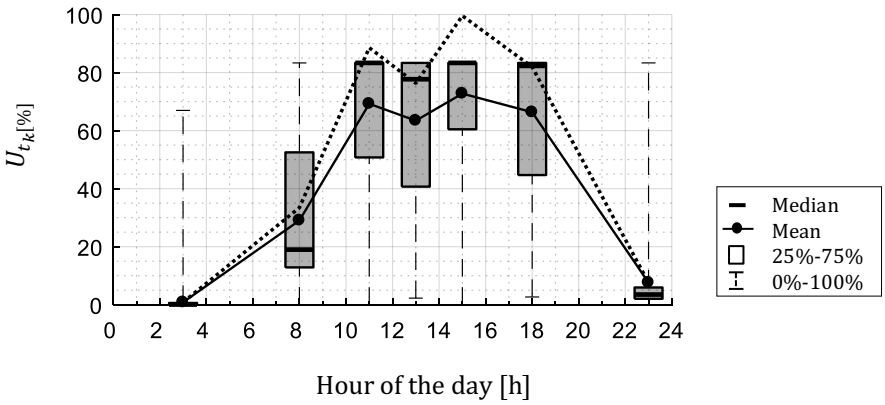


Figure 6.16 – Resource usage in different simulation intervals.

Figure 6.17 illustrates the efficiency of the proposed resource allocation model, showing that the average efficiency is more than 97% higher than the fixed allocation strategies during off-peak hours. The reason is that the model limits the BBU AICCs to their real-time demands, hence, the rest of the BBU-pool resources remain unused.

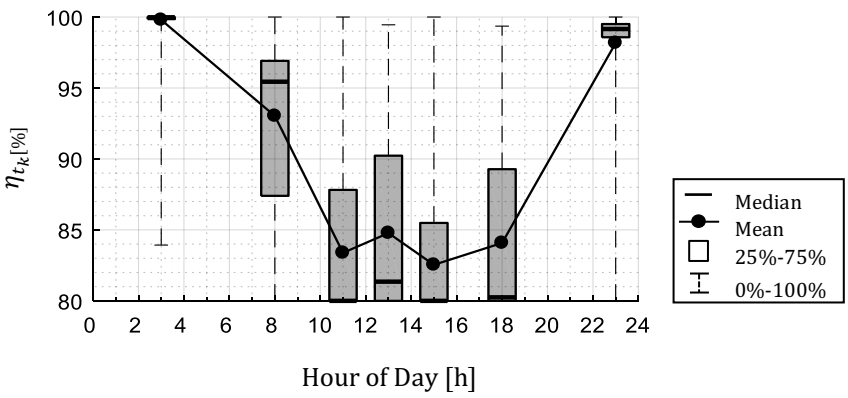


Figure 6.17 – Efficiency of computing resource allocation in different simulation intervals.

Chapter 7

Conclusions

This chapter finalizes this work by summarizing the main conclusions obtained and pointing out aspects to be developed in future work.

7.1 Framework and Novelty

C-RANs emerged in response to the need for higher data rates and capacity in upcoming mobile network generations: BBUs of BSs are decoupled from the radio units (RRHs), software-based BBUs are then consolidated in the servers of a data center, known as BBU-pool. C-RAN is a critical enabling technology of 5G providing higher data rates and lower network latencies. In C-RAN, utilization is improved and fewer resources are required compared to the sum of stand-alone BBU demands. However, a critical challenge of C-RAN is the data center's power consumption. Since computing resources are the most energy-intensive entities in data centers, it is worthwhile to apply efficient resource management strategies to maximize their utilization and reduce idle ones.

Designing efficient resource management strategies is a complicated process for cloud providers. Due to the variety of network services, user arrival rates and channel conditions, BBU resources demand fluctuate significantly throughout the day. On the one hand, a BBU computing capacity should suffice peak demands, while on the other hand, provisioning fixed resources based on peak requirements leads to idle resources in the rest of the day. As a result, an efficient resource management strategy in a BBU-pool should allocate the computing capacity dynamically, in accordance with the BBUs' instantaneous demand, while efficiently handling the resources in the case of a shortage.

This thesis focuses on computing resource allocation in C-RAN. A game-based optimization algorithm was developed to distribute the computing resources among BBUs in a BBU-pool whereby resources utilization is maximized. The model allocates computing resources on-demand, based on the instantaneous requests of BBUs, using a game-theory bargaining approach. In case the available resources are not sufficient to fulfil all instantiation requests, BBUs are prioritized to ensure the adequate QoS, low-priority ones being always guaranteed a minimum computing resource to avoid them to crash. Considering both QoS and BBU RCCs as real-time parameters, i.e., given based on TTIs, is essential not only in 4G deployments but also for the upcoming service-oriented 5G and ensures that the BBU-pool is provisioned with an optimum configuration, consistent with BBU demands.

The proposed model manages resources in two stages in the first step, BBUs' traffic demands being evaluated. Taking as inputs network and user parameters at a specific time instant, the estimation of BBUs' demands is based on a well-defined model proposed in the literature. The results are then fed into the computing resource allocation step in order to find the optimal resource allocation to BBUs. The two-fold solution maximizes both BBU-pool computing resource utilization and BBUs' processing speed. In the next time instant, the resource management process is re-instantiated over new input parameters.

The novelty of the proposed scheme is the consideration of the limits of the BBU-pool computing resources and the prioritization of BBUs in bottlenecks based on the characteristics of their ongoing services and QoS constraints. At the same time, the model guarantees all BBUs with a minimum computing resources to avoid crashing; furthermore, contrary to existing works, the proposed model has a low complexity and provides fairness of resource allocation and system efficiency, which makes it applicable in practical implementations.

7.2 Main Results

To evaluate model performance, an approach has been taken to emulate a typical day of operation in cellular networks in a scenario in which a BBU-pool includes 7 BBUs offering heterogeneous services with tidal traffic flows in a tidal channel condition. BBUs' instantaneous demands and their minimum guaranteed ones were estimated for the simulated network and the BBUs' bargaining powers and the average of the ongoing service weights were calculated as the first step of the proposed resource management model. Afterwards, the optimal resource allocation was found and the performance of the model was evaluated in terms BBU fulfilment level, fairness, resource usage and efficiency of the resource allocation.

In order to have a closer look at the model's performance, a single time instant (selected arbitrarily) is taken at first, and the model performance is evaluated accordingly. In the next step, the performance of the model is evaluated over time for 10 minutes of the simulated network traffic. The achieved results confirm that the proposed model efficiently manages resources in the case of congestions. Although none of the BBU demands can be fully met in these cases, due to the resource shortages, the allocator provides the minimum guaranteed demands to all BBUs and distributes the rest of the available resources among BBUs with respect to their bargaining power, i.e., priority, of each one, so that BBUs with higher bargaining powers are allocated with more resources. Moreover, none of BBUs' AlCCs exceed their demands. Results also confirm that 100% of the resources are fairly distributed among BBUs during the congestions, fairness being defined as the closeness of fulfilment level of BBUs to the weight of their ongoing services.

The comparison of the proposed model's performance with equal and demand proportional resource allocation schemes, which can be found in the literature as common allocation approaches, confirms that the proposed scheme shows a higher performance. There is no wastage in the proposed model during congestions and it uses the available resources entirely in these cases. Moreover, unlike the other two schemes, the proposed model shrinks the capacity share of the lower priority BBUs in the bottlenecks to compensate for the higher priority BBUs' resource shortages. This is why the high prioritized BBUs' demands are fulfilled 13% more in the proposed scheme than the other ones, confirming that it considers QoS while distributing resources among BBUs.

Besides, in order to analyze the impact of available computing capacity of the BBU-pool on the model's performance, an experiment has done wherein the resource allocation phase is repeated with the available computing capacity of BBU-pool varying within [0.4, 83] TOPS, the model performance being assessed for each run separately. The result shows that when the BBU-pool's AvCC is small, only the minimum guaranteed computing capacity is allocated to the BBUs due to the resource shortage. When AvCC increases, the proportional AlCC of BBUs with higher bargaining powers increases more than those with lower bargaining powers, since they have a higher priority in resource distribution in the pool.

The results also confirm that the BBUs' fulfilment level grows proportionally to the average weights of their active services when AvCC increases. Regardless of demand, BBUs with higher service priorities account for higher fulfilment levels in the presence of a resource shortage and by incrementing AvCC,

100% of the demand of BBUs with higher weight are fulfilled earlier that the other ones. Moreover, the fulfilment levels for BBUs with a similar average of service weights are equal, since the proposed resource allocator keeps BBU AICCs proportional to the weight of their ongoing services.

Results also demonstrate that although increasing AvCC improves the average fulfilment level, correlation is not linear and improving the average fulfilment level from 98% to 100% requires doubling the available resources at the cost of average resource usage being cut in half indicating a great waste of resources. When AvCC is small, the available resources of the BBU-pool are entirely allocated among BBUs. By increasing AvCC, the resources usage decreases. The result is that due to the dynamicity of the network, BBUs' demands fluctuate over time, leading to situations where, in some time instants, the total demand is less than the available resources, in these cases resources not being fully utilized, since the allocator bounds the BBU AICCs to their real-time demands. When AvCC increases further, more resources remain unused, and hence, resources usage drops. This behavior shows that cloud providers should carefully consider the trade-off between BBU fulfilment levels and resource usage. An idea to decrease resource wastage can be to reduce the available computing capacity in the BBU-pool, while degrading the capacity share of the delay-tolerant services in the BBU, to compensate for real-time services resource shortage.

And finally, the effect of the user arrival rate on the model's performance is evaluated by monitoring the network behavior during the day. To this end, network traffic is simulated at 7 different hours, i.e., 03h00, 08h00, 11h00, 13h00, 15h00, 18h00 and 23h00, each lasting for a 10-minute interval. The selected hours include both peak hours and off-peak ones, the model's performance being evaluated for each one separately. The amount of the BBU RCCs, minimum guaranteed RCCs, services weights and bargaining powers are calculated in the first step for each experiment, separately and the optimal resource allocation is found afterwards.

The results show that a BBU AICC changes during the day corresponding to its RCC fluctuation, meanwhile, it never exceeds the BBU RCC. As long as the available resources are large enough, all BBU demands are served. In the presence of resource shortages, however, BBUs are prioritized according to their bargaining powers; in these cases, the resource allocator decreases the capacity share of the lower priority BBUs, i.e., with lower bargaining powers, in order to compensate for the higher priority BBUs' resource shortages.

The result also confirms that by the increment of load a BBU fulfilment level decreases, since a lower proportion of demand is served due to resource limitation. In the conditions that none of the BBU demands can be served entirely, fulfilment levels are in proportion to service weights. So, fulfilment level of BBUs with the same traffic mixture are almost the same. Moreover, the resource usage decreases in the traffic off-peak hours, since the overall demand is lower, and the allocator terminates assigning more resources to the BBUs once their demand is entirely met. By contrast, usage is increased in the peak traffic hours in line with the demand increment.

