UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Radio access network optimization with proactive resource management for 5G and beyond

Rolando Guerra-Gómez

Universitat Politècnica de Catalunya

Signal Theory and Communications

Ph.D. Doctoral Thesis

Radio Access Network Optimization with Proactive Resource Management for 5G and Beyond.

Author

Rolando Guerra-Gómez

rolando.guerra@upc.edu

Supervisor

Dra. Silvia Ruíz Boqué

silvia.ruiz@upc.edu

Wireless Communications and Technologies

(WiComTec) Research Group

January 27, 2023

Abstract

The Fifth-Generation (5G) of cellular networks significantly increases the performance and flexibility of the offered services to users and service providers. The strict network requirement of 5G use cases has been supported by integrating service-based architecture in the core network, flexible radio access network architecture, and implementing numerous wireless technologies. Researchers and Mobile Network Operators (MNOs) face vast challenges not only in the definition process but also in the deployment phase.

The research community should define robust and dynamic radio network solutions to tackle the complexity and flexibility of 5G and beyond mobile network requirements. As mentioned, the radio access network architecture has been crucial in defining 5G systems. Especially the Cloud Radio Access Network (C-RAN) architecture has played a fundamental role as part of the new generation radio access network (NG-RAN) because it has the potential to support extremely dense radio network deployments while reducing costs because of the simplification of the radio units. Moreover, C-RAN enhances the network capacity by reducing the number of required resources because it centralizes the baseband functionalities in Baseband Units (BBU) pools or Central Units (CUs), sharing the same resource to manage multiple Remote Radio Heads (RRHs) or Radio Units (RUs). Moreover, Coordinated Multipoint (CoMP), enhanced Inter-Cell Interference Coordination (eICIC), and beamforming technologies could be easily implemented in the C-RAN structure, improving the 5G network performance.

However, the apparition of new use cases and network requirements forces the 5G systems to improve and go further with the initial stages of Sixth-Generation (6G) mobile networks. The network design of 5G and beyond must benefit society by being a human-centric reliable infrastructure. Moreover, future mobile networks should support immersive communication, cognition and twinning, deterministic end-to-end applications, and high-resolution sensing services. Sustainability and energy efficiency are crucial to support these services and network requirements.

Namely, It is fundamental to reduce energy consumption, resource usage, and emissions footprints to avoid excessive power consumption. The enormous increase in the number of devices, data amounts, and data rates implies an increase in the overall data traffic and required capacity, while energy reduction is not automatically guaranteed. On the other hand, the optimal management of the computational resources to satisfy current and future network requirements also becomes a challenge.

This doctoral thesis aims to address some of the challenges above. Most of the published research works employ synthetic scenarios to validate the results. A realistic C-RAN platform has been implemented, opposing these approaches. The proposed architecture considers rural and urban zones, heterogeneous deployment with macro and small cells, time-variant traffic patterns, realistic user equipments with guaranteed bit rate and best effort services, Quality of Service (QoS) constraints, a 3D ray-tracing propagation model, multiple frequency bands, different split options, among other significant features. This platform becomes fundamental in the validation of the proposed algorithms.

Additionally, 5G and beyond radio network deployment will be ultra-dense. However, optimizing the costs and energy consumption is not automatically guaranteed. For this reason, this thesis also provides a non-linear data modeling and decision-making tool to maximize cost reduction versus coverage-QoS trade-off by optimizing the active RRHs needed according to traffic demands. The cost and energy optimization are analytically expressed by modeling the complex relationships between input and output system parameters.

The optimization tool is based on a multi-objective integer linear programming model designed to reduce the network cost while maintaining suitable coverage and QoS. Results have been presented considering 3.6 and 28 GHz frequency bands and different split options. The obtaining cost reduction ranges from 30 % to 70 % depending on the scenario.

On the other hand, previous works on BBU pool resource allocation have relied on the definition of optimization problems. Most of these strategies allocate the resources assuming the instantiated computational capacity at BBU pools is fixed and equal to the maximum pool capacity. Under this assumption, the computational resources could be over-provisioned or under-provisioned, causing inefficient resource utilization or QoS degradation.

On the other hand, the design of efficient computational resource management in C-RAN environments is a challenging problem because it has to account simultaneously for throughput, latency, power efficiency, and optimization trade-offs. Most of the reviewed works assume a fixed computational capacity at BBU pools, which results in underutilized or oversubscribed resources, thus affecting the overall QoS. The resources could be dynamically

instantiated according to the required computational capacity (RCC).

For this reason, this thesis proposes a novel strategy for Dynamic Resource Management with Adaptive Computational capacity (DRM-AC) using Machine Learning (ML) techniques. Three ML algorithms have been considered in the final design after testing multiple approaches: support vector machine (SVM), time-delay neural network (TDNN), and long short-term memory (LSTM).

DRM-AC reduces the average of unused resources by 96 %, but there is still QoS degradation when RCC is higher than the predicted computational capacity (PCC). To further improve, two new strategies are proposed and tested in a realistic scenario: DRM-AC with pre-filtering and DRM-AC with error shifting, reducing the average of unsatisfied resources by 98 % and 99.9 % compared to the DRM-AC, respectively.

Contents

Acronyms

SBS	Small Base Station	25
SD	Standard-Definition	102
SDHC-CRAN	Software-defined Hyper-Cellular C-RAN	17
SDN	Software-Defined Networking	35
SDVRAN	Software-Defined and Virtualized RAN	18
SFC	Service Function Chain	77
SINR	Signal-to-Noise-plus-Interference-Ratio	34
Soft-RAN	Software-Defined for Radio Access Networks	15
SP	Service Provider	77
SRRH	small-RRH	8
SVM	Support Vector Machine	12
TCO	Total Cost of Ownership	24
TDNN	Time Delay Neural Network	12
TDP	Traffic Demand Point	32
UE	User Equipment	2
UPF	User Plane Function	19
uRLLC	ultra-Reliable Low Latency Communication	22
V2X	Vehicle to Everything	22
VoIP	Voice over IP	7
VoNR	Voice/Video over New Radio	102
XR	Extended Reality	21

List of Figures

List of Tables

Chapter 1

Introduction

Future mobile networks will face a new paradigm in which billions of things, humans, vehicles, robots, and drones will coexist. They will deal with challenging use cases such as holographic telepresence and immersive communications, which will produce more strict requirements [2].

Beyond Fifth-Generation (5G) networks must help to handle those issues. Especially, Sixth-Generation (6G) will be a self-contained ecosystem of Artificial Intelligence (AI) with numerous features such as intelligent connected management, reduction of energy footprint, and programmability. AI will be used to enhance network performance. It will help to maintain the cost-effectiveness of envisioned complex 6G services, such as the interaction between human-digital-physical worlds, to automate some level of decision-making processes, and to achieve a zero-touch approach [2].

Besides, the evolution of 5G continues toward 5G Advanced to expand its usage by supporting new use cases and verticals. In this sense, machine learning (ML) will also play a significant role in the network optimization to

support reduced capability devices and network energy efficiency [3].

The 3rd Generation Partnership Project (3GPP) includes the Cloud Radio Access Network (C-RAN) architecture in the standardization of the 5G Radio Access Network (RAN). Furthermore, it must also be considered in the definition of future mobile networks because C-RAN has the potential to support extremely dense mobile networks and to reduce the number of required Baseband Units (BBUs) by 75 % compared to the traditional RAN architecture [4], efficiently enhancing the network capacity.

Fig. 1.1 shows an essential diagram of C-RAN architecture. The Remote Radio Heads (RRHs) transmit the In-phase and Quadrature (IQ) signals from User Equipment (UE) through the fronthaul link to the BBUs. The BBU pools concentrate and virtualize the resources to handle dynamically many RRHs. As a result, BBU pools aggregate data traffic from different types of cells to the backhaul link and favor the rise of the multiplexing gain.

However, the optimization of the network deployment, computational resources, energy footprint, and power consumption is not automatically guaranteed. Besides, beyond Beyond Fifth-Generation (B5G) standards must be a step further because they should offer Tbps of data rate, sub-ms of latency, zero-touch management, and proactive decision-making. Additionally, future mobile networks should support immersive communication, cognition and twinning, deterministic end-to-end applications, and high-resolution sensing services.

Sustainability is crucial to support these services and network requirements, becoming fundamental to reducing energy consumption, resource usage, and emissions footprints.

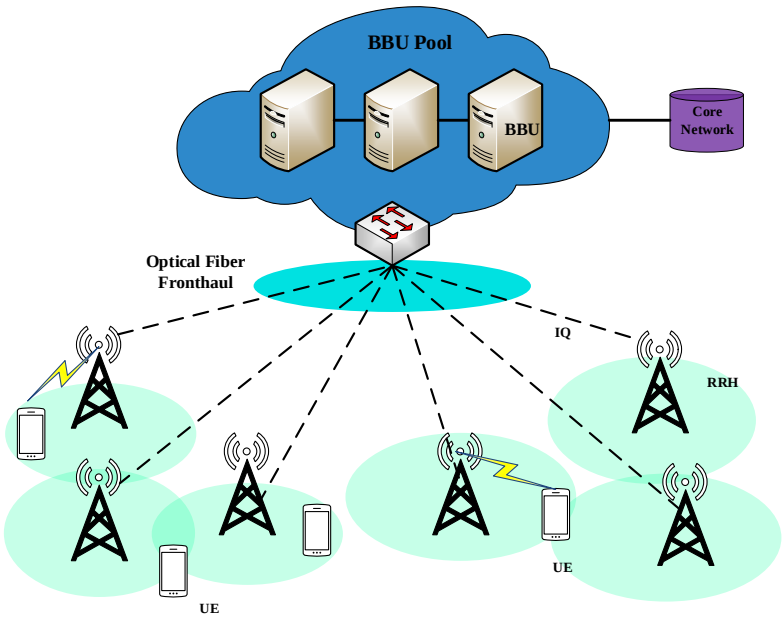


Figure 1.1: Cloud radio access network architecture

Additionally, managing computational resources at Central Units (CUs) to satisfy the traffic demand of the RRHs or Radio Units (RUs) becomes a challenging critical aspect. Previous works on BBU pool resource allocation have relied on the definition of optimization problems such as mixed-integer linear programming (MILP) or Multi-Objective Optimization (MOO). These strategies allocate the resources assuming that the instantiated computational capacity at BBU pools is fixed and equal to the maximum BBU pool capacity. Under this assumption, the computational resources could be over-provisioned or under-provisioned, causing inefficient resource utilization or Quality of Service (QoS) degradation, respectively.

This thesis contributes to solving these issues, combining the flexibility of virtualization and the availability of machine learning techniques to predict computational demands and applying optimization techniques to reduce the energy and cost footprint of 5G and beyond radio access networks.

The resources could be instantiated dynamically according to an anticipated computational capacity demand. For this reason, this work proposes the integration of Dynamic Resource Management (DRM) with a prediction of the required computational capacity based on ML techniques, which allows defining a DRM with adaptive capacity (DRM-AC) to avoid under-utilization of the computational resources and to maintain QoS.

On the other hand, this work also presents an optimization framework with two purposes. Firstly, the efficient deployment of 5G and B5G radio networks on C-RAN ecosystems. Secondly, the activation/deactivation of the RRHs to maintain the coverage and QoS while minimizing the network cost. The proposed algorithm selects the optimal distribution of RRHs from candidate locations. The algorithm includes the possibility of implementing cooperation strategies between cells, automatically establishing the cells that should cooperate to satisfy the traffic demand of a specific zone in the map. Section 1.1 summarizes the global and specific objectives that guided the research.

1.1 Research Objectives

As previously described, this doctoral thesis tackles multiple challenges and open issues of current and future radio access networks. In this direction,

research questions have been formulated, which guide and motivate the whole research:

1. How to implement a realistic C-RAN simulation platform that handles the trade-off between the flexibility of mobile networks and computational simplicity?

2. How to efficiently deploy current and future radio access networks to reduce costs and energy consumption?

3. How to efficiently manage the computational resources of the radio access network? Is it possible to proactively instantiate the computational resources necessary to increase efficiency while keeping QoS?

Numerous objectives have been appointed to solve these research problems, classifying them into global and specific objectives. The global goals of this doctoral thesis are:

1. To perform a comprehensive and systematic literature review on related works and technologies (e.g., 5G, C-RAN, ML, QoS, service generation, and mathematical optimization) to identify challenges and open issues.

2. To design a realistic C-RAN platform that represents the non-uniformity of mobile networks to obtain a suitable validation platform.

3. To test and compare multiple strategies of fronthaul design, not only to establish the BBU-RRH connections of the C-RAN platform but also to provide a comprehensive analysis to the Mobile Network Operators (MNOs).

4. To formulate and implement an efficient RRH deployment algorithm that reduces the cost and energy footprint of the massive deployments of current and future networks.

5. To propose novel approaches to efficiently manage the centralized computational resources, introducing machine learning techniques to proactively instantiate the required computational resources.

Different specific objectives or tasks have been implemented to accomplish the global goals:

- To perform a literature review on radio access network architectures.
- To analyze multiple proposals of simulation platforms, emphasizing their advantages and limitations to identify the fundamental features that should represent mobile network deployments.
- To define and implement a realistic C-RAN simulation platform that is used to validate the performance of the proposals.
- To study, test, and compare different approaches of fronthaul design to resume multiple options of BBU-RRH connections.
- To review the research related to radio access network deployment
- To formulate a nonlinear mathematical algorithm to optimize the RRH deployment to reduce costs and energy consumption.
- To analyze the literature on the management of computational resources and machine learning.

- To design a dynamic resource management strategy that allocates the computational resources of the BBU pools considering a maximum and fixed capacity, which is used as a benchmark to validate the advantage of the resource management strategies with adaptive capacity.

- To implement novel strategies of dynamic resource management with adaptive capacity.

1.2 Contributions

1. Unlike most analyzed works, a realistic scenario of a C-RAN environment over Vienna City is used to validate the results. UEs generating Voice over IP (VoIP), video streaming, File Transfer Protocol (FTP), or web browsing services have been modeled. The deployment of the C-RAN follows a non-uniform distribution.

2. A formulation and software implementation of decision-making rules to optimize the number of required active RRHs under different traffic patterns in a heterogeneous environment is presented. It is done analytically by obtaining a complete set of nonlinear equations. The complex relationships between the input and output parameters of the system in a mobile network environment are modeled to optimize the RRH network deployment. In general terms, the proposed optimization algorithm reduces Capital Expenditures (CAPEX), Operating Expenditures (OPEX), and the energy footprint.

3. We present an analysis of the consequences of different split options; the

splits introduce a cost ratio between Base Stations (BSs) types, especially when heterogeneous network deployments with Macro and Small BSs are considered. The cost ratio between macro-RRHs (MRRHs) and small-RRHs (SRRHs) increases when they contain more baseband functionalities. Thus, this cost ratio achieves the maximum value in distributed RAN (option 1) scenarios, while it is minimum in full C-RAN (option 8). Based on this, the impact of different split options on the deployment cost and coverage-QoS of the radio network is detailed.

4. As far as our knowledge, the optimization of the RRH deployments of B5Gs in terms of energy and cost-saving considering: flexible radio access network, both frequency ranges (sub 6GHz and mm-wave), coverage and capacity constraints, realistic propagation models, and heterogeneous mobile networks with MRRHs and SRRHs is not available in the literature.

5. Opposing the approaches followed by the reviewed literature, we propose to proactively instantiate just the required computational capacity at the BBU pools, to improve the resource usage ratio. Namely, the work focuses on optimizing the computational resources at BBU pools according to the required computational demand. The combination of a previously designed dynamic resource management with a strategy to forecast the necessary computational capacity to reduce the amounts of underutilized resources while keeping the required QoS. The proposal is called DRM-AC.

6. The key performance indicator that describes the QoS when DRM-AC

is analyzed is the percentage of satisfied resources. As a result, the QoS is degraded if the instantiated computational resources at a BBU pool are insufficient to satisfy the required computational capacity. The DRM-AC instantiates the resources based on predicting the necessary computational resources. However, errors during the forecasting process might produce a QoS degradation (under-provisioned case). We propose two novel schemes, DRM-AC with prefiltering (DRM-AC-PF) and DRM-AC with error shifting (DRM-AC-ES), to tackle this issue.

7. A significant contribution of this doctoral thesis is the combination of the previously mentioned proposals in a flexible simulation platform. The proposed platform supports multiple upgrades, which pave the road to a wide range of future works. This tool undoubtedly could help researchers and MNOs to improve their network planning and management, controlling the balance between QoS and cost reduction.

1.3 Publications

The research of this doctoral thesis has been periodically published. This section summarizes the publications in conferences and journals.

1. Rolando Guerra-Gómez, Silvia Ruiz, M. García-Lozano, and Joan Olmos, "Using COST IC1004 Vienna scenario to test C-RAN optimization algorithms," in COST IRACON, Dublin, Ireland, Jan. 2019.

2. Rolando Guerra-Gómez, Silvia Ruiz, M. García-Lozano, and Joan Olmos, "A weighted-sum multi-objective optimization for dynamic re-

source allocation with QoS constraints in realistic C-RAN," in COST IRACON, Oulu, Finland, May 2019.

3. Rolando Guerra-Gómez, Silvia Ruiz, M. García-Lozano, and Joan Olmos, "Predicting Required Computational Capacity in C-RAN networks by the use of different Machine Learning strategies," in COST IRACON, Gdańsk, Poland, September 2019.

4. Rolando Guerra-Gómez, Silvia Ruiz, M. García-Lozano, and Joan Olmos, "Dynamic Resource Allocation in C-RAN with Real-Time Traffic and Realistic Scenarios," in 2019 15th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob).

5. Rolando Guerra-Gómez, S. R. Boqué, M. García-Lozano, and J. O. Bonafé, "Machine Learning Adaptive Computational Capacity Prediction for Dynamic Resource Management in C-RAN," in IEEE Access, vol. 8, pp. 89130-89142, 2020, doi: 10.1109/ACCESS.2020.2994258.

6. Rolando Guerra-Gómez, S. R. Boqué, M. García-Lozano, and J. O. Bonafé, "Machine-Learning based Traffic Forecasting for Resource Management in C-RAN," 2020 European Conference on Networks and Communications (EuCNC), 2020, pp. 200-204.

7. Rolando Guerra-Gómez, S. R. Boqué, M. García-Lozano, and U. Saeed, "Energy and Cost Footprint Reduction for 5G and Beyond With Flexible Radio Access Network," in IEEE Access, vol. 9, pp. 142179-142194, 2021, doi: 10.1109/ACCESS.2021.3120765.

8. Rolando Guerra-Gómez, Silvia Ruiz, M. García-Lozano, and U. Saeed, "Flexible Radio Access Network Optimization with Cell Coordination," in 1st INTERACT: Intelligence-Enabling Radio Communications for Seamless Inclusive Interactions, Bologna, Italy, Feb 8–11, 2022.

1.4 Thesis Outline

The structure of this thesis is divided into seven chapters. Chapter 2 presents a literature review in which multiple research works with different objectives and contributions have been investigated. The fundamental aim of this chapter is to analyze strengths and weaknesses to identify challenges and open issues. Some of which have been addressed by the proposals of this thesis.

After the analysis, there is room for improvement in multiple research directions, such as energy and cost footprint reduction of the mobile networks, fronthaul design and optimization, and design of strategies to manage the centralized computational resources, among others. Additionally, It has been identified that most of the analyzed works employ synthetic scenarios that could not represent the complexity of the mobile networks. These simplifications may reduce the significance of the results.

Furthermore, ML and AI have been identified as enables of future mobile networks. For this reason, chapter 3 presents a resume of machine learning techniques. A general description of the main categories of machine learning approaches ML has been delivered. Section 3.2 describes unsupervised and supervised learning, reinforcement learning, deep learning, and federated learning. Multiple approaches have been tested and compared, and the

best results were obtained with Support Vector Machine (SVM), Time Delay Neural Network (TDNN), and Long Short-Term Memory (LSTM). For this reason, a mathematical background of these strategies is presented in sections 3.3, 3.4, and 3.5.

On the other hand, chapter 4 introduces features of the proposed simulation platform, which is a fundamental contribution of this thesis. The platform is designed in Matlab. It tries to represent the complexity of mobile networks to improve the quality of the validation.

The platform employs realistic models in each layer of the C-RAN architecture. UEs, Guaranteed Bit Rate (GBR), and Best-effort services at the packet level have been modeled in the user plane. The air interface has been represented using a 3D ray-tracing model that provides all the correlations and spatial consistencies. These features allow accounting for QoS and designing resource management strategies, among other advantages, with a high level of flexibility.

Once the simulation platform has been explained, chapter 5 presents the mathematical formulation of the proposed algorithms. Firstly, section 5.2 introduces the mathematical description of four strategies to design and analyze the fronthaul connections. Section 5.3 introduces the DRM, as well as three variants of DRM-AC. Section 5.4 details the proposed non-linear optimization model to optimize the RRH deployment.

These models are integrated into the simulation platform. They increase flexibility and facilitate the optimization of deployment and network management.

The performance of these models is analyzed in chapters 6 and 7. Chapter

6 presents the results of testing a radio network deployment using the proposed optimization algorithm. The algorithm reduces the number of required active RRH, offers acceptable coverage, and satisfies the QoS requirements.

On the other hand, chapter 7 discusses the results associated with the proposed DRM-AC. It reduces the underutilized resources by 96 % compared to the DRM with fixed computational resources.

However, it degrades the QoS when the predicted computational resources are insufficient to satisfy the demand. This issue is solved by proposing two novel strategies: DRM-AC-PF and DRM-AC-ES. They reduce unsatisfied resources by 98 % and 99.9 % compared to the DRM-AC and the number of underutilized resources by 75 % and 70 % compared to the DRM, respectively.

Finally, chapter 8 summarizes the general conclusions and future works.

Chapter 2

Literature Review

2.1 Introduction

This chapter presents a literature review to analyze state of the art. Multiple research works with different objectives and contributions have been investigated. The fundamental aim of this chapter is to analyze the main strengths and weaknesses of the related works. It helps to identify multiple challenges and open issues, some of which have been addressed by the proposals of this thesis.

2.2 Evolution of RAN architectures towards 5G

The authors of [5] describe the RAN architecture evolution. The concept of C-RAN was proposed in 2011 [6]. In 2013, Software-Defined for Radio Access Networks (Soft-RAN) and Open-RAN structures were presented in [7]

and [8], respectively. Soft-RAN is a flexible programmable architecture that decouples the control and data planes. This structure enables a centralized management layer in a software-defined network controller entity to manage the resources more efficiently.

In 2014, the Heterogeneous Cloud Radio Access Network (H-CRAN) was defined by [9] as an alternative to overcome the fronthaul capacity limitations of C-RAN. It consists of a Macro Base Station (MBS) and inside its coverage area, RRHs. MBS is connected to the BBU pool using a backhaul link, while the RRHs use the fronthaul links, as Fig. 2.1 shows. The functions of the control plane are only implemented in the MBS, while RRHs manage the data traffic. Consequently, H-CRAN split the control plane from the data plane to reduce the overhead through the fronthaul link, enhancing the C-RAN capabilities.

In 2015, function splitting was defined to overcome the fronthaul capacity limitation, splitting the baseband processing tasks between RRHs and cloud BBU. This method can overcome the additional transmission delay of the fronthaul link, especially where the distance between RRH and the cloud center is large. However, the disadvantage of this solution is financial cost increment since each RRH should have baseband processing capabilities, also called Remote Radio System (RRS).

On the other hand, authors of [11] proposed a Hierarchical H-CRAN architecture. This strategy combines H-CRAN and function splitting to overcome the fronthaul capacity limitation. The architecture is composed of a control MBS and RRHs, which have function splitting capability. Although the fronthaul limitation is addressed, the overall C-RAN advantages cannot

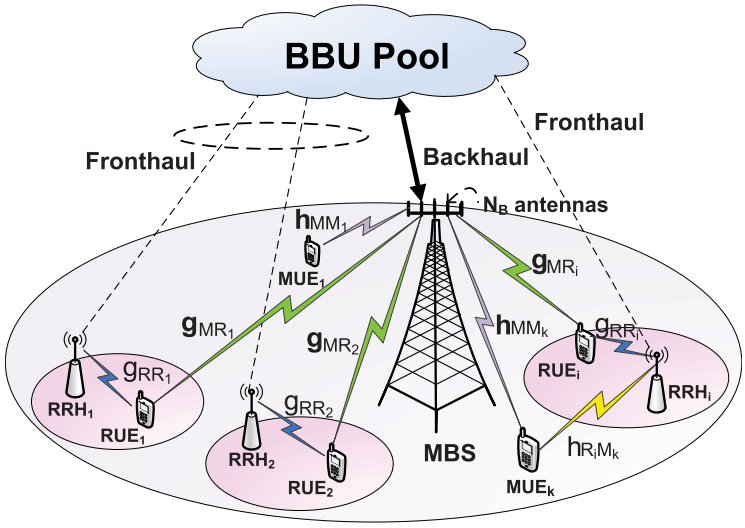


Figure 2.1: Heterogeneous C-RAN architecture [10]

be achieved, which is critical in high-density scenarios.

Additionally, authors of [12] presented a Software-defined Hyper-Cellular C-RAN (SDHC-CRAN) in 2016. A cloud-based software-defined RAN with physical decoupling and the ability to turn off RRHs during low-traffic hours. Fog Radio Access Network (F-RAN) is proposed in [13], where RRHs are equipped with caching capability to decrease the latency of popular contents.

In 2017, the authors of [14] introduced the Hierarchical Software-Defined RAN (HSD-RAN) architecture that is shown in Fig. 2.2. They proposed, instead of virtualizing all BSs in a single centralized controller, to form multiple groups concerning the BS geographic locations, which are assigned to the local controllers. The connections between the clusters and their associated controllers are established via fronthaul links. A high-level controller coordinates control plane decisions among the local controllers [14]. Moreover,

authors in [15] proposed an integrated architecture for Software-Defined and Virtualized RAN (SDVRAN) with fog computing.

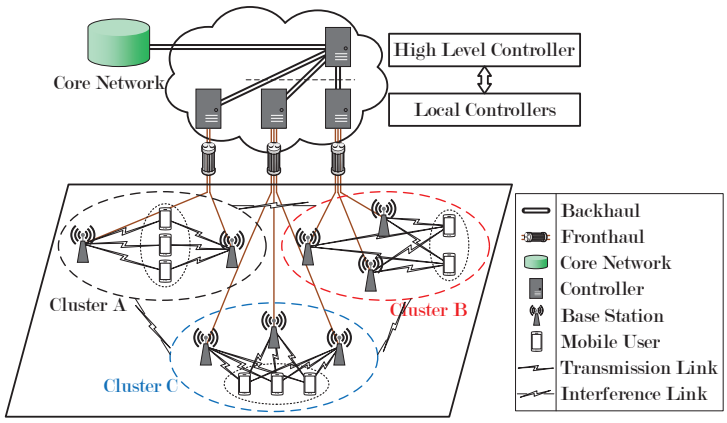


Figure 2.2: Hierarchical software-defined RAN architecture [14]

Recently, authors in [5] proposed a density-aware C-RAN design named Heterogeneous Virtualized Software-Defined C-RAN (HVSD-CRAN). The architecture manages two different scenarios in terms of network density: high-density and low-density modes. Fig. 2.3 shows the architecture, where the radio access layer is split into two parts depending on the mode. The low-density mode involves deploying RRSs that manage the UE data/control signals. This mode does not exploit the C-RAN advantages, but it is not a critical aspect due to low-density scenarios are not highly demanding. On the contrary, high-density mode scenarios demand all the advantages of fully centralized architecture. For this reason, the radio access layer of this mode is implemented by an H-CRAN strategy where the control messages are sent through a coverage layer (MBS) and data messages through a traffic layer with caching capability. The BBU cloud layer consists of a set of BBU

processing servers and a virtualized layer where a slicing controller manages the slicing resource allocation. Core and application layers complete the structure.

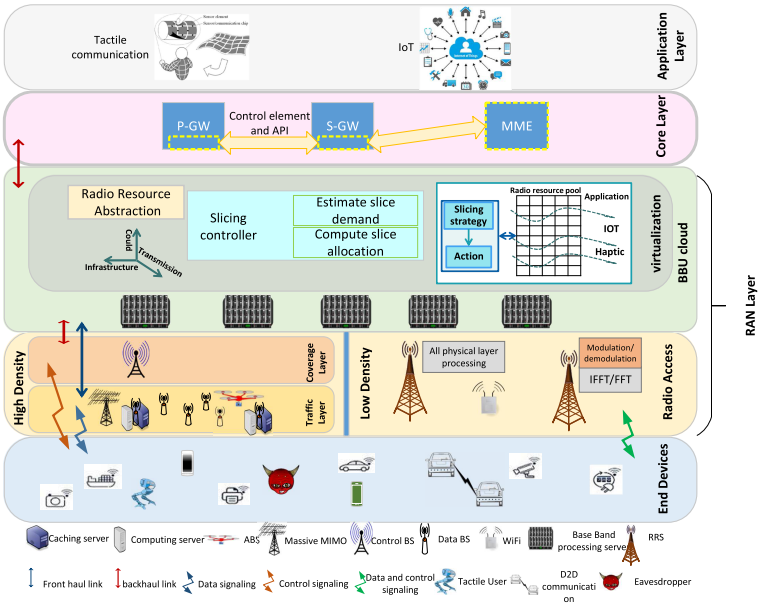


Figure 2.3: HVSD-CRAN architecture [16]

After the contribution of multiple researchers, 3GPP has introduced the 5G RAN or New Generation RAN (NG-RAN) [17]. Fig. 2.4 shows the general architecture of a 5G network. The 5G core network has been simplified to represent the control and user plane by the Access & Mobility Management Function (AMF) and User Plane Function (UPF), respectively. On the other hand, the NG-RAN is represented by the Next Generation NodeBs (gNBs), and New Generation evolved NodeBs (ng-eNBs), interconnected using the X_n interface. At the same time, the connections with the core are established using the NG interface.

