Communications multi-utilisateurs dans les réseaux d'accès radio centralisés : architecture, coordination et optimisation

Dora Boviz

université PARIS-SACLAY

CentraleSupélec

NNT : 2017SACLC035

Thèse de doctorat de l'Université Paris-Saclay préparée à CentraleSupélec

Ecole doctorale n°580
Sciences et technologies de l'information et de la communication
Spécialité de doctorat: Réseaux, information et communications

par

Mme Dora Boviz

Communications multi-utilisateurs dans les réseaux d'accès radio centralisés: architecture, coordination et optimisation

Thèse présentée et soutenue à Nozay, le 19 juin 2017.

Composition du Jury :

M.	Pierre Duhamel	Directeur de recherche CentraleSupélec	Président du jury
M.	Lars Dittmann	Professeur DTU	Rapporteur
M.	Raymond Knopp	Professeur EURECOM	Rapporteur
Mme	E. Veronica Belmega	Maître de conférences ENSEA	Examinatrice
M.	Rami Langar	Professeur UPEM	Examinateur
M	Eric Renault	Maître de conférences Télécom SudParis	Examinateur
M.	Sheng Yang	Professeur adjoint CentraleSupélec	Directeur de thèse
M.	Laurent Roullet	Directeur de recherche Nokia Bell Labs	Encadrant, membre invité

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

(In the name of God, the most Gracious, the most Merciful)

Acknowledgments

As this thesis would not have been realized without the help and support that I have received from many people, I would like to start the manuscript by thanking all of them, even if I cannot cite here all the names.

First, I would like to express my special gratitude to Mr. Laurent Roullet, my industrial supervisor, for the opportunity of realizing this thesis and for his encouragement, guidance and valuable comments about my work. I am also deeply thankful to Prof. Sheng Yang, my academic supervisor, for his support and advice all along the thesis. The experience that I have acquired by working with them will probably be a major advantage for my career. I would like to thank all the PhD committee members for accepting to evaluate my thesis and for the valuable comments and suggestions about it.

Secondly, I would like to thank Bell Labs France and the ANRT (Association Nationale de la Recherche et de la Technologie) for the financial support that allowed me to work on this thesis. I am also grateful to Dr. Vinod Kumar and Mr. Laurent Thomas for accepting me for the research engineer position to work on this thesis. I am very thankful for the funding and especially for the valuable collaborations with the partners from HARP and IDEFIX research projects. Special thanks to Prof. David Gesbert, Prof. Bruno Clerkx, Prof. Tharmalingam Ratnarajah, Dr. Salah Eddine Elayoubi, Dr. Aleksandra Checko, Dr. Nivine Abbas, Matteo Artuso and all the other participants to these projects with whom I had the possibility to work together. I would also like to thank my PhD student colleagues in the Telecom Department in Supélec, Chao, Zheng, Asma and Meryem in particular.

I would like to express my great appreciation to all the team in Bell Labs, it was a real pleasure to work there and be part of the SDMN team. I would like to thank Calvin for the work together, for all the things I could learn from him and for being available at any time. Many thanks to Aravinth with whom we progressed together during the whole thesis, for the discussions and the collaboration in various projects and to Nessrine for the conversations about the thesis and many other things. Special thanks to Alberto for always being there for the team, whether it is for a technical discussion or for a coffee and to Elena

who was there every time I needed to talk to someone. I would like to thank my office mates during the thesis, Afef, Véronique Sabine, Ivaylo and Jakob for the serious and less serious conversations. Thanks to all the people from Bell Labs with whom I have worked during the thesis, especially to Karim, Kader, Imran, Amira, Illyne, Bessem, Lionel, Amine, Bruno and Yacine and to all the others from the SDMN team that I don't have the place to list here.

I would like to thank my parents, grandparents, sisters and other relatives who have always been proud of my work and encouraged me in my studies. I am particularly grateful to my grandfather and my mother for their support since my childhood. Special thanks to my husband, Mourad, for his continuous support and for being patient during the extra hours that I spent working and to my parents-in-law for their prayers for my success. Thanks to my little son, Adam, for his unending joy that gave me the strength to make it through. Above all, I am grateful to God for everything He provided me in this life and for allowing me to successfully finish this thesis.

Contents

List of Figures

List of Tables

Part 1

Introduction

1.1 Context and motivations

1.1.1 Cloud RAN and multi-user communications

The main goal of this thesis has been to fill the gap between theoretical research on Cloud Radio Access Network (C-RAN) and the intention of network equipment constructors to enable real C-RAN deployment providing significantly higher performance than traditional distributed Radio Access Network (RAN) infrastructures. In fact, in an ideal case, C-RAN can be considered as a single Base Station (BS) with many distributed antennas, thus it operates as a massive Multiple Input Multiple Output (MIMO) BS without the limitation of correlation between the antennas. This architecture is very promising for network capacity improvement. Unfortunately, even with today's advanced computer technology this ideal case is not far from being a science fiction, mostly for practical reasons – including backward compatibility –, cell-based structure is still maintained. Very similar to the MIMO concept multi-cell techniques enable also performance improvement, but very often efficient operation requires centralized processing [1]. While multi-cell and multi-antenna techniques have been investigated, the architecture that enables to realize them [2, 3] has also been defined. Deployment constraints and technical challenges were pointed out and several studies took practical limitations into account in order to evaluate C-RAN performance in a more realistic configuration [4–8]. However, each of them addressed at most a few constraints, while the aim of this thesis is to realize an end-to-end study of C-RAN operation enabling multi-cell signal processing.

Considering the overall Quality of Service (QoS) and network performance along with implementation, architecture and efficient operation of every C-RAN component allows to propose in this thesis a practical framework to meet C-RAN gain expectations in real-word deployments [9]. A review of both theoretical performance of multi-user multi-cell techniques under realistic constraints and stud-

ies on C-RAN architecture details adapted to various technological limitations outlined the path towards a tradeoff where we can benefit from multi-cell cooperation in an actual C-RAN infrastructure. Numerous characteristics, parameters, definitions, protocols and physical constraints had to be considered. At the same time, uncountable novel technologies that were not used previously in mobile networks served as enablers to make centralized multi-cell signal processing a reality. Network Function Virtualization (NFV) techniques, software-defined networking, vitrualization, micro-services, General Purpose Processor (GPP)-based platforms, Ethernet fronthaul all contribute to integrate multi-cell MIMO while keeping the benefits of C-RAN. Besides these concepts, we have also define a practical receiver structure providing a tradeoff between computational complexity and transmission rate. To deal with the major constraint of C-RAN, i.e., the fronthaul transport, we consider novel architectures as well as optimized operation aiming to use just as much fronthaul rate as needed for providing good QoS.

Even though the idea of C-RAN was defined before 4G mobile network deployments and 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) standards included the elements necessary for multi-cell processing, most of the enabler technologies were not ready to be used. Also, the gain in LTE use cases would have been limited, since its interference management capability is efficiently exploited only in dense network deployments with many users. The explosion of mobile data traffic expected with the arrival of connected everything era [10] will require that future mobile networks serve much more users at significantly higher rate. To cope with this demand, three dimensions are available: frequency, spectral efficiency and space. In the space dimension, more Access Points (APs) will be installed, i.e., network deployments become denser. Obviously, it will provide higher data rates, but also create more interference, that makes C-RAN indispensable. To progress following spectral efficiency dimension, multi-user techniques such as Non-Orthogonal Multiple Access (NOMA) will be used in 5G. C-RAN also allows to extend these techniques among cells for further improvement. The usage of larger frequency bands also allowing to accommodate more traffic and more users, we take into account this characteristic in order to describe the solutions proposed for C-RAN in the 5G context.

1.1.2 Industrial PhD project

Besides describing the technical and scientific background of the thesis, it can of interest to present the context where the doctoral research has been carried out. French national association for research and technology (ANRT) supports companies in supervising Ph.D. projects jointly with academic research laborato-

ries, and this thesis is the result of such a collaboration between *Laboratoire des Signaux et Systèmes* at Centralesuplélec and Bell Labs, the industrial research entity that today belongs to the group Nokia. In the beginning of the thesis, the Bell Labs was part of Alcatel-Lucent, that merged in 2016 with Nokia. Both the former Alcatel-Lucent and Nokia after the fusion are present in many segments of the telecommunication industry, spanning from optical fiber technology, through fixed access to mobile networks and various research topics applied in these fields are studied in Bell Labs.

Following the agreement between the academic and the industrial research laboratories, the major part of the supervision was realized at Bell Labs, where most of the work was carried out within the Software-Defined Mobile Networks department. The group's research mainly focuses on the definition of the architecture and processing mechanisms for future mobile networks, as well as the study of evolved features such as machine learning and advanced signal processing that should be included in next generation networks. The research carried out in the telecommunications department of *Laboratoire des Signaux et Systèmes* concerns a wide range of subjects: information theory, wireless communications, advanced features in wireless networks. The joint supervision realized within these two groups allowed to adopt in this thesis an approach gathering theoretical and practical aspects. For instance, performance metrics and signal processing methods that we have used were chosen to possibly fit in real network deployments. We have also dedicated a part of the realized work to defining, evaluating and prototyping architectural solutions enabling to implement in real systems the results presented in the thesis. Being able to consult on some specific questions experienced Bell Labs researchers and other experts in Alcatel-Lucent or Nokia was very useful to build a consistent research project compatible with real deployments planned for future networks.

From 2013 to 2015, collaboration with academic and industrial partners through the European project *High capacity network Architecture with Remote radio heads and Parasitic antenna arrays (HARP)* also allowed to have a broad comprehension about C-RAN and closer insights to the newest results from recognized researchers in the field of the thesis. Some contributions to the project are also included in the present document. Another collaborative project from which we had the opportunity to benefit during the thesis was founded by the French *National Research Agency (ANR)* and was entitled *Intelligent DEsign of Future mobile Internet for enhanced eXperience (IDEFIX)*. Our contribution mainly concerned the realization of a C-RAN platform that can host applications for cooperative RAN management possibly realized by other partners in the project. All these valuable collaborations helped a lot to sketch the different ideas presented in this thesis and position our work with respect to the trends

of telecommunications industry and academic research.

1.2 Outline of the thesis

After this first part introducing the thesis, in each of the following three parts we describe background, context and related works in addition to the novel contributions.

In Part 2, we present the C-RAN architecture, its evolution and benefits. The background of current and future mobile networks is also described, including multi-cell processing techniques. We also propose an architectural solution for the major challenge of C-RAN deployments, i.e., the low-rate fronthaul.

Then, in Part 3, an efficient low latency control solution for C-RAN is presented. We use Software Defined Networking (SDN) adapted to the context of RAN to enable multi-cell processing. We also describe the elements and the operation of our prototype implementing the proposed design.

After having studied practical aspects of multi-cell processing in C-RAN, in Part 4, we consider optimization of uplink multi-cell transmissions in a realistic system model. A practical multi-user processing scheme with low-complexity receive processing is described. Then, we investigate the possibility of increasing throughput by opportunistic association of users benefiting from multi-cell processing. In the second half of Part 4, we focus on optimal fronthaul rate allocation using a metric that takes into account the cost of fronthaul usage. Two optimization schemes are described, in the first one we assume perfect channel knowledge in order to evaluate the benefit of fronthaul allocation in an ideal case. In the second scheme, we study a more realistic model where we approximate the throughput using only channel statistics. Finally, we provide concluding remarks an future research directions in Part 5.

1.3 Contributions

An overview of the main contributions described in the thesis is shown in Figure 1.1, the number of each contribution listed below is placed also in the figure. The aim of this section is two-fold: positioning the novel results presented in this thesis with respect to the related work and clearly enclose the personal contributions in collaborative research works.

(I) As a first result, we have defined a centralized architecture for C-RAN that enables multi-cell physical layer processing. It is based on the SDN concept originally used in fixed network deployments, and we have investigated how it can be adapted for RAN applications. Following the focus of this thesis,

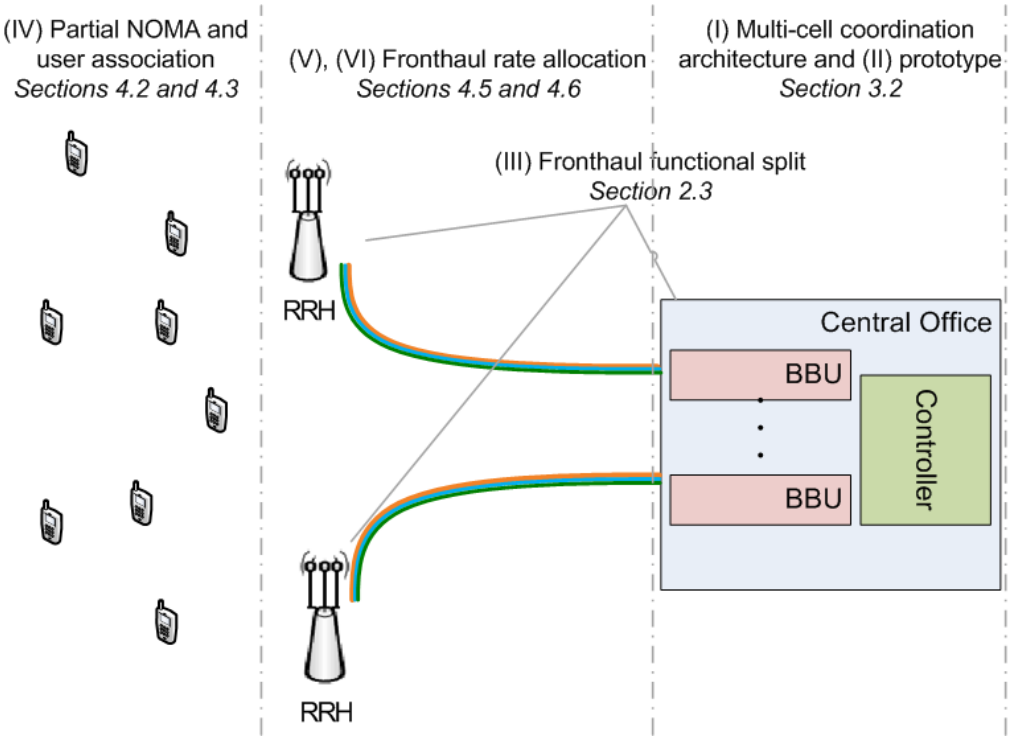


Figure 1.1 – Overview of the contributions described in the thesis

we have proposed an architecture where the case of multi-cell multi-user joint processing can be realized in a practical deployment of C-RAN.

(II) According to the SDN based architecture that we have specified, a real-time prototype has been developed using open-source software base-band units, an adapted SDN controller accommodating usage in RAN and multi-cell coordination applications. The main contribution in the realization of this platform was the definition of the general architecture, particularly the operation and constraints of multi-cell coordination and also the evaluation of the feasibility of the uplink joint reception along with the definition of the coordination algorithm for the uplink case. The major part implementation was realized in collaboration with other team members, that effort resulted in a proof-of-concept for our SDN-enabled C-RAN solution.