7.3 Key Contributions

This dissertation is structured in seven chapters. In Chapter 1, a brief historical overview of the evolution of wireless technologies is given, the motivation and the main goals set for the dissertation are pointed out, the novelty and main contributions are mentioned, and a list of published work and internal reports is presented, being summarized as the following publications:

- 1 book chapter,
- 1 international journal paper,
- 3 international conferences,
- 6 technical documents in IRACON meetings.

The chapter concludes with a detailed description of the structure of the dissertation.

Chapter 2 gives an overview of 4G and 5G networks, which form the fundamentals of this thesis, including an overview of the network architectures and radio interfaces, QoS, coverage and radio capacity, and critical principles of C-RAN and virtualization with a focus on BBU-pool virtualization and related approaches. One also explains how the concept of game theory is used to solve a resource allocation problem, in general, and mentions the state of the art related to computing resource management in the C-RAN area.

Chapter 3 discusses the proposed computing resource management model considering a single snapshot of the network, including two main steps:

1. estimating the instantaneous computing capacity demand of the BBUs in the pool,
2. developing a game-based optimization algorithm, accordingly, in order to distribute the available computing resources among BBUs in a BBU-pool whereby resource utilization is maximized.

The chapter also presents the evaluation metrics defined to assess the proposed model and the model implementation details. At the end, a canonical scenario is defined, and the model performance is assessed accordingly.

Chapter 4 provides an extension of the model proposed in the previous chapter by addressing time-varying traffic and demand, and proposes a real-time computing resource allocation framework. The chapter discusses the proper time interval between two successive resource allocations and defines the metrics used to evaluate the proposed model in a real-time framework. Moreover, the details of the simulator implementation and its assessment are discussed.

By defining a reference scenario, the proposed computing resource allocation model's performance is analyzed in Chapter 5. To this end, BBUs' real-time demands are estimated first and optimal resource allocations are achieved accordingly. The evaluation metrics are assessed separately, for both a single snapshot and a time interval of the network traffic.

Chapter 6 compares the performance of the proposed resource allocation model with other resource allocation schemes. Moreover, the effect of the model's input parameters variation on its performance is analyzed.

Finally, the current chapter concludes the thesis. The framework and the novelty of the work is

summarized first, the principal results and achievements and the presented work's key contributions are mentioned afterwards. Potential improvements and directions for future works are also provided in the final section.

7.4 Future Works

In addition to what is proposed and assessed in this dissertation, there are several studies that can be considered as future works. Some potential topics can be proposed as follows:

The proposed computing resource management model maximizes the BBU-pool computing resource utilization while prioritizes BBUs in the shortages according to the weight of their active services. Services' weights are defined as fixed, being driven from the Priority Level that 3GPP has assigned to an individual service. The proposed weighting policy is fully compatible with a QoS maximization goal, as the Priority Level is a characteristic by which 3GPP specifies QoS requirements and determines the packet forwarding treatment. However, the prioritization policy can be improved by considering Packet Error Loss Rate and Packet Delay Budget, besides the Priority Level that 3GPP has assigned to an individual service. There is also a potential to define a dynamic weight to the services considering the delays imposed on packets and lost ones.

Another research direction could be the application of the proposed model in network slicing. Network slicing plays a critical role in the forthcoming 5G standard, network resources being shared among slices and a portion of them being allocated to each slice so that the specific requirements of given vertical applications are met. The proposed computing resource management model can fairly allocate resources among network slices in the critical situation in which the network does not have enough resources to fully satisfy slices' demands. It would also be an added value to joint this thesis' work (which is focused on computing resource management) with other available studies on radio resource management in order to propose an end-to-end model of 5G network slicing.

The work also can be extended to a joint design of cloud and edge processing. C-RAN is an impractical solution for many delay-sensitive applications because of the long distance between the device and the cloud center. Moreover, the proliferation of smart Internet of Things devices causes excessive load on the backhaul, between massive devices and BBU-pool servers. An alternative approach is to offload some of the computing tasks from cloud servers to the network edge. Considering service priorities, their delay budget, and the distance between the device and cloud centers, the proposed computing resource management model can be extended to an efficient joint cloud-edge resource management model.

Annex A.

Convexity Proofs

The convexity proofs of the defined bargaining game's utility function and solution set are described in this annex.

As mentioned before, GNBS is suitable for solving the problem and guarantees to find the optimal solution if the BBUs' utility functions and defined solution set is convex and closed. In this annex, their convexity proofs are presented in what follows:

Solution set: Since it is obvious that $S_{t_k}^{FS}$ defined in (3.28) is closed, only the convexity is approved: From the definition of a convex set [BoVa04], $S_{t_k}^{FS}$ is convex if and only if the line segment between any two points in $S_{t_k}^{FS}$ lies in $S_{t_k}^{FS}$. Indeed:

$$\forall\, \mathbf{C}_{t_k[N_B\times 1]}^{Al_1}, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2}:\ \mathbf{C}_{t_k[N_B\times 1]}^{Al_1}, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2} \in S_{t_k}^{FS} \implies \mathbf{C}_{t_k[N_B\times 1]}^{Al_3} \in S_{t_k}^{FS} \tag{A.1}$$

where:

$$\boldsymbol{C}_{t_k[N_B\times 1]}^{Al_3} = \theta\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_1} + (1-\theta)\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2} \quad : \quad 0 \le \theta \le 1 \tag{A.2}$$

$Proof$:

$$\mathbf{C}_{t_k[N_B\times 1]}^{Al_1}, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2} \in S_{t_k}^{FS} \overset{(3.28)}{\Longrightarrow} 0 \le C_{b,t_k[\mathrm{GOPS}]}^{Al_1} \le C_{b,t_k[\mathrm{GOPS}]}^{R} \text{ and } 0 \le C_{b,t_k[\mathrm{GOPS}]}^{Al_2} \le C_{b,t_k[\mathrm{GOPS}]}^{R}$$

Multiplying inequalities by nonnegative reals θ and $1-\theta$ and taking sum of the results, we have:

$$0 \le C_{b,t_k[\mathrm{GOPS}]}^{Al_3} \le C_{b,t_k[\mathrm{GOPS}]}^{R} \qquad : b = \{1,2,\dots,N_B\} \tag{A.3}$$

in the same way:

$$\mathbf{C}_{t_k[N_B\times 1]}^{Al_1}, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2} \in S_{t_k}^{FS} \overset{(3.28)}{\Longrightarrow} \begin{cases} \sum_{b=1}^{N_B} C_{b,t_k[\mathrm{GOPS}]}^{Al_1} \le C_{BP\ t_{k[\mathrm{GOPS}]}}^{Av} \\ \sum_{b=1}^{N_B} C_{b,t_k[\mathrm{GOPS}]}^{Al_2} \le C_{BP\ t_{k[\mathrm{GOPS}]}}^{Av} \end{cases}$$

and multiplying inequalities by nonnegative reals θ and $1-\theta$ and taking sum of the results, we have

$$\sum_{b=1}^{N_B} C_{b,t_k[\mathrm{GOPS}]}^{Al_3} \le C_{BP\ t_{k[\mathrm{GOPS}]}}^{Av} \tag{A.4}$$

Based on (A.3) and (A.4) it is concluded that $\mathbf{C}_{t_k[\mathrm{N_B}\times 1]}^{Al_3} \in S_{t_k}^{FS}$. Therefore, the feasible solution set $S_{t_k}^{FS}$ is convex.

■

Utility function: On the other hand, from the definition of a convex function [BoVa04], function $\mathcal{U}_{b,t_k}: \mathbb{R}^{N_B} \to \mathbb{R}$ is convex if its domain, which is feasible solution set $S_{t_k}^{FS}$, is a convex set and also, for all $\mathbf{C}_{t_k[N_B\times 1]}^{Al_1}, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2} \in S_{t_k}^{FS}$ and α with $0 \le \alpha \le 1$, we have:

$$\mathcal{U}_{b,t_k}\left(\alpha\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_1} + (1-\alpha)\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2}\right) \le \mathcal{U}_{b,t_k}\left(\alpha\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_1}\right) + \mathcal{U}_{b,t_k}\left((1-\alpha)\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2}\right) \tag{A.5}$$

$Proof$:

As based on the definition of the utility function, (3.22), we have:

$$\begin{aligned} \mathcal{U}_{b,t_k}\left(\alpha\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_1} + (1-\alpha)\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2}\right) &= \frac{\alpha\, C_{b,t_k[GOPS]}^{Al_1} + (1-\alpha)\, C_{b,t_k[GOPS]}^{Al_2}}{C_{b,t_k[GOPS]}^{R}} \\ &= \mathcal{U}_{b,t_k}\left(\alpha\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_1}\right) + \mathcal{U}_{b,t_k}\left((1-\alpha)\, \boldsymbol{C}_{t_k[N_B\times 1]}^{Al_2}\right) \end{aligned} \tag{A.6}$$

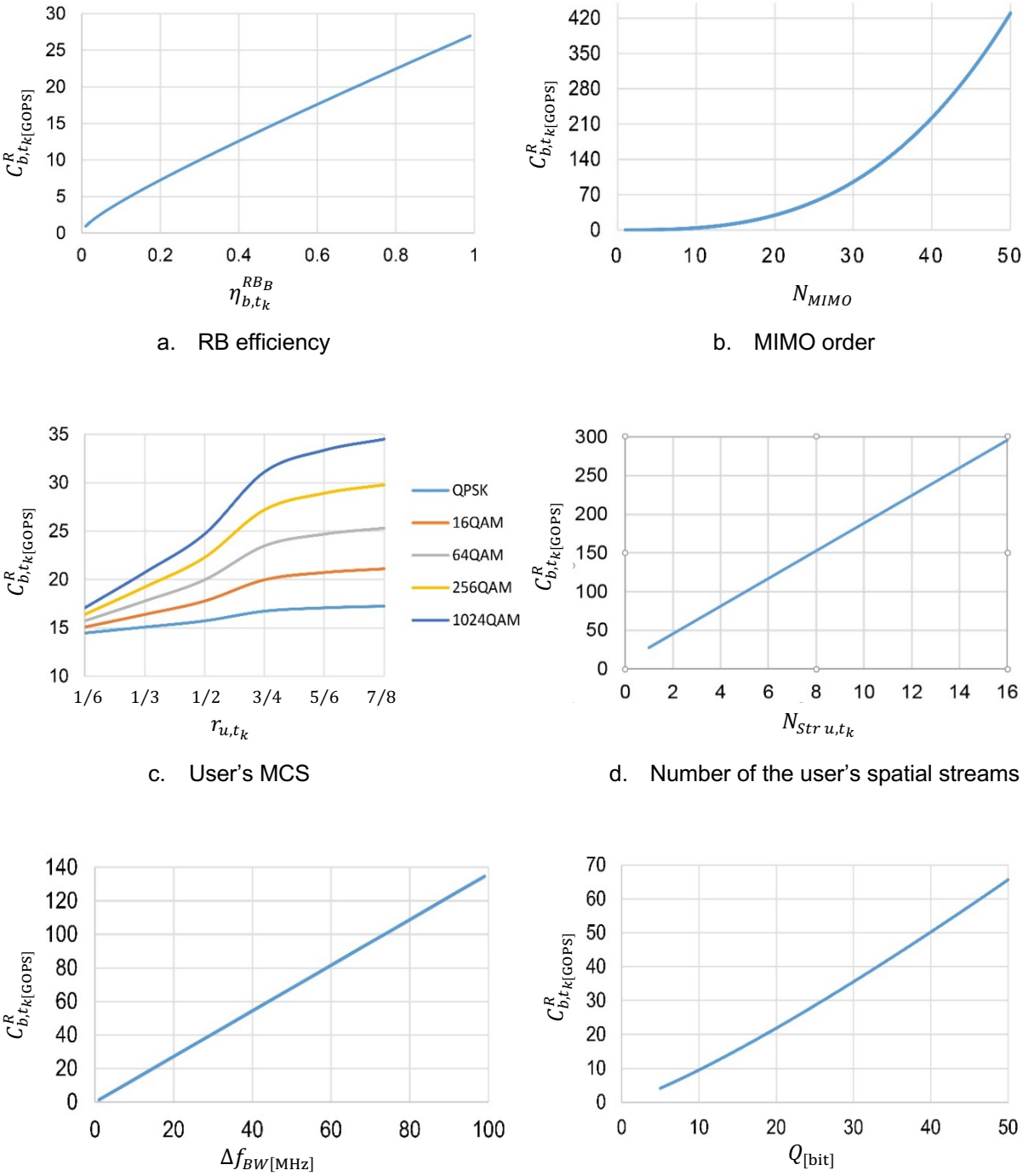
the function $\mathcal{U}_{b,t_k}$ is convex.