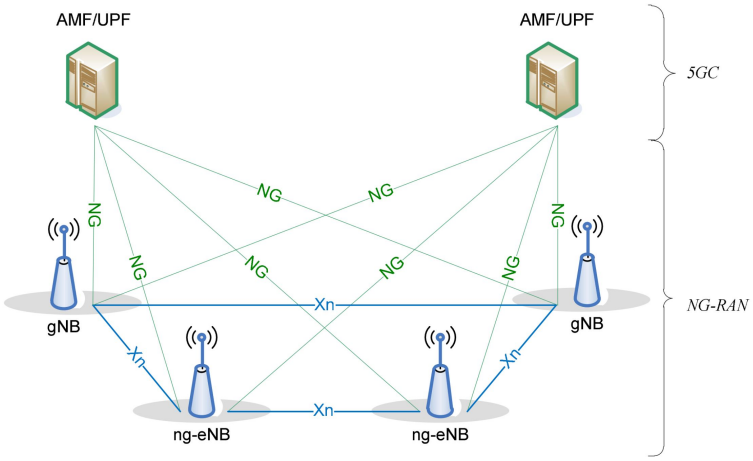


Figure 2.4: NG-RAN general architecture [17].

Furthermore, 3GPP has also defined multiple split options of the radio access network [18]. Fig. 2.5 shows a general NG-RAN scheme that details a gNB with CU-Distributed Unit (DU) split. As a result, the radio access network could include fronthaul (RU-DU), middlehaul (CU-DU, F1 interface), and backhaul links (CU-5GC, NG interface). Additionally, DU and CU could be collocated, avoiding the middlehaul.

Recently, [19] proposed and analyzed a service-based RAN, and the authors expect that it will enable MNOs to create fast and efficient service provisioning pipelines. The development of various open-source tools, libraries, and components will help accelerate the integration, deployment, and use of service-based RAN.

As previously described, multiple research works have contributed to the evolution of the RAN architectures. Works [5] and [20] provide additional details about these structures and other proposals. Mainly, the authors of [20] present a comprehensive survey on multiple RAN architectures such

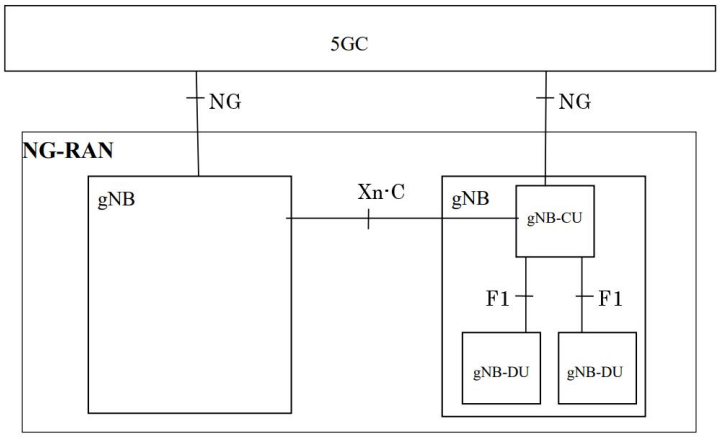


Figure 2.5: NG-RAN diagram with CU-DU split [18]

as cloud-RAN, heterogeneous cloud-RAN, virtualized cloud-RAN, and fog-RAN. They also compare the architectures from diverse perspectives, such as energy consumption, OPEX, resource allocation, spectrum efficiency, and network performance.

2.3 Evolution of 5G system towards 6G

Besides, the evolution of 5G continues through 5G Advanced towards 6G to expand its usage by supporting new use cases and verticals. In this sense, AI/ML will play a significant role in supporting Extended Reality (XR), reduced capability devices, and network energy efficiency [3]. Fig. 2.6 shows a tentative evolution plan of 3GPP networks presented in [1].

Since its introduction in release 15, 5G has targeted several main use cases, such as enhanced Mobile Broadband (eMBB). The 5G system provides superior network performance in terms of capacity; enables many new use

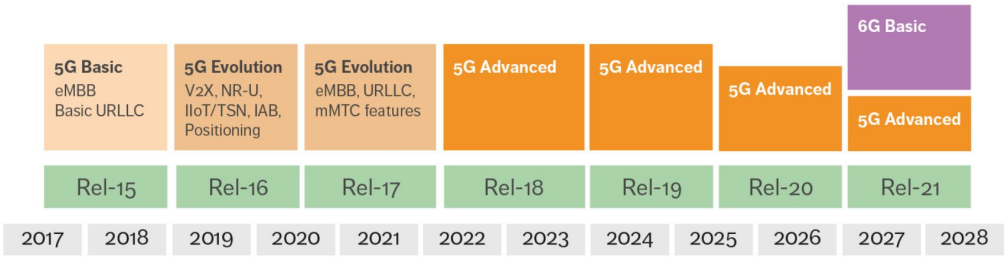


Figure 2.6: Tentative evolution diagram to 6G [1]

cases and support for new verticals compared to previous generations of 3GPP systems [3].

Release 16 contains a set of improvements to the 5G system and multiple features introduced in previous releases, such as those related to mission-critical and conversational services. It mainly enhances the NR interface and cooperation with Long Term Evolution (LTE). Numerous configurations of carrier aggregation and 256QAM are introduced to increase the bit rate. The main objective of this release is to make the 5G system suitable for several verticals (e. g. Vehicle to Everything (V2X), automated factories, time-sensitive networks, and public safety). To this end, this release introduces several enhancements to ultra-Reliable Low Latency Communication (uRLLC), non-public networks, slicing, and positioning services. In addition, release 16 also upgrades coexistence with non-3GPP networks, and network optimization [21].

On the other side, some improvements to 5G systems have been introduced in release 17. It covers roaming and non-roaming approaches, policy control and charging, among other general features such as location and

emergency services. However, this release is mostly dedicated to consolidating and enhancing concepts of previous releases as services to the industry, 5G support to Internet of Things (IoT), proximity communications in the context of V2X, and mission-critical services. The radio access network is upgraded to support these services. The introduction of 1024QAM modulation in downlink is an outstanding example. Additional details of these improvements can be found in [22].

According to [1], release 18 is a significant evolution of the 5G system, which motivated 3GPP to classify this release as 5G_Advance. Mainly, it will introduce numerous improvements related to artificial intelligence and extended reality, which result in intelligent network solutions to support multiple use cases. Fig. 2.6 represents the starting point of 6G in 2027. However, the requirement of 6G will be previously defined in parallel with 5G_Advanced evolution.

2.4 Energy Efficiency and Cost-Saving

Energy efficiency, power consumption, and cost-saving optimization are critical challenges for creating green communication environments. In recent years, numerous research works have been focused on addressing these challenges in C-RAN, such as [4, 23–27].

The potential of C-RAN to reduce power consumption and cost is evaluated in detail in [24]. A real case scenario was built accounting for different service traffic profiles. The results show that C-RAN enables the reduction of signal processing resources four times, significantly impacting cost reduction

and power saving.

The authors of [4] also presented an analysis of cell traffic profiles and the impact of the multiplexing gain. This analysis demonstrated the capabilities of C-RAN to improve cost-saving and energy efficiency. Furthermore, the authors presented considerations to optimize green deployments in terms of the Total Cost of Ownership (TCO). The performance is analyzed utilizing the ratio between the BBU pool and fiber per meter cost. The results show that values greater than 3 produce considerable cost reduction in a partially centralized C-RAN. In contrast, values greater than ten are needed to centralize all the RRHs. The authors conclude that the advantage is higher in smaller and denser scenarios (100 km2) than in larger systems (400 km2).

Authors in [25] survey the energy efficiency strategies in C-RAN environments. The authors classified these techniques into three general groups: RRHs on/off, renewable energy sources, and resource allocation optimization. RRHs on/off techniques analyze the capability of the network to activate or deactivate RRHs depending on the demand to reduce power consumption. The renewable strategy aggregates alternative energy sources to the RRHs to increase energy efficiency. The last method refers to design optimization problems for resource allocation schemes to reduce power consumption.

Reducing power consumption in mobile communications is an attractive area for many researchers. In [28], algorithms that control the RRHs on/off (sleep or active modes) status to reduce power consumption were designed, and the transition between these states was addressed considering the number of UEs per RRHs. This approach produced another challenge because high mobility environments with high fluctuations generate a pin-pong effect

(consecutive on/off transitions).

On the other hand, the authors of [27] propose two algorithms to optimize power consumption and handover. The first algorithm selects a suitable level of RRH switching. The second algorithm determines a suitable range before changing the state of the Small Base Station (SBS). A genetic algorithm is used to optimize the two parameters, and a Markov chain models the transitions between active and sleep modes.

2.5 Machine Learning and Deep Learning

Machine and deep learning techniques have been widely used in many research fields. Significantly, they have been employed in multiple tasks of mobile communications, such as traffic classification, traffic load management, and cluster formation [29–37].

Authors in [32] propose a centralized resource allocation scheme using online learning, which addresses interference mitigation, maximizing energy efficiency while maintaining the QoS requirements challenge in H-CRAN for 5G networks. Resource blocks (RBs) and transmission power are allocated and subjected to inter-tier interference and capacity constraints. The resource allocation is performed at a dedicated controller integrated with the BBU pool, and the MBS act as brokers between the controller and the RRHs for control exchange. The considered online learning model was a stochastic approximation method that solves the Bellman equation associated with the discrete-time markovian decision process.

The authors of [35] used a Random Forests algorithm to design a learning-

based resource allocation scheme for 5G systems. The algorithm is a multi-class classifier to predict the modulation and coding scheme.It aims to reduce the signal overhead in the network. Results show that due to the reduction in signaling, the proposed algorithm performs better in high user-density scenarios than in Channel State Information (CSI) schemes.

A reinforcement learning-based resource allocation strategy is proposed in [36]. The algorithm consists of two stages. First, a neural network model with LSTM cells was employed to predict the user position. LSTM is a Recurrent Neural Network (RNN) potent in predicting time series. Consequently, a reinforcement learning strategy based on the mobility pattern previously estimated is used to maximize the network throughput.

2.6 Resource management strategies

To face the naturally fluctuating traffic between day and night, weekdays and weekends, residential, commercial, and mixed areas dynamic resource allocation algorithms have been proposed in many research works [5, 23, 38–43]. However, resource allocation in C-RAN faces many challenges that need attention because dynamic resource management strategies for wireless communications are complex to design and implement. User mobility, radio channel variations, coverage, interference, frequency reuse, power control mechanism, and QoS requirements are some of the most critical factors contributing to wireless complexity in resource allocation. For these reasons, optimized solutions for resource allocation to ensure adequate resource utilization are required.

Authors in [38] survey the literature on clustering algorithms applied to C-RAN architectures, evaluate the resulting configuration of BBU pools, and present different techniques for RRH clusterings, such as multi-objective optimization clustering and bin-packing approach. The authors conclude that clustering can enhance the performance of the network. However, it is space for more analysis to select the best technique depending on the metric to optimize.

In [5] have been proposed an adaptive architecture for C-RAN with two operation modes according to the average user density: High and Low-density modes that will coexist in real 5G networks. The authors presented a Multi-Objective Resource Allocation (MORA) to optimize data rate and power consumption in the high-density mode. On the other hand, total cost and delay become the objective functions in the low-density mode.

A resource allocation strategy is implemented in a centralized architecture in high-density mode. However, a few RRHs with baseband processing capability are deployed in the low-density mode, where the authors proposed a distributed resource allocation strategy to reduce latency and cost.

On the other hand, an optimization approach is used in [23] to decompose the resource allocation problem into three sub-problems: hybrid energy management, data requesting, and power allocation. Authors consider that each RRH is equipped with an Energy Harvesting (EH) module, including a solar panel or wind turbine, a rechargeable battery, and a data buffer, as only renewable energies are not enough RRHs are also connected to the power grid.

Letter [40] addresses the problem of maximizing the total throughput of

the network via joint user association and power allocation in H-CRAN, accounting for QoS requirements. A generalized Stackelberg game approach was applied to this problem. A combination of centralized and distributed techniques was designed to achieve the solution. Significantly, the authors solved the user association problem using a centralized strategy while employing a distributed approach in the power allocation scheme.

A framework to optimize user association, radio resource allocation, and power allocation in H-CRAN is also proposed by [44]. In this case, the optimization problem is formulated to maximize the overall rate while considering RRH constraints, interference threshold for macro RRHs associated devices, and QoS constraints. The authors employed a matching game and Lagrange dual-decomposition to optimize the transmitted power.

Additionally, the authors in [45] formulated a joint user association and resource allocation problem to provide a better QoS to the IoT devices in the downlink of a fog network. They take into account the demand of QoS imposed by uRLLC and eMBB services. A matching game approach is also used to initiate a stable association between IoT users and Fog infrastructure.

A hybrid approach for RRH clustering based on game theory is presented in [41]. The authors address the BBU-RRH association problem in a decentralized manner intending to reduce power consumption. At the same time, the adequate number of active acpBBU is calculated using a centralized strategy. The authors propose two approaches: the first relies on the best response algorithm, and the second is based on a reinforcement learning method. The results show comparable performance to the fully centralized approach. The authors of [42] proposed an RRH clustering scheme that optimizes the power

consumption and re-association rate of the UEs performing handover. In this scheme, a non-cooperated game is used to solve the problem.

Authors in [43] have proposed two strategies for RRH clustering: first a centralized approach where a coalitional game is formulated. However, as this process is an exhaustive search that explores all possible solutions it is intractable for high-density scenarios. For this reason, a distributed heuristic approach was proposed based on a merge and split algorithm adapted from image processing theory. The algorithm consists of two actions: coalitions are merged if the resulting coalition has greater utility. Similarly, coalitions are divided if the sum of the utility of each resulting part is greater than the utility of the joint coalition.

The work in [46] proposes a multi-objective optimization strategy to maximize throughput and minimize power consumption. It uses the Pascoletti and Serafini methods to cluster RRHs from various locations and allocate them to BBU pools. This strategy outperforms the traditional greedy approach. Given the NP-completeness of the problem, [47] poses it as a bin-packing approach and proposes a heuristic algorithm. Similarly, [48] uses the well-known metaheuristic Tabu Search. These three techniques are reviewed in [38] and studied comparatively, showing similar performance. However, results are presented in a synthetic scenario with 15 RRHs randomly distributed. The maximum amount of RRHs that a BBU pool could handle is used as BBU pool capacity.

A multi-objective optimization problem for RRH clustering that minimizes the network transmission delay and power consumption is introduced in [49] by organizing RRHs in disjoint clusters to reduce the number of active

BBUs. Weighted-sum and ϵ–constraint methods are used to solve the problem. The considered network topology is seven hexagonal cells with a radius of 500 m. The paper in [40] addresses the problem of maximizing the total throughput of the network via joint user association and power allocation in C-RAN, accounting for QoS requirements. The validation scenario is one MBS, 30 RRHs, and 80 UEs in a square area of 500×500 m^2.

The works mentioned above propose promising resource management strategies. However, the computational capacity available at BBU pools and adaptive capacity to avoid under-provisioned and over-provisioned networks have not been considered. Moreover, for simplicity, results are evaluated using synthetic scenarios, which do not consider the complexity of mobile networks.

2.7 ML for traffic forecasting

Machine and deep learning techniques have been widely used in several research fields and wireless communications to optimize traffic classification and load management [34, 50, 51].

In [50], a multitask learning architecture using deep learning is presented. This work aims to analyze the accuracy of deep learning architectures in mobile traffic forecasting. The authors employ a dataset of Telecom Italia to predict minimum, average, and maximum traffic loads. Different deep learning models are tested, such as RNN, 3D-Convolutional Neural Network (CNN), and a combination of RNN and CNN. Results show that RNN-CNN can extract geographical and temporal traffic features.

The research [34] aims to apply data analysis techniques to support network operators to maximize resource usage during the planning stages. The authors investigate the prediction accuracy using artificial Neural Networks (NNs), especially a Multilayer Perceptron (MLP), and SVM. Results show that SVM outperforms acMLP prediction capabilities.

Although the works mentioned above lie in the analysis of the performance of ML strategies in traffic forecasting tasks, they do not apply forecasting to optimize the network resources. For this reason, how to use them in optimizing resource management algorithms in 5G and C-RAN environments remains an issue.

Mo et al. [51] propose a deep learning algorithm based on LSTM cells that predicts network resource requirements at the optical switch where each BBU pool is connected (e.g., Reconfigurable Optical Add/Drop Multiplexer (ROADM)). The authors consider a region of New York City where 9 ROADM nodes cover 400 km^2, and each ROADM routes 64 RRHs. Unlike [50] and [34], this work employs forecasting to optimize the network. It predicts an increase in the demand 30 minutes in advance to reallocate the additional traffic to another BBU pool. However, the instantiated resources at BBU pools remain underutilized in low-demand situations. Moreover, QoS with delay restrictions could be affected due to the reallocation to a farther BBU pool.

2.8 RAN deployment optimization

Multiple research works have focused on defining strategies to deploy and optimize 5G and B5G networks [52–58].

A MILP algorithm is proposed in [52] to minimize the network deployment cost and latency of a C-RAN with Mobile Edge Computing (MEC) nodes. Moreover, they present a heuristic algorithm because of the complexity of the MILP approach. The main goal of this paper is to optimize the MEC node placement and the C-RAN deployment. Although the proposed strategies are novel and could be of interest because they implement a joint optimization considering MEC and C-RAN, the analysis is limited to the transport network and the placement of BBU pools and MEC nodes, without considering the RRH deployment and the mobile network demand plane.

In [53], the authors propose an energy-effective radio network deployment where the system could select a subset of RRHs according to the traffic demand simulated using Traffic Demand Points (TDPs), which concentrate the data rate of a specific zone to satisfy the QoS requirements of the potential UEs. However, the problem is divided into two sub-optimal problems: RRH-TDP association and RRH selection, which could reduce the possibility of finding the optimal solution for the network deployment. On the other hand, the authors consider two synthetic scenarios to validate the results. The first scenario depicts a dense square region of $250\,\mathrm{m} \times 250\,\mathrm{m}$ with two micro-RRHs and seven pico-RRHs, while the second represents an area of $500\,\mathrm{m} \times 500\,\mathrm{m}$ with three micro-RRHs and 13 pico-RRHs.

Besides, the authors in [54] recently proposed a hybrid fronthaul solution

based on fibers and Free-Space Optics (FSO) to minimize the deployment costs in dense urban scenarios. They formulated and compared two Integer Linear Programmings (ILPs): joint and disjoint approaches. The disjoint method splits the problem into two sub-problems: the RRH placement and the fronthaul deployment. At the same time, the joint strategy only solves one optimization problem to deploy the whole C-RAN. The authors conclude that the joint approach is better than the disjoint strategy regarding deployment cost. Although they propose a fascinating solution to reduce the deployment cost of a C-RAN with hybrid fronthaul, there is room for improvement by introducing realistic UEs to model the traffic demand, validating the results under realistic RRH possible locations.

On the other hand, the authors in [55] propose a MOO problem for small cell planning, which considers fiber and wireless backhaul technologies and two types of BSs. The MOO aims to determine the optimum type and location of the deployed BSs. The authors propose a joint cell and fiber backhaul planning algorithm employing heuristic techniques. This work is also of interest because it focuses on the last standard of Passive Optical Networks (PONs), called Next-Generation Passive Optical Network 2 (NG-PON2), for the fiber deployment of the backhaul; however, it does not consider a C-RAN environment.

In [56], the authors design a joint optimization framework considering the costs of the mobile network and its fronthaul in a C-RAN ecosystem. Deployment cost is analyzed under different scenarios; they also extend the work to consider three optical fronthaul technologies: Common Public Radio Interface (CPRI), Physical Layer Split (PLS), and Analogue Radio-over-

Fiber (ARoF). Although the authors provide a detailed fronthaul analysis, traffic profiles are simplified to reduce model complexity. Hotpots are considered to generate traffic without accounting for different services and UEs. On the other hand, the radio network deployment is simplified by introducing a fixed coverage radio per RRH instead of modeling the Signal-to-Noise-plus-Interference-Ratio (SINR) using a suitable propagation channel model.

Additionally, the authors in [57] propose an optimization problem that minimizes the number of RRHs in a C-RAN context. Following the same approach as [53], they use the concept of TDPs to simulate the traffic demand, where the TDPs are allocated at the center of the demand zones. The algorithm starts with all the possible RRHs and connects each TDP to the nearest RRH. Then, the proposed algorithm turns off some of the RRHs at each iteration until the percentage of unsatisfied TDPs exceeds 0.1 %. However, for the sake of simplicity, the traffic demand of each TDP is established without UE and service modeling, considering only a capacity constraint, and the RRH-TDP association is based on a minimum distance approach.

In [58], the authors propose a framework to improve resource efficiency at the BS level. They employ a joint optimization problem to efficiently allocate the resources of the network slices, the cell-slice association, and the UE-BS connections. They include SINR requirements and different slice services in the network optimization problem. This work demonstrates that realistic scenarios with UEs and services can be modeled.

To the best of our knowledge, no published papers include cell cooperation in radio network deployment algorithms, saving energy, and reducing costs at different frequency ranges and split options in a realistic scenario.

The works mentioned in this section propose promising radio network deployment strategies. However, there is room for improvement because they skip traffic generation by considering only demanding points without accounting for services and UEs. Moreover, the RRH coverage and the RRH-TDP association are simplified, which results in synthetic scenarios, usually with a small number of cells, that do not reflect the complexity of mobile networks.

2.9 Challenge and open issues

As C-RAN has been standardized as part of the NG-RAN, it must address the radical evolution in flexibility, security, and performance to support uRLLC, eMBB, and Massive Machine Type Communications (mMTC) services. Multiple parameters must be enhanced, such as latency, throughput, resource allocation, handover, energy efficiency, power consumption, and cost-saving. Optimizing those parameters demands much effort from the research community and the combination of some of the most promising technologies, such as Software-Defined Networking (SDN) and Network Function Virtualization (NFV). This complexity is a big challenge. In this section, a description of open challenges and issues the research community is facing are summarized.

High fronthaul capacities needed

Fronthaul links between BBUs and RRHs must have high bandwidth capability with low delay and cost requirements. The fully centralized architecture demands the highest fronthaul bandwidths because the signal is completely

processed at the BBU pool resulting in considerable overhead. Functionality split options have been defined by [59] to reduce the fronthaul bandwidth requirements. However, the potential of C-RAN, depending on the number of centralized functionalities, is reduced; [60] carries out a detailed analysis of this situation.

RRH clustering (BBU-RRH mapping)

Designing real-time RRH clustering is a real challenge. BBU-RRH mapping methods with efficient BBU coordination algorithms with minimal overhead are highly complex problems. They should optimize multiple parameters, such as load balancing, multiplexing gain, inter-cell interference, throughput using Coordinated Multipoint (CoMP), handover frequency, energy efficiency, or power consumption. Many authors are dedicating efforts to overcoming this challenge [[38, 39, 42, 61–63].

Security and management of network slicing

Another significant challenge in C-RAN is security in terms of user privacy and isolation between slices. As resources are shared between BBUs, breaking user privacy and accessing secured data is possible. In addition, as C-RAN has to support services of different Mobile Virtual Network Operators (MVNOs) using network slicing over the same infrastructure, robust isolation among slices is a significant challenge. Hence, providing reliable, cost-effective, and QoS-guaranteed network slices under C-RAN architecture is a challenge in 5G. Works [64–66] aim to overcome these issues.

Resource management

The required density of RRHs to provide high data rates incurs high computational complexity due to the huge amounts of data related to signal processing, resource allocation, and RRHs/BBUs coordination. This complexity is a big challenge facing the establishment of scalable networks. Resource allocation strategies, which determine the allocation of centralized computation resources, fronthaul capacity, radio spectrum, and power allocation is still a challenge. Some of the works that are related to this challenge have been presented in [25, 26, 40, 44, 67].

One of the fundamental challenges is how to assign isolated resources efficiently to the different virtual operators. Resources allocation can be based on multiple criteria, e.g., bandwidth, data rate, power, interference, pre-defined contracts, channel conditions, traffic load, or a combination of these parameters. Coordination and communication protocols have to be well designed to be used between the Infrastructure Provider (InP) and the MVNOs [68].

Resource management often has to solve an optimization problem based on a set of constraints, the exhaustive search strategy is not suitable because it demands high computational complexity in high-density scenarios. For this reason, resource allocation using optimization techniques such as game theory, graph theory, matching theory, and heuristic techniques to minimize the high computational complexity of solving the combinatorial optimization problems is one of the current aims of the research community [16, 44, 69].

Introduction of adaptive machine learning techniques to achieve a proactive network capable to adapt to data demands (e.g., IoT demands) that

fluctuate over time and places, while optimizing the available resources is an important challenge [25, 32, 37, 50, 70]. Due to that, InPs rent the maximum peak of capacity demanded by each service provider or mobile network operator regardless of the instantaneous capacity that really is needed.

Energy efficiency, power consumption and cost saving

Increasing the energy efficiency of mobile communication systems while the cost is reduced has been an important research field in recent years. The integration of different technologies (e.g. SDN, NFV, MEC) to build the 5G networks creates a new challenge: How to manage the high flexibility and capacity demanded by the network while energy efficiency, power consumption, and cost are enhanced. Many authors have proposed Green C-RAN deployments to address this challenge [4, 24–27, 71]. For instance, MEC and caching, energy-efficient designs, multi-dimensional resource management, and physical layer security have been identified as major challenges.

2.10 Conclusions

This chapter has presented a rigorous study of the related works. It summarizes the advantages and disadvantages of the analyzed works. After the analysis of numerous promising works, it is possible to conclude that there is room for improvement in multiple research directions, such as energy and cost footprint reduction of the mobile networks, fronthaul design, and optimization, optimization deployment of each plane of the network (for instance BBU and RRH placement), and the design of strategies to efficiently manage

the centralized computational resources. Additionally, It has been identified that most of the analyzed works employ synthetic scenarios that could not represent the complexity of the mobile networks. These simplifications may reduce the significance of the results.

Chapter 3

Machine Learning: Brief Overview

3.1 Introduction

This chapter presents a summary of machine learning techniques. In section 3.2, a description of the main categories of ML approaches has been introduced. Unsupervised and supervised learning, reinforcement learning, deep learning, and federated learning are briefly described. Additionally, sections 3.3, 3.4, and 3.5 describe in detail the machine learning techniques that have been considered in the thesis. However, multiple approaches have been tested and compared, obtaining better results with SVM, TDNN, and LSTM.

3.2 Machine Learning Categories

The enormous increase in the mobile traffic demand combined with the complexity of the current and future heterogeneous mobile networks produce the necessity of efficient network management and orchestration strategies. For this reason, ML and AI are being widely considered to introduce cognitive capabilities to the B5G networks, as discussed in the literature review.

This section presents a general overview of ML techniques, as well as a detailed mathematical analysis of the models that have been employed in this thesis. Fig. 3.1 shows the multiple categories of machine learning techniques.

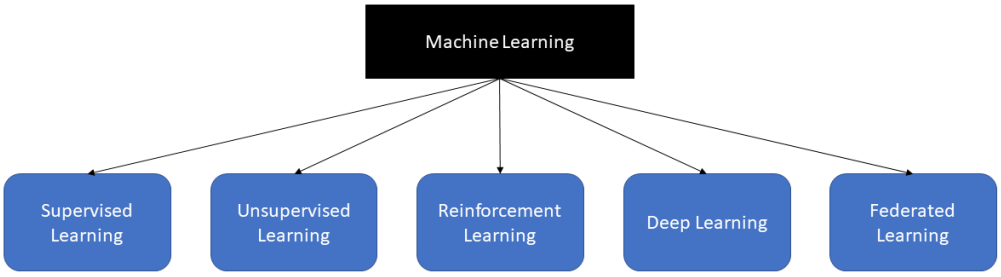


Figure 3.1: Categories of machine learning techniques

Supervised learning

Supervised learning is a type of machine learning approach that utilizes a labeled database for prediction or classification. Namely, these algorithms take as input training data to learn specific features. This procedure is called the training process. Once the algorithms finish the training process, they are ready to test their performance in a testing database which commonly

is not part of the training procedure. Fig. 3.2 shows a block diagram that represents the structure of a supervised learning algorithm.

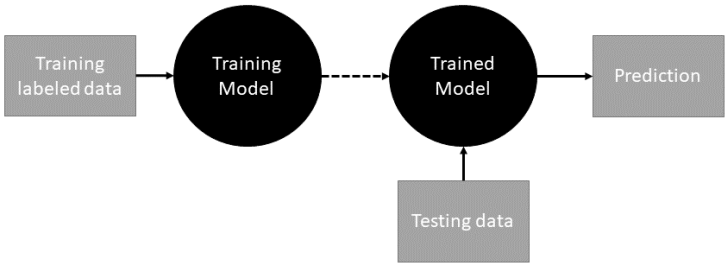


Figure 3.2: Block diagram of supervised learning

In general, algorithms in this category include classification and regression analysis, which have been considered in multiple proposals for mobile network management and characterization of data traffic profiles [72]. Some of the most employed algorithms or strategies could be SVM, linear regression, logistic regression, naive Bayes, decision trees, and neural networks, among others.

Unsupervised learning

The fundamental difference between unsupervised learning algorithms and supervised approaches is that they do not need a labeled database. This characteristic is helpful when unlabeled databases are considered in clustering or classification groups. Especially this branch of machine learning algorithms is commonly used to detect hidden patterns in data. Some of the most popular unsupervised algorithms are Principle Component Analysis, K-means clustering, and KNN (k-nearest neighbors), among others.