(III) Besides the architecture for low-latency multi-cell control, we have also designed a solution for low-rate fronthaul that allows to benefit from centralized processing in C-RAN deployments. Our contribution was to define the placement of base-band processing functions between the distributed remote units and the central unit that enables multi-cell transmissions both on the uplink and the downlink while significantly reducing fronthaul rate requirements with respect to state-of-the-art solutions.

(IV) We have also investigated how to adapt multi-user techniques that have been proposed for single-cell use cases and take into account practical aspects to the multi-cell C-RAN case. This study resulted in the evaluation of user association for uplink multi-cell NOMA where smaller user groups are formed for joint transmission, allowing low-complexity reception particularly important in the multi-cell scenario. Though the efficiency of this strategy that we call partial NOMA has been confirmed, optimization of user groups seems not to fit common use cases.

(V) For the sake of taking into account the cost of fronthaul usage in the definition of the benefit of multi-user uplink transmission in C-RAN, we have proposed an analytical modeling of this cost and described the behavior of this function in possible fronthaul deployment scenarios.

(VI) To increase the efficiency of the tradeoff between the overall throughput of multi-user transmissions in C-RAN, we have defined a method for optimizing the allocation of fronthaul rates for for multi-antenna Remote Radio Heads (RRHs) for uplink received signal using the partial NOMA strategy. We have elaborated a realistic system model and evaluated the benefit of the optimization following the considered objective function. In a first case, the knowledge of real-time channel realizations was assumed, in a second one we have used an approximation of the throughput based on channel covariances.

Though the work described in this thesis is partially the result of valuable contributions, advice and explanations from many colleagues, the ideas presented in the thesis aiming to improve most of practical aspects of multi-user techniques in C-RAN are the result of a personal work, as well as the development of the results presented and their evaluation.

1.4 List of publications

The scientific publications that I have prepared or to which I have contributed during the thesis are the following.

Journal articles

- D. Boviz and S. Yang, "Fronthaul allocation for uplink multi-user transmissions in cloud RAN with statistical CSI," *(in preparation)*
- D. Boviz, C. S. Chen, and S. Yang, "Effective design of multi-user reception and fronthaul rate allocation in 5G cloud RAN," *Journal on Selected*

Areas of Communications, special issue on Deployment and Performance Challenges for 5G, 2017 *(accepted, under revision)*

- M. Artuso, D. Boviz, A. Checko et al., "Enhancing LTE with cloud-RAN and load-controlled parasitic antenna arrays," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 183–191, 2016

Conference papers

- D. Boviz, C. S. Chen, and S. Yang, "Cost-aware fronthaul rate allocation to maximize benefit of multi-user reception in C-RAN," in *IEEE Wireless Communications and Networking Conference (WCNC)*, (San Francisco, USA), Mar. 2017

- D. Boviz and Y. El Mghazli, "Fronthaul for 5G: low bit-rate design enabling joint transmission and reception," in *IEEE Global Telecommunications Conference (Globecom), 5G RAN Design Workshop*, December 2016

- D. Boviz, N. Abbas, G. Aravinthan, C. S. Chen, and M. A. Dridi, "Multi-cell coordination in cloud RAN: architecture and optimization," in *International Conference on Wireless Networks and Mobile Communications (WINCOM)*, (Fez, Morocco), pp. 271–277, Oct. 2016 *(invited paper)*

- D. Boviz and S. Yang, "Optimal fronthaul capacity allocation and training for joint uplink receiver in C-RAN," in *European Wireless (EW2016)*, (Oulu, Finland), pp. 415–420, May 2016

- D. Boviz, G. Aravinthan, C. S. Chen, and L. Roullet, "Physical layer split for user selective uplink joint reception in SDN enabled Cloud-RAN," in *2016 Australian Communications Theory Workshop (AusCTW)*, (Melbourne, Australia), pp. 83–88, Jan. 2016

Others

- D. Boviz, M. A. Dridi, N. Abbas, and G. Aravinthan, "Multi-cell coordination in cloud RAN enabled by SDN (demonstration)," in *IEEE Wireless Communications and Networking Conference - Student Outreach Program*, (San Francisco, USA), Mar. 2017

- D. Boviz, G. Aravinthan, N. Trabelsi, and L. Roullet, "C-RAN fronthaul enhancements using software defined networking," in *Cloud- and fog-based PHY communications in 5G - GDR ISIS Workshop*, (Paris, France), Nov. 2015

- Contribution to the deliverable of the ANR project IDEFIX D6: Testbed - Multi-cell Coordination in Cloud RAN , January 2017
- Contribution to Final (Y3) report of the EU project HARP, presented to European Commission in January 2016
- Contribution to the deliverable of the EU project HARP D7.4: Final end-to-end demo , presented to European Commission in November 2015
- Contribution to the deliverable of the EU project HARP D6.4: Eth2CPRI prototype implementation , presented to European Commission in November 2015
- Contribution to the deliverable of the EU project HARP D5.2.1: Cooperative techniques with perfect CSI, presented to European Commission in November 2014

Part 2

Cloud RAN and multi-cell techniques

2.1 Background

2.1.1 Cloud RAN architecture

Mobile networks provide data connectivity to users through base stations deployed in a distributed manner and communicating with centralized servers ensuring service management. C-RAN architecture proposed in [2] remodeled the role of network components by offloading signal processing from the base stations to a server unit in charge of base-band processing for several cells. In this section we give the different variants of C-RAN architecture and describe the numerous advantages of C-RAN as well as the actual challenges that we are facing while moving toward fifth generation (5G) mobile radio network deployments.

The details of C-RAN were described in [3], where technical and economical advantages are given besides requirements and research challenges that have to be solved. In [9], savings in Capital Expenditure (CAPEX) and Operating Expenditure (OPEX) enabled by C-RAN are highlighted, authors conclude that in realistic deployment scenarios Total Cost of Ownership (TCO) can be reduced by 14% thanks to reduced energy consumption, deployment and maintenance costs and network performance can be increased up to 30% for LTE networks by using multi-cell techniques such as Coordinated Multi-Point (CoMP) and real-time Inter-cell Interference Cancellation (ICIC).

2.1.1.1 C-RAN types

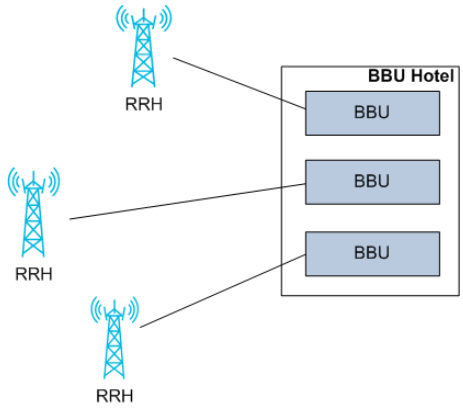
The basic idea of deploying smaller radio access points for cost saving reasons already appeared with small cell networks, but the C-RAN concept shifted the paradigm by keeping the service offered by the BSs, and splitting it between a

lightweight RRH located close to the antennas and a BBU that can be installed in the same location for several cells. The RRH performs at least Radio Functions (RF) and Analog-Digital Conversion (ADC)/Digital-Analog Conversion (DAC), and basically acts as a relay for the BBU where digital signal processing is realized. Mobile network equipment vendors proposed flexible solutions enabling this separation [21] and introducing the framework of practical C-RAN deployments.

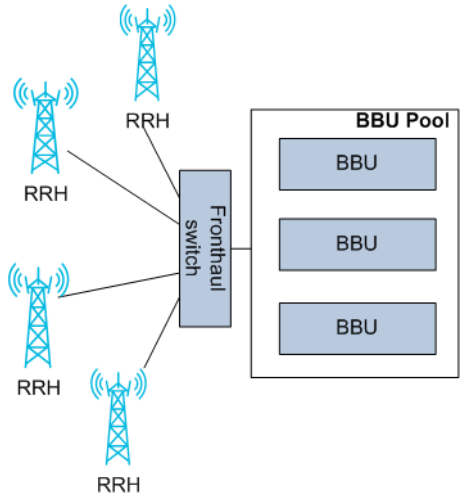
As a first step toward Cloud RAN, a centralized architecture where BBUs are deployed in a same location was considered. This solution, called BBU hoteling allows to partially leverage TCO savings, but network performance is not improved since BBUs are not interconnected, thus multi-cell cooperation is not enabled when we use this architecture. Figure 2.1a illustrates this configuration, where the CO contains the BBUs associated to a given cell-site where the RRHs are deployed.

In a more evolved solution, processing is dynamically allocated to the BBUs which are still deployed on dedicated computing units. This architecture, depicted in Figure 2.4b, is usually called centralized RAN. It allows to benefit from multiplexing gain (also called pooling gain) between different cell-sites, when the processing charge is balanced between them [22]. By controlling the association of UEs to the cells within the same C-RAN cluster, we can avoid processing overload even with if the processing capacity of the CO is lower than the total capacity of the individual BBUs would be. This way, cost savings can be realized on the deployed equipments and also on operational costs.

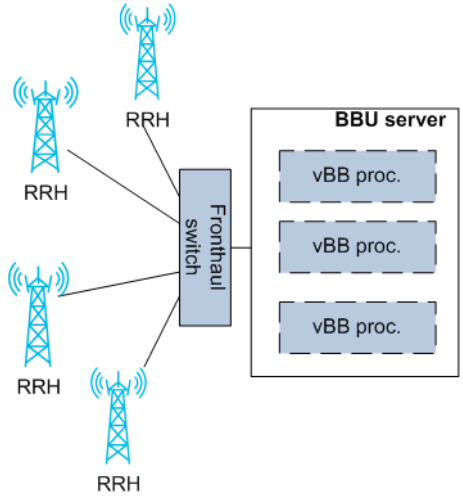
The most advanced option is the Cloud RAN architecture where BBU processing is executed in a virtualized central server with general purpose hardware [23]. This solution facilitates multi-cell processing by enabling low-latency information exchange between cells, without using the inter-cell communication protocol X2 defined for distributed RAN architecture. We show this architecture in Figure 2.1c.



(a) The Base-band Unit (BBU) hotel solution. Independent BBUs are co-located in the Central Office (CO), called BBU hotel in this case.



(b) The Centralized RAN solution. Independent BBUs are co-located in the CO, processing load is dynamically balanced between them by the fronthaul switch.



(c) Cloud RAN architecture with virtualized base-band processes running on general purpose hardware platform.

Figure 2.1 – Cloud RAN architecture options

2.1.1.2 C-RAN performance gains

The main advantages of C-RAN belong to one of the following categories:

- easier deployment and maintenance,
- savings related to centralized architecture,
- network performance improvement.

Compared to the installation of all-in-one BSs at the cell sites connected to a network core server usually located far from the BS, deploying a RAN infrastructure with an intermediate aggregation node – the BBU-pool – between the distributed radio front-end and the core is easier. Furthermore, realizing the major part of the base-band (BB) processing of several cells at the same location makes the repairs and upgrades simpler. Thus C-RAN by its intrinsic structure significantly reduces the TCO of the RAN.

Centralization of BBUs brings also the so-called pooling gain that can be of two types. The first one is related to the energy savings that we get by aggregating cell sites with different traffic profiles (e.g., residential, office) [24]. The second one is obtained by applying advanced load balancing mechanisms to minimize the total computational load within the BBU-pool. These advantages result also in lower CAPEX and OPEX.

In addition to being more energy- and cost-effective, C-RAN enables or facilitates various types of multi-cell cooperative processing. This kind of improvement brought by C-RAN is the main focus of the work realized for this thesis. As opposed to the above cited two advantages, here, we adapt a proactive strategy by introducing Physical Layer (PHY)- and Medium Access Control Layer (MAC)-level cooperation between the cells processed in the same BBU-pool. Our aim is to provide better QoS to the end-users, in order to likely improve the Mobile Network Operators (MNOs)' incomes thanks to the better service offered.

2.1.1.3 Deployment challenges

Besides its advantages, the deployment of C-RAN raises several technological challenges that have to be handled in order to maximize the above cited benefits. The most obvious limitation is due to the communication links between the RRHs and the BBU-pool the more often denoted by the term fronthaul – that we will use in the following – and sometimes, when some BB functions are offloaded to the RRHs or Remote Units (RUs), it is called midhaul. The links connecting the BBUs to the core network are termed backhaul. We will discuss

in detail fronthaul technologies and various solutions to deal with rate and delay constraints in 5G RAN in Sections 2.2.2 and 2.3.

Secondly, the control of the BBUs included in the same BBU-pool has to satisfy RAN processing time constraints and needs to ensure scalability and flexibility in order to efficiently benefit from centralized processing. The emerging concept of SDN is a promising candidate to realize advanced control in C-RAN, but it needs to be adapted to the RAN requirements. We describe the integration of SDN control in C-RAN and the related proof-of-concept platform in Part 3.

Another significant issue of C-RAN implementation is efficient low-latency communication between the BBUs dedicated to different cells. In fact, despite their centralized location, BB processing of each cell remains logically separated, thus communication mechanisms have to be added in order to enable multi-cell cooperation. In a virtualized deployment, the compatibility of standard IT methods for inter-process and inter-thread messaging has to be verified according to real-time constraints.

As the BB processing is realized jointly by the RRH and the central unit, synchronization in time and frequency between these units is required. Initially, Common Public Radio Interface (CPRI) protocol used for transferring In-phase/Quadrature (I/Q) symbols between the RRHs and the BBU-pool ensured this synchronization, but with the Ethernet based fronthaul solution a new method is needed. Indeed, with converged fronthaul infrastructure, i.e., switched network shared by several services, transmission time is variable. Delays depend not only on the location of the RRH, but also vary in time following network usage conditions. Time and frequency synchronization can be then realized using the IEEE 1588 protocol [25] that can be further enhanced by the SyncE protocol improving frequency precision.

2.1.2 Physical layer processing in mobile networks

Mobile radio communications have required more and more complex signal processing at each level of the network stack, in particular in the physical layer. Data transmission introduced already in 2G needed to be improved with the increasing usage of the internet, in order to provide access to online services on mobile devices. To serve more users at a higher rate and with larger bandwidth, new transmission methods were introduced in each network generation. As it has already proven its efficiency in other wired and wireless technologies, Orthogonal Frequency Division Multiplexing (OFDM) has been selected in 4G standard to operate mobile broad-band connectivity. MIMO transceivers introduced also in LTE required new and often more complex transceivers.