Annex B.

RCC Variations

A BBU's RCC relative to the variation of effective parameters are presented in this annex.

As mentioned in Section 3.3.2, a BBU's RCC is achieved using (3.17). The effect of the input parameters on a BBU's RCC variation is depicted in Figure B.1.



a. RB efficiency

b. MIMO order

c. User's MCS

d. Number of the user's spatial streams

e. Channel bandwidth

f. Quantization resolution

Figure B.1 – A BBU's RCC variation relative to the variation of the effective parameters.

Annex C.

UE Speeds and Corresponding Doppler Frequency Shifts in FDD Operating Bands

In this annex, UE Speeds and corresponding maximum Doppler shifts in some FDD operating Bands are listed.

As mentioned in Section 4.2.1, the speed range that LTE supports, is split into five intervals that for each one an average speed value is considered:

- Very-low-speed, (e.g., pedestrian, 5 km/h),
- Low-speed, (e.g., cyclist, vehicular urban, 5 km/h),
- Mid-speed, (e.g., vehicular sub-urban, 90 km/h),
- High-speed, (e.g., vehicular rural, 120 km/h),
- Very-high-speed, (e.g., high-speed train, 500 km/h).

Accordingly, the values of coherence time and maximum Doppler shift are calculated in accordance with (4.1) and (4.2) for each speed classes and supported operating bands. The UE speeds and related maximum Doppler shifts, with respect to the supported carrier frequencies, are listed in Table C.1.

Table C.1 – Maximum Doppler shift corresponding to speed classes and operating bands.

Speed Operating Band			Speed Class [km/h] Maximum Doppler Shift [Hz]				
			5	50	90	120	500
Band 1	UL	1920	8.9	88.95	160.11	213.48	889.5
		1980	9.17	91.73	165.11	220.15	917.3
	DL	2110	9.78	97.75	175.96	234.61	977.53
		2170	10.05	100.53	180.96	241.28	1005.33
Band 2	UL	1850	8.57	85.71	154.27	205.7	857.07
		1910	8.85	88.49	159.28	212.37	884.87
	DL	1930	8.94	89.41	160.94	214.59	894.14
		1990	9.22	92.19	165.95	221.26	921.93
Band 3	UL	1710	7.92	79.22	142.6	190.13	792.21
		1785	8.27	82.7	148.85	198.47	826.96
	DL	1805	8.36	83.62	150.52	200.69	836.23
		1880	8.71	87.1	156.78	209.03	870.97
Band 4	UL	1710	7.92	79.22	142.6	190.13	792.21
		1755	8.13	81.31	146.35	195.13	813.06
	DL	2110	9.78	97.75	175.96	234.61	977.53
		2155	9.98	99.84	179.71	239.61	998.38
Band 5	UL	824	3.82	38.17	68.71	91.62	381.75
		849	3.93	39.33	70.8	94.4	393.33
	DL	869	4.03	40.26	72.47	96.62	402.59
		894	4.14	41.42	74.55	99.4	414.18

Table C.1(contd.) – Maximum Doppler shift corresponding to speed classes and operating bands.

Band 6	UL	830	3.85	38.45	69.21	92.29	384.53
		840	3.89	38.92	70.05	93.4	389.16
	DL	875	4.05	40.54	72.97	97.29	405.37
		885	4.1	41	73.8	98.4	410.01
Band 7	UL	2500	11.58	115.82	208.48	277.97	1158.21
		2570	11.91	119.06	214.31	285.75	1190.64
	DL	2620	12.14	121.38	218.48	291.31	1213.8
		2690	12.46	124.62	224.32	299.1	1246.23
Band 8	UL	880	4.08	40.77	73.38	97.85	407.69
		915	4.24	42.39	76.3	101.74	423.9
	DL	925	4.29	42.85	77.14	102.85	428.54
		960	4.45	44.48	80.06	106.74	444.75
Band 9	UL	1749.9	8.11	81.07	145.93	194.57	810.7
		1784.9	8.27	82.69	148.84	198.46	826.91
	DL	1844.9	8.55	85.47	153.85	205.13	854.71
		1879.9	8.71	87.09	156.77	209.02	870.93
Band 10	UL	1710	7.92	79.22	142.6	190.13	792.21
		1770	8.2	82	147.6	196.8	820.01
	DL	2110	9.78	97.75	175.96	234.61	977.53
		2170	10.05	100.53	180.96	241.28	1005.33
Band 11	UL	1427.9	6.62	66.15	119.07	158.77	661.52
		1452.9	6.73	67.31	121.16	161.55	673.1
	DL	1475.9	6.84	68.38	123.08	164.1	683.76
		1500.9	6.95	69.53	125.16	166.88	695.34
Band 20	UL	832	3.85	38.55	69.38	92.51	385.45
		862	3.99	39.94	71.88	95.84	399.35
	DL	791	3.66	36.65	65.96	87.95	366.46
		821	3.8	38.04	68.46	91.29	380.36
Band 21	UL	1447.9	6.71	67.08	120.74	160.99	670.79
		1462.9	6.78	67.77	121.99	162.66	677.74
	DL	1495.9	6.93	69.3	124.74	166.33	693.03
		1510.9	7	70	126	167.99	699.97

Table C.1(contd.) – Maximum Doppler shift corresponding to speed classes and operating bands.

Band 22	**UL**	**3410**	15.8	157.98	284.36	379.15	1579.8
		3500	16.21	162.15	291.87	389.16	1621.49
	DL	**3510**	16.26	162.61	292.7	390.27	1626.12
		3600	16.68	166.78	300.21	400.28	1667.82
Band 23	**UL**	**2000**	9.27	92.66	166.78	222.38	926.57
		2020	9.36	93.58	168.45	224.6	935.83
	DL	**2180**	10.1	101	181.79	242.39	1009.96
		2200	10.19	101.92	183.46	244.61	1019.22
Band 24	**UL**	**1626.5**	7.54	75.35	135.64	180.85	753.53
		1660.5	7.69	76.93	138.47	184.63	769.28
	DL	**1525**	7.07	70.65	127.17	169.56	706.51
		1559	7.22	72.23	130.01	173.34	722.26
Max			16.68	166.78	300.21	400.28	1667.82
Min			3.66	36.65	65.96	87.95	366.46

Annex D.

Traffic Models

The services' traffic profiles are described in this annex.

- **File transfer:** The parameters listed in Table D.1 are for DL [NGMN08]. For UL, the same traffic model shall be used.

Table D.1 – File transfer Traffic Parameter (extracted from [NGMN08]).

Parameter	Statistical Characterization
File Size	Truncated Lognormal Distribution, $Mean = 1.996\ \mathrm{MB}$, $Standard\ deviation = 0.7\ \mathrm{MB}, Minimum = 100\mathrm{B}$; $Maximum = 5\mathrm{MB}$, PDF: $f_x = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}, x > 0, \sigma = 0.35, \mu = 14.45.$

- **Email:** The parameters listed in Table D.2 are for DL. For UL, the same traffic model shall be used.

Table D.2 – Email Traffic Parameter.

Parameter	Statistical Characterization
File Size	Truncated Lognormal Distribution, $Mean = 1.256\ \mathrm{MB}$, $Standard\ deviation$=0.38 MB, $Minimum = 10\mathrm{B}$, $Maximum = 3\mathrm{MB}$, PDF: $f_x = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}, x > 0, \sigma = 0.3, \mu = 14.$

- **Web-browsing using Hypertext Transfer Protocol (HTTP):** A webpage consists of a main object and embedded objects (e.g., pictures, advertisements etc.). After receiving the main page, the web-browser will parse for the embedded objects. The main parameters to characterize web-browsing are: main object size, embedded object size, number of embedded objects, reading time, and parsing time for the main page, being listed in Table D.3.
- **Video Streaming/Calling:** Each frame of video data arrives at a regular interval determined by the number frames per second. Each frame is decomposed into a fixed number of slices, each transmitted as a single packet. The size of these packets/slices is modeled to have a Truncated Pareto Distribution. The video encoder introduces encoding delay intervals between the packets of a frame. These intervals are modeled by a Truncated Pareto Distribution. Distributions listed in Table D.4, assume a source video rate of 1.5 Mbps.
- **VoIP Satisfied User Criterion and Traffic Model:** Table D.5 provides the relevant parameters of the VoIP traffic that shall be assumed in the simulations. The main purpose of this traffic model is not to favor any codec but to specify a model to obtain results which are comparable. The details of the corresponding traffic model are described in what follows.

Table D.3 – Web Browsing Traffic Parameters (based on [NGMN08]).

Parameter	Statistical Characterization
Main Object Size	Truncated Lognormal Distribution, $Mean = 11\,055$ B, $Standard\ deviation = 25\,395$ B, $Minimum = 100$B, $Maximum = 2$MB, PDF: $f_x = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}, x > 0, \quad \sigma = 1.37, \quad \mu = 8.37$
Embedded Object Size	Truncated Lognormal Distribution, $Mean = 8\,237$ B, $Standard\ deviation = 47\,307$ B, $Minimum = 50$B, $Maximum = 2$MB, PDF: $f_x = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}, x > 0, \quad \sigma = 2.36, \quad \mu = 6.17$
Number of Embedded Objects per Page	Truncated Pareto Distribution, $Mean = 7.59$, $Standard\ deviation = 10.36$, $Maximum = 53 = m$, PDF: $f_x = \frac{\alpha k^\alpha}{x^{\alpha+1}}, k \leq x < m,$ $f_x = \left(\frac{k}{m}\right)^\alpha, \quad x = m, \quad \alpha = 1.1, \quad k = 2, \quad m = 53$
Reading Time	Exponential Distribution $Mean = 30$ s PDF: $f_x = \lambda e^{-\lambda x}, x \geq 0, \ \lambda = 0.033$
Parsing Time	Exponential Distribution $Mean = 0.13$ s PDF: $f_x = \lambda e^{-\lambda x}, \quad x \geq 0, \quad \lambda = 7.69$

The model is assumed updated at the speech encoder frame rate with the duration of 20 ms. In the model, the probability of being in inactive state I is P_I, that is:

$$P_I = P_{AI}/(P_{AI} + P_{IA}) \tag{D.1}$$

where:

- P_{AI}: the probability of transitioning from active speech state A to the inactive or silent state I while in state A,
- P_{IA}: the probability of transitioning from state I to state A while in state I.