Reinforcement Learning

Another branch of machine learning systems is reinforcement learning. This strategy is a change of paradigm regarding the previously discussed techniques because the system learns iteratively. It is not based on a dataset. A reinforcement learning algorithm contains an agent and an environment, as shown in Fig. 3.3.

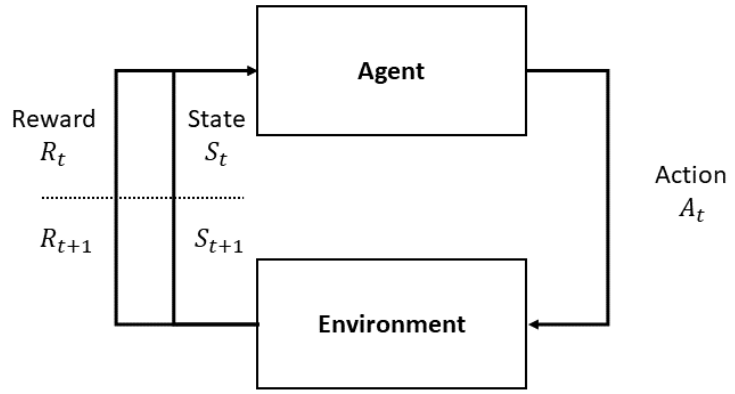


Figure 3.3: Reinforcement learning block diagram

In general terms, the system could be in a set of possible states (S_t). The agent executes an action (A_t) at each system state and observes the response of the environment (S_t) and also an associated reward R_t. This reward is fundamental and should be carefully designed to represent the effect of the action on the environment. It should reward the desired results while introducing a penalty for negative actions. The agent iterative searches for possible state-action pairs (policies) to make the decisions. After the iterative process, the algorithm converges for decision-making with higher long-term rewards. Reinforcement learning helps to solve multiple resource

management issues in mobile communication systems [73].

Deep Learning

Fig. 3.4 shows an example of a deep learning architecture. In general terms, the architecture of a deep learning model contains an input layer, hidden layers, and an output layer. The system is trained, and the knowledge is saved in the hidden layers. However, the deep learning theory is not limited to this architecture. Multiple strategies such as CNNs, 3D-CNNs, RNN, and LSTM have been considered as has been mentioned in chapter 2.

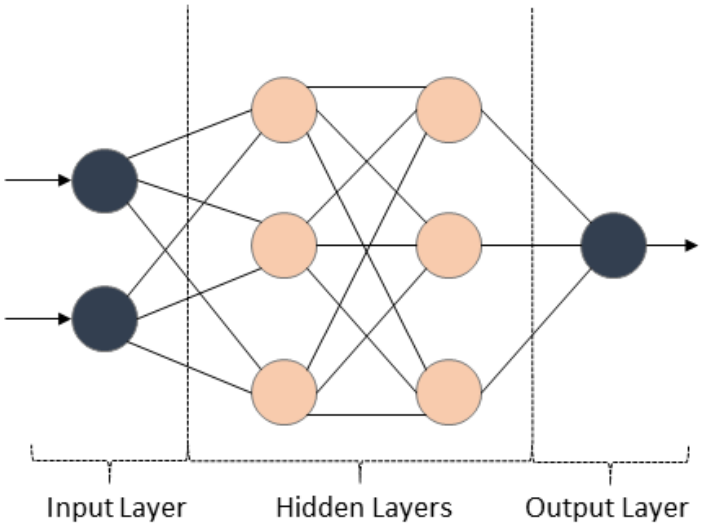


Figure 3.4: Example of Deep Learning Architecture

Federated Learning

Federated Learning (FL) is a modern machine learning approach that has motivated researchers to study multiple novel applications. Unlike traditional

machine learning approaches, FL splits the training process into multiple local models. Consequently, the models are trained using the decentralized technique, which is a fundamental advantage in mobile networks because of the strict regulations about data privacy; it is not practical to concentrate the customer data in a centralized location [74]. Fig. 3.5 shows a federated learning model diagram.

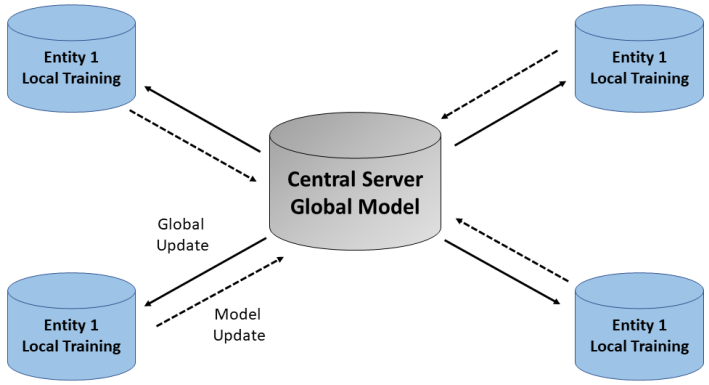


Figure 3.5: Federated learning block diagram.

3.3 Support Vector Machine

SVM theory was first proposed in [75]; since then, it has been widely used in classification and regression tasks of different scientific and engineering fields. The original idea focuses on element classification. Let us assume a simple case to illustrate how it works. Fig. 3.6 shows a set of training samples that belong to two classes (circles and squares).

The aim is to find the best hyperplane (dotted line in Fig. 3.6) for

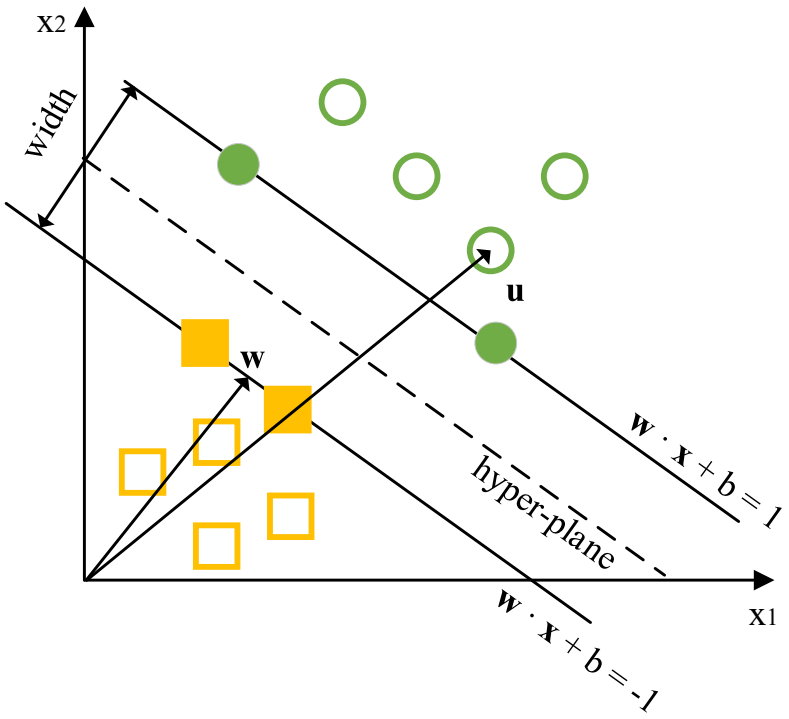


Figure 3.6: Basic classification example of SVM.

classifying the data. The algorithm uses optimization theory to maximize the width of the street. Let us assume that $\boldsymbol{\omega}$ is a vector perpendicular to the hyperplane, and $\boldsymbol{u}$ is a vector that points to an unknown observation. The decision rule used to decide if $\boldsymbol{u}$ belongs to the circle class is presented in (3.1)

$$\boldsymbol{u} \cdot \boldsymbol{\omega} + b \geq 0 \tag{3.1}$$

It means that if the projection of $\boldsymbol{u}$ onto the perpendicular line of the hyperplane is greater than the distance from the origin to the hyperplane, then the sample is on the right side (circle), where $b \in \mathbb{R}$ and $\cdot$ is the scalar product. However, as the idea is to find the line that maximizes the width of the street, let's take the square ($\boldsymbol{x}^{\mathbf{s}} \in \mathcal{S}$) and circle examples ($\boldsymbol{x}^{\mathbf{c}} \in \mathcal{O}$)

into (3.2) and (3.3), respectively, which guarantee that all the samples on the dataset are out of the street. $\mathcal{S}$ and $\mathcal{O}$ represent the set of squares and circles, respectively.

$$\boldsymbol{\omega} \cdot \boldsymbol{x}^{\mathbf{s}} + b \leq -1 \tag{3.2}$$

$$\boldsymbol{\omega} \cdot \boldsymbol{x}^{\mathbf{c}} + b \geq 1 \tag{3.3}$$

Equations (3.2) and (3.3) are joint into (3.4), introducing the variable y_i

$$y_i(\boldsymbol{\omega} \cdot \boldsymbol{x_i} + b) \geq 1 \tag{3.4a}$$

$$y_i = \begin{cases} +1, & \boldsymbol{x_i} \in \mathcal{O} \\ -1, & \boldsymbol{x_i} \in \mathcal{S}, \end{cases} \tag{3.4b}$$

where $\boldsymbol{x_i}$ represents the vector of the i^{th} training sample. It is possible to compute the width of the street ($\mathbb{W}$) by taking one example per class over each boundary due to they hold the equality condition in (3.4a); the result is shown in (3.5)

$$\mathbb{W} = (\boldsymbol{x}^{\mathbf{c}} - \boldsymbol{x}^{\mathbf{s}}) \cdot \frac{\boldsymbol{\omega}}{|\boldsymbol{\omega}|} = \frac{2}{|\boldsymbol{\omega}|} \tag{3.5}$$

where $|\cdot|$ denotes the Euclidean norm. SVM aims to maximize (3.5) subject to (3.4) to obtain the best hyperplane for classifying data. After solving this optimization problem using a Lagrangian function, it is possible to realize that the solution depends only on the samples, and vector $\boldsymbol{\omega}$ is a linear combination of those samples [75].

The work [76] extended this strategy to address regression tasks. In this case, the idea is to find a linear function $f(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{\omega} + b$ that fits the training data. The optimization problem is formulated to minimize the difference (error) between the predicted value extracted from $f(\boldsymbol{x})$ and the

observation of the regression. The mathematical process is detailed in [76]. Equation (3.6) shows the fitting function

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_\mathrm{t}} \alpha_i \boldsymbol{x_i} \cdot \boldsymbol{x} + b \tag{3.6}$$

Where a_i and b are real values obtained after the training process where the optimization problem is solved, N_t is the number of samples in the training dataset.

The previous analysis of SVM strategies assumes that it is possible to classify or predict data based on a linear hyperplane or a linear fitting function. However, in many applications, linear approaches can not process the data. In those cases, finding a linear function that describes the data is not suitable. A transformation (Φ) over the data plane to solve this problem is applied; this method is called the kernel trick. After the conversion, it is possible to use a linear approach in a higher-order space to fit or classify data. The fitting function after applying the kernel is shown in (3.7).

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_\mathrm{t}} \alpha_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b \tag{3.7}$$

where $K(\boldsymbol{x_i}, \boldsymbol{x}) = \Phi(\boldsymbol{x_i})\Phi(\boldsymbol{x})$ depicts the kernel function.

3.4 Time-Delay Neural Network

Artificial NNs have been widely used during the last years to solve different machine-learning problems, even regression and time series forecasting tasks. TDNN is a combination of typical NN architecture and an input layer that reshapes sequence time series data into parallel (shift register), employing

a set of delays (N) to use the previous time steps as features of the NN. The learning process takes place in the hidden layers of the neural network. Fig. 3.7 shows the TDNNs structure, and the basic block diagram of a neural network entity (also called a neuron). Equation (3.8) shows the behavior of a single neuron. The inputs are multiplied by the weights ($\boldsymbol{W}$), and a bias (b) is added before applying the activation function (f_{a}) to compute the output. The knowledge is in the weights and biases of the neurons in hidden layers.

$$N_{\mathrm{o}} = f_{\mathrm{a}}(\boldsymbol{W} \cdot \boldsymbol{X} + b) \tag{3.8}$$

where X is the input vector, and N_{o} is the output.

3.5 Long Short-Term Memory

Traditional NNs have outstanding prediction performance based on the status of input variables. However, they are not able to remember sequential data. RNNs try to address this issue using a feedback loop to create a hidden state where the information of previous time steps is stored. RNNs predict the subsequent output based on the current input and the hidden state. Fig 3.8(a) shows a basic structure of a recurrent neural network unit.

The hidden state of the RNN is upgraded recursively, using the same approach of a neural network (see (3.8)) but considering the previous hidden state (h_{t-1}) as another input. Equation (3.9) shows the process to upgrade the hidden state (h_t)

$$h_t = f_{\mathrm{a}}(\boldsymbol{W} \cdot [\boldsymbol{h_{t-1}}, \boldsymbol{X_t}] + b) \tag{3.9}$$

where $\boldsymbol{W}$ is the weight vector, f_{a} the activation function, and $[\boldsymbol{h_{t-1}}, \boldsymbol{X_t}]$

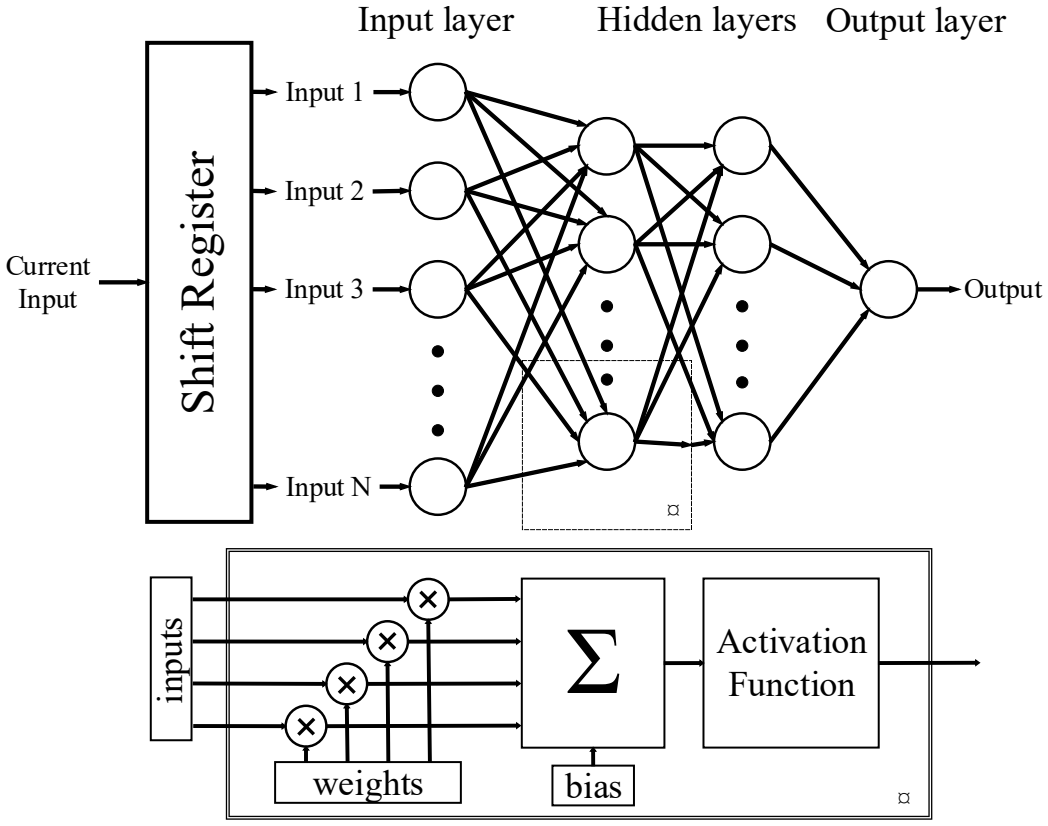
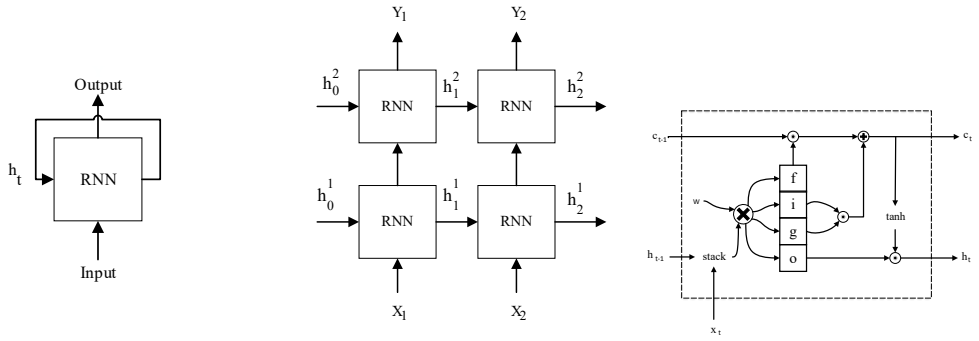


Figure 3.7: General scheme of a time-delay neural network for time series forecasting with N previous time instants.

denotes the concatenation or stack operation between the previous hidden state and the current input, respectively. The scheme of an RNN shown in Fig 3.8(a) could be unrolled to create deeper designs, such as the multilayer RNN in Fig. 3.8(b), where the hidden states of the first layer are inputs of the second layer.

Those architectures face the vanishing gradient problem that was solved by [77], defining a different kind of RNN called LSTM. Moreover, LSTM improves long-term predictions.



(a) basic scheme of an RNN. (b) multilayer architecture. (c) diagram of an LSTM entity.

Figure 3.8: General deep learning architecture with LSTM cells

The structure of an LSTM entity is shown in Fig 3.8(c). The critical aspect of an LSTM unit is the cell state denoted by c_t. LSTM units could remove and aggregate information to the cell state. Those processes are regulated by gates that combine a neural network and a pointwise multiplication; it controls the amounts of information at the output of the gate. The output of the neural network of each cell is often obtained using a sigmoid activation function, which allows quantifying the portion of the information that could pass through the gate with a coefficient from zero to one. As the output is a pointwise product, zero indicates no signal to the output, and one represents that the whole signal remains in the output.

First, the forget (f) gate decides what information to remove from the cell state. Consequently, the input (i) and gate (g) gates decide what information aggregate to the cell state. Finally, the output gate (o) decides what information goes to the output. The whole process of the LSTM is

summarized in (3.10),

$$f_t = \sigma(\boldsymbol{W_f} \cdot [\boldsymbol{h_{t-1}}, \boldsymbol{X_t}] + b_{\mathrm{f}}) \tag{3.10a}$$

$$i_t = \sigma(\boldsymbol{W_i} \cdot [\boldsymbol{h_{t-1}}, \boldsymbol{X_t}] + b_{\mathrm{i}}) \tag{3.10b}$$

$$o_t = \sigma(\boldsymbol{W_o} \cdot [\boldsymbol{h_{t-1}}, \boldsymbol{X_t}] + b_{\mathrm{o}}) \tag{3.10c}$$

$$g_t = \tanh\left(\boldsymbol{W_g} \cdot [\boldsymbol{h_{t-1}}, \boldsymbol{X_t}] + b_{\mathrm{g}}\right) \tag{3.10d}$$

$$c_t = f_t * c_{t-1} + i_t * g_t \tag{3.10e}$$

$$h_t = o_t * \tanh c_t, \tag{3.10f}$$

where $\boldsymbol{W_k}$ and b_{k} are the weights and the bias of the neural network in gate k, respectively. The activation functions of the gates are σ or tanh, which represent the sigmoid and hyperbolic tangent functions, respectively; $*$ operation denotes the pointwise product.

3.6 Conclusions

A resume of the machine learning categories has been introduced. The approaches mentioned in this chapter demonstrate that a wide range of options exists to optimize future network performance. Especially the strategies that have been widely employed in mobile network deployment and optimization. The evolution of these techniques and the future network requirements will open the door to multiple research works.

Chapter 4

Simulation Platform

4.1 Introduction

As mentioned, most of the analyzed works use synthetic scenarios to validate their contributions. However, these scenarios are not suitable to represent the enormous complexity of mobile networks. For this reason, a big effort is done to define a realistic C-RAN platform to test the different optimization algorithms proposed in the thesis.

An initial version of the mobile network deployment over Vienna city is defined by [78]. It contains the site locations, the parameters assigned to the BSs, and the propagation model. This version considers a Fourth-Generation (4G) heterogeneous radio access network deployment with a traditional RAN architecture. The research community has widely used it to validate multiple optimization algorithms [79, 80].

An upgraded version of this scenario, oriented to implement C-RAN in a flexible simulation platform, is introduced using Matlab. Section 4.2 de-

scribes the features of the platform, such as UEs, best effort and GBR services, and the resource demand estimation strategy.

This platform has been used to test BBU-RRH association algorithms and different resource management strategies to allocate the available resources at BBU pools or CUs. Additionally, an optimization of the RRH deployment is introduced to reduce not only the CAPEX and OPEX but also the energy footprint of B5G networks.

4.2 Simulation platform description

The map covers an area of 455 km^2 with a perimeter of 86 km. The blue and green points over the map in Fig. 4.1 represent the MRRHs and possible BBU pools, respectively, while the red points depict the SRRHs.

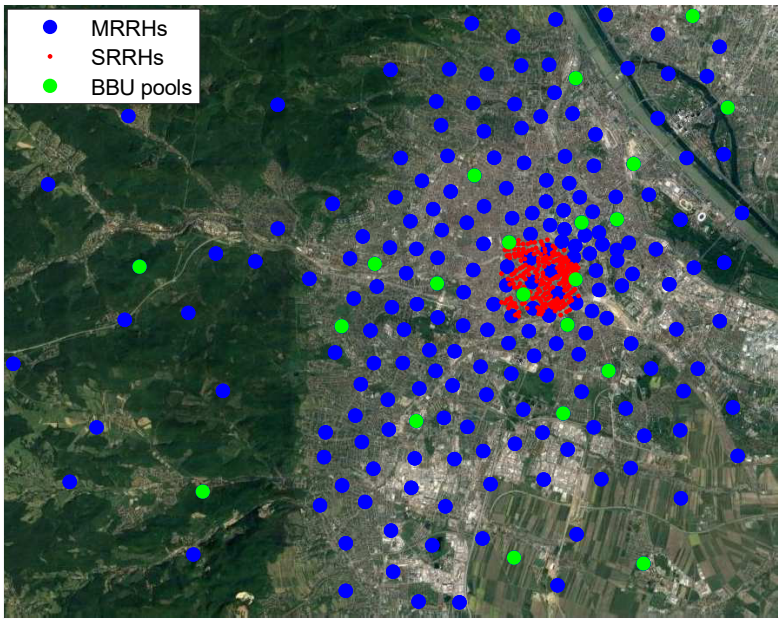


Figure 4.1: Full Vienna city map

The MBSs in the initial version of the scenario were sectorized into 3 or 2 cells. Each sector of the MBSs is considered a MRRH in the proposed structure. As a result, the design includes 628 sites, distributed in 233 MRRHs and 221 SRRH, representing a total of 849 RRHs and 21 BBU pools (see Table 4.1).

It is essential to mention that when fully centralized C-RAN deployment (split option 8) is considered, MRRHs and SRRHs contain the same functionalities. However, the platform has the feature of representing different split options. For this reason, the notation macro/small RRH is employed, but in the case of C-RAN option eight, the notations RRH or RU are suitable. However, this thesis keeps the macro/small notation because it allows the establishment of different parameters for each. For instance, frequency bands, antenna gain, and transmitted power.

Table 4.1: Distribution of the whole deployment

Parameters	**Value**
Dimension	455 km^2
Sites	444
MRRHs (sites)	628(233)
SRRHs (sites)	221(221)
BBU pools	21
RRHs	849

As can be seen in Fig. 4.1, the scenario considers the whole city of Vienna, where there are rural and urban zones, which allows the analysis of a wide range of data traffic intensities. C-RAN efficiently manages ultra-

dense deployments, such as the metropolitan area that is shown in Fig. 4.2.

Figure 4.2: Metropolitan area of Vienna city.

This city zone represents a heterogeneous deployment, being useful to test the performance of the proposed algorithms and the C-RAN platform itself because it depicts the densest traffic region where the capacity of the network may not be sufficient to satisfy the demand.

The general network distribution of this zone is summarized in table 4.2. The location of BBU pools intentionally matches with MRRH coordinates because of the availability of infrastructure and resources at these locations.

Table 4.2: Deployment of the Metropolitan Area.

Parameters	**Value**
Area (km^2)	25
Sites	228
MRRHs (sites)	51(17)
SRRHs (sites)	221(211)
BBU pools	3
RRHs	272

4.2.1 Traffic Generation

As has been mentioned, mobile networks face dynamic environments with data traffic load fluctuations according to the type of zone and the hour of the day. For this reason, different strategies to generate the traffic profile are utilized.

The simplest strategy considers three types of cells (office, residential and mixed). In this case, the traffic profiles are modeled by multiple Gaussian functions; controlling the mean and deviation is possible to adapt the data traffic demand to realistic profiles. An example of this traffic generation strategy is presented in Fig. 4.3.

On the other hand, the metropolitan area presented in Fig. 4.2 allows for a detailed analysis of C-RAN based on the instantaneous computational capacity required to manage the traffic of the RRHs. Hence, instead of using the data traffic per hour profile, realistic UEs and services and consequently traffic models per service have been introduced. This strategy is fundamental

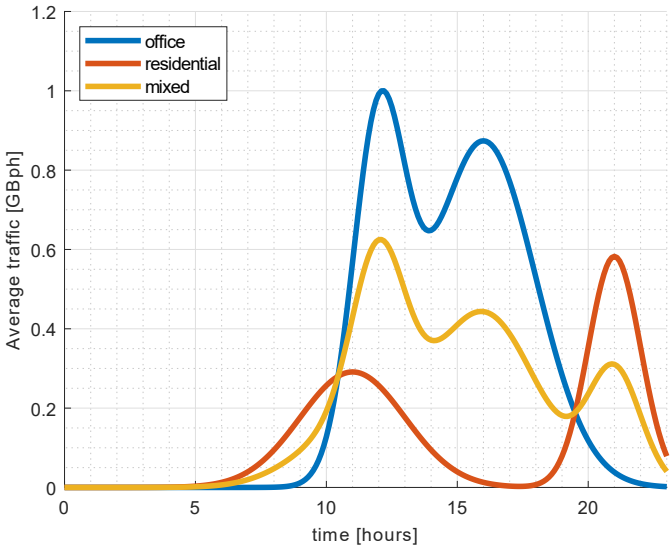


Figure 4.3: Traffic profile for office, residential and mixed cells.

to account for QoS constraints, service priorities and efficiently manage the resources at the BBU pools.

Conversational, streaming, and interactive services have been modeled based on the packet-level models defined in [81], [24] and [4] and summarized in Table 4.3.

Parameters such as packet size and interval of arrival have been extracted from [24] and [82]. Each UE generates only one service simultaneously or remains in idle status with a probability P_{idle}, which is used to control the traffic intensity and to emulate the time interval where the simulation is running. The values of P_{idle} are selected for each time interval and RRH type (office, mixed, and home) following the traffic distribution shown in Fig. 4.3.

Furthermore, the weight (w) is used by the scheduler to guarantee the

Table 4.3: Service modeling parameters

Services	w	**Size**	**Time interval**	**Duration (s)**	**Traffic mix (%)**
VoIP	83	Packet: 40 B	20 ms	Exp(120)	25
Video	59	Packet: [20-250] B	100 ms	Exp(300)	25
Web	36	Page: mean = 315 kB	Exp(30)	Exp(400)	30
FTP	36	File: mean = 2MB	Exp(180)	–	20

QoS. The scheduler assigns higher priority to VoIP and Video due to their delay constraints. Finally, the traffic mix parameter describes the percentage of active sessions per service and RRH.

The session duration follows an exponential distribution, except for FTP services, where total duration depends on the size of the packet to be transmitted and the UE throughput [81]. The time interval between consecutive packets is fixed at 20 ms and 100 ms for VoIP and video streaming services, respectively, and follows an exponential distribution for non-real-time services.

Initially, the UEs have been placed at random positions but realistic UEs coordinates can be uploaded. The UEs are served by the RRH that provides the highest SINR in the downlink. The high interference that the MRRH could cause to the UEs connected to a SRRH are avoided assuming that

MRRHs and SRRH operate in different frequency bands.

Furthermore, the random processes in the service generation produce enough traffic fluctuations to represent the variability of the required computational capacity at BBU pools. For this reason and because beamforming and handover are not the focus of the analysis, the UEs remain static during the simulation time to keep the simplicity of the system model. Then, the SINR per UE remains constant and is calculated using (4.1):

$$\begin{aligned} \text{SINR [dB]} = P_{\text{RRH}}[\text{dBm}] + G_{\text{RRH}}[\text{dB}] + G_{\text{UE}}[\text{dB}] \\ - L[\text{dB}] - 10\log(N[\text{mW}] + I[\text{mW}]), \end{aligned} \tag{4.1}$$

where P_{RRH} is the power transmitted by the RRH, G_{RRH}, and G_{UE} are the RRH and UE antenna gains respectively, L is the path-loss from the RRH to the UE, and N and I are the UE thermal noise, and the interference received power respectively.

To calculate the required computational resources to manage the traffic of the RRHs, the Modulation and Coding Scheme (MCS), as well as the number of needed Physical Resource Blocks (PRBs) to transmit a packet, should be known. The mapping between MCS and SINR is summarized in Table 4.4 and has been obtained using [83], which presents a link-abstraction model based on mutual information at the modulation symbol level. The number of PRBs required to transmit a packet is extracted from [84].

Table 4.4: Mapping between SINR and MCS.