Higher network layers that are common with other access technologies can support high data rates, thus, physical layer processing of mobile networks was a bottleneck for better performance. With the PHY chain defined for LTE, we have been able to reach up to 300 Mbps transmission rate on the Downlink (DL) with Single Input Single Output (SISO). 5G requirements go further by targeting 10 Gbps peak rates and 100 Mbps available anywhere, anytime [26]. In the following we describe the currently used LTE Advanced (LTE-A) physical layer functions and new techniques likely to be used in 5G.

2.1.2.1 UL and DL PHY processing chain in LTE

As we have already mentioned, in LTE, OFDM is implemented for frequency modulation. We can find a general evaluation of its application in wireless networks in [27]. It allows, along with Turbo Coding and Binary Phase Shift Key (BPSK) or Quadrature Amplitude Modulation (QAM) schemes, to realize high transmission rates. Both Time Division Duplex (TDD) and Frequency Division Duplex (FDD) modes are possible, though FDD is the most common, bandwidth used is between 1.4 MHz and 40 MHz. Up to 8 antennas can be used in LTE-A systems. A detailed description of LTE structure and transmission protocol can be found in [28]. PHY processing is basically symmetric, each transmitter function has its receiver equivalent. In the following, we focus on DL and Uplink (UL) PHY processing on the network side. Note that on the UL, a single-carrier variant of OFDM, Single-Carrier Frequency Division Multiple Access (SC-FDMA) is used, requiring minor modifications in the transmission chain at the User Equipments (UEs), though major functions remain the same.

As some contributions of this thesis are based on OFDM physical layer structure, we briefly describe the functional blocks of the LTE PHY processing chain shown in Figure 2.2.

CRC In LTE, Hybrid Automatic Repeat reQuest (HARQ) mechanism based on Cyclic Redundancy Check (CRC) error detection coding is implemented. The transmitter adds a set of check bits to the data bits and the receiver verifies the integrity of the received data. If the verification fails, in HARQ, a repeat request is sent to the transmitter.

Turbo Coding Inherited from 3G systems, this powerful forward error correction coding is implemented in LTE PHY. Turbo codes allow to approach the Shannon channel capacity, though in practical systems, rate constraints limit their performance. The encoding is based on parallel concatenated convolutional codes, produced by two 8-state encoders and combined by interleaving [29]. The coding rate used is 1/3. The decoding is realized also by two components, the

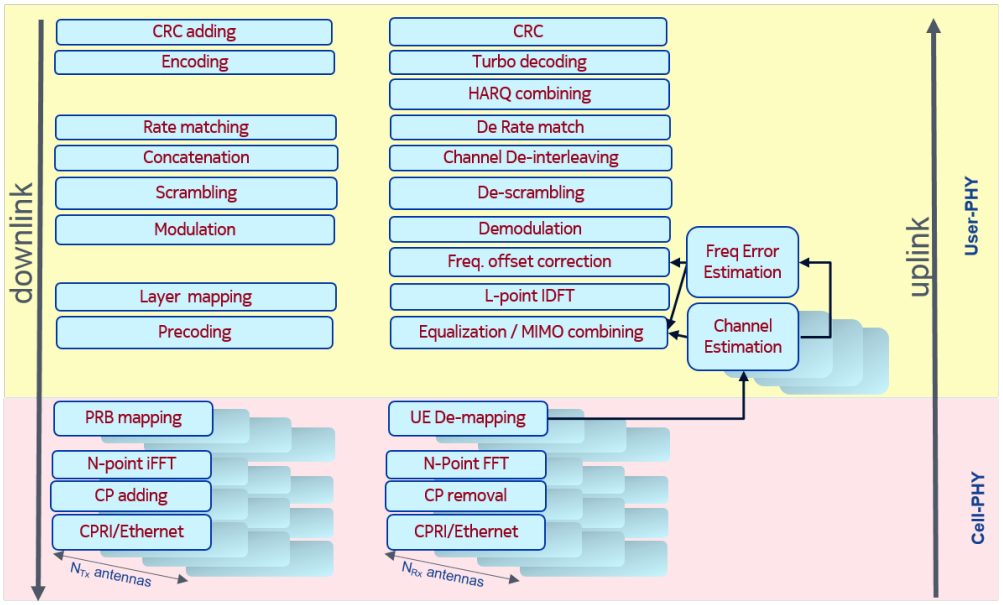


Figure 2.2 – Physical layer processing in LTE

first one performs an adapted version of BCJR algorithm and the second one uses the Viterbi-algorithm. The output of the decoder is a soft decision representing for each bit the probability of being 0 or 1, it can be represented by 8 or 16 bits, though lower precision can affect system performance particularly at low Signal-to-noise-ratio (SNR).

HARQ combining If the received block corresponds to a further transmission requested for the same data due to a negative result of the CRC, this function combines the previous and the current blocks to produce an improved output to be decoded. In fact, HARQ can be adaptive or non-adaptive, in the first case, a new Modulation and Coding Scheme (MCS) required for the retransmission is sent along with the non-acknowledgment message (NACK), thus the coded bitstream can be different for the same data at different HARQ rounds.

Rate matching Following the characteristics of the physical channel, in the rate matching process we select the bits that needed to be transmitted from the encoded bitstream. At the receiver, it can complete the received data to adapt it to the block size required for decoding.

Concatenation, de-interleaving At the transmitter, the rate matching process already includes per-block interleaving, thus we produce a single bitstream by concatenation of the outputs of the rate matching block. In reception, we

re-arrange the bitstream as required for the decoder.

Scrambling This step serves as a protection of accumulating successive errors, the order of bits is changed according to a schema know by both the transmitter and the receiver. It also provides better statistical distribution of the transmitted sequence.

Modulation / Demodulation Modulation maps the coded bits to symbols that can be – after further operations – represented in the transmitted radio signal. In LTE, the modulation scheme goes from the lowest-rank BPSK, translating each bit to a phase value, to 256-QAM on the DL, and 64-QAM on the UL. Symbol size ranges respectively between 1 and 16 (8 for the UL). In QAM, two values describe a symbol, one for the so-called In-phase (I) component, and another in the orthogonal direction for the Quadrature (Q) component. The sequence of coded bits is transformed to symbols according to the size allowed by the selected modulation order, and each symbol corresponds to a point in the constellation. The constellation points are placed following Grey coding, so that the difference between symbols corresponding to neighbor points is minimal. The position of this point is described by the I/Q values. In reception, I/Q symbols represent the actual received signal that contains some perturbation mainly due to channel noise and interference. Furthermore, they are conveyed using a finite number of bits, called symbol width, which can affect the precision of the value, thus the demodulation. The aim of the demodulation is to find the data symbol that was originally sent, based on the received I/Q values. We can find the most fitting constellation point e.g., by Zero Forcing (ZF) or Minimum Mean Square Error (MMSE) methods.

Frequency error estimation and frequency offset correction Due to shortcomings in the transmission and reception electronics, and alteration of the signal in the wireless channel, it can be affected by a frequency deviation. These functional blocks aim to quantify and compensate this error.

Layer mapping and IDFT Layer mapping consists in multiplexing the I/Q symbols according to diversity parameters (e.g. MIMO). Receiver Inverse Discrete Fourier Transform (IDFT) is needed on the UL to compensate the Discrete Fourier Transform (DFT) spread added in SC-FDMA in order to avoid high Peak-to-average power ratio (PAPR).

Precoding, equalization and MIMO combining In some cases where Channel State Information (CSI) is available at the transmitter, we can ap-

ply precoding on the signal to be transmitted to compensate the effect of the channel. If several antennas are available at the transmitter, precoding is also used over the set of antennas to maximize diversity gain. On the receiver side, the reverse processing is performed for channel equalization using the computed channel estimates (c.f. the next paragraph). With receive antenna diversity the signal coming from different antennas is combined, e.g., by Maximum Ratio Combining (MRC), or Interference Rejection Combining (IRC). In a single-user case MRC provides good performance by taking the weighted average of the signal received by each antenna. In the presence of interfering signals, IRC retrieves better the desired signal by eliminating the interference component via projecting the received signal in the direction which is orthogonal to the interfering signal [30]. Though, the knowledge of channel gains of the interfering users are needed.

Channel estimation The first function that we realize in the receiver on a per-user basis the estimation of the channel gain in the given the transmission. In LTE, channel estimation is performed using pilot symbols that are known both by the transmitter and the receiver. In fact, the received signal r is given by the product of the sent signal s and the channel gain h plus the channel noise (assumed Additive White Gaussian Noise (AWGN)) : $r = hs + n$, where s, h, n and r are complex values. During the training phase when pilot symbols are sent, the receiver knows s. Thus we can find the channel estimate $\hat{h}$ by realizing $\hat{h} = \frac{s^* r}{|s|^2}$, i.e., by zero-forcing. Other filters with slightly higher complexity, such as MMSE can give more accurate channel estimates particularly when the received signal is affected by interference. In a multi-user scenario when several transmitters send different data on the same Physical Resource Blocks (PRBs), orthogonal pilot sequences are attributed to each transmitter in order to preserve the quality of the channel estimation. With a high number of co-channel users, this can result in significant pilot overhead, though in practical systems the number of users is limited. In LTE Release 10, it has been specified to use cyclic shift of De-Modulation Reference Signal (DMRS) [31] allowing to allocate almost uncorrelated pilots to up to 12 UEs on the UL. We will discuss the implementation of UL multi-user channel estimation with cyclic shift in Subsection 3.2.4.2. Independently of the number of users, DMRS are multiplexed in a predetermined manner inside the PRBs allocated for data transmission and channel estimates are interpolated in time and frequency.

PRB mapping and de-mapping Following resource allocation decisions received from the MAC, the symbols to be transmitted are multiplexed over the resource grid shown in Figure 2.3. In the time dimension, a PRB is composed of

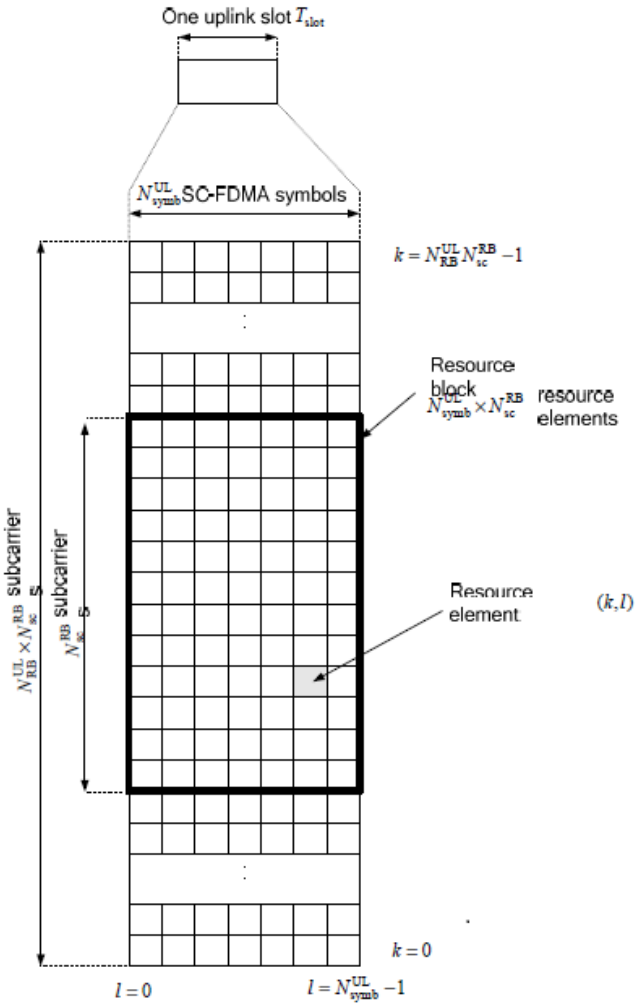


Figure 2.3 – Uplink resource grid

7 symbols that constitute a subframe (or slot) with 1 ms duration. Frequency-wise, 12 subcarriers are included in a PRB. It is possible to allocate several PRBs to the same user. Similarly, at the reception, following the known resource allocation pattern, signals are de-multiplexed to continue in the processing chain dedicated to each UE.

Frequency modulation: IFFT and FFT When every symbol has its place, we modulate the signal in frequency to place them on the required subcarrier. This is realized by Inverse Fast Fourier Transform (IFFT), allowing to convert the signal from frequency-domain representation to time-domain one that can be sent then over-the-air. Furthermore, frequency-domain symbols are handled in base-band, i.e., the carrier frequency is not considered, while IFFT transposes them to the right subcarrier frequency. In a symmetric manner, the receiver realizes the Fast Fourier Transform (FFT) to convert time-domain received signal

to base-band one in the frequency domain. FFT and IFFT computation requires the number of frequency components to be a power of 2, thus we complete the input with zeros to have the required size and remove them at the receiver. The challenge that the application of Fourier transforms arises in a real-time like radio communications is the computational complexity that scales with the size of the FFT or IFFT. In some implementations, FFT/IFFT is replaced by DFT/IDFT that allows to reduce zero-padding, but have higher computational complexity for the same size.

CP We add a Cyclic Prefix (CP) to each OFDM symbol created by the IFFT in order to avoid inter-symbol interference between adjacent symbols caused by time delay in the multi-path channel. It consists in adding a prefix of length L to each symbol, so that while they transit in the channel, the convolution with it does not result in an overlap between successive symbols [27]. Using CP allows to consider OFDM subcarriers as orthogonal channels after reception. However, it decreases the effective payload, thus CP length has to be correctly dimensioned. In LTE, two CP sizes are possible, normal CP is used for shorter channel length with less difference between different channel paths, and extended CP is needed with large distances when propagation time difference between paths can be more significant.

CPRI/Ethernet encapsulation At this point, we are at the limit of the PHY processing and communication with the RRH has to take place. For this, the solution mostly used in first generation C-RAN deployments is CPRI, designed for the purpose of transporting radio signal over other media. In evolved C-RAN we plan to use standard Ethernet connection between the BBU-pool and the RRHs, in this case, the data has to be encapsulated in Ethernet packets to be transported in the network independently of the technology used for fronthaul connection.

2.1.2.2 Multi-antenna systems

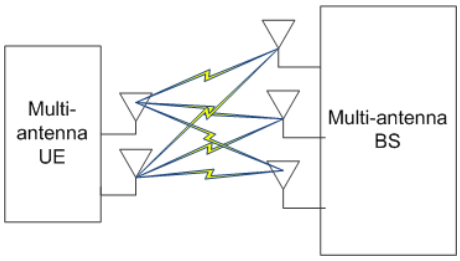
MIMO transmissions are today used in several wireless standards such as WiMAX, WiFi 802.11n and LTE. The theory of multi-antenna nodes was developed in the 90s, the interest of spatial multiplexing was initially described in [32] and the benefit of multipath propagation of signals in MIMO OFDM systems was identified in [33]. The basic idea is that using an adapted space-time coding, we can benefit from channel diversity when both the transmitter and the receiver have several antennas. This allows to create – similarly to what OFDM does in the frequency dimension – independent parallel channels where we can transmit several datastreams simultaneously. A simple example of space-time code

is Alamouti's one [34]. Bell Labs researchers have also defined codes for layered transmission [35] performing iterative MMSE reception.