Table D.4 – Video Streaming Traffic Parameters (modified [NGMN08]).

Parameter	Statistical Characterization
Inter-Arrival time between the beginning of each frame	Deterministic, 100ms (based on 10 frames per second)
Number of packets (slices) in a frame	Deterministic, 14 packets per frame
Packet (slice) size	Truncated Pareto Distribution $Mean = 1\,272\text{ B},\ Standard\ deviation = 257\text{ B},$ $Maximum\ = m = 1\,500\text{B}$ PDF: $f_x = \frac{\alpha k^\alpha}{x^{\alpha+1}}, k \le x < m,$ $f_x = \left(\frac{k}{m}\right)^\alpha, x = m, \quad \alpha = 1.2, \quad k = 800\text{B}$
Inter-arrival time between packets (slices) in a frame	Truncated Pareto Distribution $Mean =\ 6.01\ ms, Standard\ deviation = 3.59\ ms,$ $Maximum\ = m = 13\ ms$ PDF: $f_x = \frac{\alpha k^\alpha}{x^{\alpha+1}}, k \le x < m,$ $f_x = \left(\frac{k}{m}\right)^\alpha, x = m, \quad \alpha = 1.2, \quad k = 2.5\text{ms}$

Table D.5 – Voice traffic parameters (extracted from [NGMN08]).

Parameter	Statistical Characterization
Codec	RTP AMR 12.2, Source rate 12.2 kbps
Encoder Frame Length	20 ms
Voice Activity Factor	50% ($P_{IA} = 0.004, P_{AA} = 0.996$)
SID Payload	Modeled 15 B (5 B + header) SID packet every 160 ms during silence
Protocol Overhead with Compressed Header	10 bits + padding (RTP-pre-header) 4 B (RTP/UDP/IP), 2 B (RLC/security), 16 bits (CRC)
Total Voice Payload on Air Interface	40 B (AMR 12.2)

And the probability of being in state A is:

$$P_A = P_{IA}/P_{AI} + P_{IA} \tag{D.2}$$

The voice activity factor , F_A, is given by:

$$F_A = P_A = P_{IA}/(P_{AI} + P_{IA}) \tag{D.3}$$

The probability that a talk period, τ_{TS}, has duration n speech frames is given by:

$$P_{\tau_{TS}=n} = P_{TS}(n) = P_{AI}(1 - P_{AI})^{n-1} \quad : \qquad n = 1,2, \ldots \tag{D.4}$$

Correspondingly, the probability that a silence period has duration n speech frames is given by:

$$P_{\tau_{SP}=n} = P_{SP}(n) = P_{IA}\big((1-P_{IA})\big)^{n-1}: \qquad n = 1,2,\ldots \tag{D.5}$$

The mean talks spurt duration μ_{TS} (in speech frames) is given by:

$$\mu_{TS} = E(\tau_{TS}) = \frac{1}{P_{AI}} \tag{D.6}$$

while the mean silence period duration μ_{SP} (in speech frames) is given by:

$$\mu_{SP} = E(\tau_{SP}) = \frac{1}{P_{IA}} \tag{D.7}$$

The distribution of the time period τ_{AE} (in speech frames) between successive active state entries is the convolution of the distributions of τ_{SP} and τ_{TS} . This is given by:

$$P_{\tau_{AE}=n} = P_{AE}(n) = \frac{P_{IA}}{P_{IA}-P_{AI}}\, P_{AI}(1-P_{AI})^{n-1} + \frac{P_{AI}}{P_{AI}-P_{IA}}\, P_{IA}(1-P_{IA})^{n-1} \quad : \quad n = 1,2,\ldots \tag{D.8}$$

Since the state transitions from state A to state I and vice versa are independent, the mean time μ_{AE} between active state entries is given simply by the sum of the mean time in each state. That is:

$$\mu_{AE} = \mu_{TS} + \mu_{SP} \tag{D.9}$$

Annex E.

Simulator's Assessment Results

Results obtained for simulator assessment are presented in this appendix. Initially, the results related to the simulator's transitory interval are presented. After that, the results related to the sensitivity to the number of simulations, are shown.

E.1 Simulator's Transitory Interval

The values collected from simulation over time for BBU RCC and AICC, user satisfaction level, efficiency and Cost saving are graphically represented in Figure E.1 to Figure E.3, respectively.

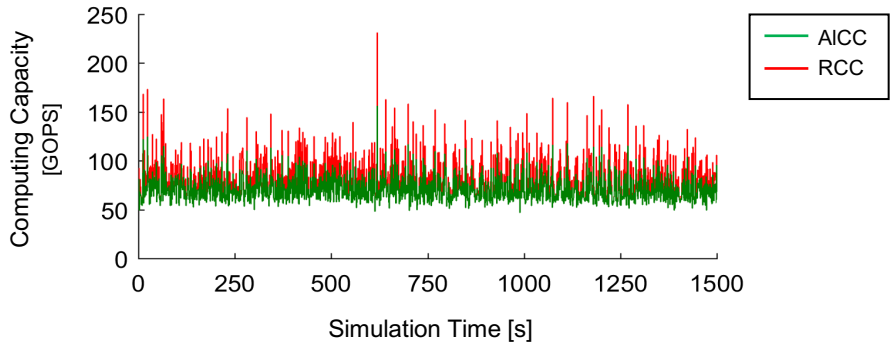


Figure E.1 – BBU1's RCC and AICC per second.

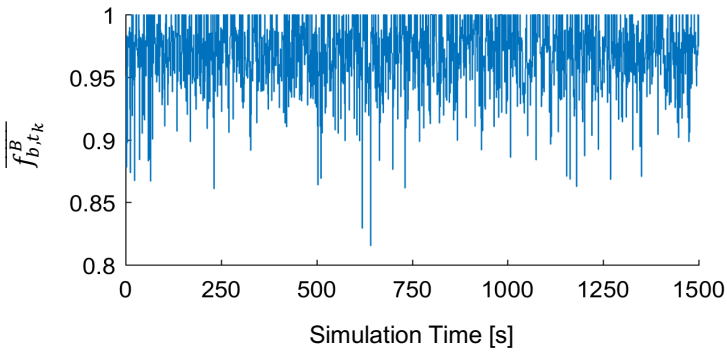


Figure E.2 – BBU1's average fulfilment level per second.

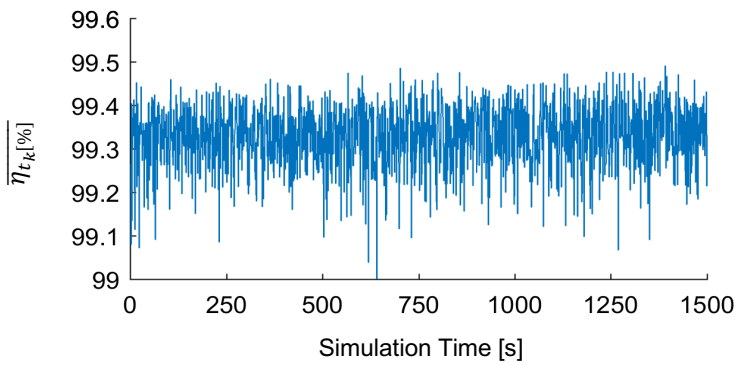


Figure E.3 – Average efficiency per second.

Moreover, Table E.1 lists the relative deviation percentage given by (4.11) for the RB efficiency in the simulation. The relative deviation for each of n millisecond simulations is achieved by comparing X^{aprox} with the average of all values collected for the total set of simulations as X^{ref}.

Table E.1 – Values and deviation for a BBU's RB efficiency per second in various simulation durations.

Sim. Duration [s]	DL			UL		
	Average per Simulation [%]	$\Delta_{[\%]}$	Standard Deviation per Simulation [%]	Average per Simulation [%]	$\Delta_{[\%]}$	Standard Deviation per Simulation [%]
60	17.86	0.02	6.9	10.85	0.09	7.13
120	18.52	0.06	6.58	10.65	0.07	6.44
180	17.96	0.03	6.04	10.08	0.01	5.96
240	17.83	0.02	5.83	10	<0.01	5.91
300	17.71	0.01	5.7	9.76	0.02	5.71
360	17.54	<0.01	5.64	9.94	<0.01	5.7
420	17.45		5.57	9.98		5.67
480	17.47		5.46	10		5.63
540	17.41		5.47	10.22	0.03	5.65
600	17.46		5.55	10.1	0.01	5.58
660	17.53		5.59	10.19	0.02	5.78
720	17.56	0.01	5.59	10.28	0.03	5.85
780	17.61	0.01	5.61	10.23	0.03	5.84
840	17.63	0.01	5.64	10.2	0.02	5.79
900	17.53	<0.01	5.65	10.12	0.02	5.75
960	17.56	0.01	5.67	10.08	0.01	5.74
1020	17.6	0.01	5.75	10.04	0.01	5.77
1080	17.66	0.01	5.85	10.07	0.01	5.76
1140	17.69	0.01	5.84	10.02	0.01	5.75
1200	17.64	0.01	5.78	10	<0.01	5.76
1260	17.56	0.01	5.72	9.99		5.78
1320	17.5	<0.01	5.69	10.01	0.01	5.79
1380	17.5		5.68	10.02	0.01	5.76
1440	17.47		5.67	9.98	<0.01	5.73
1500	17.45	0	5.66	9.96	0	5.68

E.2 Sensitivity Analysis as a Function of the Number of Simulations

The results for different sets of simulations for the average RCC, average AICC, and efficiency are depicted graphically in Figure E.4 to Figure E.6, as a function of the number of simulations performed.

To quantify the variation achieved in the observations, the deviation percentage relative to the average of all simulation values, computed from (4.11) for each set of simulations, are presented in Table E.2 to Table E.3.

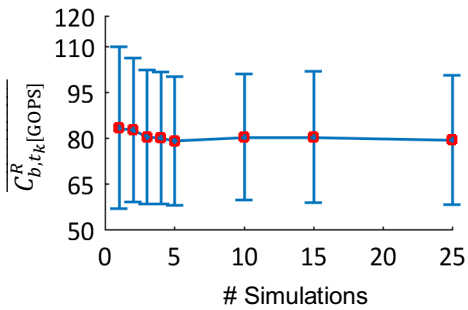


Figure E.4 – Average RCC for n number of simulations.