SINR [dB]	**Modulation order (M)**	**code rate (ρ)**
< -5	QPSK (2)	0.076
$[-5, 1]$	QPSK (2)	0.3
$[1, 3.1]$	QPSK (2)	0.44
$[3.1, 6.1]$	QPSK (2)	0.59
$[6.1, 9]$	16QAM (4)	0.48
$[9, 13]$	16QAM (4)	0.6
$[13, 16]$	64QAM (6)	0.65
> 16	64QAM (6)	0.85

4.2.2 Resource Demand Estimation

The introduction of UEs and realistic services generation at the packet level allows estimating the network traffic load in terms of the Required Computational Capacity (RCC) at BBU pools.

The RCC is defined as the minimum amount of computational operations necessary to implement physical layer functions at the BBU pool, such as channel coding, modulation, MIMO precoding, and Orthogonal Frequency-Division Multiplexing (OFDM) symbol mapping. The RCC is calculated based on the strategy proposed by [85] and modified by [81] to introduce parallel processing. The strategy uses a Long-Term Evolution (LTE) reference scenario, where the RCC and a set of scaling factors that describe how the RCC evolves to other scenarios are tabulated. Those scaling factors depend on the network parameters and the physical function to be imple-

mented. Equation (4.2) describes this method.

$$
\begin{aligned}
C &= \sum_{i \epsilon \mathcal{I}} C_i^{\text{ref}} \prod_{x \epsilon \mathcal{X}} \left(\frac{x_{\text{act}}}{x_{\text{ref}}} \right)^{s_{i,x}} \\
\mathcal{X} &= \{B_{\text{w}}, N_{\text{a}}, Q, M, \rho, N_{\text{s}}\},
\end{aligned}
\tag{4.2}
$$

where C represents the RCC of the desired scenario, C_i^{ref} is the processing capacity needed to address the function i in the reference scenario in Giga operations per second (GOPS). Subscripts act and ref depict actual scenario and reference scenario respectively, $s_{i,x}$ is the scaling factor of the function i and parameter $x \in \mathcal{X}$. The set $\mathcal{X}$ contains the operating bandwidth (B_{w}), the number of antennas (N_{a}), the quantization resolution (Q), the modulation order (M), the code rate (ρ) and the number of streams ($N_{\text{s}} \leq N_{\text{a}}$). Finally, set $\mathcal{I}$ contains the PHY functionalities that have been shown in Table 4.5 and Fig. 4.4.

Fig. 4.4 shows the protocol stack and the split options, which were considered to face the extreme traffic demand of the fronthaul links. As the resources are centralized at BBU pool entities and the functionalities are virtualized, it is possible to split those functions into two groups. The functions that may be implemented by user sessions, processed independently and in parallel are called user-processing functions (UFs), such as channel coding and modulation. The functions that are common to all users in the same carrier component/cell and could not be split by user sessions, such as OFDM modulation, are denoted as common-processing functions (CFs). Table 4.5 summarizes the reference computational capacity, as well as the scaling factors of the considered PHY functions.

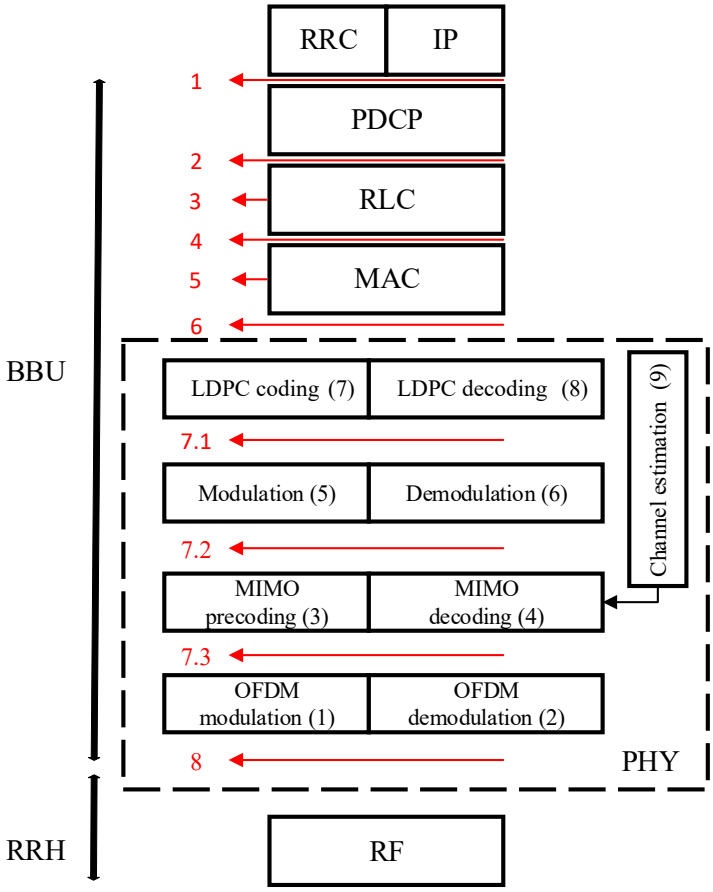


Figure 4.4: 3GPP Protocol stack split options.

Function indexes are the identifiers of the PHY functionalities, which have been shown in Fig. 4.4. The total RCC of a BBU is calculated by (4.3):

$$C_{r,t} = \sum_{i=1}^{N_{\mathrm{CF}}} C_{r,i,t}^{\mathrm{CF}} + \sum_{u=1}^{N_{r,t}} \sum_{j=1}^{N_{\mathrm{UF}}} C_{r,u,j,t}^{\mathrm{UF}}, \tag{4.3}$$

where $C_{r,t}$ is the RCC to handle the RRH r at time t, $C_{r,i,t}^{\mathrm{CF}}$ is the capacity associated with the common functions i needed to handle the RRH r at time t, and $C_{r,u,j,t}^{\mathrm{UF}}$ is the capacity to run the UF j of the active UE u through the RRH r. N_{CF} and N_{UF} are the amount of CFs and UFs respectively, while

Table 4.5: Scaling factors ($s_{i,x}$) for function i and RCC of the reference scenario (C_i^{ref}) (based on [81, 85]).

Function index i	$C_{i,ref}$	B_{w}	N_{a}	Q	M	ρ	N_{s}
1 (CF)	1.3	1	1	1.2	-	-	-
2 (CF)	2.7	1	1	1.2	-	-	-
3 (UF)	1.3	1	1	1.2	0	0	1
4 (UF)	5.3	1	2	1.2	0	0	0
5 (UF)	1.3	1	0	1.2	1.5	1.5	1
6 (UF)	2.7	1	0	1.2	1.5	1.5	1
7 (UF)	1.3	1	0	1.2	1	1	1
8 (UF)	8	1	0	1.2	1	1	1
9 (UF)	3.3	1	1	1.2	0	0	1

$N_{r,t}$ is the number of active UEs in RRH r at time instant t.

4.3 Conclusions

As has been previously analyzed, most of the research works drastically simplify the scenarios of validation. It is impossible to strictly simulate a mobile network because it is an extremely complex system with high time variations. However, it is important to keep a trade-off between the simplification of the scenario and the quality of the validation. For instance, the results of a deployment of a 5G radio access network, which should represent an ultra-dense network deployment should not be validated with a synthetic scenario of few cells.

This chapter presented a realistic scenario in the city of Vienna, which tries to represent the complexity of mobile networks. The presented platform employs realistic models in each layer of the C-RAN architecture. In the user plane, UEs, GBR, and Best-effort services at the packet level have been modeled. The air interface has been represented using a 3D ray-tracing model that provides all the correlations and spatial consistencies. Additionally, an estimation of the required computational capacity at BBU pools has been introduced. These features allow accounting for QoS and designing of resource management strategies among other advantages, with a high level of flexibility.

Chapter 5 aggregates to the platform multiple optimized features as a direct consequence of the research proposed in this thesis. The platform is a powerful tool for researchers and mobile network operators to validate numerous upgrades and it is open to new contributions.

Chapter 5

Mathematical Models

5.1 Introduction

This chapter presents the mathematical formulation of the proposed algorithms. Firstly, section 5.2 introduces the mathematical description of four strategies to design and analyze the fronthaul connections. Section 5.3 introduces the DRM, as well as three variants of DRM-AC. Finally, section 5.4 details the mathematical model of the proposed non-linear optimization model to optimize the RRH deployment.

5.2 BBU-RRU Association

One of the key points of the C-RAN deployment is the design of the fronthaul links, the connections between BBU pools and RRHs. This section presents a detailed analysis of several connection strategies attending to different optimization criteria that could help the MNOs when planning the network

: minimum delay, load balancing based on traffic or number of RRHs, and multiplexing gain optimization. The mathematical notation presented in this section is introduced in [86].

5.2.1 Minimum delay (MD)

The minimum delay algorithm takes into account the distance to establish the connections between RRHs and BBU pools. To minimize the delay, the algorithm selects for each RRH the nearest BBU pool following (5.1).

$$s_i = \{j \mid d_{ij} \leq d_{max} \cap \ d_{ij} = min(d_i)\} \tag{5.1}$$

where d_i is a vector that contains the distance from the RRH i to each BBU pool, d_{max} is the maximum allowed fronthaul distance, $min(\cdot)$ operator returns the minimum value and s_i is the BBU pool selected to connect to RRH i.

5.2.2 Load balancing (LB)

The load-balancing algorithms can use two different metrics: the number of RRHs already assigned and the capacity handled by BBU pools. The i^{th} RRH is connected to BBU pool c following (5.2).

$$c = \{j \mid d_{ij} \leq d_{max} \cap C_j = min(C)\} \tag{5.2}$$

where C is a vector that depending on the version used contains the number of RRH connected to each BBU pool or the capacity handled per

BBU pool. C_j is the capacity of the less loaded BBU pool (j) that is selected for the algorithm to establish the connections.

5.2.3 Multiplexing gain

This algorithm balances different types of traffic profiles to improve the multiplexing gain. The connections are established following two steps, described by (5.3) and (5.4). First, the algorithm connects RRH i^{th} to BBU m using (5.3) where $max(\cdot)$ denotes maximum operator.

$$\begin{aligned} m = \{j \mid & d_{ij} \leq d_{max} \\ & \cap\ MG_j < max(MG_{jn}) \\ & \cap\ MG_j = min(MG)\} \end{aligned} \tag{5.3}$$

If $m = \emptyset$ the algorithm uses the second condition to establish the connection (5.4), where the RRH is connected to the BBU pool with the highest multiplexing gain. The algorithm repeats this process until each RRH is connected to the network.

$$m = \{j \mid d_{ij} \leq d_{max} \cap\ MG_j = min(MG)\} \tag{5.4}$$

In (5.3) and (5.4) MG is a vector that contains the multiplexing gain of each BBU pool, MG_j is the multiplexing gain of the BBU pool j and MG_{jn} is a vector that stores the achievable multiplexing gain after connecting each possible RRH, computed as:

$$MG_j = \frac{\sum_{k=1}^{N_{RRH,j}} C_{RRH,k}[GBph]}{C_j[GBph]} \tag{5.5}$$

where $N_{RRH,j}$ is the number of RRHs connected to the j^{th} BBU pool, $C_{RRH,k}$ is the peak traffic through the k^{th} RRH and C_j is the traffic handled by the j^{th} BBU pool.

5.3 Dynamic Resource Management (DRM)

In this section, the dynamic resource allocation problem with QoS constraint is presented. The aim is to optimize the allocated capacity at each BBU pool considering the required computational capacity, the priority of running services, and the maximum capacity available at the BBU pool. The mathematical model of the DRM was presented in [87].

Let's assume that the coverage area of a specific region is served by a set of $R = \{1, \ldots, N\}$ RRHs, managed by a BBU pool. The required computational capacities to handle each RRH are $C_{tk} = \{C_{1,tk}, \ldots C_{N,tk}\}$, which are computed using (4.3). The goal is to maximize the allocated computational capacity, which is described by the set $ACC_{tk} = \{ACC_{1,tk}, \ldots, ACC_{N,tk}\}$.

The problem could be modeled using a game-theoretical approach where RRHs are connected to BBUs that are competing for computational resources at each transmission time interval (TTI). The allocated resources must not surpass the total capacity of the BBU pool (M), as expressed in (5.6). We call this condition C_1.

$$C_1 : \sum_{i \in R} ACC_{i,tk} \leq M \quad \forall\, tk \tag{5.6}$$

The weight of each service establishes priorities by aggregating QoS constraints. We denote the average service weights at each RRH as $\overline{w}_{tk} =$

$\{\overline{w}_{1,tk}, \ldots, \overline{w}_{N,tk}\}$. Bargaining power is defined in (5.7), where the average service weights at each RRH act as a fitness parameter.

$$\mathrm{C}_2: \ B_{i,tk} = \frac{\overline{w}_{i,tk} C_{i,tk}}{\sum_{j \in R} \overline{w}_{j,tk} C_{j,tk}} \tag{5.7}$$

BBU allocated resources must not be greater than the required computational capacity (5.8).

$$\mathrm{C}_3: \ ACC_{i,tk} \leq C_{i,tk} \quad \forall\, i \in R, \forall\, tk \tag{5.8}$$

Then the underlying optimization problem to perform the proposed strategy is formulated as:

$$\begin{aligned} &\underset{ACC_{tk}}{\text{maximize}} \quad \sum_{i \in R} B_{i,tk} ACC_{i,tk} \\ &\text{subject to}: \quad \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 \end{aligned} \tag{5.9}$$

The problem becomes a weighted-sum MOO problem, where BBUs-RRHs running higher priority services are prioritized because the allocated resources are weighted by the bargaining power factors.

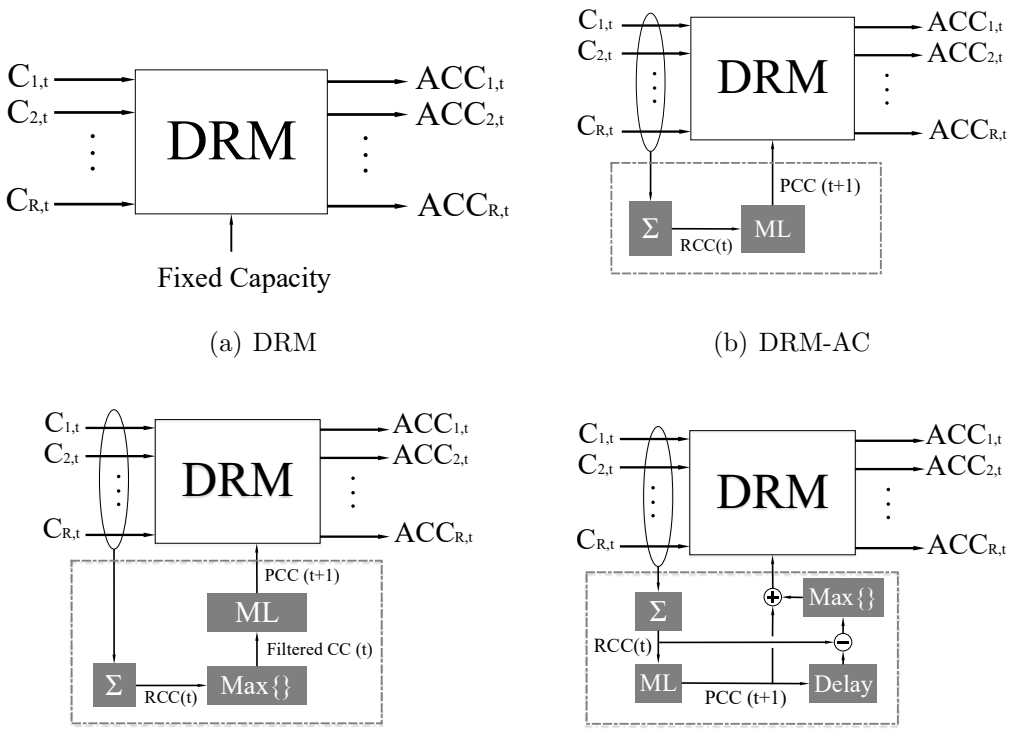
Problem (5.9) is solved by CVX tool [88] iteratively during the simulation period.

5.3.1 DRM with adaptive capacity (DRM-AC)

Fig. 5.1(a) shows the general scheme of the DRM, where $C_{r,t}$ depicts the required computational capacity to handle the RRH r at time t (computed using (4.3)). Moreover, $\mathrm{ACC}_{r,t}$ represents the allocated computational capacity, where $r \in [1, \mathrm{R}]$, being R the amount of RRHs connected to the BBU

pool under analysis.

The DRM allocates the resources available at the BBU pool to manage each RRH with service priority as well as QoS constraints. This strategy is analyzed in [89]. However, the instantiated computational capacity at the BBU pool is fixed (see 5.1(a)). It causes QoS degradation (under-provisioned) or inefficient resource usage (over-provisioned). To tackle this issue, we propose to dynamically instantiate resources using the schemes shown in 5.1(b), 5.1(c), and 5.1(d).



(a) DRM

(b) DRM-AC

(c) DRM-AC-PF

(d) DRM-AC-ES

Figure 5.1: Block diagrams of the dynamic resource management strategies

Fig. 5.1(b) shows the block diagram of the DRM-AC, which is also described in [90]. A machine learning entity is introduced. Its mission is to predict the required computational resources at the BBU pools based on the current network load. An aggregation block computes the RCC(t) based on the current demand of each RRH ($C_{r,t}$), which depicts the database of the ML block to predict the computational capacity at the next time step PCC($t+1$). The analysis of multiple machine learning techniques to tackle the prediction is presented in [91].

However, negative errors in the prediction produce QoS degradation. Two approaches to address this issue are proposed, as will be detailed afterward. These solutions and their performance are published in [92].

- Filtering the data before the training process using a sliding window method and applying the maximum operation. Fig 5.1(c) shows the general diagram of this approach that has been called DRM-AC-PF.

- Establishing a margin amount of computational resources equal to the maximum error in a previous time window. Fig 5.1(d) depicts the block diagram to implement this strategy, which is a DRM-AC-ES.

ML block contains the machine-learning algorithm to predict the computational capacity. On the other hand, the delay block is a memory that stores the input value for the next iteration. The Max{} block depicts a non-linear filter; it computes the maximum, sliding a window through the input data. The output of the Max{} block is equal to the maximum of the θ previous time steps.

DRM-AC-PF employs the Max{} block to filter the RCC and the ML block to predict the computational capacity in terms of the envelope of the RCC. On the other hand, the DRM-AC-ES predicts the computational resources based on the RCC; it makes use of a delay block to save the previous Predicted Computational Capacity (PCC) for calculating the error. Finally, it applies a Max{} filter to the error, which is aggregated to the PCC as a marginal amount of computational operations to the predicted computational capacity.

5.4 RRH optimization deployment

This section presents the mathematical model developed to optimize the number and distribution of the active RRHs required to minimize the deployment cost while simultaneously maximizing the coverage and satisfying QoS requirements, considering multiple BBU–RRH split options. The mathematical description is introduced in [93] and [94]. The model also allows the consideration of cooperation among RRHs.

Let $\mathcal{R}$ be the set of candidate RRHs and their locations. Information provided by the MNO about the already deployed cellular networks should be provided (4G and 5G). In general, locations with feasible access to the power grid and line of sight propagation should be considered to improve the propagation conditions.

Two types of RRHs can be used, MRRHs and SRRHs, which have a deployment cost of C_{MRRH} and C_{SRRH}, respectively. A binary vector $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_{|\mathcal{R}|}\}$ indicates what kind of cell could be deployed at each possible

location, where the notation $|\mathcal{R}|$ denotes the number of elements of the set $\mathcal{R}$.

Equation (5.10) shows the definition of the elements of η. The set $\mathcal{R}$ is subdivided into the sets $\mathcal{M}$ and $\mathcal{S}$ that represent MRRHs and SRRHs respectively, such that $\mathcal{M} \cup \mathcal{S} = \mathcal{R}$.

$$\eta_r = \begin{cases} 1 & \text{if } r \in \mathcal{M} \\ 0 & \text{if } r \in \mathcal{S} \end{cases} \tag{5.10}$$

On the other hand, UEs have been modeled to generate the traffic demand and also have been represented mathematically by the set $\mathcal{U}$. Each UE is subscribed to a unique Service Provider (SP) and, for the sake of simplicity, each SP is associated with one service. Thus, each UE generates only one service and is associated with one Service Function Chain (SFC) of its SP. Those services could be GBR or non-GBR (Best Effort) services such as High-Definition (HD) video streaming and FTP, respectively.

Especially, the GBR services must guarantee a minimum bit rate to each UE, which is denoted as $D_u^{\min}$, where $u \in \mathcal{U}$. This parameter is selected by the SP according to the minimum QoS that should be assured for each service. In order to provide the minimum bit rate to the GBR-UEs, a minimum SINR should be maintained (denoted as $\gamma_u^{\min}$), which can be estimated employing Shannon's equation (5.11).

$$\gamma_u^{\min} = 2^{\frac{D_u^{\min}}{B_u}} - 1 \qquad \forall u \in \mathcal{U} \tag{5.11}$$

where B_u depicts the bandwidth assigned to the UE u.

The geographical area under analysis is divided into TDPs, where each

of them aggregates and concentrates the data rates of the UEs inside it. Let $\mathcal{Z}$ be the set of demand zones or TDPs in the region and D_z the demand bit rate of the zone $z \in \mathcal{Z}$.

The algorithm should select the optimum RRH distribution that reduces the deployment cost while increasing the coverage. For this reason, the binary decision vector $\boldsymbol{\rho} = \{\rho_1, \rho_2, \ldots, \rho_{|\mathcal{R}|}\}$ has been defined to indicate the RRH distribution. ρ_r is a binary variable that indicates if the candidate RRH $r \in \mathcal{R}$ is activated or not, see (5.12).

$$\rho_r = \begin{cases} 1 & \text{if } r \in \mathcal{R} \text{ is selected as RRH} \\ 0 & \text{otherwise} \end{cases} \tag{5.12}$$

Besides, a binary decision matrix $\boldsymbol{X}$ of dimension $|\mathcal{R}| \times |\mathcal{Z}|$ is employed to represent the association between RRHs and TDPs. The elements of $\boldsymbol{X}$ are represented by the binary variable $x_{r,z}$, which is defined as (5.13):

$$x_{r,z} = \begin{cases} 1 & \text{if } z \in \mathcal{Z} \text{ is served by } r \in \mathcal{R} \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

Next, constraint (5.14) takes into account cooperation among RRHs to improve the network capacity (e.g. Joint Transmission (JT) or any other CoMP technique) by the introduction of the integer μ, that limits the number of RRHs that can cooperate to increase the bit rate while mitigating the interference of the zones. It guarantees that each zone is served by a maximum of μ RRHs. Its minimum value is $\mu = 1$ when cooperation techniques are not allowed. MNOs should select μ before the optimization process and

Table 5.1: Glossary of terms of the RRH selection algorithm

Terms	Sets	Input Parameters	Binary Variables	Description
$\mathcal{R}$	✓			set of possible RRHs and their locations
$\mathcal{M}$	✓			set of possible MRRHs and their locations
$\mathcal{S}$	✓			set of possible SRRHs and their locations
$\mathcal{Z}$	✓			set of zones or TDPs
$\mathcal{U}$	✓			set of UEs
ρ_r			✓	RRH distribution, 1 if $r \in \mathcal{R}$ is selected as RRH, 0 otherwise
$x_{r,z}$			✓	1 if RRH $r \in \mathcal{R}$ manages the traffic demand of zone $z \in \mathcal{Z}$, 0 otherwise
C_{SRRH}		✓		deployment cost of an SRRH
C_{MRRH}		✓		deployment cost of a MRRH
η_r		✓		Identifier of RRH type, 1 if r is a MRRH, 0 otherwise
σ		✓		ratio between MRRH and SRRH costs
$D_u^{\min}$		✓		required bit rate of the UE $u \in \mathcal{U}$
$\gamma_u^{\min}$		✓		SINR required by the UE $u \in \mathcal{U}$ to satisfy $D_u^{\min}$
$\gamma_z^{\min}$		✓		minimum SINR to satisfy at zone $z \in \mathcal{Z}$
κ^{adj}		✓		SINR factor to consider the mitigation of the interference
γ_z		✓		perceived SINR at zone $z \in \mathcal{Z}$
$P_{r,z}$		✓		received power at $z \in \mathcal{Z}$ from $r \in \mathcal{R}$
P_r		✓		transmission power of the RRH $r \in \mathcal{R}$
G_r		✓		antenna gain of the RRH $r \in \mathcal{R}$
G_{UE}		✓		antenna gain of the UEs
$L_{r,z}$		✓		transmission loss from the RRH $r \in \mathcal{R}$ to the TDP at zone $z \in \mathcal{Z}$
L^{RRH}		✓		loss introduced at the RRHs (e.g., transmission lines and connectors losses)
L^{UE}		✓		loss introduced at the UEs (e.g., transmission lines and coupling losses)
L^{FD}		✓		loss introduced by the fading effects
$L_{r,z}^{\text{PL}}$		✓		path loss from the RRH $r \in \mathcal{R}$ to the TDP at zone $z \in \mathcal{Z}$
$\text{F}_2^{\min}$		✓		minimum normalized coverage–QoS
B_u		✓		bandwidth of the UE $u \in \mathcal{U}$
D_z		✓		traffic demand (bit rate) of the TDP at zone $z \in \mathcal{Z}$
D_r		✓		achievable throughput at RRH $r \in \mathcal{R}$
μ		✓		number of allowed simultaneous connections of each UE to the RRHs
ξ_r		✓		resource usage ratio of the RRH $r \in \mathcal{R}$
$\xi^{\max}$		✓		maximum resource usage ratio of the RRHs
N		✓		noise power

according to the available cooperation technology of the considered network.

$$\sum_{r \in \mathcal{R}} x_{r,z} \leq \mu \qquad \forall z \in \mathcal{Z} \tag{5.14}$$

Additionally, a key point is the establishment of a relationship between the decision variables ρ_r and $x_{r,z}$ to ensure that, if a possible RRH $r \in \mathcal{R}$ is selected, it must serve at least one zone $z \in \mathcal{Z}$; and if a RRH is associated to a zone z, it must be active. Equations (5.15) and (5.16) account for these conditions, respectively.

$$\sum_{z \in \mathcal{Z}} x_{r,z} \geq \rho_r \qquad \forall r \in \mathcal{R} \tag{5.15}$$

Moreover, equation (5.16) is also a capacity constraint. It ensures that a selected RRH has enough capacity to satisfy the demand of its associated TDPs or zones.

$$\begin{aligned} &\xi_r \leq \xi^{\max} \rho_r && \forall r \in \mathcal{R} \\ &\xi_r = \sum_{z \in \mathcal{Z}} \frac{x_{r,z} D_z}{D_r \sum_{r' \in \mathcal{R}} x_{r',z}} \end{aligned} \tag{5.16}$$

In (5.16) D_r represents the achievable throughput at the RRH r. This parameter depends on the RRH configuration, e.g., Multiple-Input Multiple-Output (MIMO) order, modulation order, and bandwidth. Besides, the real variable ξ_r depicts the traffic load of the GBR services through the RRH r. The MNOs establish a partition of the resources between the GBR and best effort services by controlling the parameter $0 \leq \xi_r \leq \xi^{\max} \leq 1$. For instance, if $\xi^{\max} = 0.8$ it means that 80% of the radio resources of the RRHs could be dedicated to satisfying the traffic of the slices with GBR services. The remainder 20% of the RRH capacity is reserved for the non-GBR traffic. The

parameter $\xi^{\max}$ should be selected according to the demand for the different types of services.