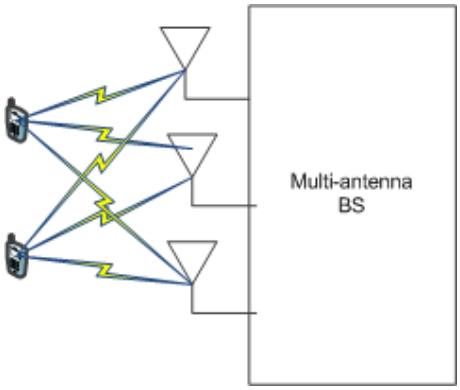
Besides full MIMO systems where both the transmitter and the receiver has multiple antennas, we can consider the case of Multi-User MIMO (MU-MIMO) (or virtual MIMO), where a multi-antenna BS communicates with users having one or more antennas. In C-RAN, there is a third option that we commonly call *network MIMO*, where the BBU pool with several RRHs serves the users. In practice, this configuration is not a straightforward extension of single-cell MIMO processing, as we have mentioned previously, even in a centralized architecture, multi-cell cooperation mechanisms are needed. We will describe Joint Transmission (JT) and Joint Reception (JR) – that cover indeed the network MIMO case – in Section 2.1.3. The various configurations of MIMO systems are depicted in Figure 2.4.

A major advantage of MIMO is to enable beamforming, i.e., directional transmission towards the intended receiver using spatial precoding over the antennas. By concentrating the transmission power in a given direction, the SNR increases. It allows also to minimize interference to other users placed in a different direction. Complex precoding weights applied to the analog signals are more precise, but require to be applied for each antenna individually, while digital beamforming is realized when the symbols are spread over the antenna array. With a high number of antennas, it can be more convenient to use the second option, however it can limit the capacity gain intended by using massive MIMO.

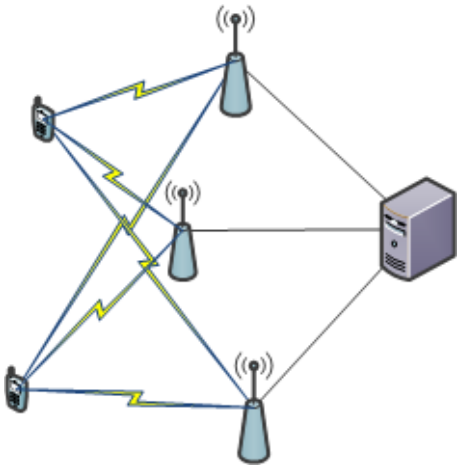
In the past few years, massive MIMO also called large-scale antenna system was considered in a high number of theoretical or practical research papers, e.g., [36–38]. In fact, it is a very promising technology for 5G systems, since with a careful design it can significantly increase cell capacity and also energy efficiency. Major challenges in its practical deployment are the pilot contamination that can decrease the accuracy of channel estimates, UL/DL non-reciprocity of the channel and difficult hardware implementation. The massive MIMO concept assumes a single stand-alone BS with very high computing capacity, thus it is complementary to the C-RAN approach where smaller access points with a limited number of antennas are used and the benefit is from processing several cells together.



(a) Single-user MIMO transmission with multiples antennas both at the UE and the BS.



(b) Multi-user MIMO scheme: several UEs with one antenna communication with a multi-antenna BS.



(c) Network MIMO configuration where UEs are served by distributed APs with one or several antennas connected to a centralized base-band processing unit.

Figure 2.4 – Various MIMO configurations

2.1.2.3 5G New Radio design

As in this thesis we aim to propose solutions for future mobile networks, we have to study the evolution of the RAN and the PHY processing with respect to the one defined for LTE that we have described in Subsection 2.1.2.1. Today, 5G requirements and service types are clearly defined, and each service has its own implementation and deployment challenges. Due to spectrum scarcity below 6 GHz, higher frequency bands such as millimeter-wave are considered to be very promising for 5G deployments [39]. Moving to higher frequencies results in different propagation conditions that affect the network design and require new techniques in the RF, e.g., for beamforming and ADC/DAC. However, their impact on the PHY processing should be limited since OFDM-based solutions perform well in the new frequency bands too. Obviously, parameters such as CP, or the subcarrier spacing have to be adapted, but the overview of the signal processing functions is likely to be similar [40].

While a consensus is reached in 3GPP about keeping OFDM as the basis for 5G Enhanced Mobile BroadBand (eMBB) waveform, several variants of it have been proposed. A structural, operational and performance analysis of contestant waveforms that are possibly applicable for 5G is provided in e.g., [41, 42]. Authors in [43] enumerate various waveform candidates in the perspective of applying them in a mm-wave system, they suggest that DFT-spread OFDM would be the most suitable. DFT-spread OFDM along with Single-Carrier Orthogonal Frequency Division Multiplexing (SC-OFDM) is considered in 3GPP as the most relevant propositions [44], their performance evaluation is still ongoing particularly regarding PAPR, out-of-band emissions and spectral efficiency [45, 46].

All these promising results show that OFDM still has its place in 5G New Radio, especially for eMBB, and only minor changes in the PHY processing chain would be needed. Based on this conclusion and for practical considerations to demonstrate our results in real implementations, in this thesis we take for the examples the PHY processing as it is standardized in LTE, but we aim to solve challenges that appear in the context of 5G. Our results can be transposed in 5G RAN, once the implementation details are fully defined.

2.1.3 Multi-cell processing in LTE and beyond

The issue of Inter-cell Interference (ICI) limiting network performance and QoS of cell-edge users has been considered in uncountable research works on mobile radio communications. Also, lots of different solutions has been proposed to deal with this limitation. There are several ways to cope with ICI:

- *Interference avoidance*, when interferers are coordinated in a way to not to disturb each other's transmissions any more. This concept is used e.g., in

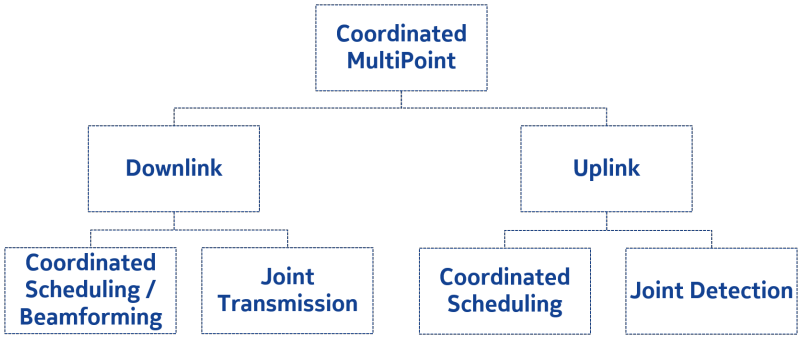


Figure 2.5 – Classification of CoMP techniques in LTE

Coordinated Scheduling (CS), Enhanced ICIC (eICIC) and Dynamic Point Blanking (DPB) techniques.

- *Interference alignment* aims to benefit from spatial degrees of freedom to reduce the interference efficiently, by performing Coordinated Beamforming (CB) for example.
- *Interference mitigation* where we apply receive techniques eliminating the interference component of the received signal, with our without using information about the interference channel.
- *Interference exploitation* introduced in [47] comprises all the techniques where we let interference happen and realize specific signal processing at the receiver to retrieve the interference-free signal. JR methods include Successive Interference Cancellation (SIC) and multi-user MMSE.

In many cases, methods of first three types give good performance results, however, the last approach mentioned seems to be the most adapted to centralized network architectures. In fact, the possibility of sharing user and control data among several cells can be very beneficial for turning interference to useful signal. However, many challenges have to be solved to enable practical deployments.

In LTE, CoMP, present since Release 10, aims to provide better QoS through multi-cell cooperation. We show in Figure 2.5 the classification of these techniques. Both in DL and UL, there are two categories: coordination (CS/CB) and joint processing (JT and JR). Coordination requires only to share control-plane information between cells, while for joint processing, the data-plane sharing is also needed. As its name suggests, CoMP requires communication between several APs located in a distributed manner on the cell-sites. There are basically two ways of inter-cell communication [1]: in distributed RAN (i.e., legacy deployments with distributed BSs including RF and the complete BB processing

for the given cell) BSs are inter-connected and mutually exchange information, or in C-RAN, communication happens among BBUs hosted inside the same CO and each BBU is linked to an RRH which serves only to relay the signal between the users and the CO. These two possibilities are depicted in Figure 2.6 [1].

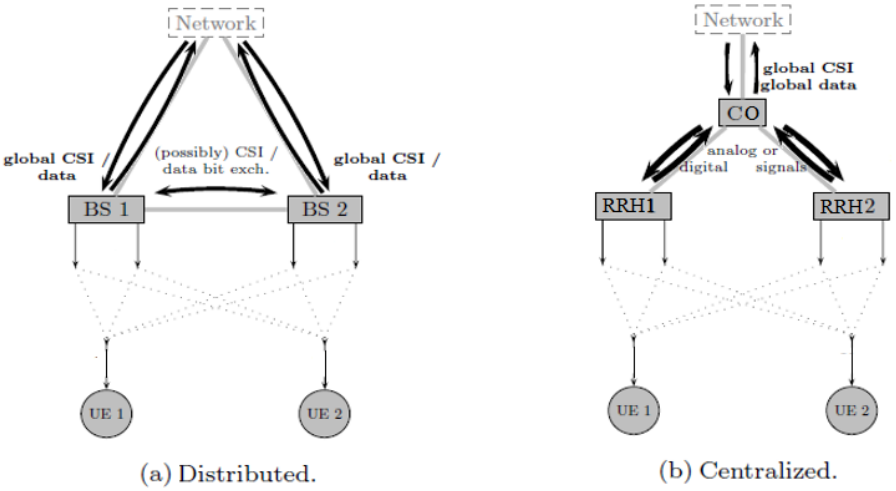


Figure 2.6 – Architecture types for multi-cell cooperation.

For communication between BSs, the X2 protocol was defined in [48] in order to provide a standard interface for information exchange. Although it is convenient to exchange small amounts of data, e.g., CSI, transmitting user data can result in high network load. Furthermore, sending information between locations separated by a long distance can introduce significant delay, which is not compatible with real-time operation. For CoMP techniques based on coordination, realizing relatively localized cooperation in a distributed manner is possible and beneficial. However, with large number of cooperating BSs increases not only the communication delay, but also the complexity of coordination algorithms, that makes it unsuitable in practical systems.

The centralized architecture still allows multi-cell coordination, and by its structure, it enables to satisfy rate and delay requirements of inter-cell communications needed for multi-cell joint processing. As we have mentioned in Section 2.1.1, in C-RAN, BB processing is performed in the same server for several cells, thus it is easy to exchange even a high amount of data with low latency, as needed for JT and JR. Often, the structure of CS/CB algorithms requires a centralized controller entity, even if BB processing can be performed in the BSs. Even for algorithms that support distributed implementation, computation and communication overhead can be reduced by centralization. Though centralization requires communication between each distributed RRH and the CO, which

is ensured via the fronthaul links. If this scheme is simpler than distributed cooperation, but the cooperating cell-sites are connected to a single centralized unit, the performance of joint processing depends on the quality of fronthaul transmissions.

CoMP allows to decrease the interference by coordination and benefit from spatial diversity with JT and JR, however it does not include explicitly interference exploitation. To enable this feature, a combined approach is needed, where neighboring cells schedule their users in a cooperative manner to the same PRBs, and multi-user transceiver techniques are used in the BBUs. These techniques are similar to the ones used in MIMO transceivers, as despite the distributed placement of the antennas, in C-RAN, they can potentially operate jointly among the RRHs. Furthermore, having a few antennas per RRH and jointly process several cells allows to reduce the correlation between the antenna elements with respect to a single massive MIMO BS [49] and benefit from channel diversity thanks to a richer propagation environment. Though, transposing MIMO techniques to C-RAN is not straightforward, since the internal structure of the CO keeps for practical reasons, the separation of processing per cell and requires dynamic and scalable implementation. These challenges and proposed solutions will be discussed in Part 3.

2.2 Preliminaries

2.2.1 Uplink multi-cell JR

Densification of APs planned in 5G in order to provide high-rate connectivity everywhere to a large number of eMBB users. In such a configuration, it happens more often that a UE is located in the coverage area of several APs that, in C-RAN, have their BB processing performed in the same CO. The basic idea of multi-cell JR is to schedule cell-edge users on the same PRB and use a MIMO receiver across the cells so that we fully exploit received signals through this cooperation. It can be also seen as an extension of UL NOMA to a multi-cell configuration enabled by C-RAN.

For geometric reasons, in most of models, as well as in real network deployments, overlap between cell coverage areas happens most often between two cells. Three or more cells overlap with much less probability, for this reason, in the following we consider only cooperation between 2 cells. This configuration is depicted in Figure 2.7. For simplicity, we consider in this section single-antenna RRHs and demonstrate JR through the example of 2 cooperating UEs (each of them having also one antenna). This model will be extended in Part 4.

Note that even if users can generate interference for cells that are not in their

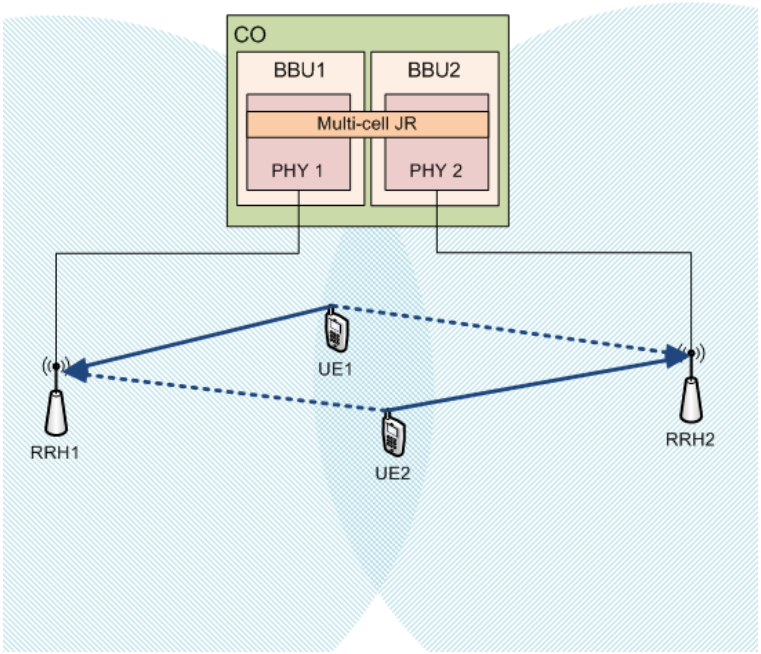


Figure 2.7 – UL JR of two cell-edge users.

first-step neighborhood, due to higher distances, attenuation of the signal is more important, so it would be difficult to exploit that interference. Since interference from many users is added, it can affect the transmission, but each interference component individually is small. Furthermore, as we will see in the following, some MIMO receiver techniques are more efficient when received signals have similar power.