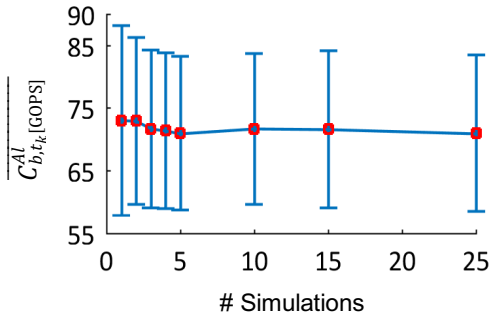


Figure E.5 – Average AICC for n number of simulations.

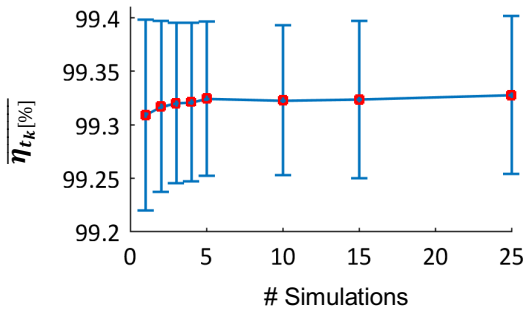


Figure E.6 – Average efficiency for n number of simulations.

Table E.2 – Values and deviation for the BBU's RB efficiency in several numbers of simulations.

# Sim.	Average of $\eta_{b,t_k}^{RB_B}$ per second					
	DL			UL		
	Average [%]	$\Delta_{[\%]}$	Standard Deviation [%]	Average [%]	$\Delta_{[\%]}$	Standard Deviation [%]
1	17.86	0.02	6.9	10.85	0.09	7.13
2	18.52	0.06	6.58	10.65	0.07	6.44
3	17.96	0.03	6.04	10.08	0.01	5.96
4	17.83	0.02	5.83	10	<0.01	5.91
5	17.71	0.01	5.7	9.76	0.02	5.71
10	17.46	<0.01	5.55	10.1	0.01	5.58
15	17.53	<0.01	5.65	10.12	0.02	5.75
25	17.45	0	5.66	9.96	0	5.68

Table E.3 – Values and deviation for the BBU's RCC, AICCs and the resource allocation's efficiency in several number of simulations.

# Sim.	Average of $\overline{C^{R}_{b,t_k}}$ per second			Average of $\overline{C^{Al}_{b,t_k}}$ per second			Average of $\overline{\eta_{t_k}}$ per second		
	Average [GOPS]	$\Delta_{[\%]}$	Std. Dev. [GOPS]	Average [GOPS]	$\Delta_{[\%]}$	Std. Dev. [GOPS]	Average [%]	$\Delta_{[\%]}$	Std. Dev. [%]
1	83.29	0.05	26.53	99.31	<0.01	0.09	>0.99	<0.01	
2	82.62	0.04	23.57	99.32		0.08			
3	80.36	0.01	21.89	99.32		0.08			
4	80.01	0.01	21.68	99.32		0.07			
5	79.09	<0.01	21.07	99.32		0.07			
10	80.29	0.01	20.58	99.32		0.07			
15	80.29	0.01	21.47	99.32		0.07			
25	79.42	0	21.21	99.33	0	0.07		0	<0.01

Annex F.

Traffic Generation

Results obtained for evaluation of the simulator's generated samples are presented in this annex.

F.1 Generated Samples' Histogram

In this subsection the samples generated by the simulator are compared with their associated PDFs. The comparison results are presented separately for each single PDF that is used by the simulator.

- **User arrival rate:** Given PDF for the users' arrival rate is a mixture of two normal distributions for both residential and business areas, each with distinct parameters that are defined in, Section 4.4.2. Frequency of the generated samples are presented in Figure F.1 for residential area. The figures histogram is for 50 000 users in the BS per 24 hours in total. The first distribution sample mean is 10:01AM with the standard deviation of 161min. The second distribution sample mean is 6:02PM with the standard deviation of 141min.

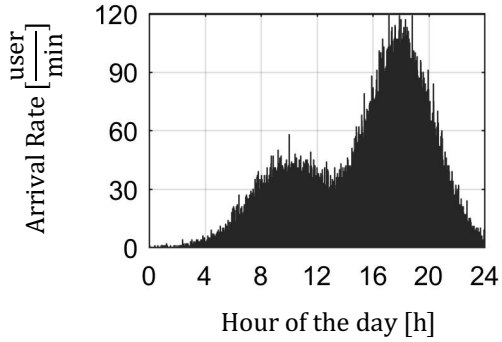


Figure F.1 – User arrival rate in residential areas.

With the same considerations, Figure F.2 presents the user arrival rate for business area. The first distribution sample mean is 11:01AM with the standard deviation of 94min. The second distribution sample mean is 3:01PM with the standard deviation of 95min.

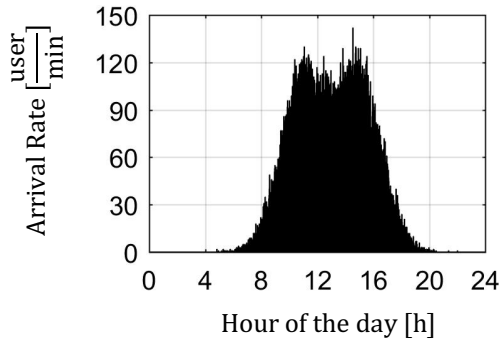


Figure F.2 – User arrival rate in business areas.

- **File size in file transfer service:** Figure F.3 shows the given file size PDF for file transfer service. The samples generated in the simulation are presented in Figure F.4.

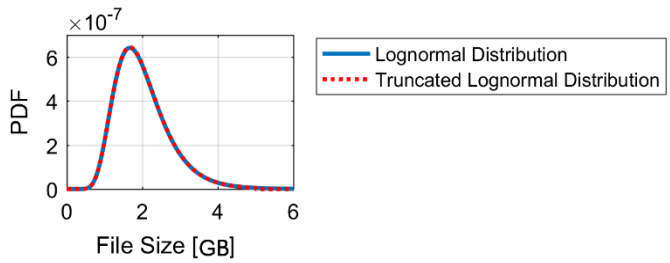
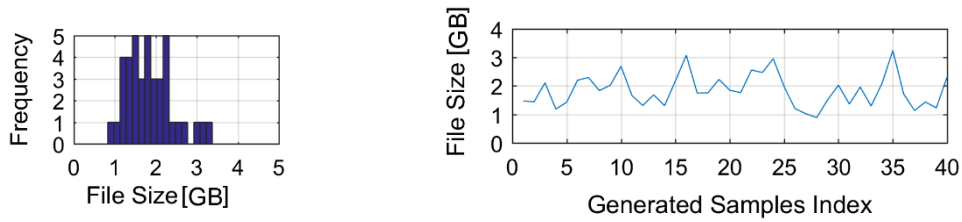


Figure F.3 – PDF of the file size in file transfer service. Truncated $logNormal(14.45, 0.12), mean = 1.996\text{MB}, standard\ deviation = 0.7\text{MB}$.



(a) Frequency (b) Size

Figure F.4 – Generated sample for file size in file transfer service. $Sample\ size = 40$. $sample\ mean = 1.989\text{MB}$, $sample\ standard\ deviation = 0.69\text{MB}$.

- **Email file size:** Figure F.5 shows the email file size PDF. Based on the given PDF, the samples that are generated in a simulation are presented in Figure F.6.

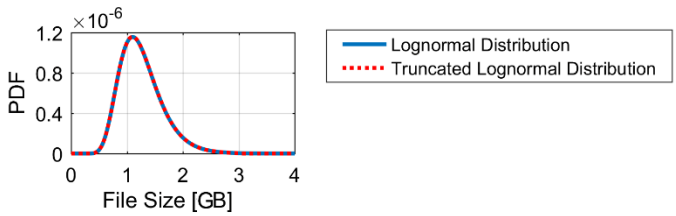
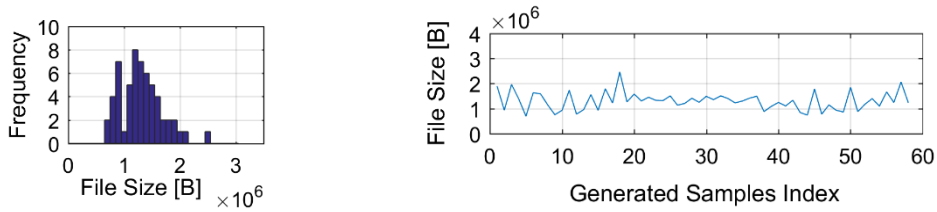


Figure F.5 – PDF of the email file size service. Truncated $logNormal(14, 0.09), mean = 1.256\text{MB}, standard\ deviation$=0.38MB.



(a) Frequency (b) Size

Figure F.6 – Generated sample for file size in email service. $Sample\ size = 59$, $sample\ mean = 1.269\text{MB}$, $sample\ standard\ deviatoin = 0.383\text{MB}$.

- **Web browsing service duration:** Figure F.7 shows the PDF of web browsing service duration.

The sample that are generated accordingly, are presented in Figure F.8.

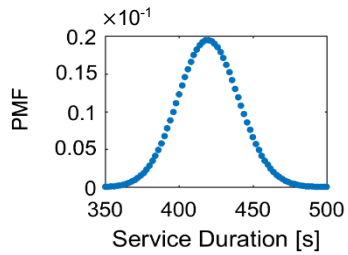
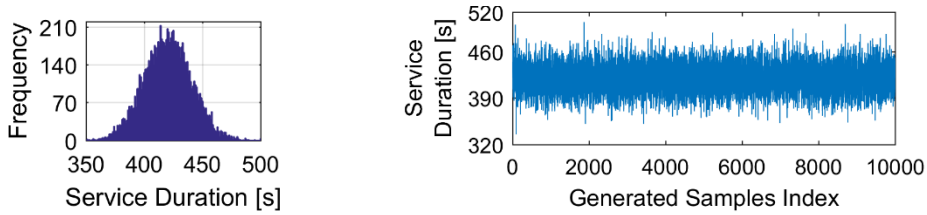


Figure F.7 – PMF of the web browsing duration. $Poisson(420),\ mean = 420\text{s},\ standard\ deviation = 20.49\text{s}$.



(a) Frequency (b) Duration

Figure F.8 – Generated sample for service duration in web browsing service. $Sample\ size = 10\ 000, sample\ mean = 420\text{s}, sample\ standard\ deviation = 20.58\text{s}$.

- **Web browsing main object size:** Figure F.9 shows the web browsing main object size PDF. Based on the given PDF, samples generated in a simulation are presented in Figure F.10.