As it has been mentioned, to guarantee the QoS, it is important that the UEs experience a SINR greater than a minimum ($\gamma_u^{\min}$). However, due to the enormous amounts of UEs that are expected in 5G networks, it is not scalable to define a constraint that maintains an independent SINR requirement for each UE in the optimization algorithm. For this reason, equation (5.17) guarantees that the SINR constraint is accomplished while relaxing the specifications by moving them to an upper level (zone plane).

$$\gamma_z \geq \gamma_z^{\min} \kappa^{\text{adj}} \quad \forall z \in \mathcal{Z} \tag{5.17}$$

In (5.17), γ_z represents the perceived SINR at the TDP or zone $z \in \mathcal{Z}$, while $\gamma_z^{\min}$ depicts the minimum SINR that must be kept at TDP of the zone z, which is taken equal to the required SINR of the UE with highest demand in the zone z. This approach is highly restrictive and it does not take into account techniques such as enhanced Inter-Cell Interference Coordination (eICIC). For this reason, the factor κ^{adj} is introduced, which should be adjusted by the MNO to consider the mitigation of interference by dynamic resource allocation techniques.

$$\sum_{r \in \mathcal{R}} x_{r,z} P_{r,z}^{\text{Rx}} \geq \gamma_z^{\min} \kappa^{\text{adj}} (\sum_{r \in \mathcal{R}} \rho_r (1 - x_{r,z}) P_{r,z}^{\text{Rx}} + N) \qquad \forall z \in \mathcal{Z} \tag{5.18}$$

Equation (5.17) is transformed into (5.18) by introducing an estimated value for γ_z. The left side of (5.18) represents the useful received power at TDP z, while the expression in brackets of the right term models the interference

plus noise power. The parameter N represents the average UE thermal noise power. It is important to mention that according to equation (5.17), all the active RRHs that are not serving the considered TDP are a source of interference, which means that the mobile network is designed with a frequency reuse factor equal to unity. However, operating at the same frequency band in the entire network will produce a high level of interference. For this reason, the proposed algorithm considers that MRRHs and SRRHs operate at different frequency bands. In this work, the MRRHs operate at 2.6 GHz while the SRRHs are able to operate at multiple frequency bands (for instance, 3.6 GHz and 28 GHz). As a result, equation (5.18) is split into (5.19) and (5.20)

$$\sum_{r\in\mathcal{M}} x_{r,z}P_{r,z}^{\mathrm{Rx}} \geq \gamma_z^{\min}\kappa^{\mathrm{adj}}(\sum_{r\in\mathcal{M}} \rho_r(1-x_{r,z})P_{r,z}^{\mathrm{Rx}} + N) \qquad \forall z \in \mathcal{Z} \tag{5.19}$$

$$\sum_{r\in\mathcal{S}} x_{r,z}P_{r,z}^{\mathrm{Rx}} \geq \gamma_z^{\min}\kappa^{\mathrm{adj}}(\sum_{r\in\mathcal{S}} \rho_r(1-x_{r,z})P_{r,z}^{\mathrm{Rx}} + N) \qquad \forall z \in \mathcal{Z} \tag{5.20}$$

where, as mentioned above, $\mathcal{M}$ and $\mathcal{S}$ are subsets of $\mathcal{R}$ that contain the sets of MRRH and SRRH respectively, such that $\mathcal{R} = \mathcal{M} \cup \mathcal{S}$. The parameter $P_{r,z}^{\mathrm{Rx}}$, which represents the received power at TDP z from the RRH r, is calculated by using the link budget equation (5.21)

$$P_{r,z}^{\mathrm{Rx}} = \frac{P_r^{\mathrm{Tx}}G_rG^{\mathrm{UE}}}{L_{r,z}} \tag{5.21}$$
$$L_{r,z} = L^{\mathrm{RRH}}L^{\mathrm{UE}}L^{\mathrm{FD}}L_{r,z}^{\mathrm{PL}}$$

where P_r^{Tx} and G_r denote the transmission power and the antenna gain of

the RRH r, respectively. G^{UE} represents the UE antenna gain, while $L_{r,z}$ takes into account the radio link losses. Namely, L^{RRH} and L^{UE} account for the losses due to connectors, transmission lines, and other mismatches at the RRH and the UE respectively. Besides, $L^{\mathrm{PL}}_{r,z}$ represents the path-loss from the RRH r to the TDP z. Finally, L^{FD} is a random variable modeling the slow fading. It is important to notice that this parameter should be eliminated if the considered channel model already takes into account the shadowing effects.

Additionally, equation (5.22) ensures that if a RRH r is serving the zone z, the received power ($P^{Rx}_{r,z}$) is greater or equal than the sensitivity of the UE-receivers ($\mathrm{P}^{\mathrm{Rx}}_{\mathrm{min}}$).

$$x_{r,z} P^{\mathrm{Rx}}_{r,z} \geq \mathrm{P}^{\mathrm{Rx}}_{\mathrm{min}} \tag{5.22}$$

As it has been stated, the proposed algorithm allows for RRH cooperation with the introduction of the parameter μ. On the other hand, constraint (5.14) does not limit the cooperation between MRRH and SRRH, which introduces a high complexity to the UEs because they would have to operate simultaneously at different frequency bands, for instance in a dual connectivity operation mode. For this reason, constraint (5.23) guarantees that the cooperation is limited to a specific frequency band (known as inter-site aggregation). In other words, cooperation in a specific zone is carried out by only one type of RRH.

$$\sum_{r \in \mathcal{M}} x_{r,z} \leq 0 \quad \vee \quad \sum_{r \in \mathcal{S}} x_{r,z} \leq 0 \quad \forall z \in \mathcal{Z} \tag{5.23}$$

where $\vee$ stands for the logical disjunction (logical OR operation), guaranteeing that the zone $z \in \mathcal{Z}$ is not served by RRHs of different classes.

The goal of the proposed algorithm is to select the optimum distribution of the RRHs, minimizing the radio network deployment cost while simultaneously maximizing the coverage and satisfying QoS requirements. The deployment cost is computed as in equation (5.24).

$$\begin{aligned} \mathrm{F}_1 &= \sum_{r \in \mathcal{R}} \rho_r (\eta_r C_{\mathrm{M}} + (1 - \eta_r) C_{\mathrm{S}}) \\ \mathrm{F}_1 &= C_{\mathrm{S}} \sum_{r \in \mathcal{R}} \rho_r (\eta_r \sigma + (1 - \eta_r)) \\ \sigma &= \frac{C_{\mathrm{M}}}{C_{\mathrm{S}}} \end{aligned} \tag{5.24}$$

where σ represents the ratio between the MRRH and SRRH costs, C_{M} and C_{S}, respectively. This parameter is useful to consider different kinds of scenarios; for instance, to represent heterogeneous mobile network deployments ($\sigma > 1$). Furthermore, it is used in this work to consider different BBU–RRH split options because it modifies the cost ratio between MRRHs and SRRHs. On the other hand, it allows the normalization of the RRH costs by considering $C_{\mathrm{S}} = 1$. The MNOs should carefully select the cost ratio (σ) according to the cost of the considered network devices.

The coverage–QoS is estimated by the number of served zones, computed as in equation (5.25).

$$\mathrm{F}_2 = \sum_{z \in \mathcal{Z}} u[\sum_{r \in \mathcal{R}} x_{r,z} - 1] \tag{5.25}$$

$$\mathrm{F}_2 \geq \mathrm{F}_2^{\min} |\mathcal{Z}| \tag{5.26}$$

$$u[n] = \begin{cases} 1 & \text{if} \quad n \geq 0 \\ 0 & \text{if} \quad n < 0 \end{cases}$$

where $u[\cdot]$ denotes the Heaviside sequence and the operator $|\cdot|$ stands for the cardinal of the considered set. Moreover, the MNO has the flexibility to establish the minimum coverage-QoS that must be provided by controlling the parameter $0 \leq \mathrm{F}_2^{\min} \leq 1$ in constraint (5.26). Finally, the optimum radio network deployment algorithm is formulated as an integer optimization problem in equation (5.27).

$$\begin{aligned} &\underset{\rho_r, x_{r,z}}{\text{minimize}} \quad \mathrm{F}_1, -\mathrm{F}_2 \\ &\text{subject to :} \quad (5.10) - (5.16), (5.19) - (5.23), (5.26) \end{aligned} \tag{5.27}$$

Table 5.1 summarizes the sets, variables, and parameters of the proposed algorithm.

5.4.1 Integer Linear Optimization Problem

The algorithm formulated in section 5.4 is a non-linear integer programming model. The non linearity is introduced by the constraints (5.16), (5.19), (5.20), (5.23) and the coverage function (5.25). In order to solve the proposed algorithm employing a linear optimization solver, a reformulation of these expressions to linear equations is needed. In this section, the problem is transformed into an ILP model. The presented mathematical manipulations were published in [93].

The linearization of the previously mentioned expressions uses theorem 1:

Theorem 1 *Lets assume* $\mathcal{D} \subseteq \mathbb{R}^n$, $f : \mathcal{D} \to \mathbb{R}$, $M \in \mathbb{R} : M \neq 0; M \geq \max\{f(\phi)|\phi \in \mathcal{D}\}$ *and* δ *a binary variable such that* $\delta \in \{0, 1\}$. *Then, the following expressions are equivalent,*

i: $\delta = 0 \implies f(\phi) \leq 0$

ii: $f(\phi) - M\delta \leq 0$

The linearization of the constraint (5.16) implies the modification of the term $\frac{x_{r,z}}{\sum_{r' \in \mathcal{R}} x_{r',z}}$. Following the constraint (5.14), the denominator of this term is an integer ($k \in \mathbb{N} | 0 \leq k \leq \mu$), bounded by the maximum number of RRHs that could cooperate to serve a zone (μ). So, it is possible to reformulate the expression as it is described in (5.28)

$$\frac{x_{r,z}}{\sum_{r' \in \mathcal{R}} x_{r',z}} = \begin{cases} \sum_{k=1}^{\mu} k^{-1} x_{r,z} \delta_{k,z} & \text{if} \quad k \neq 0 \\ 0 & \text{if} \quad k = 0 \end{cases} \tag{5.28}$$

where $\delta_{k,z}$ are binary variables indicating if the TDP z is served by k RRHs, as described in (5.29)

$$\delta_{k,z} = 1 \implies \sum_{r \in \mathcal{R}} x_{r,z} = k \quad \forall z \in \mathcal{Z}, k \in [0, \mu] \tag{5.29}$$

However, the equations (5.28) and (5.29) are also non-linear expressions that should be linearized. The product of binary variables in (5.28) is substituted by another binary variable such that $x_{r,z}\delta_{k,z} = \psi_{r,z,k}$, which is equivalent to (5.30).

$$\psi_{r,z,k} = 1 \iff x_{r,z} + \delta_{k,z} = 2 \tag{5.30}$$

After this mathematical procedure, the equations (5.29) and (5.30) could be converted to linear expressions employing the *Theorem* 1. Equation (5.31)

shows the linear equivalent expressions of the constraint (5.16).

$$\sum_{z\in\mathcal{Z}} D_z \sum_{k=1}^{\mu} k^{-1}\psi_{r,z,k} \leq \xi^{\max}\rho_r D_r \tag{5.31a}$$

$$\psi_{r,z,k} \leq x_{r,z} \tag{5.31b}$$

$$\psi_{r,z,k} \leq \delta_{k,z} \tag{5.31c}$$

$$\psi_{r,z,k} \geq \delta_{k,z} + x_{r,z} - 1 \tag{5.31d}$$

$$\sum_{r\in\mathcal{R}} x_{r,z} \leq \mu - \delta_{k,z}(\mu - k) \tag{5.31e}$$

$$\sum_{r\in\mathcal{R}} x_{r,z} \geq k\delta_{k,z} \tag{5.31f}$$

$$\sum_{k=0}^{\mu} \delta_{k,z} = 1 \tag{5.31g}$$

For the sake of simplicity, the domain of the indexing subscripts r, z, k has been made explicit only when it is different from the defined domain.

The other non-linear constraints are (5.19), (5.20) and (5.23). As it has been explained, each TDP could be served by k RRHs, where $k = 0$ means there is no RRH that can serve TDP z satisfying the constraints. This situation has not been considered by constraints (5.19) and (5.20), which do not hold for this special case. For this reason, the binary variables $\beta_z^{\mathcal{M}}$ and $\beta_z^{\mathcal{S}}$, that are used to indicate if the zone z is served by MRRHs or SRRHs, are introduced in (5.32) and (5.33).

$$\beta_z^{\mathcal{M}} = 1 \Longleftrightarrow \sum_{r\in\mathcal{M}} x_{r,z} \leq 0 \tag{5.32}$$

$$\beta_z^{\mathcal{S}} = 1 \Longleftrightarrow \sum_{r\in\mathcal{S}} x_{r,z} \leq 0 \tag{5.33}$$

Considering this approach, the constraints could be rewritten as it is shown

in (5.34)

$$\sum_{r\in\mathcal{T}} x_{r,z}P^{\mathrm{Rx}}_{r,z} + L\beta^{\mathcal{T}}_{z} \geq \tag{5.34}$$
$$\gamma^{\min}_{z}\kappa^{\mathrm{adj}}(\sum_{r\in\mathcal{T}} \rho_r P^{\mathrm{Rx}}_{r,z} - \sum_{r\in\mathcal{T}} x_{r,z}P^{\mathrm{Rx}}_{r,z} + N) \qquad \forall z \in \mathcal{Z}$$

where $\mathcal{T}$ is equal to $\mathcal{M}$ or $\mathcal{S}$ to represent the constraints (5.19) and (5.20) respectively. An alternative is to write both constraints in the same expression. It is important to notice that the non-linear expression $\rho_r x_{r,z}$ has been substituted by $x_{r,z}$, because the constraint (5.16) ensures that if $x_{r,z} = 1$ then $\rho_r = 1$, which is equivalent to $\rho_r x_{r,z} = x_{r,z}$. The parameter $L \in \mathbb{R}$ is a large number to hold the constraint (5.34) when the zone z is not served by this type of RRH.

On the other hand, the constraint (5.23) is an inclusive disjunction that could be rewritten combining (5.32) and (5.33) with the expression $\beta^{M}_{z}+\beta^{S}_{z} \geq 1$. *Theorem* 1 has been employed to obtain the linear expressions of (5.32) and (5.33), which are shown in (5.35).

$$\sum_{r\in\mathcal{M}} x_{r,z} \leq \mu(1-\beta^{\mathcal{M}}_{z}) \tag{5.35a}$$
$$\sum_{r\in\mathcal{M}} x_{r,z} \geq \epsilon(1-\beta^{\mathcal{M}}_{z}) \tag{5.35b}$$
$$\sum_{r\in\mathcal{S}} x_{r,z} \leq \mu(1-\beta^{\mathcal{S}}_{z}) \tag{5.35c}$$
$$\sum_{r\in\mathcal{S}} x_{r,z} \geq \epsilon(1-\beta^{\mathcal{S}}_{z}) \tag{5.35d}$$
$$\beta^{M}_{z} + \beta^{S}_{z} \geq 1 \tag{5.35e}$$

Besides, the coverage is estimated in equation (5.25), as the number of served zones. Following this definition, the multi-objective optimization problem

(5.27) should minimize the deployment cost while maximizing the coverage. However, maximizing the number of served zones is equivalent to minimizing the zones without service. This approach allows for the reuse of the binary variable $\delta_{0,z}$ that indicates if the zone z is not served. Then, constraint (5.26) is expressed as (5.37) and the underlying linear coverage–QoS function is shown in equation (5.36).

$$\sum_{z \in \mathcal{Z}} \delta_{0,z} = \mathrm{F}_3 = |\mathcal{Z}| - \mathrm{F}_2 \tag{5.36}$$

$$\sum_{z \in \mathcal{Z}} \delta_{0,z} \leq \mathrm{F}_2^{\min} |\mathcal{Z}| \tag{5.37}$$

This strategy reduces the complexity of the proposed algorithm by eliminating the additional variables in the linearization of the Heaviside sequence $u[\cdot]$ in (5.25).

Finally, the ILP model of the proposed algorithm is summarized in equation (5.38). The multi-objective optimization problem is solved by employing the weighted–sum method, where the weights ω_1 and ω_3 should be carefully selected by the MNO in order to obtain an optimal point on the Pareto Front. Additionally, the subscript n in equation (5.38a) means that the objective functions have been normalized to guarantee that their values are in the same range. In this case, each function has been divided by its maximum value.

The maximum of F_1 and F_3 are $C_{\mathrm{S}}(\sigma\,|\mathcal{M}|+|\mathcal{S}|)$ and $|\mathcal{Z}|$, respectively.

$$\underset{\rho_r, x_{r,z}}{\text{Min}} \qquad \omega_1 \mathrm{F}_{1n} + \omega_3 \mathrm{F}_{3n} \tag{5.38a}$$

subject to :

$$\sum_{r\in\mathcal{R}} x_{r,z} \leq \mu \tag{5.38b}$$

$$\sum_{z\in\mathcal{Z}} x_{r,z} \geq \rho_r \tag{5.38c}$$

$$x_{r,z} P^{\mathrm{Rx}}_{r,z} \geq \mathrm{P}^{\mathrm{Rx}}_{\min} \tag{5.38d}$$

$$\sum_{z\in\mathcal{Z}} D_z \sum_{k=1}^{\mu} k^{-1} \psi_{r,z,k} \leq \xi^{\max} \rho_r D_r \tag{5.38e}$$

$$\psi_{r,z,k} \leq x_{r,z} \tag{5.38f}$$

$$\psi_{r,z,k} \leq \delta_{k,z} \tag{5.38g}$$

$$\psi_{r,z,k} \geq \delta_{k,z} + x_{r,z} - 1 \tag{5.38h}$$

$$\sum_{r\in\mathcal{R}} x_{r,z} \leq \mu - \delta_{k,z}(\mu - k) \tag{5.38i}$$

$$\sum_{r\in\mathcal{R}} x_{r,z} \geq k\delta_{k,z} \tag{5.38j}$$

$$\sum_{k=0}^{\mu} \delta_{k,z} = 1 \tag{5.38k}$$

$$\sum_{r\in\mathcal{T}} x_{r,z} P^{\mathrm{Rx}}_{r,z} + L\beta^{\mathcal{T}}_z \geq$$
$$\gamma^{\min}_z \kappa^{\mathrm{adj}} (\sum_{r\in\mathcal{T}} \rho_r P^{\mathrm{Rx}}_{r,z} - \sum_{r\in\mathcal{T}} x_{r,z} P^{\mathrm{Rx}}_{r,z} + N) \tag{5.38l}$$

$$\sum_{r\in\mathcal{M}} x_{r,z} \leq \mu(1 - \beta^{\mathcal{M}}_z) \tag{5.38m}$$

$$\sum_{r \in \mathcal{M}} x_{r,z} \geq \epsilon(1 - \beta_z^{\mathcal{M}}) \tag{5.38n}$$

$$\sum_{r \in \mathcal{S}} x_{r,z} \leq \mu(1 - \beta_z^{\mathcal{S}}) \tag{5.38o}$$

$$\sum_{r \in \mathcal{S}} x_{r,z} \geq \epsilon(1 - \beta_z^{\mathcal{S}}) \tag{5.38p}$$

$$\beta_z^M + \beta_z^S \geq 1 \tag{5.38q}$$

$$\sum_{z \in \mathcal{Z}} \delta_{0,z} \leq \mathrm{F}_2^{\min} |\mathcal{Z}| \tag{5.38r}$$

$\rho_r, x_{r,z}, \delta_{k,z}, \psi_{r,z,k}, \beta_z^M, \beta_z^S$ binary variables

5.5 Conclusions

The optimization algorithms described in this chapter have been integrated into the C-RAN platform detailed in 4. These models not only improve the flexibility of the software planning tool by adding more parameters that can be controlled or used by the MNOs to get extra information; they facilitate the optimization both, along the network design and management phases, respectively. The performance analysis of the proposed algorithms is presented in chapters 6 and 7.

Chapter 6

Radio access network deployment study

6.1 Introduction

This chapter presents the results of testing a radio network deployment using the proposed optimization algorithm. These results demonstrate how the algorithm reduces the number of required active RRH. Besides, it offers acceptable coverage and satisfies the UE requirements in terms of QoS. This is crucial because it entails a considerable reduction in the network cost, with the consequent improvements in energy-saving, necessary when a large number of cells are deployed as is the case of 5G and beyond.

The algorithm could adapt to variations in traffic patterns and load, recalculating the set of RRHs that should be active for each case. Without the optimization procedure, the MNO would activate all the RRHs, not benefiting from the resource, cost, and energy-saving improvements. The results

presented in this chapter have been gradually published [89, 93, 94].

6.2 Fronthaul deployment analysis

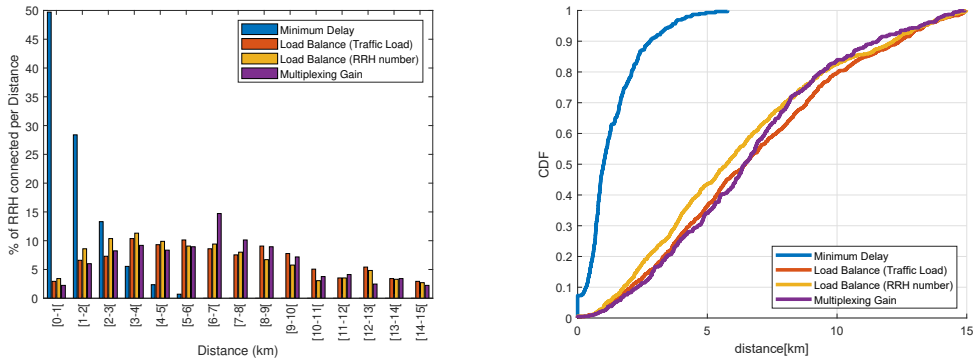
The C-RAN scenario described in section 4.2 offers multiple opportunities to test and validate different optimization algorithms. The fronthaul deployment is a crucial step in the design of current and future mobile networks. In this section, four different fronthaul planning strategies have been considered and analyzed: minimum delay, load balancing based on traffic or number of RRHs, and multiplexing gain algorithm, which has been adapted from [95].

This study is introduced not only to establish the BBU-RRH connections of the proposed C-RAN but also to analyze different strategies that could be used to optimize the fronthaul deployment.

The maximum fronthaul distance was fixed at 15 km in order to satisfy the delay requirement when optical fiber links are considered. However, this is a flexible parameter that should be carefully selected by the network designer taking into account the type of services or slices that will be running in the network and the QoS that should be provided.

As it has been presented in section 5.2, the minimum delay algorithm minimizes the fronthaul distance connecting each RRH to the nearest BBU pool in order to reduce the round trip time. Load balancing algorithms establish the connections balancing the capacity or the number of connected RRHs per BBU pool in the network. Finally, the multiplexing gain algorithm mixes different types of traffic in each BBU pool to achieve a good performance of the overall network. It is important to mention that the study of

the fronthaul connections considers the full map of Vienna city, which has been presented in Fig. 4.1. For sake of simplicity, the traffic per hour scheme presented in 4.3 has been considered.

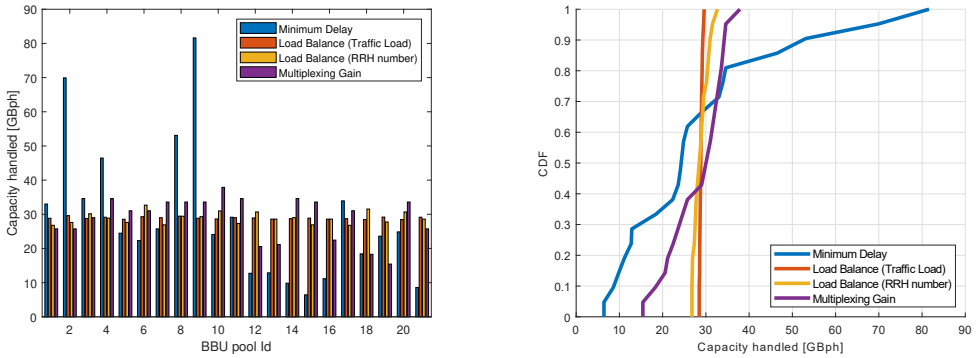


(a) (b)

Figure 6.1: Distribution of the RRH connections a) intervals of 1 km b) cumulative distribution function

Fig. 6.1(a) exhibits the distribution of the fronthaul links, while Fig. 6.1(b) shows the Cumulative Distribution Function (CDF) of each strategy. As expected, with the minimum delay strategy most of the RRHs are connected close to the BBU pool, while for the other strategies some RRHs are connected with the maximum fronthaul distance. Minimum delay design not only minimizes the latency but also reduces the CAPEX of the fronthaul because all the RRHs are connected with fronthaul distances below 6 km. The rest of the strategies exhibit similar performance in terms of delay and fronthaul cost.

A fundamental metric to increase flexibility is network balancing. Fig. 6.2 shows the distribution of the capacity handled by the BBU pools. For



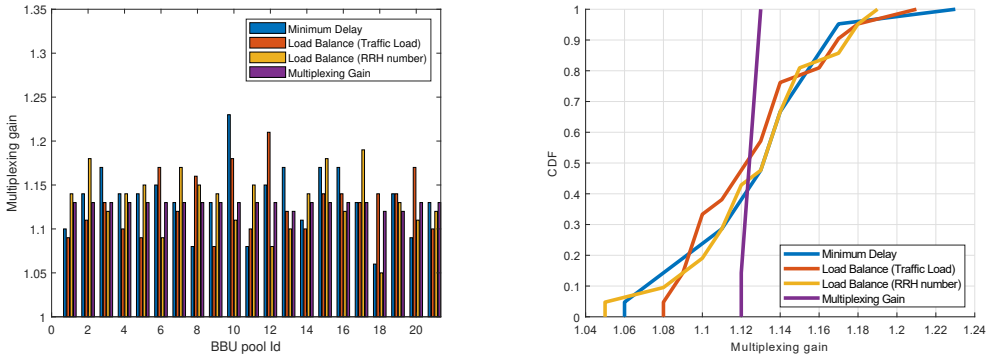
(a) (b)

Figure 6.2: Distribution of the capacity in terms of traffic load per BBU pool a) capacity handled per BBU pool b) CDF

load balancing planning strategies the capacity handled and the number of RRHs per BBU pool are almost constant around 40 RRHs and 28 GBph, which is more robust to face dynamic network variations. On the opposite, minimum delay and multiplexing gain strategies exhibit wider CDFs, hence the worst performance, because there are overloaded BBU pools while others are underutilized.

On the other hand, as it is possible to see in Fig. 6.3, the multiplexing gain strategy achieves almost constant values of multiplexing gain per BBU pool while the rest of the methods experience lower values for some BBU pools. The performance of this strategy is strongly connected to the traffic profiles handled by the network.

As it has been mentioned above, one of the most important requirements of 5G and beyond systems is the delay. The maximum fronthaul distance to satisfy the 1 ms delay constraint of applications such as virtual reality can be



(a) (b)

Figure 6.3: Distribution of the multiplexing gain per BBU pool a) MG per BBU pool b) CDF

estimated and some authors obtain distances between 20-40 km using optical fiber [96]. Furthermore, the cost to deploy a C-RAN is strongly related to the cost of the optical fiber [4], for this reason in large-scale scenarios, such as the full deployment over Vienna city, mobile operators will centralize the resources of only a certain percentage of the total number of base stations to reduce the CAPEX.

Fig. 6.4 shows the performance of the proposed C-RAN deployment in terms of the allowed maximum fronthaul distance. When the maximum fronthaul distance is decreased the percentage of RRHs that are sharing resources in BBU pools also decreases, which results in a degradation of the multiplexing gain because operators have to allocate additional resources to these RRHs. However, the cost to deploy the C-RAN is also reduced, becoming more attractive for small-size networks in dense environments such as the metropolitan area of the city. Mobile operators or Infrastructure providers

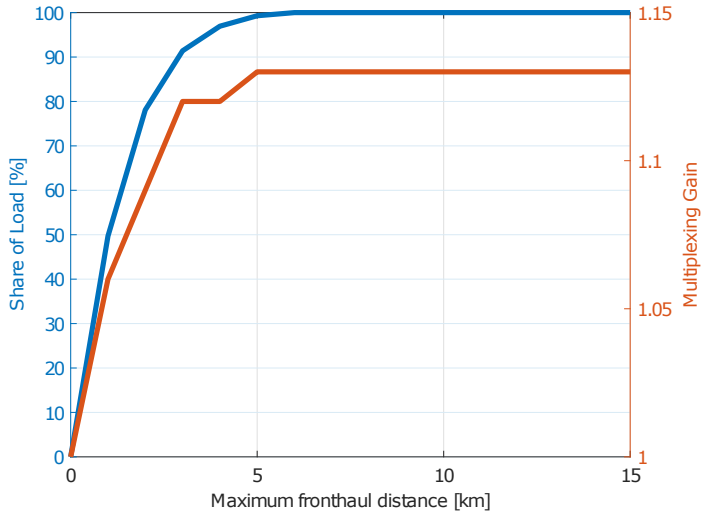


Figure 6.4: C-RAN deployment in terms of the fronthaul distance: in blue the % of shared RRHs, in red the multiplexing gain.

have to take into account this tradeoff in order to reduce the investment.

Table 6.1 summarizes the results obtained by each planning strategy, where $\mathbf{d_{max}}$ depicts the maximum fronthaul distance while $\mathbf{\Delta RRHs}$, $\mathbf{\Delta C}$, and $\mathbf{\Delta MG}$ represent the maximum deviation of the number of RRHs, traffic capacity, and multiplexing gain among BBUs, respectively. Table 6.1 shows how the algorithms guarantee the highest balance of their parameter for each central unit. For instance, the maximum deviation in the number of RRHs is one, when the load balancing based on the RRHs strategy is considered.

6.3 Analysis of the RRH deployment

This section analyzes the performance of the optimization algorithm described in section 5.4 considering as a test platform, a region inside of the

Table 6.1: Resume of the fronthaul connections results

Algorithms	$\mathbf{d_{max}}$ **[km]**	**ΔRRHs**	**ΔC [GBph]**	**ΔMG**
MD	***6***	104	75.18	0.17
LB traffic	15	7	***1.14***	0.13
LB RRHs	15	***1***	5.89	0.14
MG	15	32	22.45	***0.01***

metropolitan area of Vienna. Section 6.3.1 describes the selected region and presents the main features and modifications introduced in this validation.

6.3.1 Simulation conditions

Fig. 6.5 shows the region employed in this section. It has a map resolution of 5 m with 205×291 points, which is equivalent to an area of 1025×1455 m^2.

The strategy described in 5.4 aims to optimize the RRH plane. Fig.6.6 represents the hierarchical architecture, with the three layers used as part of the proposed optimization. The details of each layer and their interrelations are introduced as follows:

RRH plane

RRHs parameters have been chosen to describe a realistic 5G radio network deployment. Transmitted powers of MRRHs and SRRHs are $P_r = 43$ dBm and $P_r = 24$ dBm respectively. When both, MRRH and SRRH, operate at FR1 (sub 6 GHz) the antenna gains are 18 dBi and 2 dBi respectively. When FR2 is considered (mmWave, 28 GHz) the antenna gain of the SRRHs is increased to 12 dBi, because more elements can be added at the antenna

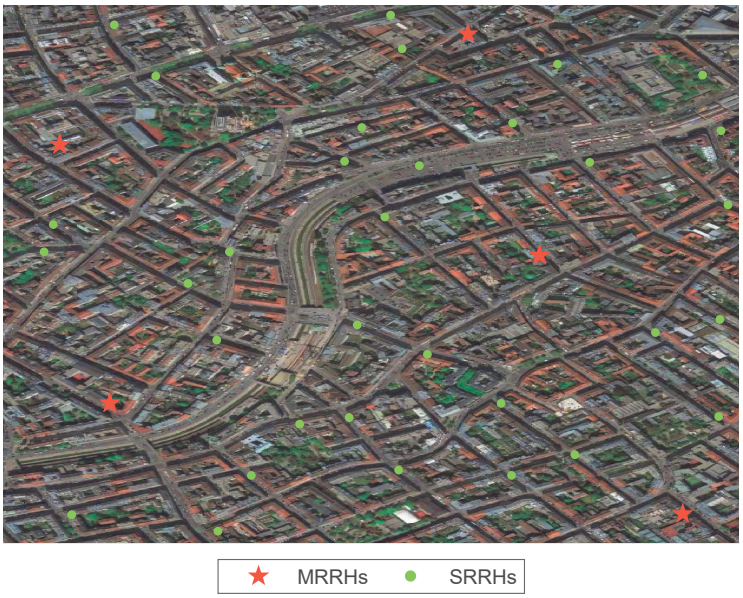


Figure 6.5: Possible RRH locations on the considered Scenario

array, according to [97].