Users broadcast a signal x_i, with $i \in \{1, 2\}$ (forming the vector $\boldsymbol{x} = (x_1, x_2)^T$) with unit power. Both receivers get the sum of the signals transported via point-to-point the wireless channels represented by the channel coefficients h_{ij}, with $i, j \in \{1, 2\}$ which form the channel matrix $\boldsymbol{H} = \begin{pmatrix} h_{11} & h_{21} \\ h_{12} & h_{22} \end{pmatrix}$.
We denote by $\boldsymbol{h}_i = (h_{i1}, h_{i2})^T$ for $i \in \{1, 2\}$ the channel vector of user i, thus we can write $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2)$. The received signal at the RRH j can be written as

$$y_j = h_{1j}x_1 + h_{2j}x_2 + z_j \text{ with } j \in \{1, 2\} \tag{2.1}$$

which gives in a the matrix formulation

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}, \tag{2.2}$$

where $\boldsymbol{z}$ is the AWGN vector with covariance $\sigma_z^2 \boldsymbol{I}_2$.

We assume the channel gain is perfectly known at the receiver. In the following, we describe a few examples of MIMO detection methods that can be used to estimate the sent symbol (i.e., QAM constellation point) of each user. The vector of the retrieved signals is denoted by $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2)^T$.

Zero-forcing (ZF) Without multi-cell processing on the receiver side, the SISO channel equalization technique conventionnaly use is ZF. It consists in computing the following estimates

$$\hat{x}_i = \frac{h_{ii}^* y_i}{\mid h_{ii} \mid^2}. \tag{2.3}$$

We can easily see, that the interference component with a power equivalent to the power of the useful signal corrupts the estimated signal.

Maximum Ration Combining (MRC) As we have already mentioned in Subsection 2.1.2.1, MRC is often used for multi-antenna combining in single-cell MIMO systems. It estimates the signal of user i based on both components of the received signal and the channel gains between user i and both receive antennas.

$$\hat{x}_i = \frac{h_{i1}^* y_1 + h_{i2}^* y_2}{\mid h_{i1} \mid^2 + \mid h_{i2} \mid^2}. \tag{2.4}$$

This method takes advantage of multi-antenna diversity, however, it does not take into account that the received signals are linear combinations of x_1 and x_2.

Successive Interference Cancellation (SIC) It worth mentioning, that in a scenario where users do not transmit with similar power of pathloss is different for each user, we can combine the above single-user detection techniques with SIC in order to estimate more accurately the signal of each user. In fact, if we arrange users by transmission power, the one with the highest power creates the most of interference for other users, and it is also more likely that single-user detection gives a correct results, since the term sent by that user dominates in the received signal. For our two-by-two example, it would result in a two-step detection where, assuming $\mid h_{11} \mid^2 > \mid h_{22} \mid^2$, we compute first $\hat{x}_1$ as in 2.4

$$\hat{x}_1 = \frac{h_{11}^* y_1 + h_{12}^* y_2}{\mid h_{11} \mid^2 + \mid h_{12} \mid^2}, \tag{2.5}$$

then, we estimate x_2 based on the received signal from which we subtract the estimated interference: $\boldsymbol{y}' = \boldsymbol{y} - \boldsymbol{h}_1 \hat{\boldsymbol{x}}_1$,

$$\hat{x}_2 = \frac{h_{21}^* y_1' + h_{22}^* y_2'}{\mid h_{21} \mid^2 + \mid h_{22} \mid^2}, \tag{2.6}$$

This interference-aware receiver technique improves the throughput thanks to accurate symbol detection [50] in the configuration where interfering users transmit with different powers, e.g., when multi-cell cooperation is considered in the whole cell, not only in the cell-edge region. However its performance depends on the correct selection of the first user, if this selection is wrong or channel gains are similar, symbol detection results can we affected by error propagation. In addition, inherently to its structure, it is executed sequentially, thus introduces computational delay, especially when there are several interfering users.

Maximum Likelihood (ML) For a general case of multi-user detection where we do not have any *a priori* knowledge of power allocation of pathloss, the optimal method to estimate the signal of every user is Maximum Likelihood (ML) detection [1]. It is performed by computing

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{t}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{t}\|^2 \tag{2.7}$$

that finds estimate where the symbol detected for each user is the closest one to the corresponding constellation point in terms of Euclidean distance. The drawback of this method is its high computational complexity when applied to multi-antenna and multi-user system, since it exponentially increases with the number of antennas and the number of users.

Multi-user MMSE A computationally more accessible method allowing to exploit interfering signals is MMSE. It computes symbol estimates based on a detection matrix that takes into account channel estimates on each link and also noise covariance. The MMSE filter exploits for each user the useful component of each received signal, while reducing the effect if interference signals on the estimated symbol. It consists in computing the MMSE equalizer matrix

$$\boldsymbol{W} = (\boldsymbol{H}\boldsymbol{H}^H + \sigma_z^2\boldsymbol{I})^{-1}\boldsymbol{H} \tag{2.8}$$

and apply it in

$$\hat{\boldsymbol{x}} = \boldsymbol{W}^H\boldsymbol{y}. \tag{2.9}$$

In fact, it separates user signals x_1 and x_2 based on the received signals at each antenna which are their linear combinations. Note that the computation of (2.9) can be performed simultaneously for each user with the knowledge of every received signal and the row of $\boldsymbol{W}$ corresponding to the given user. [1]

[1] This technique can be improved by combining with interference cancellation [1], however computational delay would slightly increase, and the performance of MMSE in terms of Block Error Rate (BLER) is enough to satisfy the requirements of higher layers for mobile data

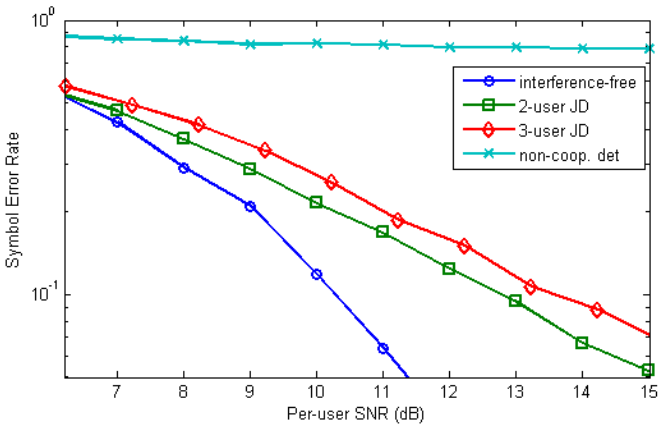


Figure 2.8 – Performance of multi-user MMSE detection compared to single-user ZF (non-cooperative detection). JD (Joint Detection) is realized by multi-user MMSE as defined in equations (2.8) and (2.9). In the interference-free case, we consider that only one user transmits without any interfering signal.

We show in Figure 2.8 that multi-user MMSE enables accurate reception in a scenario with interference, where ZF cannot be used. With multi-user reception we increase overall spectral efficiency, as only 1 PRB is used for the uplink transmission of several users. These results confirm that the same multi-user MMSE receiver can also be used in the less frequent case of cooperation between 3 cells. However, we observe that as the number of users transmitting on the same PRB increases, we need to transmit at slightly higher SNR to reach a given error rate. In fact, at the cost of increasing the transmission power by about 2.5 dB, we can double the spectral efficiency of a multi-user transmission.

We can see that among the enumerated receiver techniques that are enabled by C-RAN, multi-user MMSE can satisfy practical constraints and provide the receiver performance needed for the targeted eMBB service. More insights about the receiver design are given in Subsection 3.2.4.3.

2.2.2 Handling fronthaul limitation

As we have mentioned in Section 2.1.3, a major limitation for efficient C-RAN deployment is fronthaul connectivity between the RRHs and the CO. In this section, we describe first why fronthaul rate limitation impacts 5G C-RAN design given the available technologies, then we present the various functional split options between the RRH and the CO, aiming to reduce the required fronthaul rate.

---

traffic.

2.2.2.1 Fronthaul requirements and available technologies

The main motivation of reducing transmission rates between the RU and the CO is that the cost of Ethernet transport infrastructures required for transferring modulated symbols in a 5G scenario would be very high. Higher data-rate 5G transmission (up to 10 Gbps [26]) is possible by allocating large bandwidth e.g., in mm-wave frequency bands. RRHs should also comprise many antennas, especially where massive MIMO is deployed. Thus, the amount of data to be transferred on the fronthaul links – with a low latency – would increase significantly due to higher bandwidth, more antennas and more users.

Several technologies are available for fronthaul links, as of today, microwave and optical fiber are the mostly considered candidates, but in 5G systems millimeter-wave fronthaul and its co-existence with RAN transmissions is also studied. The advantage of wireless fronthaul is to be easy to deploy w.r.t. optical fiber, however, it can achieve high transmission rate mainly in Line-of-Sight (LoS) configurations over short distances, which limits the use cases where it can be deployed. Highly dependent on the deployment conditions, wireless fronthaul links can in some cases result in bad quality transmissions that degrade RAN performance.

In urban mobile network deployments, fiber infrastructure is often already available and it can provide relatively high capacity connection between the RUs and the CO, which can be located at up to 20 km from the RU.[2] Fiber links with Wavelength-Division Multiplexing (WDM) make possible the increase of available fronthaul capacity compared to dark fiber, but optical equipments realizing Dense WDM are too expensive to be deployed as dedicated fronthaul infrastructures. Sharing metropolitan fiber infrastructures can be a cost-efficient solution if we are able to control transport latency. To realize low-latency and reliable data transfer between the RU and the CO, data-flow routing through the fronthaul Ethernet network can be controlled using Software Defined Networking techniques. Nevertheless, with such a shared network infrastructure, the lowest possible data-rate usage for fronthaul has to be targeted in RAN design, to ensure its efficient usage.

2.2.2.2 Fronthaul quantization of UL signals

In the previous section we have defined a dynamic functional split between the RUs deployed on the cell sites and the CO where a part of BB processing for many cells takes place. This split allows to reduce fronthaul rate requirements

[2]The distance of 20 km enables a fronthaul Round Trip Time (RTT) below 1 ms, which is needed in LTE to satisfy HARQ timing constraint. If future mobile network specifications include a less constraining HARQ protocol, the distance between RUs and the CO can be increased in theory, however, to ensure low transmission latency, it is not expected to be much longer than 20 km.

by placing a major part of PHY functions in the RUs, but executes multi-cell JR and user-PHY functions of some selected UEs in the CO. For these users, we still need to transfer I/Q symbols via the fronthaul links. In the current section, we study in the light of rate-distortion theory the impact of representing these originally analog values over a finite number of bits.

Indeed, in digital communication system it is necessary to quantize analog radio signals to process them, in all radio APs, ADC/DAC is included besides RF processing, regardless the RAN architecture used. In C-RAN, the importance of the degree of quantization becomes more significant, since it determines the fronthaul transmission rate. Though, it is obvious that too much quantization affects the quality of the transmission. Rate-distortion theory defines more precisely this concept [51] by establishing the minimum quantization rate needed to represent it accurately a signal with a given probability distribution.

In several works, reduction of the fronthaul rate in C-RAN joint processing scenarios by the mean of adapted compression schemes was investigated (see [52] and references within). In particular, on the uplink, each RRH compresses its received signal following a distributed scheme that takes into account the correlation with its neighbors, then it forwards the compressed signal to the CO which performs joint decompression of the signals received from the whole set of RRHs. A different fronthaul compression optimization approach consists in joint design of the fronthaul quantization scheme and the beamform precoders at the UEs having several transmit antennas, in order to minimize the effect of quantization and allow to forward I/Q samples at a lower rate [53]. A given compression method results in a rate-distortion function ensuring accurate fronthaul transmission at a reduced rate. These results based on information-theoretic considerations and provide efficient signal processing methods for fronthaul compression, however, in practical systems, where short processing time is indispensable, including such a compression algorithm can not be envisaged currently. For this reason, we consider here a linear point-to-point quantization of I/Q samples forwarded via the fronthaul links.

We take the same simple system model with two users and two RRHs as in Section 2.2.1, and we add in the signals transmitted on the uplink over the fronthaul links, thus each RRH receives y_j and forward it to the CO, which gets $\hat{\boldsymbol{y}}$. Based on the received signal, the channel is also estimated in the CO in order to obtain $\hat{\boldsymbol{H}}$. The system is depicted on Figure 2.9.

Our aim is to determine the loss of information due to quantization that still allows to estimate the sent signal accurately. The distortion d_j between the signal y_j received by the RRH and the compressed signal $\hat{y}_j$ received by the CO

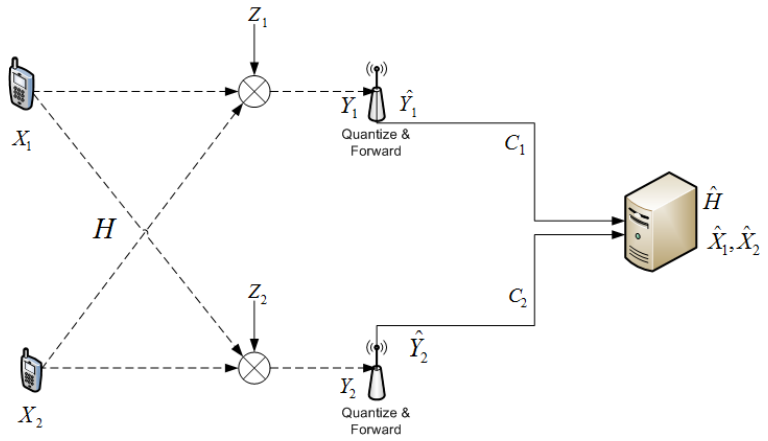


Figure 2.9 – C-RAN system model with 2 RRHs and 2 users

is defined as the squared-error distortion between y_j and $\hat{y}_j$.

$$d_j = D(y_j, \hat{y}_j) \coloneqq \mathbb{E}[\mid y_j - \hat{y}_j \mid^2 \mid \hat{\mathbf{H}}] \tag{2.10}$$

The maximal distortion happens when the received information (i.e. the information on the source signal delivered by the received one) has Gaussian probability distribution, in this case the distortion is also Gaussian and we denote its variance by $\sigma^2_{d_j}$. Then, we can use the following bound to characterize the distortion:

$$d_j \leq \sigma^2_{d_j}. \tag{2.11}$$

The minimum achievable rate of a signal quantized with distortion D is given by the mutual information between the initial (received) signal and the quantized (forwarded) one [51, Theorem 10.2.1]. If this rate is lower than the capacity of the fronthaul link, the used quantization allows an accurate transmission where the distortion does not exceed a given variance $\sigma^2_{d_{lk}}$. We can write for a point-to-point link lk:

$$\begin{aligned} r_j &\leq c_j \\ r_j &\geq \min_{p_{\hat{y}_j \mid y_j} : D \leq \sigma^2_{d_j}} I(\mathbf{Y}_j, \hat{\mathbf{Y}}_j \mid \hat{\mathbf{H}}) \end{aligned} \tag{2.12}$$

The power of the signal y_j received at the RRH j given the channel estimate is denoted by $\sigma^2_{\mathrm{Y}_j \mid \hat{\mathbf{H}}}$.We can apply a scalar scaling on y_j so that the resulting distortion allows to quantize it at the optimal rate $r_j = c_j$. We have the following bound on $\sigma^2_{d_j}$

$$\sigma^2_{d_j} \leq \sigma^2_{\mathrm{Y}_j \mid \hat{\mathbf{H}}} 2^{-c_j} \tag{2.13}$$

We define the scaling factor $\alpha_j = \frac{\sigma^2_{\mathrm{Y}_j|\hat{\mathbf{H}}} - \sigma^2_{d_j}}{\sigma^2_{\mathrm{Y}_j|\hat{\mathbf{H}}}}$ $\forall j \in \{1,2\}$. As the inequality in (2.13) holds for any quantization satisfying (2.12) the following bound characterizes the distortion introduced by the quantization at a rate c_j.

$$\frac{\sigma^2_{d_j}}{\alpha_j} = \frac{\sigma^2_{d_j}\sigma^2_{y_j|\hat{\mathbf{H}}}}{\sigma^2_{y_j|\hat{\mathbf{H}}} - \sigma^2_{d_j}} \leq \frac{\sigma^2_{\mathrm{Y}_j|\hat{\mathbf{H}}} 2^{-c_j}}{1 - 2^{-c_j}} \quad \forall j \in \{1,2\} \tag{2.14}$$

In Part 4 we will make use of this relation to develop a performance metric for uplink transmissions in C-RAN, taking into account fronthaul quantization.