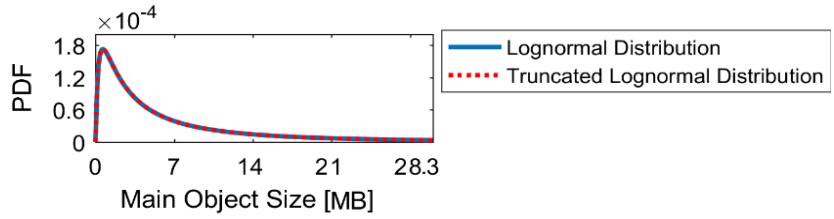
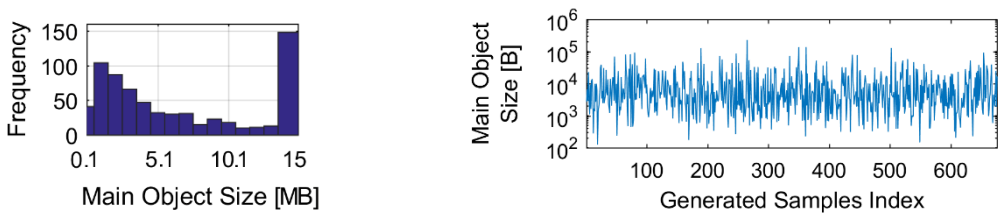


Figure F.9 – PDF of the main object size in web browsing. Truncated $lognormal(8.37, 1.88),\ mean = 11\ 055\text{B},\ standard\ deviation = 25\ 395\text{B}$.



(a) Frequency (b) Size

Figure F.10 – Generated samples for main object size in web browsing service. $Sample\ size = 637, sample\ mean = 12\ 091\text{B}, sample\ standard\ deviation = 28\ 149\text{B}$.

- **Web number of embedded objects per page:** Figure F.11 shows the PDF of number of embedded objects per page. Samples that are generated accordingly are presented in Figure F.12.

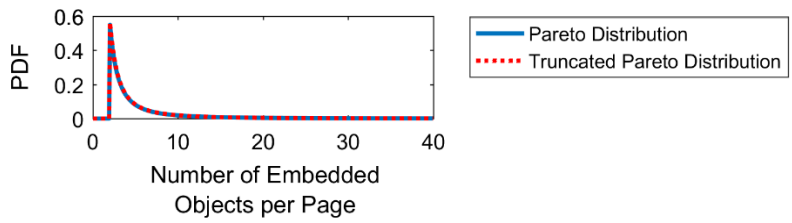
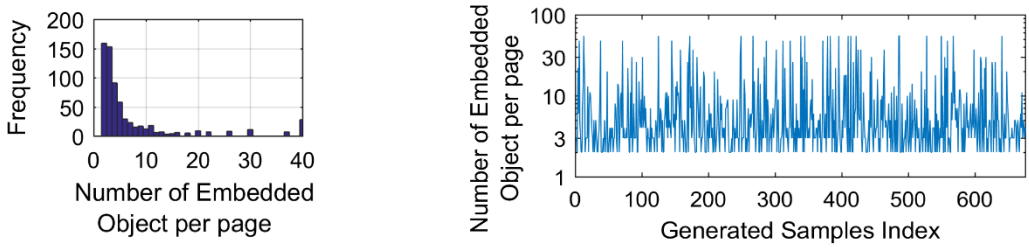


Figure F.11 – PDF of the number of embedded objects per page in web browsing service. Truncated $Pareto(1.1, 2)$, $mean = 7.59$, $standard\ deviation = 10.36$.



(a) Frequency (b) # embedded object per page

Figure F.12 – Generated sample for number of embedded objects per page in web browsing. $Sample\ size = 637. sample\ mean = 7.47, sample\ standard\ deviation = 10.38.$

- **Reading time in web browsing:** Figure F.13 shows the web reading time PDF. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.14.

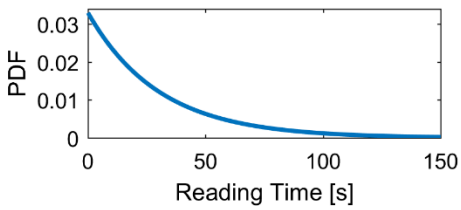
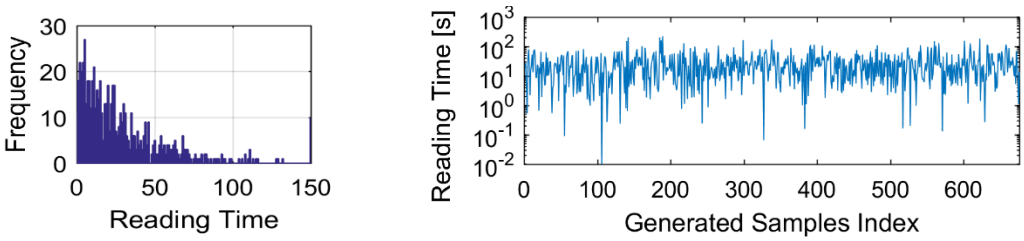


Figure F.13 – PDF of the reading time in web browsing. $Exponential\ (0.033), mean = standard\ deviation = 30s.$



(a) Frequency (b) Reading time

Figure F.14 – Generated sample for reading time in web browsing. $Samples\ size = 637.$ $sample\ mean = 30.72s$, $sample\ standard\ deviation = 31.63s.$

- **Parsing time in web browsing:** Figure F.15 shows the web parsing time PDF. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.16.

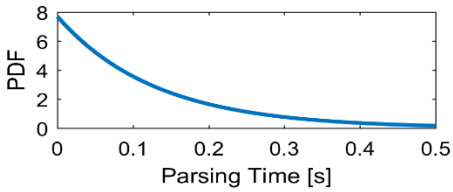
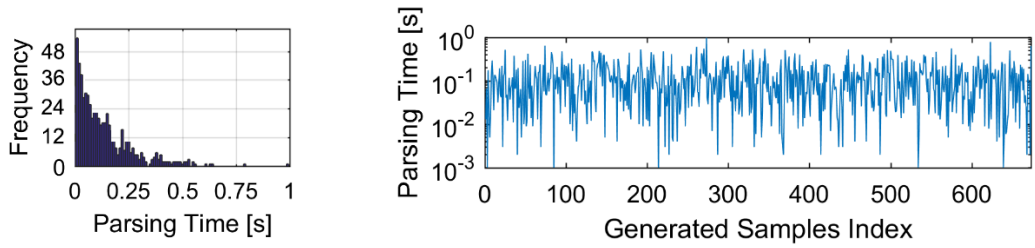


Figure F.15 – PDF of the reading time in web browsing. $Exponenial\ (7.69), mean = standard\ deviation = 0.13s.$



(a) Frequency (b) Parsing time

Figure F.16 – Generated samples for parsing time in web browsing. $Sample\ size = 637,$ $sample\ mean = 0.13s,\ sample\ standard\ deviation = 0.12s.$

- **Embedded object size in web browsing:** Figure F.17 shows the embedded object size PDF. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.18.

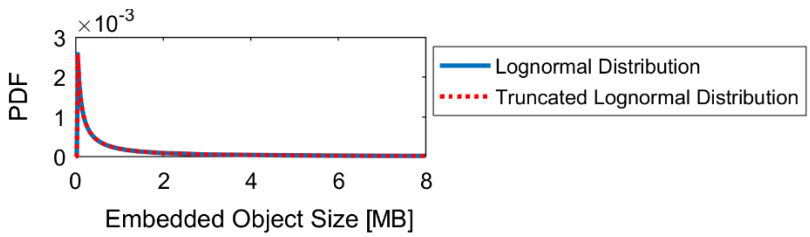
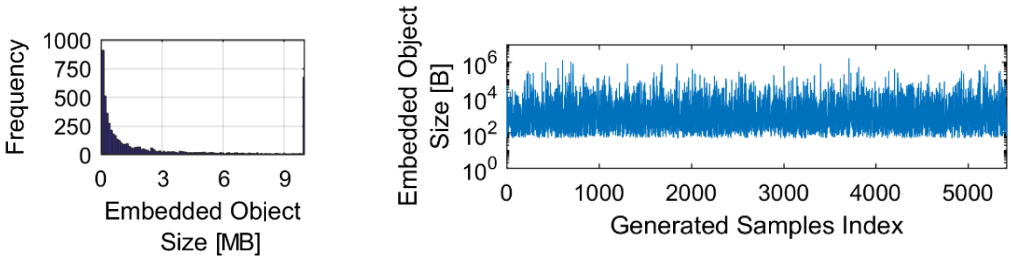


Figure F.17 – PDF of the embedded object size in web browsing service.

Truncated $logNormal(6.17, 5.57),\ mean = 8\ 237\text{B},\ standard\ deviation = 47\ 307\text{B}.$



(c) Frequency (d) Size

Figure F.18 – Generated sample for the embedded object size in web browsing. $Sample\ size = 637.$ $Sample\ mean = 6\ 373\text{B}, sample\ standard\ deviation = 26\ 488\text{B}.$

- **Video service duration:** Figure F.19 shows the PDF of video service duration. The samples that are generated accordingly, are presented in Figure F.20.

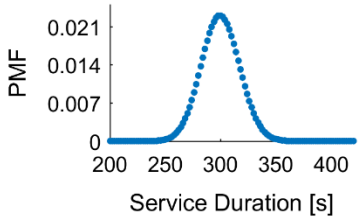
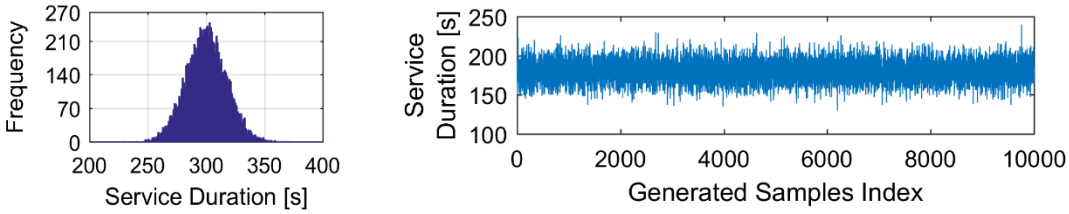


Figure F.19 – PMF of the video duration. $Poisson(300),\ mean = 300\text{s}, standard\ deviation = 17.32\text{s}$.



(e) Frequency (f) Duration

Figure F.20 – Generated sample for service duration in video service. $Sample\ size = 10\ 000$, $sample\ mean = 300.08\text{s}, sample\ standard\ deviation = 17.34\text{s}$.

- **Video packet size:** Figure F.21 shows the video packet size PDF. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.22.

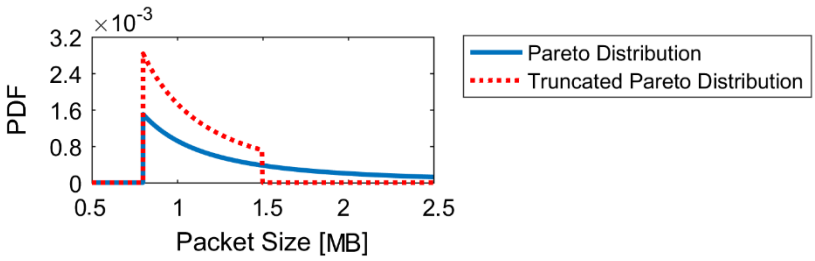
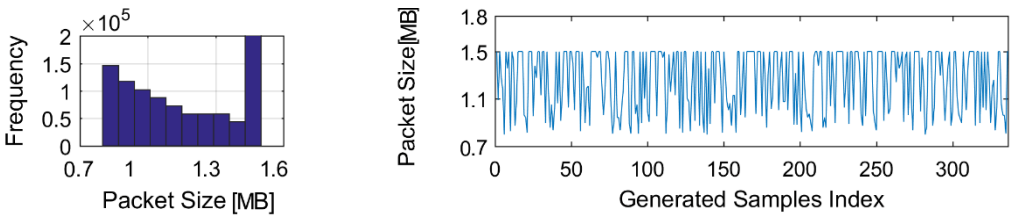


Figure F.21 – PDF of the packet size in video service. Truncated $Pareto(1.2, 800), minimum = 800\text{B}, maximum = 1\ 500\text{B}, mean = 1\ 272\text{B}, standard\ devaition = 257\text{B}$.