Besides the sensitivity and SINR constraints, the capacity constraint should also be satisfied. To this end, it is fundamental to introduce an estimation of the maximum bit rate capacity of the RRHs. To do so, it is necessary to define additional RRHs configuration parameters such as modulation, MIMO order, operating frequency band, bandwidth, and 5G numerology.

As previously discussed, to reduce the inter-cell interference MRRHs and SRRHs operate at different frequencies. In this work 2.6 GHz (FR1) is selected for the MRRH, while SRRHs could operate at 3.6 GHz (FR1) or 28 GHz (FR2). In particular, n41 (2496-2690 MHz) and n77 (3300-4200 MHz) frequency bands are considered when MRRHs and SRRHs operate at FR1, with a bandwidth of 100 MHz for each. When the SRRHs

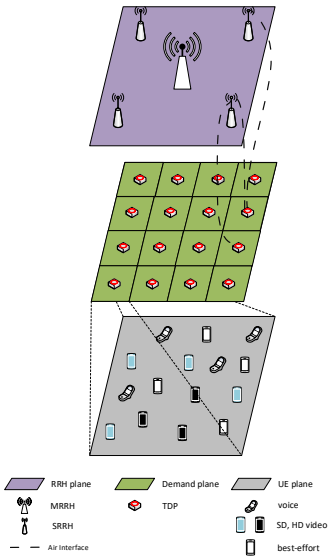


Figure 6.6: Hierarchical structure of the scenario

operates at FR2, the selected frequency band is n257 (26.50-29.50 GHz) with 300 MHz of bandwidth. A fundamental motivation to select the frequency bands is that the 3D ray-tracing propagation model has been already employed using these frequency ranges.

While the maximum modulation order considered for all the frequency ranges is 256QAM, lower values would be dynamically assigned according to the interference and propagation conditions. Regarding the MIMO, 8×8 and 16×16 are considered in FR1 and FR2 respectively. Finally, the theoretical RRHs maximum capacity is estimated according to [98], and using the values summarized in Table 6.2.

The selected section of the map is composed of 41 possible RRHs locations, with 8 of them for possible MRRHs while the remainder are possible SRRHs (see Fig. 6.5).

Table 6.2: RRH plane: configuration and maximum capacity

Frequency Range	**Bandwidth (MHz)**	**Modulation**	**MIMO**	D_r **(Gbps)**
FR1	100	256QAM	8×8	4
FR2	300	256QAM	16×16	28

Demand Plane

The whole map is divided into $\sqrt{|\mathcal{Z}|} \times \sqrt{|\mathcal{Z}|}$ homogeneous zones. This approach has the advantage that by increasing the number of divisions (so, by decreasing the area of one zone), a finer tuning is obtained at the expense of increasing the computational complexity. Once the demand plane has been split, the traffic demand of each zone should be estimated. It depends on the demand of the UEs inside the zone.

UE Plane

As has been mentioned above, each UE is associated with a slice of a specific SP. Different kinds of services have been modeled to generate traffic demand. The voice and video services on 5G networks will be delivered based on the IP Multimedia Subsystem (IMS). In general, these kinds of services are enclosed in the standardization of Voice/Video over New Radio (VoNR) [99]. In particular, three examples of these services, which have been specified by [100], are considered in this work: conversational voice, HD video, and Standard-Definition (SD) video.

The GBR of each service must be selected by the SP to guarantee a specific QoS. In this case, the conversational audio service uses Enhanced Voice Services (EVS) codec (EVS), which has different bit rates configurations with

a maximum value of 128 kbps; this is the value considered as the GBR to provide the maximum QoS to the end user. On the other hand, the video services use H.265 and EVS codecs. In the proposed scenario, two different video qualities are considered: HD video and SD video with 10 Mbps and 2 Mbps of GBR respectively, which is consistent with [101]. Table 6.3 summarizes the service parameters, where the parameter SP mix represents the percentage of UEs subscribed to each SP. A random user distribution is considered. However, the MNOs should allocate the UEs according to their historical data distribution.

Table 6.3: Service parameters of RRH deployment strategy

Service	**GBR/Best-effort**	$D_u^{\min}$ **(Mbps)**	**SP mix (%)**
voice	GBR	128 kbps	25
HD video	GBR	10 Mbps	15
SD video	GBR	2 Mbps	30
FTP	Best-effort	-	15
Web	Best-effort	-	15

Fig. 6.7 shows the demand plane with Low traffic (LT) profile that contains 30000 randomly distributed UEs in a regular divided map of 49 zones (7×7) or Traffic Demand Points (TDPs). Fig. 6.7a shows the traffic demand of each TDP in Mbps (D_z). On the other hand, Fig. 6.7b shows the number of UEs that belong to each zone. To analyze the performance of the proposed optimization algorithms, medium and high traffic profiles have also been considered, with 60000 and 300000 UEs, respectively.

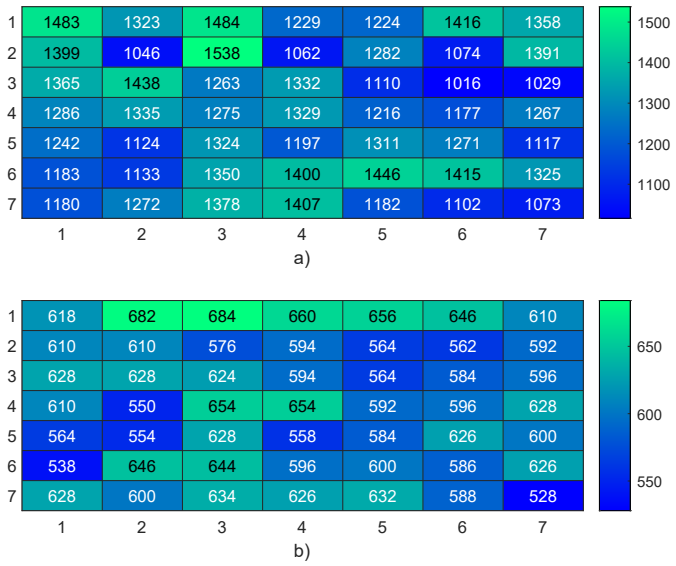


Figure 6.7: Traffic distribution of the demand plane at each zone or TDP with a total of 30000 UEs: a) traffic demand in Mbps, b) number of UEs

Propagation Model

The use of an adequate propagation model is fundamental for analyzing a realistic system; especially in a millimeter-wave 5G environment. In this platform, a 3D ray-tracing map-based propagation model is employed [78]. It is similar to the METIS model for urban macro and micro cells [102]. This channel model provides correlation and spatial consistency because it employs deterministic and physical principles accounting for scattering mechanisms, such as diffraction, scattering, and blocking.

It is important to mention that the proposed analysis considers downlink and outdoor communications. In particular, it is unfeasible to provide indoor communications at 28 GHz. Table 6.4 summarizes the parameters of the

presented validation.

The air interface establishes not only the considered propagation model but also the RRH–TDP association, which has been described by the binary variables $x_{r,z}$, including cooperation among RRH to serve the same TDP or demand zone, as represented in Fig. 6.6.

Table 6.4: Features of the scenario

Parameters	**Values**
Area (m^2)	1025×1455
Resolution (m)	5
$\lvert\mathcal{R}\rvert$	41
$\lvert\mathcal{M}\rvert$	8
$\lvert\mathcal{S}\rvert$	33
P_r (dBm)	$(43, 24)^\dagger$
G_{r} at FR1 (dB)	$(18, 2)^\dagger$
G_{r} at FR2 (dB)	12
G_{UE} (dB)	0
L^{RRH} (dB)	1
L^{UE} (dB)	1
$\xi^{\max}$	0.8
κ^{adj}	1
Propagation Model	3D ray-tracing

†The format of the data is (MRRH,SRRH)

6.3.2 Performance analysis and discussion

As 3GPP includes different split options of the protocol stack between BBU and RRH to reduce bandwidth and latency requirements, it is fundamental to consider them in the optimization process. For this reason, the following splits are analyzed: split option 8 that corresponds to a fully centralized C-RAN; split 6, where MIMO precoding and Orthogonal Frequency-Division Multiplexing (OFDM) modulation are maintained at the RRH side; and split 1 that represents a traditional architecture where all the baseband functions are allocated at the RRH (see Fig. 4.4).

Besides, a comparison of results when SRRHs operate at different frequency ranges (FR1 and FR2) is included. In particular, MRRHs operate at 2.6 GHz, while SRRHs could work at 3.6, 5, and 28 GHz. For simplicity, only the 3.6 and 28 GHz cases are presented since there are no significant differences between the results obtained at 3.6 and 5 GHz.

Finally, it is interesting to stress the algorithm considering different traffic loads. For this purpose, three data traffic options have been considered: Low, Medium, and High Traffic patterns, with 30000, 60000, and 300000 UEs, respectively.

The simulations and modeling have been carried out in MatLab. Specifically, the convex programming software CVX [88] and the Mosek solver [103] has been used to solve the optimization problem. An MSI Prestige 15 A10SC computer, with a Core i7 10th gen. CPU and 32 GB of RAM, have been used to carry out the simulations.

6.3.3 Cost reduction

As it has been explained in subsection 5.4.1, the cost is normalized using the maximum cost of the considered split option, which is $C_{\mathrm{S}}(\sigma\,|\mathcal{M}|+|\mathcal{S}|)$, where σ represents the ratio between the cost of MRRHs and SRRHs. The value of σ changes depending on the split option. For split 8 (fully centralized C-RAN), it corresponds to a value of $\sigma = 1$, whereas for split options 6 and 1 the assigned values are $\sigma = 10$ and $\sigma = 50$, respectively. However, this parameter should be adjusted according to the cost of the available devices. Cost differences associated with power amplifiers and antennas have not been considered in the value of σ, nor the additional cost of the hardware equipment when working at higher frequencies, because the purpose is to measure the impact of different split options. However, they could be easily included by changing σ values.

It is fundamental to fix the weights ω_1 and $\omega_3 = 1-\omega_1$ to solve the multi-objective optimization problem. Equation (5.38a) shows that these weights are associated with the cost and coverage-QoS optimization, respectively. The considered values for ω_1 cover the range from 0 to 1 with a step of 0.2. For instance, if $\omega_1 = 0.2$ and $\omega_3 = 0.8$, the algorithm provides more importance to coverage than cost reduction optimization.

Table 6.5 shows the comparison before and after running the optimization algorithm for the combinations of the considered parameters. The first three columns indicate the traffic pattern (Low, Medium, or High), the frequency range for the SRRH, which can be 3.6 GHz or 28 GHz, and the considered splits (8, 6, and 1). The fourth column displays the cost of the deployment before the optimization, that is, assuming that all the RRHs in the scenario

are active. Values under the Normalized Cost columns give the normalized cost factors after the optimization for different ω_1: from 0, meaning that the optimization is focusing on coverage-QoS, to 1, meaning that the optimization is focusing on cost reduction. Absolute cost values could be obtained by multiplying the normalized factor by the cost value before optimization. The final columns under the Coverage-QoS label provide the percentage of covered zones after the optimization.

Table 6.5: Resume of the optimized cost and coverage-QoS for different weights, frequency bands, split options, and traffic profiles.

Traffic	**Frequency**	**Split**	**Maximum Cost (CU)**	**Normalized Cost**						**Coverage-QoS**					
				ω_1											
				0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
LT	FR1	8	41	0.68	0.63	0.61	0.51	0.29	0.29	1	1	1	0.88	0.51	0.51
		6	113	0.65	0.31	0.22	0.21	0.20	0.12	1	0.98	0.96	0.96	0.94	0.51
		1	433	0.63	0.06	0.06	0.06	0.06	0.03	1	0.94	0.94	0.94	0.96	0.51
LT	FR2	8	41	0.68	0.34	0.29	0.27	0.07	0.07	1	1	1	1	0.57	0.55
		6	113	0.73	0.29	0.10	0.12	0.08	0.03	1	1	0.96	0.96	0.92	0.51
		1	433	0.74	0.26	0.03	0.03	0.03	0.01	1	1	0.96	0.96	0.96	0.51
MT	FR1	8	41	1	0.98	1	0.61	0.61	0.61	0.84	0.82	0.84	0.51	0.51	0.51
		6	113	1	0.38	0.29	0.29	0.22	0.22	0.84	0.69	0.67	0.67	0.51	0.51
		1	433	1	0.19	0.08	0.08	0.08	0.06	0.84	0.69	0.67	0.67	0.67	0.51
HT	FR2	8	41	0.73	0.51	0.49	0.39	0.22	0.22	0.94	0.96	0.96	0.90	0.55	0.51
		6	113	0.27	0.19	0.17	0.17	0.14	0.08	0.94	0.96	0.94	0.96	0.90	0.51
		1	433	0.07	0.05	0.05	0.05	0.05	0.02	0.94	0.92	0.94	0.94	0.96	0.51

The data from Table 6.5 can be used to extract multiple conclusions:

- Assuming that only the solutions with a final Coverage-QoS higher than 95 % are acceptable, it is possible to see that some combinations of parameters should not be used. It is the case of Medium Traffic (60000 users) at 3.6 GHz where, regardless of the split option and the considered weights, the requirements are never achieved. Even when

Coverage-QoS is prioritized ($\omega_1 = 0$), results show that all the 41 RRHs need to be active, but the maximum achieved Coverage-QoS is only 84 %. For this reason, the combination High Traffic (300000 users) at 3.6 GHz is not analyzed, as it is known in advance that this combination will never accomplish the coverage-QoS requirement.

- On the other hand, there is always a solution that guarantees a Coverage-QoS higher than 95 % in the remaining combinations, and in most cases, there is more than one solution. If this is the case, the best one in terms of cost reduction is the solution associated with the higher ω_1 value, because it is the solution that maximizes the cost reduction of the scenario, allowing for a higher number of inactive RRHs with the consequent energy-saving.

- Cost reduction is indirectly given in Table 6.5 as the complementary of the Normalized Cost value. For example, in the first row, the value is 0.68 for LT, FR1, split 1, and $\omega_1 = 0$. In this case, the algorithm provides a cost reduction of 32 %.

6.3.4 Cost vs Coverage-QoS

Table 6.5 shows the trade-off between the achieved cost reduction and the Coverage-QoS of the UEs in the scenario. Fig. 6.8 illustrates this trade-off. Each circumference represents a different Coverage-QoS percentage, starting with 50% for the most internal, meaning that only 50% of the TDPs of the scenario has been covered with the required QoS, to 100 % for the most external, meaning that all the coverage-QoS requirements have been fulfilled.

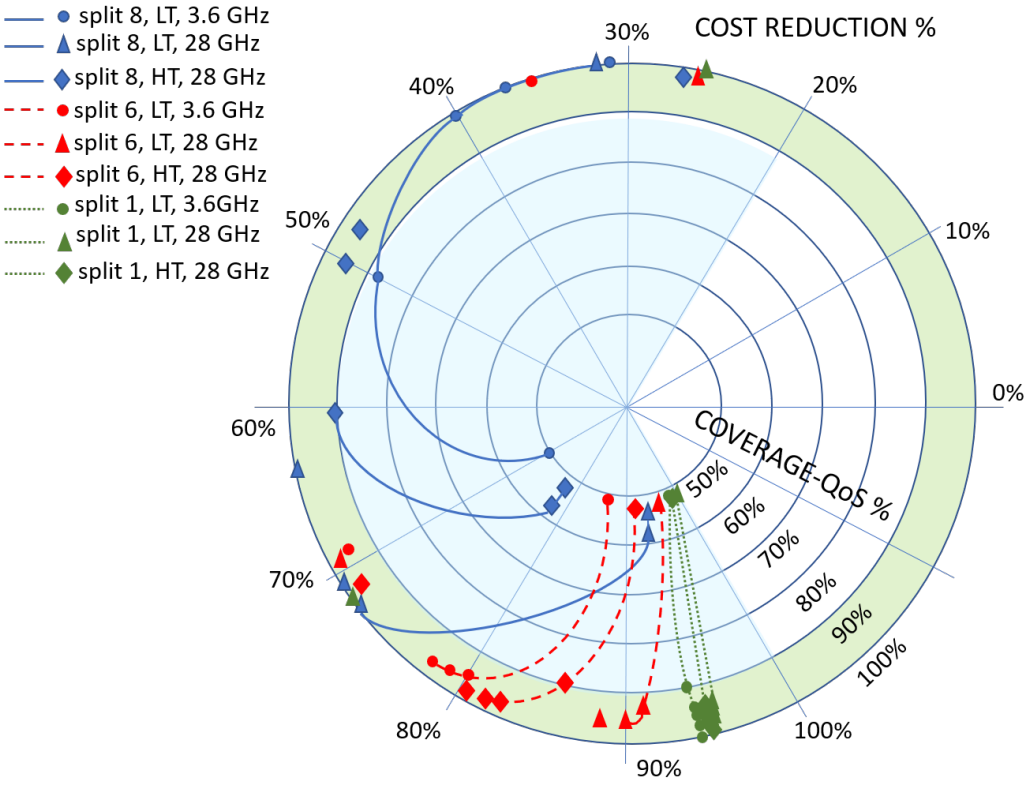


Figure 6.8: Coverage-QoS and Cost Reduction trade-off after running the optimization algorithm

On the other hand, each radial shown in Fig. 6.8 represents a different cost reduction value, ranging from 0 to 100 %, written at the edge of the radial. Remember that cost reduction is calculated with respect to the maximum cost, obtained when all the RRHs remain active and shown in Table 6.5. The colored region in green is the area where the coverage-QoS is higher than 90 %. The light blue area represents the region where the cost reduction is higher than 20 %. Each point (circles, triangles, or rhomboids) is obtained after running the optimization and represents the result for different ω_1 val-

ues. Moreover, the points closest to the most internal circumference are associated with the maximum value ω_1=1. Additionally, the distance to the center of the circumference increases as the value of ω_1 decreases, indicating that coverage-QoS is gaining priority with respect to cost reduction. The blue, red, and green lines in Fig. 6.8 connect the points with the same input simulation parameters (FR, split, and traffic level).

The blue-continuous lines represent the performance associated with split 8, while red-discontinuous and green-punctured lines represent split 6 and 1, respectively. The cases in Table 6.5 that do not achieve a good coverage-QoS after the optimization have not been represented in Fig. 6.8, as they are not considered valid solutions.

The best solution for each case (above 95 % of coverage-QoS) is represented by the symbol located at the outermost end of the line. There are other symbols of the same type showing a better Coverage-QoS, even in some cases close to 100 %, at the expense of an increasing cost. They are represented in Fig. 6.8 by the corresponding symbols, but they appear isolated (not connected to the line) to distinguish them. Despite this analysis, the MNO could select the solution point that best reflects the network requirements, addressing the trade-off between coverage-QoS and cost reduction.

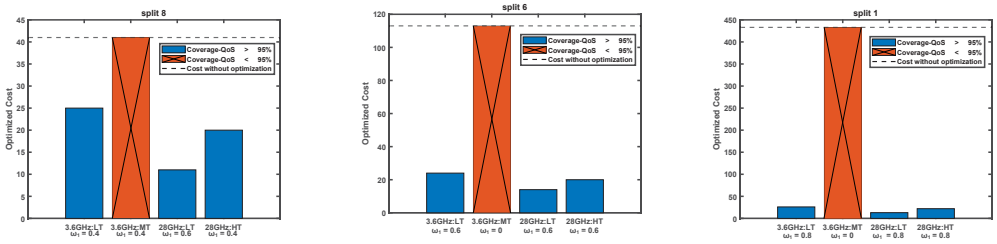
Firstly, it should be appreciated that split 1 provides the highest cost reduction (around 95 %) while offering a good coverage-QoS. This extreme cost reduction is due to the higher cost of the RRHs, as they contain all the baseband functionalities. Additionally, split 6 shows cost reductions of around 80-90 %, while split 8 exhibits cost reductions of 70-50-40 % for the different combinations of carriers and traffic. The cost is also reduced when

moving to higher frequency bands because, as wider bandwidths are assigned to the RRHs, they are able to serve more UEs. However, if some system parameters change (for example antenna gains or transmitted power), the number of RRHs needed to satisfy receiver sensitivity requirements, could increase when working at 28 GHz.

On the other hand, analyzing LT and Medium traffic (MT) cases at 3.6 GHz, it is shown that when the traffic profile is close to the maximum capacity of the whole network, the cost reduction decreases since most of the RRHs should be active to satisfy the demand. This behavior is similar at 28 GHz. However, in this case, the algorithm saves at least 20 % of the network cost because of the higher capacity of the network at FR2. The cost reduction reaches approximately 75 % when LT demand is considered, while the coverage-QoS reaches 100 %.

Fig. 6.9 complements the previous results by representing the minimum cost that guarantees a Coverage-QoS higher than 95 % after solving the multi-objective problem where Fig. 6.9(a), 6.9(b) and 6.9(c) stand for split 8, 6, and 1 respectively.

Fig. 6.9 also shows the cost values before optimizing, to facilitate the comparison. The blue bars represent valid solutions, while the red bars represent solutions that do not satisfy the 95 % of Coverage-QoS and neither reduce the cost. In terms of absolute cost deployment, optimal resource management, and computational capacity efficiency, split 8 is the best option for C-RAN networks, as has been widely shown. As the cost to deploy a new RRH is lower than with splits 6 and 1, the cost reduction when turning off an RRH is also lower; however, it is still a significant reduction. It is fundamental



(a) Split 8 (b) Split 6 (c) Split 1

Figure 6.9: Minimum cost and corresponding weights for the different splits, frequency bands, and traffic patterns

to notice that cost reduction is directly associated with an increase in the energy efficiency of the network. The lower the number of RRH required to satisfy the UEs requirements, the higher the energy-saving.

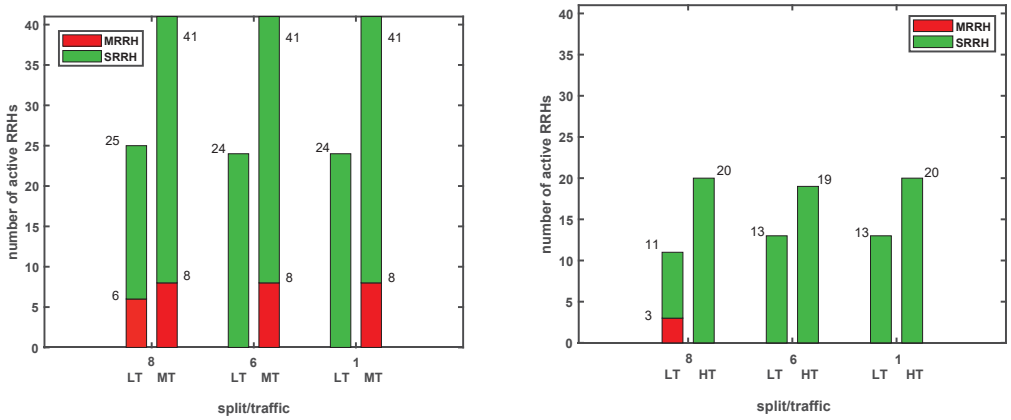
To summarize the analysis, it has been shown that the proposed algorithm is highly efficient allowing practical cost deployment reductions between 20 to 70 % depending on the traffic level (Low, Medium, High), carrier frequency used, and selected split option.

6.3.5 Active RRHs and usage ratio reduction

The proposed optimization framework is worthy for the MNO, not only in the deployment phase but also to select the RRHs that should be active to satisfy the current traffic demand or even the predicted traffic demand. This could be achieved by combining the present algorithm, or the look-up tables that can be generated after running it, with optimized AI prediction tools allowing to analyze a dynamic scenario.

Additionally, it contributes enormously to energy-saving, a key parameter

for 5G and future 6G networks. The number of required active RRHs after the optimization is shown in Fig. 6.10 for each split option, being Fig. 6.10(a) for the 3.6 GHz carrier frequency, while Fig. 6.10(b) shows the 28 GHz results. The first bar of each split corresponds to LT profile, while the second bars stand for MT or High traffic (HT), depending on the figure. Remember that the initial situation, without optimization, uses the 41 RRHs of the scenario, 8 of them MRRHs. The presented solutions correspond to the weights considered in Fig. 6.9, which guarantee a 95 % Coverage-QoS while reducing simultaneously the cost.



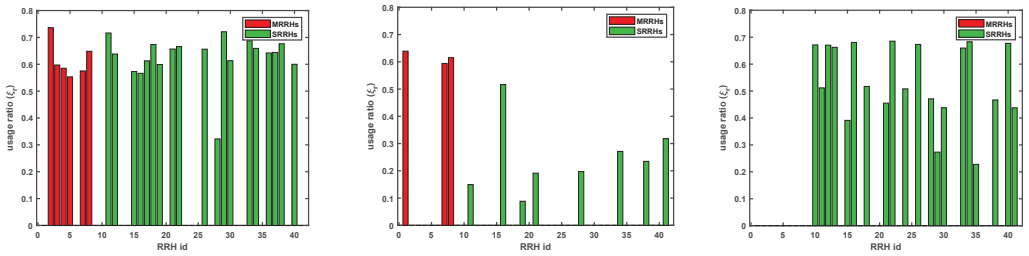
(a) 3.6 GHz

(b) 28 GHz

Figure 6.10: Number of active RRHs vs split options. Left and right bars of each split option represent LT, and MT or HT cases, respectively.

As expected, regardless of the frequency band, the number of active RRHs increases with the traffic demand. On the other hand, the distribution of MRRHs and SRRHs is detailed, showing that the algorithm prioritizes SRRHs when σ increases, to reduce the cost. The MT simulation working at

3.6 GHz needs all the RRHs of the scenario to maximize the coverage. This MT solution is represented to show that the optimization algorithm could signal when the initially assigned resources are insufficient. In this case, to find a feasible solution, MNOs should increase the number of RRHs deployed or the bandwidth allocated to them.



(a) (b) (c)

Figure 6.11: Resource usage ratio (ξ_r) of the RRHs by the GBR slices in C-RAN (split option 8): (a) at 3.6 GHz with LT profile, (b) at 28 GHz with LT profile and (c) at 28 GHz with HT profile.

It is also interesting to show that in Fig. 6.10(a) the optimized solution ends with a similar number of required active RRHs(25-24), being the main difference that optimal split 8 requires keeping six active MRRHs while in splits 6 and 1 MRRHs are not needed at all. This also explains why in Fig. 6.9 there is a significant difference in cost between the three splits: in splits 1 and 6 the cost of an MRRH is very high compared to the cost of an SRRH, so the algorithm tries to avoid the activation of MRRHs when searching for the optimal solution.

The operation at 28 GHz shows an enormous reduction in the number of required active RRHs, around one-third of them are needed in LT condi-

tions while half of them are required when HT is considered. Propagation is not the limiting factor in the scenario because the cells are close enough and transmitted power and antenna gain are high enough to satisfy the UE requirements. However, only outdoor UEs have been considered, assuming that at 28 GHz, indoor users should be served by indoor Base Stations due to the large building penetration losses. The main difference with the 3.6 GHz operation is that the bandwidth associated with each RRH is higher.

The last fundamental parameter analyzed in this work is the usage ratio, (ξ_r), which has been previously defined as the ratio between the GBR traffic load at RRH r and its maximum capacity. In the simulations, 20 % of the resources of an RRH are dedicated to best-effort services, while 80 % is for the GBR services. The usage ratio for a fully C-RAN (split 8) and for the ω_1 values given in Fig. 6.9 is shown in Fig. 6.11. The red bars from id 1 to 8 correspond to the active MRRHs, while the remaining green bars represent the active SRRHs. Each figure is for a different frequency band and traffic pattern combination. Even in the most loaded case 6.11(a) most of the RRHs still have at least 30 % of remaining capacity that could be used to attend sudden network variations as new or handover UEs as well as cooperative beamforming. In those cases where the available capacity is not enough to serve a new TDP or zone, the capacity of several RRHs could be aggregated, using cooperation techniques that will improve the coverage and efficiency of the network.

Finally, Fig. 6.12 shows an example of a radio network deployment after applying the optimization process, in particular, a C-RAN at 28 GHz with $\omega_1 = 0.6$ and LT profile. The gray markers on Fig. 6.12 depict the RRHs

Figure 6.12: Resulting radio network for a C-RAN at 28 GHz with $\omega_1 = 0.6$ and LT profile.

that have been deactivated from the original and non-optimized network deployment (see Fig. 6.5). It allows the reader to realize the advantage of optimization and to analyze the resulting network distribution.

6.3.6 Performance analysis considering cell cooperation

This subsection presents the algorithm performance when cooperation ($\mu = 2$, in the analysis) is considered. This assumption allows satisfying the aggregated traffic of a single zone through two RRHs. The computational complexity of the algorithm drastically increases when cooperation is considered. For this reason, the presented results only consider particular cases of

the scenario. However, the performance is not limited to these conditions. Especially, FR1 with LT and FR2 with HT in 6.9(a) were considered, in order to analyze both frequency bands and traffic profiles.