2.2.2.3 Functional splits in C-RAN

With the emergence of 5G research challenges, different options for splitting the base-band processing between the RRHs (or RUs) and the BBU server are being considered. Initially, in Cloud RAN architecture for LTE deployments, RRHs realized only RF and ADC/DAC. This resulted in a relatively high, but sustainable traffic on the fronthaul links that ensure the connection between the RRHs and the CO. For LTE networks, BBU functional splits are studied in [54]. To accommodate the characteristics planned for 5G deployments, reducing centralization was proposed in order to generate less FH traffic and allow higher FH latency. In fact, with massive MIMO, higher bandwidth and transmission frequency, fronthaul traffic can become too heavy for affordable communication links. Various options of functional splits between the RU and the CO were investigated in [23,55]. Next Generation Fronthaul Interface (NGFI) consortium gathering MNOs and network equipment constructors define Layer 1 and Layer 2 split possibilities and evaluate them regarding FH throughput and delay. Figure 2.10 shows the different interfaces described in [55]. The requirements in terms of fronthaul rate and latency are also evaluated for these interfaces in [55], though only the LTE use case is considered.

For the small cell use-case, FH bandwidth and latency with several split options are described in [56]. The analysis shows that data-plane traffic requirements are the same for all of the functional splits where at least PHY functions are in the RU, only latency requirements decrease with more functions executed in a distributed manner. Advantages of different split options are analyzed in [24] regarding energy and cost saving. Authors quantify multiplexing gains and show the importance of effective dimensioning of the fronthaul network to ensure good QoS. They focus on three splits: the one at the interface between the RF and base-band processing (denoted by the n^o 5 in Figure 2.10), the one where UE-cell functions are distributed in the RUs (n^o 4), and the one between PDCP and RLC layers (n^o 1') and, as a conclusion, they propose a C-RAN deployment

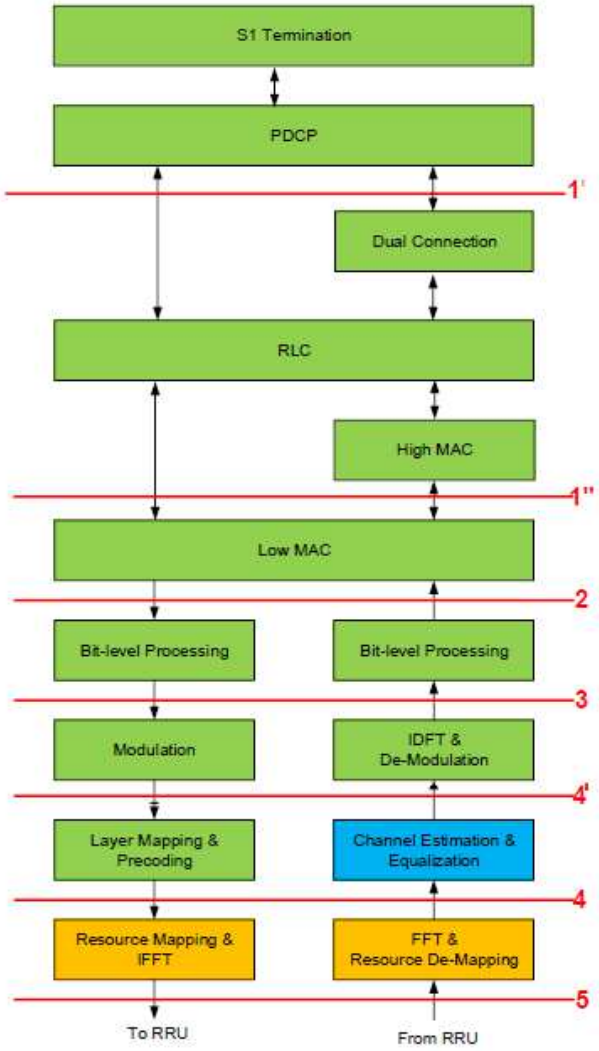


Figure 2.10 – Split options defined by NGFI.

strategy including fronthaul dimensioning. It worth noting, that in addition to data-plane fronthaul traffic considered in most of the research works on the subject, for higher splits such as n° 2, 1" and 1', the fronthaul rate needed for control messages is not negligible for the design of the fronthaul interface for C-RAN deployments.

2.3 Novel user-centric asymmetric fronthaul split

In this section we focus on the fronthaul design to ensure low transmission rate while maximizing the benefits of C-RAN, particularly regarding features requiring multi-cell cooperation. First, we describe the functional requirements of joint processing on the DL and the UL, then, we propose a dynamic fronthaul functional split satisfying those requirements.

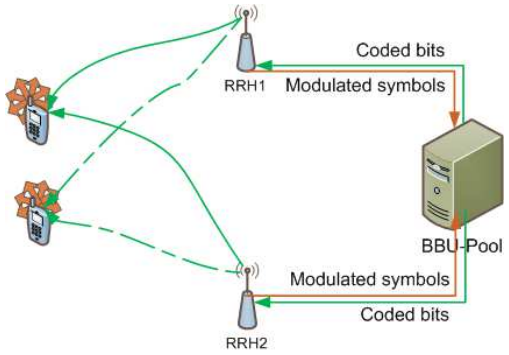


Figure 2.11 – Fronthaul transfer for joint processing

2.3.1 Functional requirements for multi-cell processing

As we have described in Section 2.1.3, CoMP techniques based on multi-cell coordination (e.g., CS and CB) require to perform scheduling and precoding in a cooperative manner for several neighboring cells, in order to realize interference avoidance or interference alignment. This coordination is realized through the MAC layer and it is compatible with all fronthaul splits below it. Yet, joint processing performed at the PHY level can depend on the level of centralization. Figure 2.11 summarizes data required to be shared for DL JT and UL JR.

Downlink Joint Transmission Realizing multi-cell transmission on the downlink means that a UE receives its data from several APs. In terms of connectivity, the UE is still attributed to one cell where happens all its control-plane traffic, but centralized processing of several cells enables coherent joint transmission to leverage diversity gain for the data-plane. This multi-cell transmission scheme does not have any impact on the reception processing of the UE – thanks to adequate precoder design –, neither it requires specific PHY functions.
For JT, the data to be sent to the UEs and the channel statistics from every participating cell have to be shared among the BBUs. In the CO, we need to compute cooperatively precoding matrices for every transmitter. Precoding matrix and uncoded bits are then forwarded to the physical layer where corresponding processing is executed. The coordination of the JT can take place in the MAC layer or in a more scalable architecture in a centralized coordinator module common to all cells connected to the CO. In both cases, executing MAC layer and upper functions in the CO is sufficient for JT.

Uplink Joint Reception Multi-user JR allows to exploit interference between users and increase spectral efficiency, but requires RAN centralization enabling low-latency communication between the BBUs. More insights about the imple-

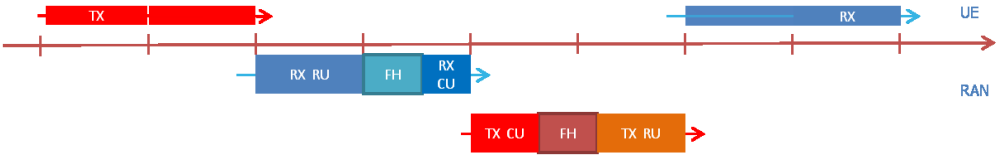


Figure 2.12 – Time budget in LTE for UL HARQ in C-RAN

mentation are provides in Subsection 2.2.1. To be able to jointly process signal received by several RRHs, modulated symbols (at interface 4') as well as uplink channel estimates of every cell on both direct and interference links need to be shared between the BBUs. Similarly to downlink JT, coordination is performed at the MAC level (locally or by a centralized coordinator) and scheduling decisions are forwarded to the UEs, which do not perceive that JR is used. Then received I/Q symbols have to be forwarded to the CO where JR is executed and the user signal is extracted from received signals of every RRH included in the JR cluster. User-PHY processing is then performed without any modification and accurate decoding can be realized. To summarize, multi-user JR requires the PHY functions above interface 4' of users implied in the joint processing to be executed in the CO.

2.3.2 Low latency HARQ

As low round trip time is required between the UE and the network to satisfy feedback time constraint required by HARQ. HARQ protocol exploits CRC to request new transmission if the received data is erroneous, thus in LTE, acknowledgment message (ACK) or NACK has to be received by the transmitter within a given time. This repeat request mechanism is expected to be included in 5G in a similar way, however the length of required RTT will depend on system parameters.

This limitation is due to the fact that the number of simultaneously operating HARQ processes is limited to 8, thus from the initiation of a subframe by its sender, it has to complete the reception of the feedback from the receiver under 8 subframes, i.e., in 8 ms for LTE. Since du to synchronization requirements we account the duration of one subframe both for transmission and reception processing and also one subframe for transiting the air interface, 4 ms are available on the network side to process and send UL HARQ reply. We illustrate in Figure 2.12 the steps performed in a C-RAN deployment for producing HARQ feedback message. Note that in this configuration fronthaul transport time is added to the processing time in the RU and the CO. Moreover, data transfer needs to be performed both in uplink and downlink directions, so the fronthaul

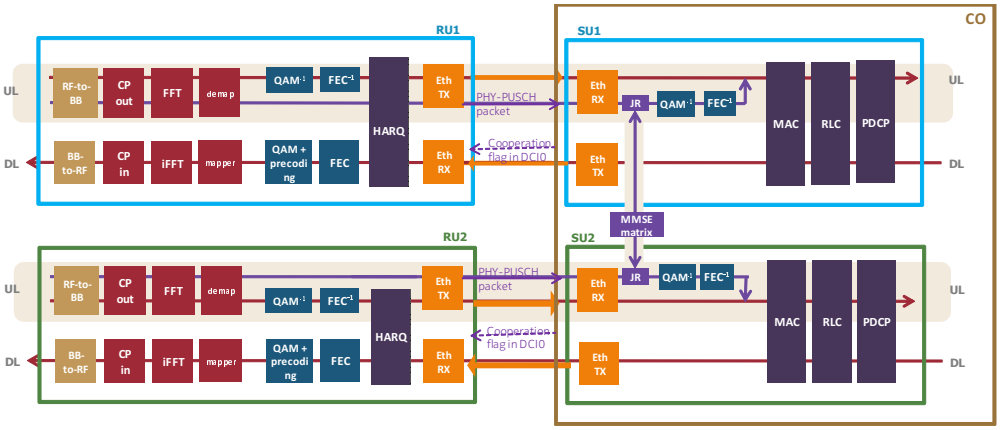


Figure 2.13 – Dynamic intra-MAC split with PRB selection

latency cannot exceed 0.5 ms. Note that in 5G shorter subframes will be used, but processing time can be longer, thus the fronthaul latency limitation will be similar.

To take into account this constraint, when high fronthaul delay is measured, we can process HARQ feedback in the RU for users that are not selected for JR, as indicated in Figure 2.13. The feedback decision can be separated from the MAC and performed in the RU directly after decoding. However, this solution does not decrease user-plane fronthaul traffic, so in case of low transport delay maintaining HARQ in the CO should be preferred in order to maintain the pooling gain. This solution allows to satisfy timing constraint for a major part of users, even if delay for JR users can happen to be longer in some cases.

For JR users, as we process user-PHY functions including decoding in the CO, it is not possible to directly apply such offloading of HARQ decision in order to reduce the RTT. For single-user MIMO reception in C-RAN, it has been proposed in [57] to estimate the channel at the RRHs and based on channel estimates compute the error probability [58] enabling to send an early feedback already from the RRH. One can consider to extend this scheme to MIMO channels following the results in [59], that should allow even for multi-user MIMO users (i.e., JR) to receive low latency feedback before completing PHY processing in the CO.

2.3.3 Dynamic split with user selection

In this subsection we describe the dynamic functional split between the RU and the CO that we have defined in order to accommodate joint processing while ensuring the low fronthaul rate.

2.3.3.1 Architecture

The ideal candidate between interfaces shown in Figure 2.10 to reduce fronthaul data rate and make it load-dependent would be the transmission of uncoded bits (interface 2) on the limit of PHY and MAC layers, since it allows to achieve a significant pooling gain in the CO. However, uplink JR requires to forward modulated symbols from the RU to the CO (interface 4'). By realizing the fronthaul split depicted in Figure 2.13, we ensure low fronthaul rate and joint processing on the UL and the DL where needed. In this dynamic split UL and DL PHY functions are executed in the RU except for UL JR users who see their user-PHY processing takes place in the CO.

UL cell-PHY functions up to resource demapping are always executed in the RU. After demapping of the PRBs, the I/Q symbols on the PRBs allocated to the UEs implied in JR should be mapped in Ethernet packets. The other PRBs corresponding to regular users which transmit individually would complete PHY processing in the RU and forwarded as decoded bits. On the downlink, sending uncoded bits with the associated control messages is enough to enable JT.