(a) Frequency (b) Size

Figure F.22 – Generated sample for packet size video service. $Sample\ size = 1\ 485\ 727$, $sample\ mean = 1\ 273\text{B}, sample\ standard\ devaition = 257\text{B}$.

- **Video packets inter-arrival time:** Figure F.23 shows the inter-arrival time PDF between video packets. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.24.

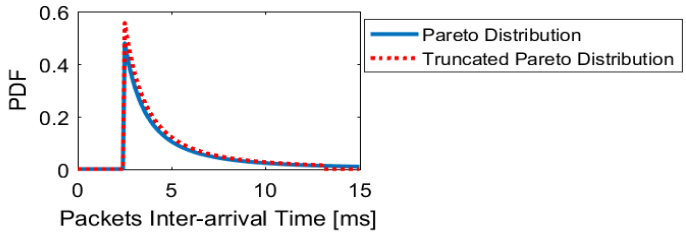
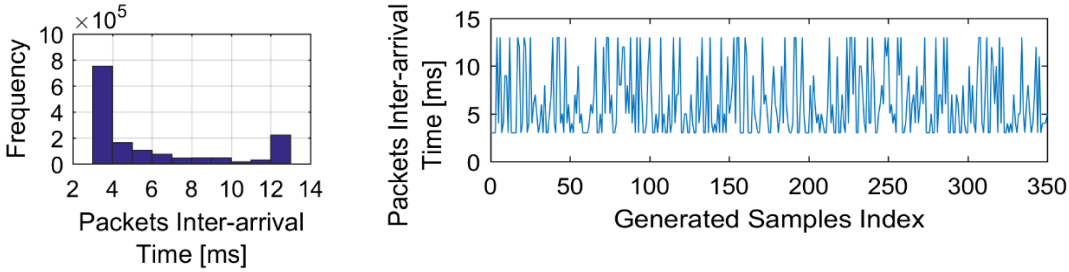


Figure F.23 – PDF of the packet inter-arrival time in video service. Truncated $Pareto(1.2, 2.5)$, $minimum = 2.5\ ms, maximum = 13\ ms, mean = 6.01\text{ms}, standard\ deviation = 3.59\text{ms}$.



(a) Frequency (b) Inter-arrival time

Figure F.24 – Generated sample for packet inter-arrival time in video service. $Sample\ size = 1\ 485\ 727. sample\ mean = 6.01\text{ms}, sample\ standard\ devaition = 3.59\text{ms}.$

- **VoIP service duration:** Figure F.25 shows the PDF of VoIP service duration. The samples that are generated accordingly, are presented in Figure F.26.

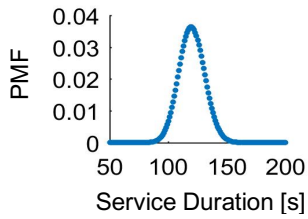
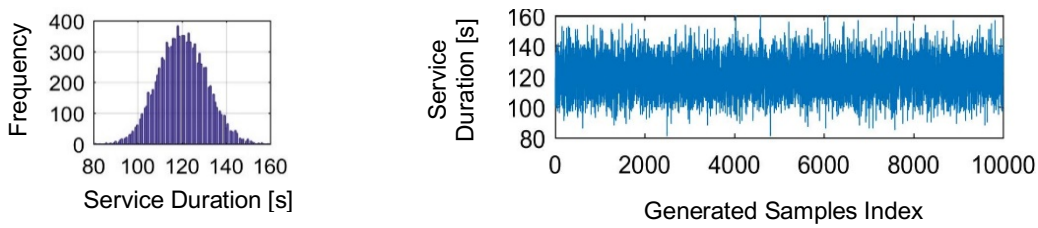


Figure F.25 – PMF of the VoIP duration. $Poisson(120),\ mean = 120\text{s},\ standard\ deviation = 10.95s$



(a) Frequency (b) Duration

Figure F.26 – Generated sample for service duration in VoIP service. $Samples\ size = 10\ 000$, $sample\ mean = 120.11s,\ sample\ standard\ deviation = 10.9\text{s}.$

- **VoIP inactive state duration:** Figure F.27 shows the VoIP inactive state duration PDF. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.28.

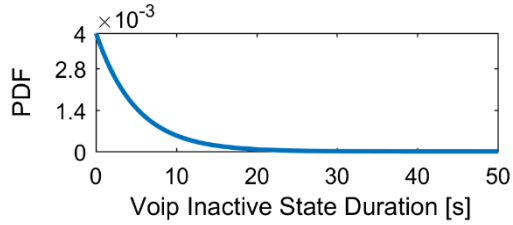
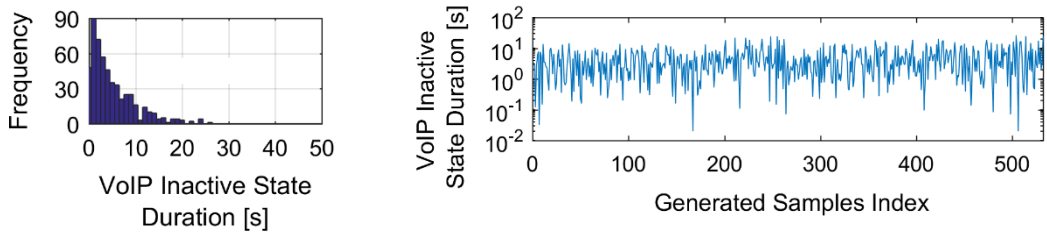


Figure F.27 – PDF of the inactive state duration in VoIP service. $Mean = standard\ deviation = 5s$.



(a) Frequency (b) Duration

Figure F.28 – Generated sample for inactive state duration for VoIP service. $Sample\ size = 532, sample\ mean = 4.96s, sample\ standard\ deviation = 4.65s$.

- **VoIP active state duration:** Figure F.29 shows the VoIP active state duration PDF. Based on the given PDF, samples that are generated in a simulation are presented in Figure F.30.

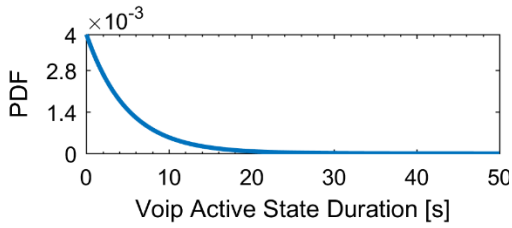
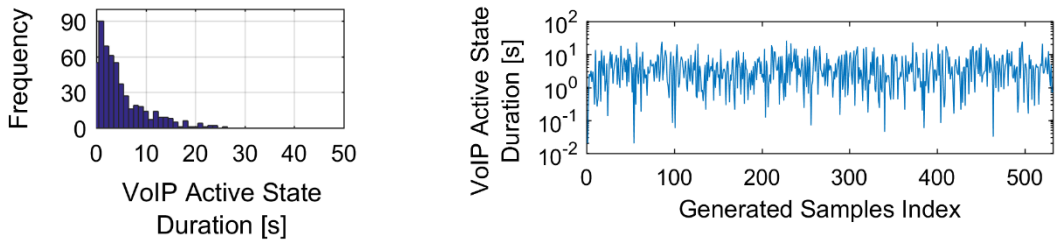


Figure F.29 – PDF of the active state duration in VoIP service. $Mean = standard\ deviation = 5s$.



(a) Frequency (b) Duration

Figure F.30 – Generated sample for active state duration for VoIP service. $Sample\ size = 532, sample\ mean = 5.13s, sample\ standard\ deviation = 4.94s$.

F.2 Relative Deviation of the Means and Standard Deviations

In this subsection, the generated samples of 35 different simulations are analyzed based on the relative deviation percentage (4.11). The randomly generated samples' standard deviation and mean, as the approximated values, X^{aprox}, are compared with the theoretic standard deviation and mean of the given PDF, as the reference values, X^{ref}. The results are presented in Table F.1.

Table F.1 – Average relative deviation of the mean and standard deviation of 35 different simulations.

Service	Parameter	Average Relative Deviation of the Mean Value [%]	Average Relative Deviation of the Standard Deviation [%]
VoIP	User Active Duration	0.03	0.04
	User Inactive Duration	0.02	0.03
	Service Duration	<0.01	0.01
Video	Packet Size	0.01	0.01
	Packets Inter-Arrival Time	0.01	0.01
	Service Duration	<0.01	0.01
Web Browsing	Main Object Size	0.02	0.09
	Embedded Object Size	0.05	0.11
	#Embedded Objects per Page	0.03	0.04
	Reading Time	0.01	0.01
	Parsing Time	0	0
	Service Duration	<0.01	0.01
File Transfer	File Size	0.05	0.09
Email	File Size	0.04	0.38

References

[3GPP09] 3GPP, *Technical Specification Group Radio Access Network; Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN),* Technical Specification TR 25.913, V9.0.0, Dec. 2009.

[3GPP16a] 3GPP, *Technical Specification Group Services and System Aspects; Policy and charging control architecture,* Technical Specification TS 23.203, V14.1.0, Sep. 2016.

[3GPP16b] *3GPP, Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers; Stage 1,* Technical Specification TS 22.891, V14.2.0, Sep. 2016.

[3GPP17] 3GPP, *Discussion on CQI and MCS table,* 3GPP TSG-RAN WG1 Meeting #91, R1-1719731, Nov. 2017. [Online]. Available: https://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_91/Docs

[3GPP18] *3GPP, Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2,* Technical Specification TS 23.501, V15.4.0, Dec. 2018.

[3GPP20a] 3GPP, *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification,* Technical Specification TR 36.331, V16.2.1, Sep. 2020.

[3GPP20b] 3GPP, *Technical Specification Evolved Universal Terrestrial Radio Access Network (E–UTRAN); Overall description; Stage 2,* Technical Specification TS 36.300, V16.3.0, Sep. 2020.

[3GPP20c] 3GPP, *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding,* Technical Specification TR 36.212, V16.3.0, Sep. 2020.

[3GPP20d] *3GPP, Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description; Stage 2,* Technical Specification TS 38.300, V16.3.0, Sep. 2020.

[3GPP20e] *3GPP, Technical Specification Group Radio Access Network; NR; Multiplexing and channel coding,* Technical Specification TS 38.212, V16.3.0, Sep. 2020.

[3GPP20f] *3GPP, Technical Specification Group Radio Access Network; NR; Base Station (BS) radio transmission and reception,* Technical Specification TS 38.104, V16.5.0, Sep.

2020.

[3GPP20g] *3GPP, Technical Specification Group Radio Access Network; NR; Physical channels and modulation,* Technical Specification TS 38.211, V16.3.0, Sep. 2020.