Table 6.6: Cost vs Coverage-QoS comparison

	$\mathbf{Cost}_1$	$\mathbf{Cost}_2$	$\mathbf{Coverage}_1$	$\mathbf{Coverage}_2$
FR1, LT	0.61	0.58	1	1
FR2, HT	0.49	0.46	0.96	0.98

Table 6.6 shows the results with and without cooperation, where subscripts 1 and 2 stand for the value of μ. It is important to mention that due to the computational complexity of the algorithm with $\mu = 2$, its execution has been finished by time, so it cannot be said that they are the optimal solutions. With a larger simulation period probably the cost would experience an additional reduction. But it is already possible to see an improvement of 3% in cost reduction with respect the non-cooperative case. Moreover, in the case of 28GHz with a high traffic profile, the coverage is 2% higher.

6.4 Conclusions

In this chapter, two different scale and realistic scenarios of the C-RAN simulation platform C-RAN have been considered. Firstly, a fronthaul deployment based on four optimization strategies and using a simple per-hour traffic profile is presented in order to establish the BBU-RRH connections. Additionally, a study of the performance of large C-RAN designs, in which only a percentage of the cells, is centralized to reduce the investment is also

presented. It may support network operators to implement an optimal design accounting for the cost of the optical fiber, the area to be covered, and the density of users.

The design of a C-RAN deployment of a region of the metropolitan area has been analyzed. The proposed algorithm is tested by using a realistic scenario that includes 41 possible RRHs in a heterogeneous deployment with MRRHs and SRRHs, UEs modeled with different services, and an accurate 3D ray-tracing propagation model. Additionally, operation at frequency bands 3.6 and 28 GHz, as well as different C-RAN split options are studied.

The overall power consumption of future mobile networks should not grow beyond what it is now for 5G. For this reason, a strategy that provides a sustainable optimal deployment not only for 5G but also for B5G radio networks has been provided. The main objectives are to reduce the footprint on energy efficiency, and the deployment and operational costs of the network while maintaining the coverage-QoS. This complex problem has been modeled, introducing a Multi-objective ILP optimization algorithm to select the optimum distribution of the RRHs in the densest zone of the city of Vienna.

It is impossible to briefly summarize the results because multiple parameters could be compared after the optimization. However, it is possible to resume some key aspects. For instance, the algorithm reduces the deployment cost while maintaining the coverage-QoS better than 95 %. Especially, at 3.6 GHz with low traffic demand, the cost reduction is around 35 %, while at 28 GHz it reaches 70 % with LT profile and almost 50 % under an HT condition.

The integration of the C-RAN platform and the algorithms described in

this chapter could help the MNOs to improve their network planning, not only the RRH deployment but also the analysis and design of the fronthaul link.

Finally, the proposed platform will be upgraded by integrating prediction tools based on AI to efficiently manage the resources centralized at the BBU pools. This approach is detailed in chapter 7.

Chapter 7

DRM-AC: Analysis and Discussion

7.1 Introduction

As has been mentioned in chapter 2, most of the previous works on BBU pool resource management have relied on the definition of optimization problems such as MILP or MOO. However, these strategies allocate the resources assuming that the instantiated computational capacity at BBU pools is fixed and equal to the maximum BBU pool capacity. The computational resources could be over-provisioned or under-provisioned under this assumption, causing inefficient resource utilization or QoS degradation.

This issue could be addressed by combining the flexibility of virtualization and the availability of machine learning techniques to predict computational demands. As the resources are virtualized, they could be instantiated dynamically according to an anticipated computational capacity demand.

To do this, a DRM with adaptive capacity (DRM-AC) was defined in the section. 5.3. This section presents an analysis of the DRM-AC performance and also the results of a classical DRM with fixed capacity used as a benchmark to establish a comparison.

Three ML algorithms have been analyzed: Support Vector Machine (SVM), Time Delay Neural Network (TDNN), and Long Short-Term Memory (LSTM). In general terms, the DRM-AC reduces the average of unused resources by 96 % in the considered scenario, but there is still QoS degradation when RCC is higher than the predicted computational capacity (PCC). However, DRM-AC-PF and DRM-AC-ES address this issue, reducing the average of unsatisfied resources by 98 % and 99.9 % compared to the DRM-AC, respectively. The presented results are a combination of multiple research works that have been gradually published [89–92].

7.2 Simulation conditions

This section describes the simulation conditions that have been considered to validate the DRM-AC algorithms. Further analysis with different parameters could be carried out if needed thanks to the flexibility of the developed platform.

The analysis presented in this section employs the metropolitan area of Vienna introduced in Fig. 4.2. On the other hand, a fully centralized C-RAN or split option 8 has been considered (see Fig. 4.4). The RRHs are only responsible for transmitting/receiving the in-phase and quadrature components of the signal to/from the BBU pool, being the remaining functionalities

centralized at the BBU pools.

As has been mentioned above, the propagation model is a 3D ray-tracing map-based model [78], which is similar to the METIS model for urban macro and micro cells [102]. The model automatically provides all the correlations and spatial consistencies employing deterministic and physical principles accounting for scattering mechanisms, such as specular reflections, diffraction, scattering by rough surfaces and objects, and blocking.

For simplicity, RRHs to BBU pool connections (fronthaul links) are established by minimizing the delay. This strategy is mathematically defined in section 5.2, while the advantages and inconveniences of this assumption are analyzed in section 6.2.

On the other hand, 7000 UEs are randomly placed in the scenario. The UEs generate conversational, streaming, and interactive services based on a packet level model used in [4, 24, 81]. The details of the service model have been presented in subsection 4.2.1, and table 4.3 summarizes the service main parameters while table 7.1 gives the considered network parameters.

The traffic load of the network is estimated based on the required computational capacity as described in 4.2.2

7.3 DRM performance evaluation

To analyze the performance of the DRM, the maximum capacity at each BBU pool is fixed at 300, 100, and 300 GOPS, respectively. The capacity distribution is intentionally asymmetric to analyze the DRM with under and over-provisioned situations. Fig. 7.1 shows the total required capacity at each

Table 7.1: Simulation parameters for DRM.

Parameters	Value
Area (km^2)	25
Sites	228
MRRHs (sites)	51(17)
SRRHs (sites)	221(211)
BBU pools	3
RRHs	272
Power (dBm)	(43,24)*
Quantization resolution (bit)	(24,16)*
RRH antenna gain (dB)	(18,10)*
Bandwidth (MHz)	20
Number of RBs	100
Total UEs	7000
UEs antenna gain (dB)	0

* The format of the data is (MBSs,SBSs)

BBU pool RCC $= \sum_i C_{i,t}$ and the total allocated capacity ACC $= \sum_i \text{ACC}_{i,t}$. ACC at BBU pools 1 and 3 equals the total required computational capacity.

It can be clearly appreciated that the capacity given to BBU pool 2 is not enough to handle the traffic demand. The key performance indicator (KPI) to quantify the QoS is defined as the ratio between the allocated capacity and the required capacity ($KPI \in [0, 1]$).

At BBU pool 2 for most of the simulation time, the required capacity is higher than the maximum capacity, which results in a degradation of the

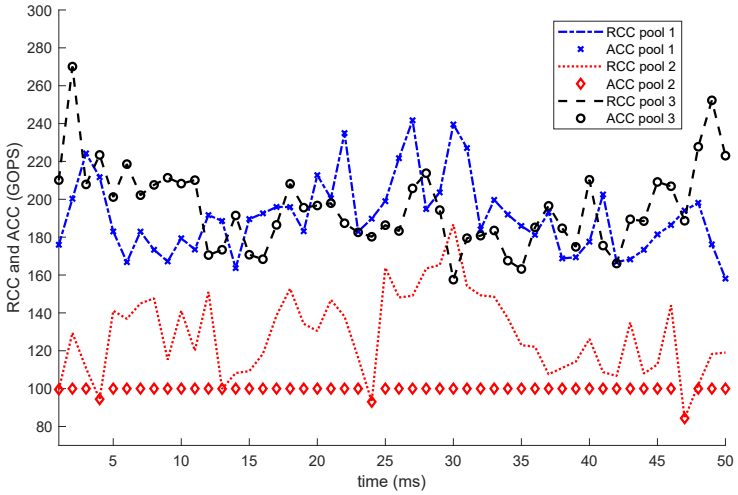


Figure 7.1: Total required and allocated computational capacity for each BBU pool.

QoS. Fig. 7.2 represents the percentage of unsatisfied resources per RRH, which is calculated as $1 - KPI$. Details are given only for BBU pool 2. It can be appreciated that many RRHs experience a high dissatisfaction level which corresponds to a low QoS.

The computational capacity of BBU pool 2 has been intentionally selected low with the aim of highlighting how the proposed algorithm is powerful enough to reveal clearly those cases that have not been appropriately designed. The influence of the bargaining power is observed in Fig. 7.2. Notice that at the same time, there are BBUs with different QoS, because the optimization algorithm is allocating more resources to the cells with high-priority services.

BBU pools 1 and 3 have enough capacity to handle the demand. However, due to this capacity being fixed there are intervals where it is underutilized as Fig. 7.3 remarks. Consequently, there is a trade-off between QoS degradation

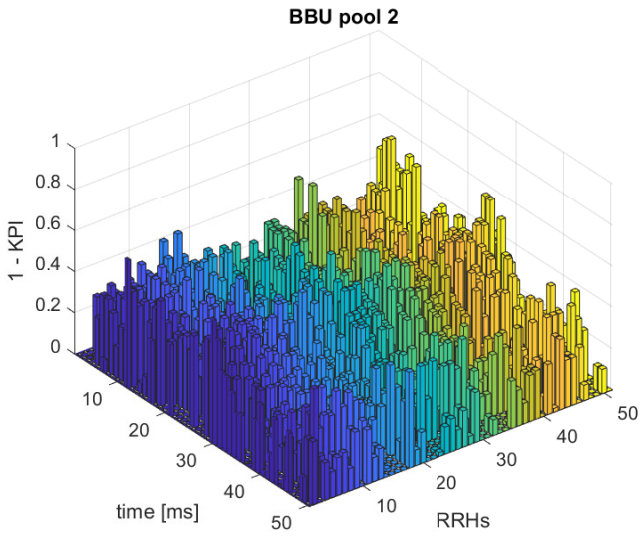


Figure 7.2: Temporal QoS per RRH in BBU pool 2.

when the computational capacity is under-provisioned and the inefficient use of the resources when the network is over-provisioned. For this reason, intelligent resource management tools based on ML approaches, where the system is able to learn from past situations to proactively predict the traffic demand, are required to optimize future dynamic infrastructure networks. The next subsections present how the proposed DRM-AC, DRM-AC-PF, and DRM-AC-ES address this trade-off in the BBU pool 1.

7.4 DRM-AC performance and ML models configuration

This section presents the configuration and comparison of the ML models as well as the analysis of the performance of the proposed adaptive resource

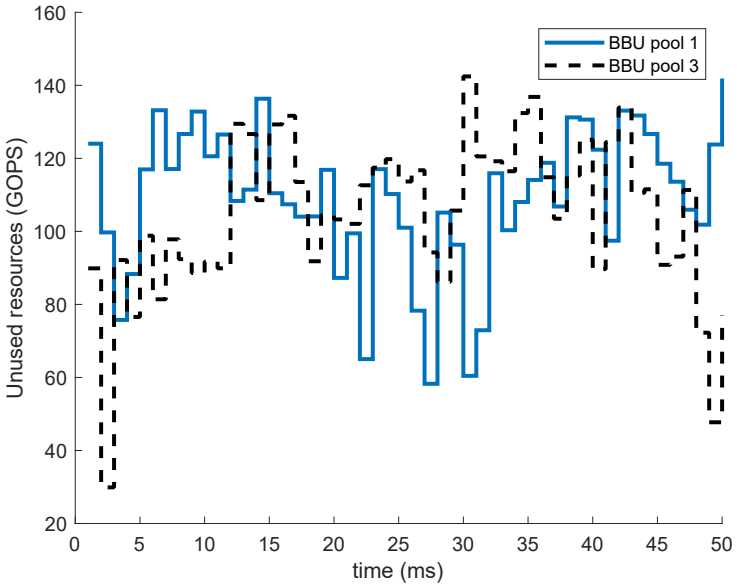


Figure 7.3: Amounts of unused resources in BBU 1 and 3.

management strategies using these models.

7.4.1 ML models and data analysis

As it has been above-mentioned, ML-based resource management tools are required to optimize the use of the resources at BBU pools. In this case, the system would be able to learn from past situations to proactively predict traffic demand. This subsection describes the database, and it establishes the simulation conditions of the supervised learning techniques (SVM, TDNN, and LSTM) in the DRM-AC.

Data Configuration

For simplicity, the analysis of the forecasting models has been limited only to BBU pool 1, and one minute of traffic database is generated. Fig.7.4 shows the database, which is split into a training set (first 80 %) and a testing set (the remaining 20 %); the dotted line indicates the boundary between those sets.

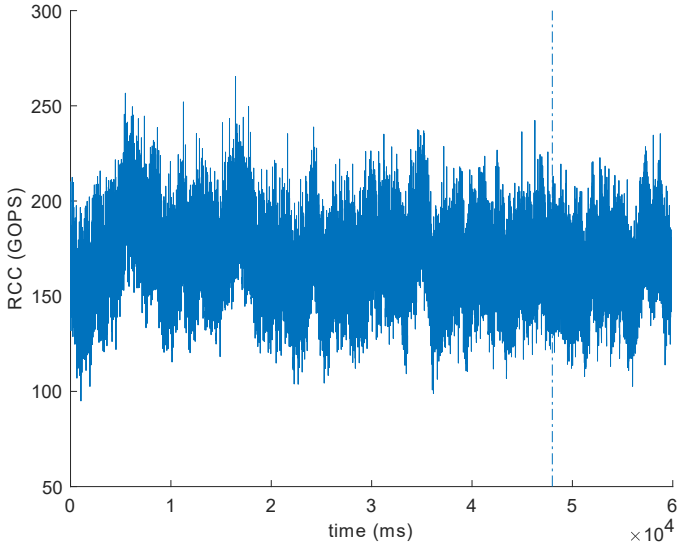


Figure 7.4: Instantaneous evolution of the RCC at BBU pool 1. Database of 60000 samples. First 80 % of the data is used as a training set and the remaining 20 % as a testing set.

Models Configuration

SVM and TDNN models predict the RCC based on a set of previous time steps. Hence, an analysis of how many previous time steps are required to predict the RCC is necessary. The first approximation is carried out

by the calculation of the sample partial autocorrelation function (PACF), represented in Fig. 7.5. PACF values are split according to their amplitudes in high and low contribution with a threshold of 10 % of the maximum value. The PACF decreases with the number of previous time steps, with the exception of some isolated values (four samples after 250 ms). The cumulative distribution function (CDF) of the high contribution values (CDF 1) is shown on Fig. 7.5, the 78 % of the values are located before 150 ms. Furthermore, the CDF of the high contribution values without concerning the isolated samples after 250 ms is also shown (CDF 2), where 97 % of the samples are before 150 ms. Based on this fact, the previous 150 ms is considered as a significant time window to adjust this parameter in SVM and TDNN.

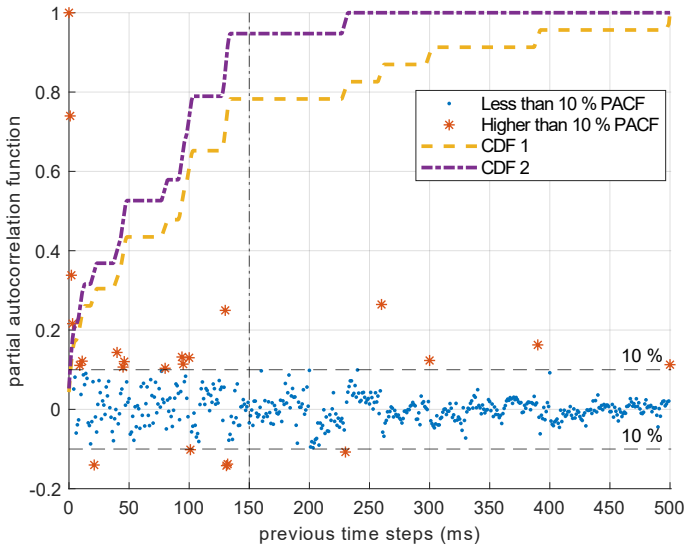


Figure 7.5: Partial autocorrelation function of the database concerning 500 previous time-steps.

After testing multiple configurations of SVM and TDNN, the best results

were obtained using SVM with a Gaussian kernel and TDNN with two hidden layers of 10 neurons and sigmoid as the activation function. Fig. 7.6 shows the root-mean-square error (RMSE) of SVM and TDNN using different amounts of previous time-steps until 150 ms. The RMSE decreases when the number of previous time-steps increases; however, after 100 ms and 130 ms in SVM and TDNN respectively, the RMSE remains almost constant. This behavior shows that the convergence of TDNN and SVM is improved by increasing the number of previous time steps until those limits. For this reason, only $\theta = 100$ ms and $N = 130$ ms previous time-steps are considered in the subsequent analysis. Nevertheless, the method based on PACF is shown to be a perfectly valid rule-of-thumb, and there would be no need to test each case.

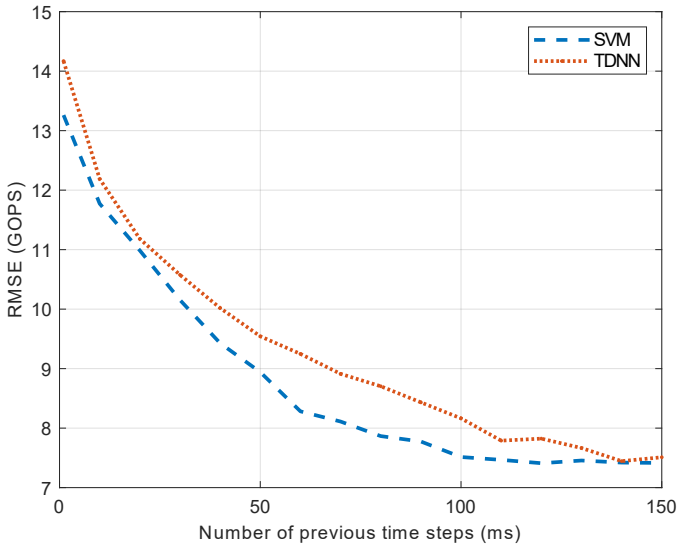


Figure 7.6: Gaussian SVM and TDNN performance in terms of the number of previous steps.

Regarding the LSTM approach, its performance does not depend on the number of previous time steps because their contribution is saved in the internal gates of the LSTM cell. However, different network architectures were tested and compared to find a suitable deep learning scheme. Table 7.2 summarizes those architectures. Two hidden layers with different numbers of LSTM cells, where the learning process takes place, are used. Following [104] recommendation, dropout layers (with a dropping probability of 0.2) are used after each hidden layer to prevent overfitting. Finally, a regression output layer is aggregated to map the output of the last hidden layer to a predicted value.

Table 7.2: Tested deep learning LSTM architectures

	Network structure : index									
	1	2	3	4	5	6	7	8	9	10
L1	Sequential input layer									
L2	Hidden layer: number of LSTM cells									
	20	40	60	80	100	120	140	160	180	200
L3	Dropout: probability of dropping out 0.2									
L4	Hidden layer: number of LSTM cells									
	10	20	30	40	50	60	70	80	90	100
L5	Dropout: probability of dropping out 0.2									
L6	Regression output layer									

Fig. 7.7 shows the performance of the network structures in Table 7.2, based on the RMSE achieved in the testing dataset (last 20 % of the data). The RMSE decreases when the number of LSTM cells increases, reaching

its minimum value for network structure number four. For this reason, this structure is selected for comparison with SVM and TDNN strategies. It has less computational cost than higher network structure labels. The RMSE under this architecture is 12.6 GOPS, which represents 7.6 % of the mean value.

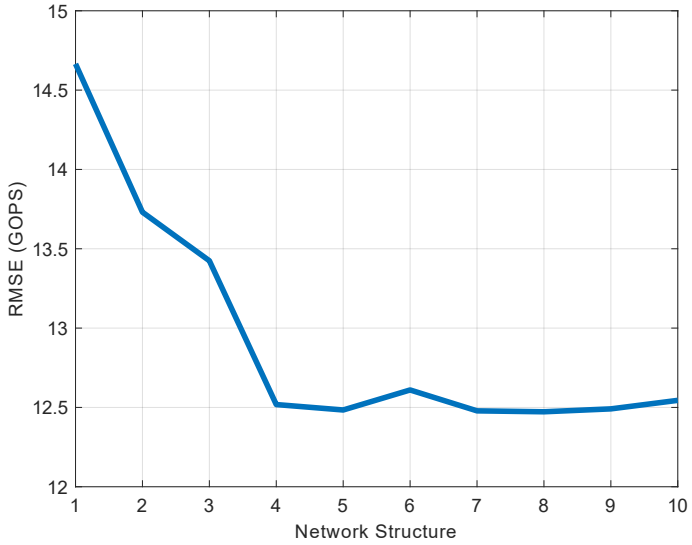


Figure 7.7: Performance (RMSE) on the testing data of the deep learning LSTM architectures in Table 7.2.

7.4.2 DRM-AC performance evaluation

This subsection analyzes the performance of the DRM-AC using the DRM with fixed capacity as a benchmark.

Fig. 7.8 summarizes the convergence of the DRM-AC using the ML approaches, which are analyzed in terms of the error distribution and to what extent the prediction is close to the perfect prediction. Fig. 7.8(a), 7.8(b) and

7.8(c) show the predicted computational capacity in terms of the real computational demand of each strategy. Most of the predicted values are close to the perfect prediction line, being the degree of dispersion indicator of the quality of the prediction strategy and the convergence of the algorithms. The maximum error of SVM and TDNN approaches is around 35 GOPS, and the RMSE is close to 7.5 GOPS, which represents a deviation of 4.5 % of the mean value of the overall dataset. The Pearson correlation coefficients (slope of the regression line) are 0.92 and 0.91 for SVM and TDNN, respectively. On the other hand, the LSTM strategy presents a RMSE = 12.6 GOPS that depicts the 7.6 % of the mean value and the Pearson coefficient is r = 0.7, which is more deviated from the perfect prediction line.

Fig. 7.8(d) shows the error distribution of each approach. Regardless of the used strategy, the error distribution is almost a Gaussian curve with zero mean. As the ML algorithms predict the required computational capacity at the BBU pool, it is important to analyze the effect of these errors. Positive errors (right side of perfect prediction line on Fig. 7.8(d)) represent the number of underutilized resources, while negative errors are the amounts of unsatisfied resources. The main objective is to minimize the underutilized resources while maintaining the QoS. Improving the prediction capacity of the machine learning strategies is not enough to address this challenge because negative errors always reduce the QoS. Table 7.3 summarizes the behavior of the three proposed strategies.

SVM and TDNN improve the performance of LSTM in 3 %. However, as it is possible to see in Fig. 7.6, the behavior of SVM and TDNN strongly depends on the number of previous time steps used in the prediction. As

mobile networks experience large fluctuations and they are not stationary processes, results obtained under the assumption of variable parameters as the number of previous time steps might be more robust. The design based on LSTM cells is an example; it obtains similar performance to Gaussian SVM and TDNN without requiring a fixed number of previous time steps. The useful information of the previous time steps is stored in the forget gates of the LSTM entities in the hidden layers.

Table 7.3: Summary of the proposed ML techniques.

ML technique	RMSE (GOPS)	RMSE (%)	Pearson coefficient
SVM	7.52	4.5	0.92
TDNN	7.45	4.47	0.91
LSTM	12.6	7.6	0.7

7.4.3 DRM-AC-PF and DRM-AC-ES performance

As it was aforementioned, the LSTM approach could be more robust to face high fluctuation environments. For this reason and without losing generality, the performance of DRM-AC-PF and DRM-AC-ES, reducing negative errors, are evaluated based on the LSTM approach.

Fig. 7.9 shows the performance of the solution applying DRM-AC-PF. The Max{} block extracts the envelope of the RCC acting as a low pass filter eliminating the fastest variations; the solid blue line represents the filtered computational capacity. The fixed capacity (300 GOPS) is also represented to remark the advantage of predicting the required computational capacity.

Negative errors cause QoS degradation. Fig. 7.10 exhibits the distribution of the errors of the proposed schemes using the same LSTM architecture. Although positive errors in DRM-AC-ES have increased, negative errors are almost eliminated. In the case of the DRM-AC-PF, the results are similar; negative errors appear only in isolated cases at the cost of increasing the positive error with respect to the original LSTM approach (LSTM DRM-AC on Fig. 7.10).

Two key performance indicators have been defined to facilitate a numerical comparison of the strategies: the mean of unused resources (MUR_+) and the mean of unsatisfied resources (MUR_-), calculated by (7.1) and (7.2), respectively.

$$\mathrm{MUR}_+ = \frac{1}{K}\sum_{j=1}^{K} e_j^+ \tag{7.1}$$

$$\mathrm{MUR}_- = \frac{1}{K}\sum_{j=1}^{K} e_j^-, \tag{7.2}$$

being K the number of time-steps in the whole database ($K = 60000$ ms), e_j^+ and e_j^- depict the absolute values of each kind of error at instant j in GOPS. Those errors are complementary because only one of them could be different from zero.

Table 7.4 shows the advantage of using each strategy in terms of MUR_+ and MUR_- key performance indicators. The DRM without adaptive capacity has an average of 138.56 GOPS/ms of unused resources. Under the considered traffic conditions and with a fixed capacity (300 GOPS) in BBU pool 1, the resources are enough to handle the instantaneous RCC ($\mathrm{MUR}_- = 0$). However, as the maximum capacity is fixed, if the RCC surpasses the max-

imum capacity at BBU pool 1, UEs would be in degradation; consequently, the MUR_{-} would increase, and the QoS would be degraded. DRM-AC reduces the MUR_{+} considerably (5.5 GOPS/ms), but the error in the prediction causes the instantiated resources to be insufficient to satisfy the demand (50 % of the time approximately). DRM-AC-PF and DRM-AC-ES strategies reduce considerably the MUR_{-} at the cost of increasing the average of unused resources but maintaining the UEs QoS.

Table 7.4: Performance summary in terms of the MUR_{+} and MUR_{-}.

Proposals	MUR_{+} (GOPS/ms)	MUR_{-} (GOPS/ms)
DRM	138.56	0
DRM-AC	5.5	4.49
DRM-AC-PF	34.08	0.072
DRM-AC-ES	41.15	0.0016

7.5 Conclusions

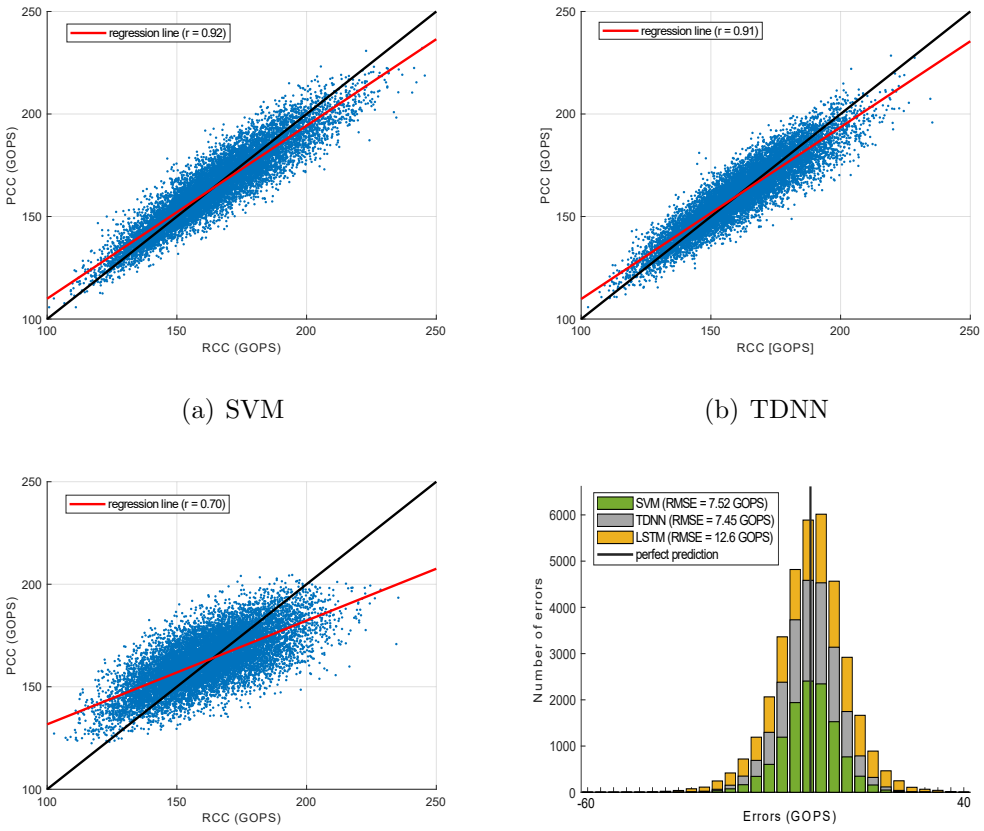
This chapter integrates ML techniques into a dynamic resource management in C-RAN to optimize the utilization of computational resources. Three ML strategies have been implemented and exhaustively compared: SVM, TDNN, and LSTM in terms of their ability to predict the instantaneous computational capacity at the BBU pools.