This flexible functional split realizes on one hand, DL/UL separation, on the other hand, it allows user-based processing chain while benefiting from joint processing in C-RAN, following the dynamic structure required for 5G [39].

Note that with the split at the interface 2 transmission of control data for each user both on the UL and the DL need to be added. This control-plane traffic is expected to be around 30%-50% of the data-plane one [55]. It contains information related to scheduling, modulation and precoding. A split at interface 4' does not require any additional control data other than a signaling of UL PRBs which need to be extracted and forwarded at that interface.

Synchronization It is indispensable for multi-cell joint reception and transmission to synchronize all the elements of the access network in time and frequency. For JT, simultaneous transmission of the signals by all the serving APs is required and for JR, signals received on the same time-slot need to be processed jointly in the CO. In current mobile network deployments timing coordination between the access network and UEs is performed using random access channel signals.

Sharing the clock between all APs allows also to get synchronous transmission both on the UL and the DL. In addition, as the a part of the base-band processing is executed in the CO, processing and transport delays need to be computed and compensated to ensure synchronization of transmission and reception on the air interface. JT requires the coordination of advances at each AP needed for coherent transmission.

Communication protocol We have defined how the proposed dynamic split can be implemented in the context of LTE, by making use of the signaling possibilities specified for the UL and the DL. Downlink Control Indication (DCI) messages destined to the UEs can contain also the information that the RUs need for UL and DL communications. The CO sends these messages for each subframe which results in a non negligible control overhead on the MAC-PHY interface. On the UL, the DCI configuration message describes for example scheduling, power control, modulation order and transmission scheme. In our dynamic split design, it makes use of a so-called *cooperation flag* to notify to the RU if a PRB needs to be forwarded at the interface 4' (see Figure 2.13). The DL configuration message includes in addition the precoding matrix indicator (PMI) computed in the MAC layer.

Both data and control messages need to be mapped in asynchronous Ethernet packets in order to be transported over the fronthaul network. Data-plane fronthaul packets containing uncoded bits or modulated symbols - for selected PRBs on the uplink - do not require a large header, no further control overhead is introduced at that level. It is common to represent of I/Q samples over 32 bits, but other sample width is also possible, even though reducing it can decrease precision.

Implementation The proposed dynamic split is compatible with both dedicated hardware-based and software-based implementation on the RU side, but for latency-sensitive applications and to satisfy the HARQ timing constraint optimization of the processing time needs to be considered in the choice of the platform used. Implementation in Field-Programmable Gate Array (FPGA) allows to make use of hardware accelerators for high complexity operations. A dynamically adaptable implementation is also possible by using FPGAs that can be partially reconfigured to accommodate computational load variation. Although, on the CO side a software based and virtualized implementation is used to benefit from C-RAN advantages.

2.3.3.2 Performance evaluation

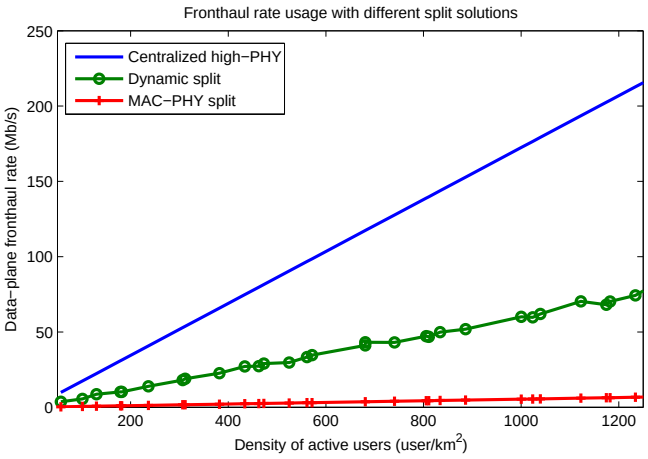
To quantify the benefit and the cost of the proposed split, we have evaluated the fronthaul rate with different split options for data-plane traffic in a 5G scenario in the configuration described in Table 2.1. Figure 2.14a illustrates the load over the fronthaul link associated to one antenna of an AP with a split following interfaces 4 (Centralized user-PHY) and 2 (MAC-PHY split) compared to the solution that we propose.

One can observe that even with high user density, fronthaul usage remains limited and affordable. In fact, only a small part of users requires UL JR ac-

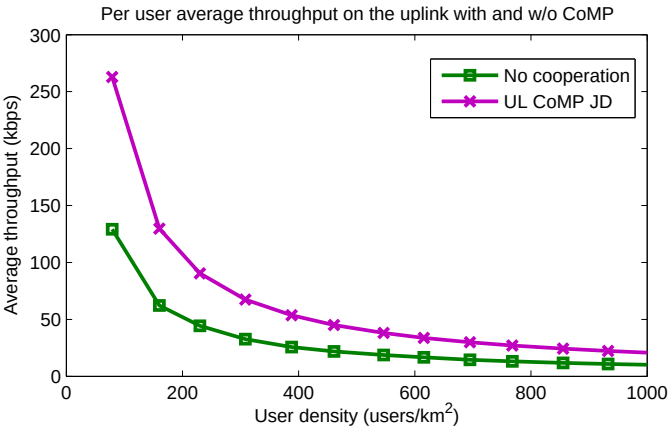
Parameter name	Value
Bandwidth	100 MHz
Number of PRBs	500
Number of subcarriers	6666
FFT size	8172
Distribution of UEs	uniform
Sample width	16 bits
Radius of cells	350 m
Distance between APs	500 m
Number of neighbor cells	8
Number of antennas per cell	1
Modulation scheme	16-QAM
Code rate	1/3
Scheduling strategy	Round Robin
Type of antennas	omnidirectional
Traffic model	Full buffer

Table 2.1 – Simulation parameters

cording to their SINR, hence split at interface 4 would introduce fronthaul load without any benefit for the major part of the users. Dynamic split with PRB selection reduces fronthaul load with respect to interface 4 split, while keeping its advantage to enable JR. Even if the increase of the data-rate compared to the MAC-PHY split is not negligible, it is a necessary tradeoff to a achieve higher spectral efficiency using physical layer joint reception. We can observe on Figure 2.14b the improvement enabled by UL JR, in fact average user throughput is 2 times higher for any cell load, since cell-edge QoS is better thanks to cooperation.



(a) Fronthaul rate reduction using dynamic split



(b) Average throughput improvement with uplink JR

	Split 5	Split 4	Split 4' JR UEs	Split 2 non-JR UEs	Split 2 all UEs
Packet frequency	$66,6\mu s$	$66,6\mu s$	1 ms	1 ms	1 ms
Data size	261,5 kbits	96 kbits	28 kbits	5.2 kbits	5.6 kbits
Rate	3922 Mbps	1440 Mbps	28 Mbps	5.2 Mbps	5.6 Mbps

Table 2.2 – Fronthaul traffic characteristics for various split options

We have also evaluated the behavior of uplink fronthaul traffic for the different split solutions in order to compare to the proposed architecture. Table 2.2 characterizes the Ethernet packets sent for a user density with 1000 users per km^2. By comparing fronthaul rates, we can see that the dynamic split which we get by combining splits at interface 4' for JR users and interface 2 for regular users results in a rate with the same order of magnitude as split 2. Note that the increase of rate w.r.t. split 2 is needed in order to guarantee the realization of multi-cell JR processing that could be executed otherwise with splits 5 and 4 operating at significantly higher fronthaul rates. Moreover, by applying the same split at interfaces from 4' to 2 data is released after performing a given PHY function for the whole subframe. With dynamic split, a part of the data is sent after demapping, another part after decoding, so the overall frequency of packets is less than 1 ms and the traffic is more regular.

Part 3

Software-defined networking for multi-cell cooperation

After having described the context of future mobile networks and the C-RAN architecture with a particular configuration that we have proposed to accommodate multi-user processing, in the current chapter we focus on the evolution of network operation regarding implementation. In the first part, we explain emerging concepts in mobile network implementation and introduce how they can be applied for multi-user processing. In the second part, we provide details about the proof-of-concept platform that we have realized to demonstrate multi-cell techniques in C-RAN via a basic end-to-end deployment.

3.1 Context: RAN "softwarization"

With the evolution of computing technologies, generic platforms are becoming convenient for operations performed earlier by dedicated hardware. Latency and reliability constraints of mobile networks can now be satisfied by simple GPPs. For 5G systems, this opens huge opportunities regarding flexibility and scalability, but also releases new challenges to be solved. Hardware-independent implementation of processing in mobile networks allows to adopt approaches previously used in the Information Technology (IT) industry and adapt them to the constraints of telecommunication services. A design enabling high performance under any condition and for all the supported services needs to be defined. In the 5G context, softwarization of the processing needs to be accompanied by the separation of control-plane and data-plane enabled by SDN. Furthermore, the control plane has to be reorganized with multiple hierarchy layers to satisfy mobile networks requirements following the novel architecture where several network slices dedicated to a given service are operating simultaneously.

3.1.1 Virtualization of base-band processing

3.1.1.1 The clouds

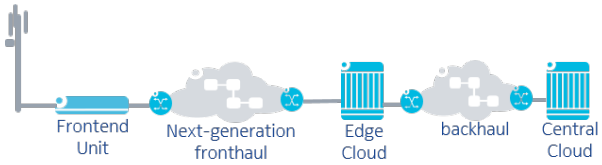


Figure 3.1 – Cloudified 5G RAN architecture

The structure of server units realizing in RAN processing is evolving from the traditional all-in-one BSs towards architecture where BB processing is distributed between several entities. The elements realizing computations in future mobile networks are depicted on Figure 3.1.

The Front-End Unit (FEU) or RU is the distributed light-weight processing unit located close to the cell-sites and realizing signal processing for them. Often, it is associated to a radio AP and it is installed in its close proximity. In some cases, it can belong to several APs, located at a relatively short distance (less then 1 km). A limited set of operations is performed in this entity, mainly the physical layer processing (see Section 2.3), thus it is composed of hardware that can efficiently execute PHY functions, such as FPGAs. Some flexibility is also needed at this level of the network to accommodate traffic variations and low-latency services, thus FPGAs offering reprogrammability are more suitable then dedicated circuits.

Connected to the network front-end through wired or wireless fronthaul, the edge-cloud is the fundamental unit of C-RAN, a major part of the BB processing of several cells is aggregated there. As it has to be able to support PHY functions with a low latency, to limit the delay due to fronthaul transport, it is usually deployed within 20 km from the FEUs. It realizes the major part of the processing using GPPs, but also includes accelerators such as Graphical Processing Units (GPUs) and FPGAs executing high complexity functions. Similarly to existing mobile network deployments, a centralized server, i.e. the central cloud realizes the core network functions. Between the edge-cloud and the central cloud, converged Ethernet network ensures the backhaul connectivity.

Cloudification of the mobile networks does not consist only in gathering processes in a single server, but also in executing them independently of the hardware constituting this server, i.e. virtualizing them. Thanks to these two aspects we can increase TCO savings enabled by C-RAN. Today's high performance computers enable real-time BB processing even in a virtualized environment [60], despite the statement considering that a process is slower running in

a virtualized environment than running on hardware platform optimized for the given function remains true, thus a new software architecture with optimized processing latency is required for virtualized deployments.

3.1.1.2 From monolithic software to microservices

As we have mentioned, in next generation mobile networks, more efficient and flexible processing will be realized on GPP by virtualized processes that come replace hardware-optimized operations running on dedicated units. In today's network deployments, engineers design complete and unchanging processing stacks for a given hardware, modifying a given function or add a new type of processing, the whole unit had to be replaced or upgraded. Virtualized processing, besides enabling more efficient operation, also allows to split the processing chain in small elements, microservices, with well-defined input and output interfaces realizing a single function. The microservice-based architecture is already efficiently used in IT services and can improve processing in mobile networks if adequately designed. Shifting to this type of architecture can greatly improve deployment practices by segmenting validation and deployment, thus accelerating the evolution of network features. It will allow service providers to have agile networks in constant improvement using *DevOps* that accelerate the delivery of new features in microservice-enabled mobile networks.

Base-band processing functions are tightly interconnected and require to transfer data between them, thus to optimally choose the limits of the microservices and avoid that communication overhead reduce their effectiveness, the amount of traffic at the interface needs to be taken into account. Besides dimensioning, it is also important to carefully choose the virtualization technique. Using Virtual Machines (VMs) to execute multiple isolated services in a single server reduces the amount of operations as well as the memory needed [61]. The VMs instantiated in a host computer execute their own Operating System (OS) and a hypervisor is in charge of managing the VM instances that have a full abstraction of the hardware platform through the kernel of the OS of the host. Recently, container-based virtualization offered a lighter alternative to deployments using VMs. Software containers allow to separate services in their own environment without requiring a hypervisor process, thus the computational overhead created by this type of virtualization is lower. A comparison of VM- and container-based virtualization in mobile networks is provided in [62]. We can deploy several containers in a host and each of them is directly connected to the kernel of the host OS. Though a separate OS can be executed inside a container, it is more suitable to the microservice architecture to dedicate each container to one application and build a service processing chain by interconnecting the

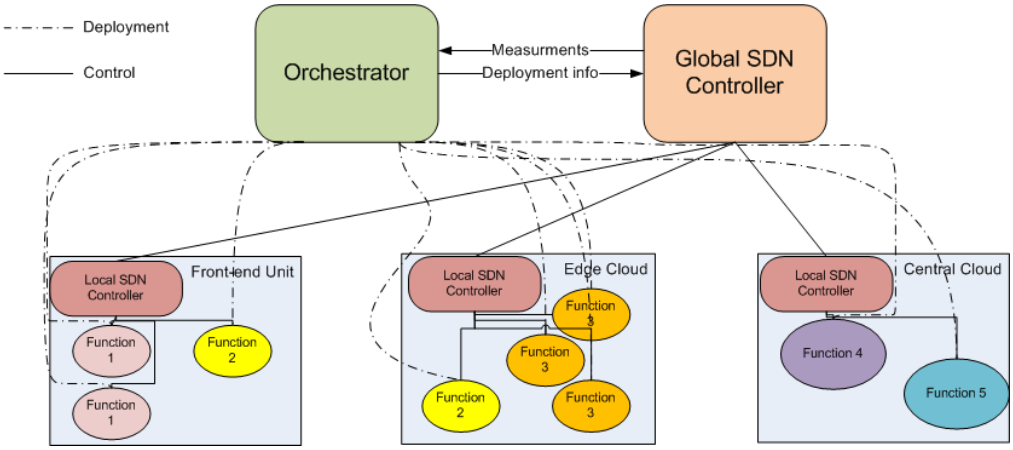


Figure 3.2 – Orchestration and control in microservice network architecture

containers. Currently, the most commonly used container-based virtualization engines are Docker [63] and Linux Containers [64].