[3GPP20h] 3GPP, *Technical Specification Universal Mobile Telecommunications System (UMTS); Quality of Service (QoS) concept and architecture,* Technical Specification TS 23.107, V16.0.0, July. 2020.

[3GPP20i] 3GPP, *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures,* Technical Specification TR 36.213, V16.3.0, Sep. 2020.

[Ahma13] S. Ahmadi, *LTE–Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*, Academic Press, San Diego, CA, USA, 2013.

[Amar15] Amari LTE 100, *Software LTE base station on PC*, [Online]. Available: http://www.amarisoft.com. [Accessed Jan. 2021].

[ARET19] I.A. Alimi, A.M. Abdalla, A.O. Mufutau, F.P. Guiomar, I. Otung, J. Rodriguez, P.P. Monteiro and A.L. Teixeira, "Energy Efficiency in the Cloud Radio Access Network (C-RAN) for 5G Mobile Networks," in A.M. Abdalla, J. Rodriguez, I. Elfergani and A. Teixeira (eds.), *Optical and Wireless Convergence for 5G Networks*, Wiley-IEEE PRESS, Hoboken, NJ, USA, 2019.

[ASBD15] I. Alyafawi, E. Schiller, T. Braun, D. Dimitrova and A. Gomes, “Critical issues of centralized and cloudified LTE–FDD Radio Access Networks*”,* in *Proc. of ICC 2015 - IEEE International Conference on Communications*, London, UK, June 2015.

[Binm91] K. Binmore, *Fun and Games: A Text on Game Theory*, D.C. Heath, Lexington, MA, USA, 1991.

[BoVa04] S.P. Boyd and L. Vandenberghe, *Convex Optimization,* Cambridge University Press, West Nyack, NY, USA, 2004.

[Carr11] P. Carreira, *Data Rate Performance Gains in UMTS Evolution to LTE*, M.Sc. Thesis, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, Oct. 2011.

[Cerw20] P. Cerwall (ed.), *Ericsson Mobility Report,* Technical report, Jun. 2020. [Online]. Available: https://www.ericsson.com/49da93/assets/local/mobility-report/documents/2020/june2020-ericsson-mobility-report.pdf.

[CCYS14] A. Checko, H.L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M.S. Berger and L. Dittmann, “Cloud Ran for Mobile Networks – a Technology Overview”, *IEEE Communications Surveys & Tutorials,* vol. 17, No. 1, Sep. 2014, pp. 405 – 426.

[Cisc19] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022,* White paper, Feb. 2019. [Online]. Available: https://www.cisco.com.

[CMRI13] China Mobile Research Institute, *C-RAN: The Road towards Green Radio RAN*, White Paper, Dec. 2013, [Online]. Available: http://labs.chinamobile.com/cran/wp–content/uploads /2014/06/20140613–C-RAN –WP–3.0.pdf.

[Corr20] L.M. Correia, *Mobile Communication Systems*, Lecture Notes, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2020.

[COST20] COST, *Action CA15104 - IRACON*. [Online]. Available: http://www.iracon.org. [Accessed Jan. 2021].

[CVX20] *CVX - Software for Disciplined Convex Programming*, Dec. 2020, [Online]. Available: http://cvxr.com. [Accessed Jan. 2021].

[DaPS11] E. Dahlman, S. Parkvall and J. Skold, *4G, LTE/LTE–Advanced for Mobile Broadband*, Academic Press, Oxford, UK, 2011.

[DaPS18] E. Dahlman, S. Parkvall and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, Academic Press, Oxford, UK, 2018.

[DeDL15] B. Debaillie, C. Desset and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," in *Proc. of VTC 2015 Spring - IEEE Vehicular Technology Conference*, Glasgow, Scotland, May 2015.

[DaWF16] M. Dayarathna, Y. Wen and R. Fan, "Data Center Energy Consumption Modeling: A Survey," *IEEE Communications Surveys & Tutorials,* vol. 18, no. 1, First Quarter 2016, pp. 732 - 794.

[Eart12] *Energy Aware Radio and neTwork tecHnologies (EARTH)*, EC FP7–ICT Project, June 2012, [Online]. Available: https://cordis.europa.eu/project/id/247733 [Accessed Jan. 2021].

[FMPS20] F. Fossati, S. Moretti, P. Perny and S. Secci, "Multi-Resource Allocation for Network Slicing," *IEEE/ACM Transactions on Networking,* vol. 28, no. 3, pp. 1311-1324, June 2020.

[FPHG14] L.S. Ferreira, D. Pichon, A. Hatefi, A. Gomes, D. Dimitrova, T. Braun, G. Karagiannis, M. Karimzadeh, M. Branco and L.M. Correia, "An architecture to offer cloud–based radio access network as a service", in *Proc. of EuCNC 2014 - European Conference on Networks and Communications*, Bologna, Italy, June 2014.

[HDGK13] B. Haberland, F. Derakhshan, H. Grob–Lipski, R. Klotsche, W. Rehm, P. Schefczik and M. Soellner, “Radio base stations in the cloud”, *Bell Labs Technical Journal,* vol. 18, No. 1, June 2013, pp. 129–152.

[HMDM19] M.F. Hossain, A.U. Mahin, T. Debnath, F.B. Mosharrof and K.Z. Islam, "Recent research in cloud radio access network (C-RAN) for 5G cellular systems - A survey," *Journal of Network and Computer Applications,* vol. 139, pp. 31-48, Aug. 2019.

[HoTo11] H. Holma and A. Toskala, *LTE for UMTS Evolution to LTE–Advanced*, John Wiley & Sons Ltd., Chichester, West Sussex, UK, 2011.

[JaCH84] R.K. Jain, D.M. Chiu and W.R. Hawe, *A quantitative measure of fairness and discrimination for resource allocation in shared systems,* DEC Technical Report TR-301, Digital Equipment Corporation, Sep. 1984. [Online]. Available: https://www.cse.wustl.edu/~jain/papers/ftp/fairness.pdf.

[KeMT98] F.P. Kelly, A.K. Maulloo and D.K.H. Tan, "Rate control for communication networks: Shadow prices proportional fairness and stability," *Journal of the Operational Research Society,* vol. 49, no. 3, Apr. 1998, pp. 237-252.

[KhLL14] S. Khakurel, C. Leung and T. Le-Ngoc, "A Generalized Water-Filling Algorithm with Linear Complexity and Finite Convergence Time," *IEEE Wireless Communications Letters,* vol. 3, no. 2, Apr. 2014, pp. 225-228.

[KoSo19] U.C. Kozat and A.C.K. Soong, "On the Impact of Slicing Granularity on the Availability and Scalability of 5G Networks," in *Proc. of ICC 2015 - IEEE International Conference on Communications*, Shanghai, China, May 2019.

[LSJY20] W. Lei, A.C.K. Soong, L. Jianghua, W. Yong, B. Classon, W. Xiao, D. Mazzarese, Z. Yang and T. Saboorian, *5G System Design, An End to End Perspective*, Springer International Publishing, Cham, Switzerland, 2020.

[LZGN16] J. Liu, S. Zhou, J. Gong, Z. Niu and S. Xu, "Statistical Multiplexing Gain Analysis of Heterogeneous Virtual Base Station Pools in Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications,* vol. 15, no. 8, Aug. 2016, pp. 5681-5694.

[MAMM16] *Massive MIMO for Efficient Transmission (MAMMOET)*, FP7–ICT–2013–11, Dec. 2016, [Online]. Available: https://cordis.europa.eu/docs/projects/cnect/6/619086/080/deliverables/001-619086MAMMOETD32renditionDownload.pdf. [Accessed Mar. 2021].

[Myer91] R.B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, MA, USA, 1991.

[MCN15] *Mobile Cloud Networking project (MCN)*, EC FP7–ICT Project, Oct. 2015, [Online]. Available: http://mobile-cloud-networking.eu/site. [Accessed Jan. 2021].

[NGMN08] Next Generation of Mobile Network (NGMN), Next *Generation Mobile Networks Radio Access Performance Evaluation Methodology*, White paper, Jan. 2008, [Online].

Available: https://www.ngmn.org/wp-content/uploads/NGMN_Radio_Access_Performance_Evaluation_Methodology.pdf.

[NGMN13] Next Generation of Mobile Network (NGMN), *Suggestions on potential solutions to C-RAN,* White paper, 2013, [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN_CRAN_Suggestions_on_Potential_Solutions_to_CRAN.pdf

[NGMN15] Next Generation of Mobile Network (NGMN), *NGMN 5G White Paper*, White paper, Feb. 2015, [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf.

[OAI14] *Open Air Interface*, EURECOM, Oct. 2014, [Online]. Available: http://www.openairinterface.org. [Accessed Jan. 2021].

[PaEl10] D.P. Palomar and Y.C. Eldar, *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, West Nyack, NY, USA, 2010.

[PLLL11] C. Peng, S.B. Lee, S. Lu, H. Luo and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *Proc. of ACM MobiCom 2011 - International conference on Mobile computing and networking*, Las Vegas, NV, USA, Sep. 2011.

[PMZW10] S. Pelley, D. Meisner, P. Zandevakili, T.F. Wenisch and J. Underwood, "Power Routing: Dynamic Power Provisioning in the Data Center," in *Proc. of ASPLOS 2010 - Conference on Architectural Support for Programming Languages and Operating Systems*, Pittsburgh, PA, USA, Mar. 2010.

[PoHT16] D. Pompili, A. Hajisami and T.X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN", *IEEE Communications Magazine,* vol. 54, No. 1, Jan. 2016, pp. 26-32.

[PSAD05] J. Pérez–Romero, O. Sallent, R. Agustí and M.A. Díaz–Guerra, *Radio Resource Management Strategies in UMTS*, John Wiley & Sons, Chichester, West Sussex, UK, 2005.

[Rapp96] T.H. Rappaport, *Wireless Communications: Principles and Practices*, IEEE Press, Piscataway, NJ, USA, 1996.

[WeGP13] T. Werthmann, H. Grob-Lipski and M. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *Proc. of PIMRC 2013 - IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, London, UK, Sep. 2013.

[WRBL14] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy and G. Fettweis, “Benefits and Impact of Cloud Computing on 5G Signal Processing”, *IEEE Signal Processing Magazine,* vol. 31, No. 6, Nov. 2014, pp. 35–44.

[Viei18] A.B. Vieira, *Analysis of 5G Cellular Radio Network Deployment over Several Scenarios*, M.Sc. Thesis, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, 2018.

[ZCLD15] Q. Zheng, Y. Chen, H. Lee, R. Dreslinski, C. Chakrabarti, A. Anastasopoulos, S. Mahlke and T. Mudge, "Using Graphics Processing Units in an LTE Base Station," *Journal of Signal Processing Systems*, vol. 78, No. 6, Jan 2015, pp. 35–47.

[ZJJL17] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu and X. Wang, "Reconfigurable optical mobile fronthaul networks for coordinated multipoint transmission and reception in 5G," *IEEE Journal of Optical Communications and Networking*, vol. 9, no. 6, June 2017, pp. 489-497.