DRM-AC reduces the underutilized resources by 96 % when compared with the DRM with fixed computational resources. However, it degrades the QoS when the predicted computational resources are not enough to satisfy

the demand. This situation appears approximately 50 % of the time due to the error following a gaussian distribution with zero mean. This issue is solved by proposing two novel strategies.

Firstly, a DRM-AC with prefiltering is proposed, where high-frequency variations in input data are removed. DRM-AC-PF extracts the envelope of the RCC, improving the learning process, and it almost eliminates QoS degradation. Secondly, DRM-AC-ES monitors the maximum error computed in past observation times. This allows estimating a marginal amount of resources to be added to the predicted computational capacity.

As a consequence, DRM and DRM-AC are outperformed. DRM-AC-PF and DRM-AC-ES reduce the unsatisfied resources by 98 % and 99.9 % compared to the DRM-AC, respectively. Moreover, they reduce the number of underutilized resources by 75 % and 70 % compared to the DRM, respectively.



(a) SVM

(b) TDNN

(c) LSTM

(d) Comparison

Figure 7.8: Performance of the DRM-AC for each ML technique. (a), (b), and (c) show the predicted computational capacity in terms of the real computational demand of SVM, TDNN, and LSTM respectively. Black lines denote perfect prediction lines, the red line depicts the regression line and r is the Pearson correlation coefficient. (d) represents the histogram of the error distribution.

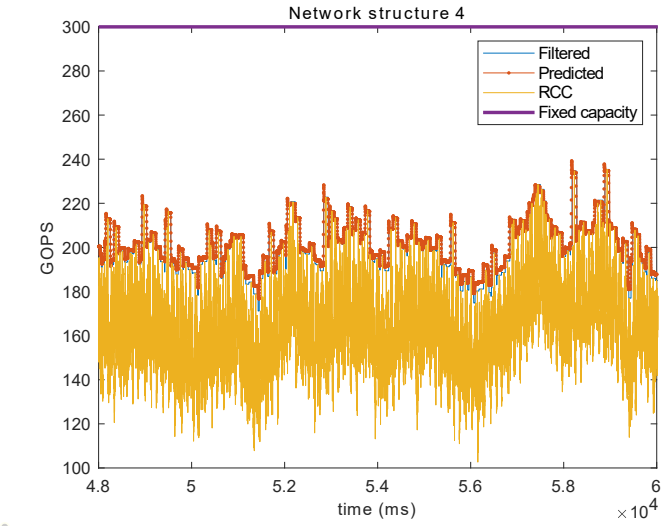


Figure 7.9: Evolution of the computational capacity at BBU pool 1 showing the fixed maximum computational capacity, the RCC during the testing dataset, the filtered RCC and the predicted computational capacity after applying DRM-AC-PF strategy.

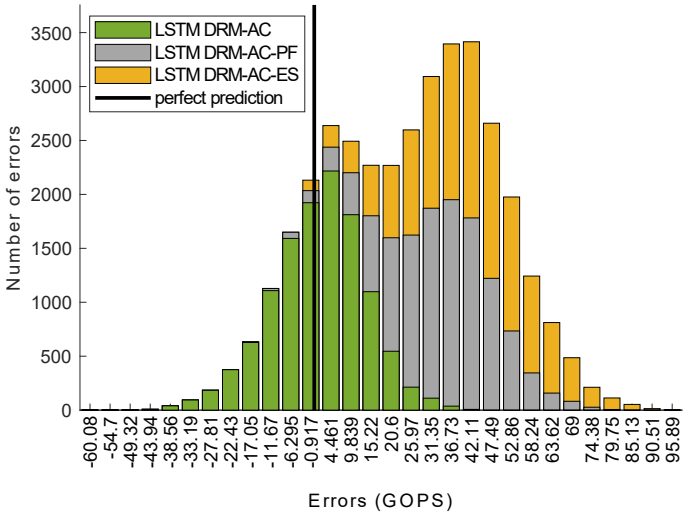


Figure 7.10: Error distribution: DRM-AC, DRM-AC-PF and DRM-AC-ES.

Chapter 8

Conclusions and Future Works

Overall power consumption of future mobile networks should not grow beyond what it is now for 5G. A strategy that provides a sustainable optimal deployment not only for 5G but also for B5G and 6G radio networks has been provided. The main objectives are to reduce the footprint on energy efficiency, and on the deployment and operational costs of the network while maintaining the coverage-QoS. This complex problem has been modeled, introducing a Multi-objective ILP optimization algorithm to select the optimum distribution of the RRHs in the densest zone of the city of Vienna.

The proposed algorithm is tested by using a realistic scenario that includes 41 possible RRHs in a heterogeneous deployment with MRRHs and SRRHs, UEs modeled with different services, and an accurate 3D ray-tracing propagation model. Additionally, operation at frequency bands 3.6 and 28 GHz, as well as different C-RAN split options are studied.

As so many parameters can be compared after the optimization, it is impossible to summarize the main results in a few words. Only mention that

the algorithm reduces the deployment cost while maintaining the coverage-QoS better than 95 %. Especially, at 3.6 GHz with low traffic demand, the cost reduction is around 35 %, while at 28 GHz it reaches 70 % with LT profile and almost 50 % under an HT condition.

This tool undoubtedly could help the MNOs to improve their network planning, providing network optimizing and controlling by allowing the operator to balance between coverage-QoS and cost reduction and consequently power consumption savings.

Moreover, this work also integrates ML techniques into dynamic resource management in C-RAN to optimize the utilization of computational resources. Three ML strategies have been implemented and exhaustively compared: SVM, TDNN, and LSTM in terms of their ability to predict the instantaneous computational capacity at the BBU pools.

The DRM-AC reduces the underutilized resources by 96 % when compared with the DRM with fixed computational resources. However, it degrades the QoS when the predicted computational resources are not enough to satisfy the demand. This situation appears approximately 50 % of the time because of the Gaussian distribution.

This issue is solved by proposing two novel strategies. First, a DRM-AC with prefiltering is proposed, where high-frequency variations in input data are filtered. DRM-AC-PF extracts the envelope of the RCC improving the learning process, and it almost eliminates QoS degradation. Second, DRM-AC-ES monitors the maximum error computed in past observation times. This allows estimating a marginal amount of resources to be added to the predicted computational capacity. As a consequence, DRM and DRM-AC

are outperformed. DRM-AC-PF and DRM-AC-ES reduce the unsatisfied resources by 98 % and 99.9 % compared to the DRM-AC, respectively. Moreover, they reduce the number of underutilized resources by 75 % and 70 % compared to the DRM, respectively.

Additionally, this research work is not limited to the presented contributions because it also opens the door for novel proposals or upgrades. Plenty of future research lines could be carried out to improve the proposed platform and algorithms. Some of the future research are summarized as follows.

Future Works

- Design an orchestration and management strategy that could efficiently guarantee an end-to-end Quality of Experience in the mobile networks. This strategy should handle the high complexity and variability of future wireless systems. For this reason, it must include AI techniques for decision-making and global management. This proposal could be tested in the proposed platform and will be a completely novel research line.

- From a practical point of view, the RRH deployment algorithm proposed in this thesis could be used in the current deployment of multiple MNOs, not only in the deployment phase but also to decide the active RRHs under certain traffic conditions. However, a study of the effect of turning off RRHs in the number of handovers and consequently in the QoS will be fundamental.

- On the other hand, a meta-heuristic solution of the proposed RRH de-

ployment optimization should be introduced because the computational cost of the solution exponentially increases with the number of RRHs and zones in the map. Simulated Annealing, genetic algorithms, and swarm optimization could be considered to compare the performance.

- The dynamic resource management with adaptive capacity presented in this thesis proposes to activate only the required computational capacity instead of considering a fixed maximum capacity at BBU pools or CUs. The models are introduced and tested considering a system-level simulation. However, the proposal should be tested on a real scenario that could be created using open-source platforms such as OpenAir-Interface or srsRAN. One of the fundamental analyses is the timing to instantiate the virtual network functions at CUs, which should be synchronized with the prediction of the computational capacity.

Bibliography

[1] I. Rahman, S. M. Razavi, O. Liberg, C. Hoymann, H. Wiemann, C. Tidestav, P. Schliwa-Bertling, P. Persson, and D. Gerstenberger, "5G Evolution Toward 5G Advanced: An overview of 3GPP releases 17 and 18," *Ericsson Technology Review*, vol. 2021, no. 14, pp. 2–12, Oct. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9904665/

[2] C. J. Bernardos and M. A. Uusitalo, "European Vision for the 6G Network Ecosystem," Zenodo, Tech. Rep. 1.0, Jun. 2021. [Online]. Available: https://zenodo.org/record/5007671

[3] Ericsson, "The 5G Advanced, an evolution towards 6G," Ericsson, White paper BNEW-22:024836, Jun. 2022.

[4] A. Checko, H. Holm, and H. Christiansen, "Optimizing small cell deployment by the use of C-RANs," in *European Wireless 2014; 20th European Wireless Conference*, May 2014, pp. 1–6.

[5] M. Baghani, S. Parsaeefard, and T. Le-Ngoc, "Multi-objective resource allocation in density-aware design of C-RAN in 5G," *IEEE Access*,

vol. 6, pp. 45 177–45 190, 2018.

[6] China Mobile, "C-RAN The Road Towards Green RAN," China Mobile, Tech. Rep. Version 2.5, Oct. 2011.

[7] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: software defined radio access network," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking - HotSDN '13*. Hong Kong, China: ACM Press, 2013, p. 25. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2491185.2491207

[8] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "OpenRAN: a software-defined ran architecture via virtualization," in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM - SIGCOMM '13*. Hong Kong, China: ACM Press, 2013, p. 549. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2486001.2491732

[9] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[10] M. Peng, H. Xiang, Y. Cheng, S. Yan, and H. V. Poor, "Inter-Tier Interference Suppression in Heterogeneous Cloud Radio Access Networks," *IEEE Access*, vol. 3, pp. 2441–2455, 2015.

[11] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: beyond baseband samples and point-to-point links," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 90–97,

Oct. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7306542/

[12] S. Zhou, T. Zhao, Z. Niu, and S. Zhou, "Software-defined hyper-cellular architecture for green and elastic wireless access," *IEEE Communications Magazine*, vol. 54, no. 1, pp. 12–19, Jan. 2016.

[13] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: issues and challenges," *IEEE Network*, vol. 30, no. 4, pp. 46–53, Jul. 2016.

[14] X. Chen, Z. Han, Z. Chang, G. Xue, H. Zhang, and M. Bennis, "Adapting Downlink Power in Fronthaul-Constrained Hierarchical Software-Defined RANs," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2017, pp. 1–6.

[15] K. Liang, L. Zhao, X. Chu, and H. Chen, "An Integrated Architecture for Software Defined and Virtualized Radio Access Networks with Fog Computing," *IEEE Network*, vol. 31, no. 1, pp. 80–87, Jan. 2017.

[16] M. Baghani, S. Parsaeefard, and T. Le-Ngoc, "Multi-Objective Resource Allocation in Density-Aware Design of C-RAN in 5G," *IEEE Access*, vol. 6, pp. 45 177–45 190, 2018.

[17] 3GPP, "5G; NR; Overall description; Stage-2," 3GPP, Tech. Rep. TS 38.300 version 15.3.1 Release 15, Oct. 2018.

[18] ——, "5G; NG-RAN; Architecture description," 3GPP, Tech. Rep. TS 38.401 version 15.4.0 Release 15, Apr. 2019.

[19] E. Zeydan, J. Mangues-Bafalluy, J. Baranda, M. Requena, and Y. Turk, "Service Based Virtual RAN Architecture for Next Generation Cellular Systems," *IEEE Access*, vol. 10, pp. 9455–9470, 2022, conference Name: IEEE Access.

[20] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," *IEEE Access*, vol. 7, pp. 70 371–70 421, 2019, conference Name: IEEE Access.

[21] 3GPP, "Technical specification group services and system aspects; summary of rel-16 work items," 3GPP, Tech. Rep. TR 21.916 version 16.2.0 Release 16, Jun. 2022.

[22] ——, "Technical specification group services and system aspects; release 17 description; summary of rel-17 work items," 3GPP, Tech. Rep. TR 21.917 version 2.0.0 Release 17, Dec. 2022.

[23] D. Zhang, Z. Chen, L. X. Cai, H. Zhou, S. Duan, J. Ren, X. Shen, and Y. Zhang, "Resource allocation for green cloud radio access networks with hybrid energy supplies," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1684–1697, Feb. 2018.

[24] A. Checko, H. Christiansen, and M. S. Berger, "Evaluation of energy and cost savings in mobile Cloud RAN," in *Proceedings of OPNETWORK Conference*, 2013, p. 8.

[25] A. Alnoman, G. H. S. Carvalho, A. Anpalagan, and I. Woungang, "Energy Efficiency on Fully Cloudified Mobile Networks: Survey, Chal-

lenges, and Open Issues," *IEEE Communications Surveys Tutorials*, vol. 20, no. 2, pp. 1271–1291, 2018.

[26] A. Younis, T. X. Tran, and D. Pompili, "Bandwidth and Energy-Aware Resource Allocation for Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6487–6500, Oct. 2018.

[27] R. F. Ahmed, T. Ismail, L. F. Abdelal, and N. H. Sweilam, "Optimization of Power Consumption and Handover Margin of Sleep/Active Cells in Dynamic H-CRAN," in *2018 11th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP)*, Jul. 2018, pp. 1–6.

[28] I. Ashraf, F. Boccardi, and L. Ho, "Sleep mode techniques for small cell deployments," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, August 2011.

[29] J. Riihijarvi and P. Mahonen, "Machine Learning for Performance Prediction in Mobile Cellular Networks," *IEEE Computational Intelligence Magazine*, vol. 13, no. 1, pp. 51–60, Feb. 2018.

[30] E. Balevi and R. D. Gitlin, "Unsupervised machine learning in 5G networks for low latency communications," in *2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC)*, Dec. 2017, pp. 1–2.

[31] M. Chen, W. Saad, C. Yin, and M. Debbah, "Echo State Networks for Proactive Caching in Cloud-Based Radio Access Networks With Mobile

Users," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3520–3535, Jun. 2017.

[32] I. AlQerm and B. Shihada, "Enhanced machine learning scheme for energy efficient resource allocation in 5G heterogeneous cloud radio access networks," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017, pp. 1–7.

[33] G. Shen, L. Pei, P. Zhiwen, L. Nan, and Y. Xiaohu, "Machine learning based small cell cache strategy for ultra dense networks," in *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct. 2017, pp. 1–6.

[34] A. Y. Nikravesh, S. A. Ajila, C. Lung, and W. Ding, "Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM," in *2016 IEEE International Congress on Big Data (BigData Congress)*, Jun. 2016, pp. 402–409.

[35] S. Imtiaz, H. Ghauch, G. P. Koudouridis, and J. Gross, "Random forests resource allocation for 5G systems: Performance and robustness study," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Apr. 2018, pp. 326–331.

[36] T. Gao, M. Chen, H. Gu, and C. Yin, "Reinforcement learning based resource allocation in cache-enabled small cell networks with mobile users," in *2017 IEEE/CIC International Conference on Communications in China (ICCC)*, Oct. 2017, pp. 1–6.

[37] L. Le, B. P. Lin, L. Tung, and D. Sinh, "SDN/NFV, Machine Learning, and Big Data Driven Network Slicing for 5G," in *2018 IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 20–25.

[38] K. Thaalbi, M. T. Missaoui, and N. Tabbane, "Short Survey on Clustering Techniques for RRH in 5G networks," in *2018 Seventh International Conference on Communications and Networking (ComNet)*, Nov. 2018, pp. 1–5.

[39] H. Taleb, M. E. Helou, S. Lahoud, K. Khawam, and S. Martin, "Multi-Objective Optimization for RRH Clustering in Cloud Radio Access Networks," in *2018 International Conference on Computer and Applications (ICCA)*, Aug. 2018, pp. 85–89.

[40] H. Dai, Y. Huang, J. Wang, and L. Yang, "Resource Optimization in Heterogeneous Cloud Radio Access Networks," *IEEE Communications Letters*, vol. 22, no. 3, pp. 494–497, Mar. 2018.

[41] K. Boulos, K. Khawam, M. El Helou, M. Ibrahim, S. Martin, and H. Sawaya, "A hybrid approach for RRH clustering in Cloud Radio Access Networks Based on Game Theory," in *Proceedings of the 16th ACM International Symposium on Mobility Management and Wireless Access*, ser. MobiWac'18. New York, NY, USA: ACM, 2018, pp. 128–132. [Online]. Available: http://doi.acm.org/10.1145/3265863.3265880

[42] K. Boulos, K. Khawam, M. E. Helou, M. Ibrahim, H. Sawaya, and S. Martin, "An Efficient Scheme for BBU-RRH Association in C-RAN Architecture for Joint Power Saving and Re-Association Optimiza-

tion," in *2018 IEEE 7th International Conference on Cloud Networking (CloudNet)*, Oct. 2018, pp. 1–6.

[43] H. Taleb, M. El Helou, K. Khawam, S. Lahoud, and S. Martin, "Centralized and distributed RRH clustering in cloud radio access networks," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, July 2017, pp. 1091–1097.

[44] M. K. Elhattab, M. M. Elmesalawy, T. Ismail, H. H. Esmat, M. M. Abdelhakam, and H. Selmy, "A Matching Game for Device Association and Resource Allocation in Heterogeneous Cloud Radio Access Networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1664–1667, Aug. 2018.

[45] S. F. Abedin, M. G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 489–502, Jan 2019.

[46] X. Chen, N. Li, J. Wang, C. Xing, L. Sun, and M. Lei, "A dynamic clustering algorithm design for c-ran based on multi-objective optimization theory," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–5.

[47] K. Boulos, M. E. Helou, M. Ibrahim, K. Khawam, H. Sawaya, and S. Martin, "Interference-aware clustering in cloud radio access networks," in *2017 IEEE 6th International Conference on Cloud Networking (CloudNet)*, Sep. 2017, pp. 1–6.

[48] K. Thaalbi, M. T. Missaoui, and N. Tabbane, "Performance analysis of clustering algorithm in a C-RAN architecture," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, June 2017, pp. 1717–1722.

[49] H. Taleb, M. E. Helou, S. Lahoud, K. Khawam, and S. Martin, "Multi-objective optimization for RRH clustering in cloud radio access networks," in *2018 International Conference on Computer and Applications (ICCA)*, Aug. 2018, pp. 85–89.

[50] C. Huang, C. Chiang, and Q. Li, "A study of deep learning networks on mobile traffic forecasting," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017, pp. 1–6.

[51] W. Mo, C. L. Gutterman, Y. Li, G. Zussman, and D. C. Kilper, "Deep neural network based dynamic resource reallocation of BBU pools in 5G C-RAN ROADM networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[52] X. Wang, Y. Ji, J. Zhang, L. Bai, and M. Zhang, "Joint Optimization of Latency and Deployment Cost Over TDM-PON Based MEC-Enabled Cloud Radio Access Networks," *IEEE Access*, vol. 8, pp. 681–696, 2020, conference Name: IEEE Access.

[53] A. Li, Y. Sun, X. Xu, and C. Yuan, "An energy-effective network deployment scheme for 5G Cloud Radio Access Networks," in *2016*

IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Apr. 2016, pp. 684–689, iSSN: null.

[54] F. Tonini, C. Raffaelli, L. Wosinska, and P. Monti, "Cost-Optimal Deployment of a C-RAN With Hybrid Fiber/FSO Fronthaul," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 7, pp. 397–408, Jul. 2019.

[55] A. L. Rezaabad, H. Beyranvand, J. A. Salehi, and M. Maier, "Ultra-Dense 5G Small Cell Deployment for Fiber and Wireless Backhaul-Aware Infrastructures," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12 231–12 243, Dec. 2018.

[56] C. Ranaweera, E. Wong, A. Nirmalathas, C. Jayasundara, and C. Lim, "5G C-RAN With Optical Fronthaul: An Analysis From a Deployment Perspective," *Journal of Lightwave Technology*, vol. 36, no. 11, pp. 2059–2068, Jun. 2018, conference Name: Journal of Lightwave Technology.

[57] D. Pliatsios, P. Sarigiannidis, I. D. Moscholios, and A. Tsiakalos, "Cost-efficient Remote Radio Head Deployment in 5G Networks Under Minimum Capacity Requirements," in *2019 Panhellenic Conference on Electronics Telecommunications (PACET)*, Nov. 2019, pp. 1–4, iSSN: null.

[58] F. Bahlke, O. D. Ramos-Cantor, S. Henneberger, and M. Pesavento, "Optimized Cell Planning for Network Slicing in Heterogeneous

Wireless Communication Networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1676–1679, Aug. 2018.

[59] 3GPP, "3GPP TR 38.801 V14.0.0 (2017-03): Study on new radio access technology: Radio access architecture and interfaces." 3GPP, Tech. Rep., 2017.

[60] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2018.

[61] K. Boulos, M. E. Helou, K. Khawam, M. Ibrahim, S. Martin, and H. Sawaya, "RRH clustering in cloud radio access networks with re-association consideration," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018, pp. 1–6.

[62] M. Khan, R. S. Alhumaima, and H. S. Al-Raweshidy, "Quality of Service aware dynamic BBU-RRH mapping in Cloud Radio Access Network," in *2015 International Conference on Emerging Technologies (ICET)*, Dec. 2015, pp. 1–5.

[63] M. Khan, Z. H. Fakhri, and H. S. Al-Raweshidy, "Semistatic Cell Differentiation and Integration With Dynamic BBU-RRH Mapping in Cloud Radio Access Network," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 289–303, Mar. 2018.

[64] Behnam Rouzbehani, Luis M. Correia, and Luísa Caeiro, "An SLA-Based Method for Radio Resource Slicing and Allocation in Virtual RANs," in *COST IRACON*, Cartagena, Spain, Jun. 2018.

[65] C. Lee, M. Lee, J. Wu, and W. Chang, "A Feasible 5G Cloud-RAN Architecture with Network Slicing Functionality," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2018, pp. 442–449.

[66] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, "Dynamic Network Slicing for 5G IoT and eMBB services: A New Design with Prototype and Implementation Results," in *2018 3rd Cloudification of the Internet of Things (CIoT)*, Jul. 2018, pp. 1–7.

[67] X. Wang, K. Wang, S. Wu, S. Di, H. Jin, K. Yang, and S. Ou, "Dynamic Resource Scheduling in Mobile Edge Cloud with Cloud Radio Access Network," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 11, pp. 2429–2445, Nov. 2018.

[68] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless Network Virtualization With SDN and C-RAN for 5G Networks: Requirements, Opportunities, and Challenges," *IEEE Access*, vol. 5, pp. 19 099–19 115, 2017.

[69] H. Taleb, M. E. Helou, S. Lahoud, K. Khawam, and S. Martin, "An Efficient Heuristic for Joint User Association and RRH Clustering in Cloud Radio Access Networks," in *2018 25th International Conference on Telecommunications (ICT)*, Jun. 2018, pp. 8–14.

[70] O. Narmanlioglu and E. Zeydan, "Learning in SDN-based multi-tenant cellular networks: A game-theoretic perspective," in *2017 IFIP/IEEE*

Symposium on Integrated Network and Service Management (IM), May 2017, pp. 929–934.

[71] K. Lin, W. Wang, Y. Zhang, and L. Peng, "Green Spectrum Assignment in Secure Cloud Radio Network with Cluster Formation," *IEEE Transactions on Sustainable Computing*, pp. 1–1, 2018.

[72] N. Nomikos, S. Zoupanos, T. Charalambous, and I. Krikidis, "A Survey on Reinforcement Learning-Aided Caching in Heterogeneous Mobile Edge Networks," *IEEE Access*, vol. 10, pp. 4380–4413, 2022, conference Name: IEEE Access.

[73] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L.-C. Wang, "Deep Reinforcement Learning for Mobile 5G and Beyond: Fundamentals, Applications, and Challenges," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 44–52, Jun. 2019, conference Name: IEEE Vehicular Technology Magazine.

[74] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Access*, vol. 8, pp. 140 699–140 725, 2020, conference Name: IEEE Access.

[75] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://link.springer.com/10.1007/BF00994018

[76] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Informa-*

tion Processing Systems 9, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. MIT Press, 1997, pp. 155–161.

[77] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[78] H. Wang, M. Garcia-Lozano, E. Mutafungwa, X. Yin, and S. Ruiz, "Performance study of uplink and downlink splitting in ultradense highly loaded networks," *Wireless Communications and Mobile Computing*, vol. 2018, p. 12, Jul. 2018.

[79] S.Ruiz, H. Ahmadi, L.M.Caeiro, N.Cardona, L.M.Correia, M.Garcia-Lozano, T.Javornik, and V.Petrini, "IRANCON reference scenarios," in *EURO-COST*, Nicosia, Jan. 2018, pp. 34–37.

[80] U. Saeed, J. Hämäläinen, M. Garcia-Lozano, and G. David González, "On the feasibility of remote driving application over dense 5G roadside networks," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2019, pp. 271–276.

[81] Mojgan Barahman, Luis M. Correia, and Lucio S. Ferreira, "A real-time computational resource management in C-RAN," in *EURO-COST*, Cartagena, Spain, May 2018.

[82] Guangyi Liu, Shen Xiadong, Jürgen Krämer, Sadayuki Abeta, Thomas Sälzer, Eric Jacks, Andrea Buldorini, and Georg Wannemacher, "Next generation mobile networks radio access performance evaluation methodology," NGMN Alliance, Tech. Rep., Jan. 2008.

[83] Joan Olmos, Silvia Ruiz, Mario Garcia Lozano, and David Martin Sacristan, "Link abstraction models based on mutual information for LTE downlink," in *EURO-COST*, Aalborg, Denmark, Jun. 2010.

[84] 3GPP, "3GPP TR 36.213 v14.3.0 (2017-05): Technical specification group radio access network; evolved universal terrestrial radio access (e-utra); physical layer procedures." 3GPP, Tech. Rep., 2017.

[85] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–7.

[86] Rolando Guerra-Gómez, Silvia Ruiz, M. Garcia-Lozano, and Joan Olmos, "Using COST IC1004 Vienna scenario to test C-RAN optimisation algorithms," in *COST IRACON*, Dublin, Ireland, Jan. 2019.

[87] R. Guerra-Gómez, Silvia Ruiz, M. Garcia-Lozano, and Joan Olmos, "A weighted-sum multi-objective optimization for dynamic resource allocation with QoS constraints in realistic C-RAN," in *COST IRACON*, oulu, Finland, May 2019.

[88] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[89] R. Guerra-Gómez, S. R. Boqué, M. García-Lozano, and J. O. Bonafé, "Dynamic resource allocation in C-RAN with real-time traffic and realistic scenarios," in *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Oct 2019, pp. 1–6.

[90] R. Guerra-Gómez, S. R. Boqué, M. García-Lozano, and J. O. Bonafé, "Machine-learning based traffic forecasting for resource management in C-RAN," in *2020 European Conference on Networks and Communications (EuCNC)*, 2020, pp. 200–204.

[91] R. Guerra-Gómez, S. R. Boqué, M. García-Lozano, and J. O. Bonafé, "Predicting required computational capacity in C-RAN networks by the use of different machine learning strategies," in *COST IRACON*, Gdańsk, Poland, Sep. 2019.

[92] R. Guerra-Gómez, S. Ruiz-Boqué, M. García-Lozano, and J. O. Bonafe, "Machine learning adaptive computational capacity prediction for dynamic resource management in C-RAN," *IEEE Access*, vol. 8, pp. 89 130–89 142, 2020.

[93] R. Guerra-Gómez, S. Ruiz-Boqué, M. García-Lozano, and U. Saeed, "Energy and cost footprint reduction for 5G and beyond with flexible radio access network," *IEEE Access*, vol. 9, pp. 142 179–142 194, 2021.

[94] R. Guerra-Gómez, Silvia Ruiz, M. Garcia-Lozano, and Umar Saeed, "Flexible radio access network optimization with cell coordination," in *1st INTERACT: Intelligence-Enabling Radio Communications for Seamless Inclusive Interactions*, Bologna, Italy, Feb. 2022.

[95] Tiago Monteiro, Luis M. Correia, and Ricardo Dinis, "Implementation analysis of cloud radio access networks architectures in small cells," in *EURO-COST*, Portugal, Feb. 2017.

[96] 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, pp. 405–426, 2015.

[97] C. Mao, M. Khalily, P. Xiao, T. W. C. Brown, and S. Gao, "Planar Sub-Millimeter-Wave Array Antenna With Enhanced Gain and Reduced Sidelobes for 5G Broadcast Applications," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 1, pp. 160–168, Jan. 2019.

[98] K. Grobe, A. Mitcsenkov, S. Krauß, F. Geilhart, and et. al., "Assessment of candidate transport network architectures for structural convergence," COMBO, Tech. Rep. D3.4, Jun. 2016.

[99] L. HUAWEI TECHNOLOGIES CO., "Vo5G Technical White Paper," HUAWEI, Tech. Rep., Jul. 2018.

[100] 3GPP, "5G; System Architecture for the 5G System (3GPP TS 23.501 version 15.3.0 Release 15)," 3GPP, Tech. Rep., Jul. 2018.

[101] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, May 2018.

[102] V. Nurmela and et. al., "Deliverable d1.4: METIS Channel Models," Mobile and wireless communications Enablers for the Twenty–twenty Information Society (METIS), Tech. Rep. ICT-317669-METIS/D1.4, Feb. 2015.

[103] E. D. Andersen and K. D. Andersen, "The Mosek Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm," in *High Performance Optimization*, ser. Applied Optimization. Boston, MA: Springer US, 2000, pp. 197–232. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4757-3216-0_8

[104] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.