3.1.1.3 NFV, network slicing and orchestration

To facilitate the adoption of microservice architecture in telecommunication applications, a specification group was created at European Telecommunication Standards Institute (ETSI) in charge of defining NFV principles and the Virtual Network Functions (VNFs). Though it does not focus exclusively on mobile networks, RAN and mobile core network virtualization are among the target use cases [65]. As it is easier compared to RAN to implement and manage VNF in the core network, advanced microservice-based solutions has been proposed for this segment. However, efficient and reliable RAN VNFs are not yet ready to be deployed in FEUs.

As 5G mobile networks will provide various types of service (i.e., eMBB, Massive Machine Type Communications (MMTC) and Ultra-Reliable and Low-Latency Communications (URLLC)), and within each service type, different configurations can be available following context-dependent requirements and constraints. Also, the network functions forming a processing chain have to be configured coherently. To ensure versatility, in 5G, it is planned to attribute the execution of a service instance within a given context to a so-called network slice which realizes the end-to-end processing for that service. A network slice includes the needed microservices, i.e., the VNFs, chained adequately and deployed over the available hardware platforms in a way that satisfies requirements of the given service. Controller instances are also present in each slice to manage control-plane operations. The vision of composing network slices from a 5G VNF repository presented in [66] would be a commode way to efficiently realize the slicing concept.

The deployment of the microservices is performed by the orchestrator that is aware about the available computing resources and network interconnections as well as about the requirements of the service realized by the network slice to be deployed. As it can be seen in Figure 3.2, it operates in collaboration with the SDN controller in charge of network configuration (for more details see Subsection 3.1.2). Note that here we consider only controllers dedicated to RAN, others ones are used for the transport network (e.g., fronthaul and backhaul) between the various processing units. Orchestrators that are able to instantiate interconnected containers realizing a specific function while conforming to a set of requirements are already available (see e.g., [67, 68]). Per-slice orchestration is the generalization of the concept defined as RAN-as-a-Service [69, 70] to 5G architecture. Setting up and end-to-end RAN processing chain on demand and operate it independently of the underlying hardware increases significantly the scalability, flexibility and efficiency of the network. Currently available orchestrator tools operate dynamically and adapt the deployment to the variations of the computing and network environment. More powerful orchestration algorithms that potentially improve the performance of mobile networks are currently studied.

3.1.2 Software-Defined RAN: related work

The separation between control plane and data plane in networks ensures high flexibility and programmability so that operators can have a complete control over the network from a centralized point [71]. This would facilitate to network operators deploying new applications and services and adjusting network policies [72]. Load balancing and scheduling of computational resources among the available platforms can also be easily realized. Software-defined networking has been used in wide variety of network environments such as enterprise networks, data centers and infrastructure-based wireless access networks.

OpenFlow has been the first protocol used for SDN. It has two key components: the controller having a complete control over all the switches in the network, i.e., it controls all network functions, and the switches which forward the first packet of each new flow to the controller for routing decision. The controller can update the forwarding table depending on some pre-defined rules or new policies. Though OpenFlow has programmability and flexibility in managing and controlling various network elements in SDN, there are some concerns regarding its efficiency with heavy traffic load. DevoFlow [73] is a modification of OpenFlow protocol which maintains flow visibility and reduces load and overhead in the controller. The number of switch flow table entries and control messages decreases by more than ten times using DevoFlow [73]. The efficient

switching and routing solution that we have defined for high performance operation in C-RAN is inspired by this example.

The importance of SDN is also to bring dynamic programmability into the control plane which will be indispensable for deploying adaptive mobile networks. OpenRoads [74] and OpenRadio [75] are examples of using SDN for RAN control and enabling flexible control especially at PHY and MAC layers. In [76], generalized SDN framework for RAN is demonstrated based on OpenDayLight controller for possible load balancing, interference management and dynamic radio resource sharing. Based on these existing works, we could define how to benefit from the flexibility, programmability and agility of SDN framework in C-RAN particularly for multi-cell processing, where coordination between BBUs is essential.

3.2 Prototype of multi-cell JR in C-RAN

The aim of this section is to describe the elements and the operation of the proof-of-concept platform realized during the thesis. It demonstrates the feasibility of multi-cell processing in C-RAN and validates the proposed architecture involving software BBUs and SDN.

3.2.1 OpenAirInterface software BBU

After having described in the previous subsection the new principles and concepts involved in the design of future mobile networks, we give in the current subsection an overview of the OpenAirInterface (OAI) [77, 78] open-source software implementing the complete LTE stack. We have used this platform to demonstrate multi-user processing in C-RAN, since its flexibility makes easy to include new, 5G oriented functions. As we have already mentioned, even though, our implementation is based on the standardized LTE stack, but aims to show in practice solutions for 5G. Though OAI [78] has a UE, a BBU (eNodeB (eNB)) and a core network (Evolved Packet Core (EPC)) component, in this thesis focusing on PHY processing and control we have used only the BBU part that we will briefly describe in the following.

3.2.1.1 Overall architecture

The implementation of the OAI software follows entirely the 3GPP specifications of LTE, and it can operate with standard compliant EPCs or a UEs of any origin. The various elements composing the OAI eNB are depicted in Figure 3.3. As we see, only a part of the components is dedicated to BB processing, the remaining part consists in functions providing connectivity to the host OS,

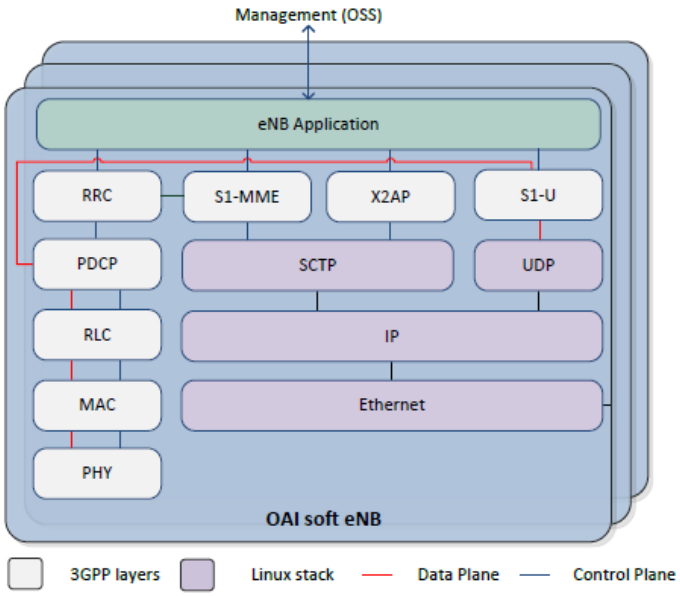


Figure 3.3 – OpenAirInterface eNB structure

to the core network and to the radio front-end. The main eNB application manages the modules in charge of the different levels of the LTE stack and ensures control-plane communication within the process. Note that even though the current version of the eNB application that we have used is a monolithic software, its modular structure facilitates its transformation to a microservice-based deployment. For a prototype platform like ours, we do not need the same flexibility as in a real network deployment, thus the lack of segmentation to microservices has not had any practical effect. Ongoing research works aim to transform it following the microservice architecture by requiring the splitting of the eNB to its components and designing a complete separated control-plane, the multi-cell processing would be then easy to add to that virtualized platform.

3.2.1.2 Simulation environments

A major advantage of the OAI software is to be able to produce several types of executable based on the same implementation. For simulation purposes, we can build the eNB process together with the UE in a simulated framework called *oaisim*, where the wireless channel is modeled. In this simulation the whole LTE stack is executed, only the radio front-end is bypassed, thus it is very convenient to test added features in various configurations. Many commonly used channel models are available and the network topology along with the transmission parameters can be configured. Though simulations are not executed in real-time due to the high processing load created by launching simultaneously the channel generator, the UE and the eNB, however they are fast enough to

observe results continuously. These full-stack simulations allowed us to test multi-cell coordination.

It is also possible to simulate only a selected PHY channel, in this case both network- and user-side PHY functions are processed, but only for the given channel. To develop and validate multi-cell joint reception, executing the physical layer processing chain only for the uplink data channel was very useful. These simulation environments executing exactly the same functions as the real-time eNB made easier the implementation of joint processing in our C-RAN prototype.

3.2.1.3 Deployment of OAI eNBs with RRHs

To establish connection between a UE and the OpenAirInterface software eNB, we need to connect it to a software-defined radio equipment in charge of the RF. We have used Universal Software Radio Peripheral (USRP) boards acting as RRHs and connected them through two different fronthaul infrastructures to the server where two BBUs were simultaneously launched and they were connected to an EPC. One was directly connected to the BBU-pool server and the other through an Ethernet connection to a distant RRH. The antenna in- and outputs of the USRP boards were connected to the UEs through wires able to transfer radio signal in order to limit the impact of interference coming from other equipments. By mixing the uplink signals of the two UEs connected, we have created a cell-edge scenario where we could control the pathloss by varying the attenuation of the radio signals. The setup of C-RAN with OAI BBUs is depicted in Figure 3.4.

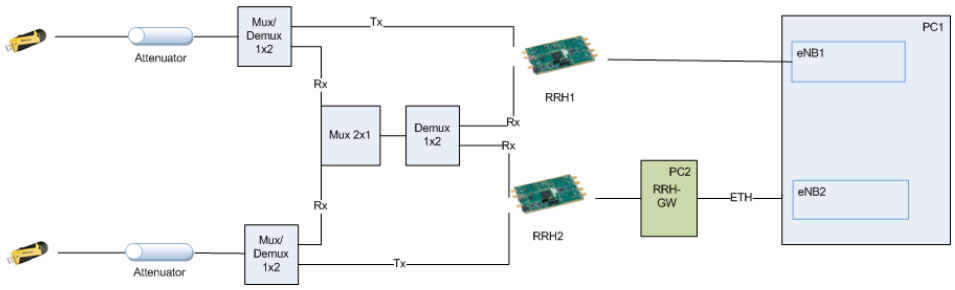


Figure 3.4 – Elements of the C-RAN prototype implementation: the USRP boards are used as RRHs, one is connected through a dedicated gateway, another directly to the BBU-pool. Commercial LTE dongles are the UEs connected.

3.2.2 SDN-based control of the RAN

3.2.2.1 SDN controller for C-RAN

As we have several interconnected and jointly operating elements in a C-RAN deployment, using a unique controller managing them can improve significantly the overall performance. However, the architecture of the controller has to enable real-time operation in the RAN. The SDN paradigm can satisfy this constraint by:

- providing an intelligent abstraction layer between the controlled elements and the controlling process,
- leaving in the distributed elements the control aspects that do not require joint processing,
- performing control by lightweight processes that are in charge of the control of one aspect only.

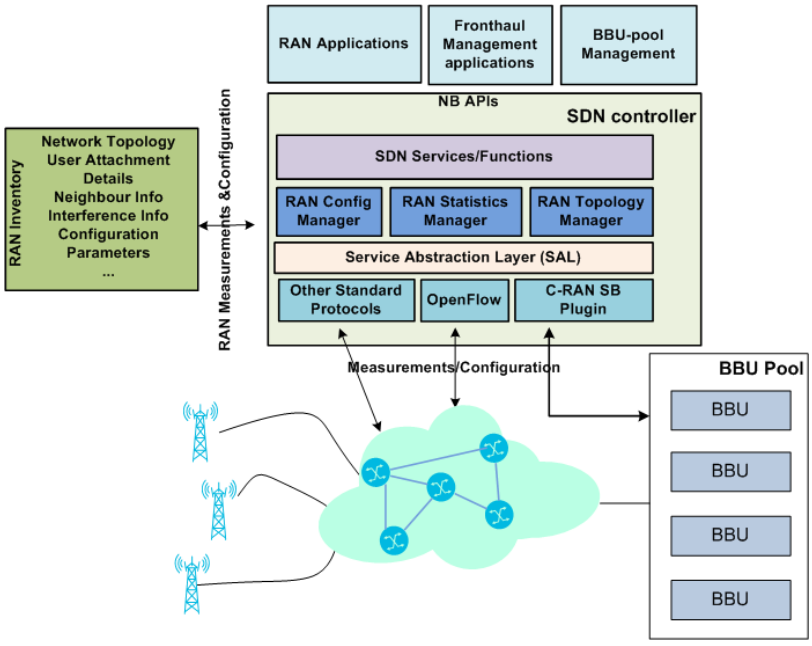


Figure 3.5 – SDN-enabled C-RAN deployment

The SDN controller for a C-RAN deployment is depicted in Figure 3.5. Signals are processed partly in the RRHs and partly in the BBUs dedicated to each cell, but able to communicate between one another inside the BBU-pool. The controller configures via control plane connections the data flows from RRHs towards the BBU-pool through the fronthaul as well as inside the BBU-pool. NorthBound (NB) applications are in charge of various types of network control involving several BBUs. Through the NB interface, they get measures and

parameters from the controller and also send commands to the network and BBU-pool. Other applications are in charge of providing QoS-based dynamic switching throughout the fronthaul infrastructure that is possibly shared among several operators and load balancing of virtualized processes. In summary, measurement and control data is centralized in the RAN inventory and made available for modules inside the controller and on the NB interface which realize end-to-end network control and optimization.

3.2.2.2 RAN optimization

In a distributed RAN architecture, cooperation between cells require to exchange a high amount of data and adds latency to the processing time. C-RAN makes possible centralized control, but implementing global multi-cell control algorithms would result in high complexity and long computation time. Thus centralizing only the part of RAN control needed for multi-cell optimization enables to use low complexity algorithms. The control- and data-plane separation offered by SDN perfectly accommodates RAN control when BBUs are centralized, since only a single communication interface is needed on SouthBound (SB), and the abstraction layer offered by the controller establishes a many-to-many communication between the control functions and the BBUs. It enables to adapt various network dynamics in their global context. In our design, the SDN controller has access to all system measurements that are stored in the associated database (see Fig. 3.5). This allows network applications connected through the NB interface to ask for the measurements needed to run the optimization algorithms. When the optimization is done, the controller sends instructions back to the network elements in charge of executing them.

3.2.3 Multi-cell coordination using SDN

After having introduced the application of SDN for mobile networks in Subsection 3.1.2, we now describe the details of the architecture defined for multi-cell coordination in C-RAN. We also provide some implementation details in order to illustrate how our design satisfies constraints of the RAN. An overview of the system with several BBUs managed using an SDN controller is depicted in Figure 3.6. The OAI eNBs are the BBUs that include the whole PHY processing, since in this prototype using lower bandwidth limited fronthaul rate was not a problem, thus in practice it was not necessary to use the user-selective fronthaul split.

The controller can be deployed in the BBU-pool server or in a remote location. Each eNB running in the same BBU-pool communicates individually with the SDN controller, sends and receives data related to its assigned cell. This

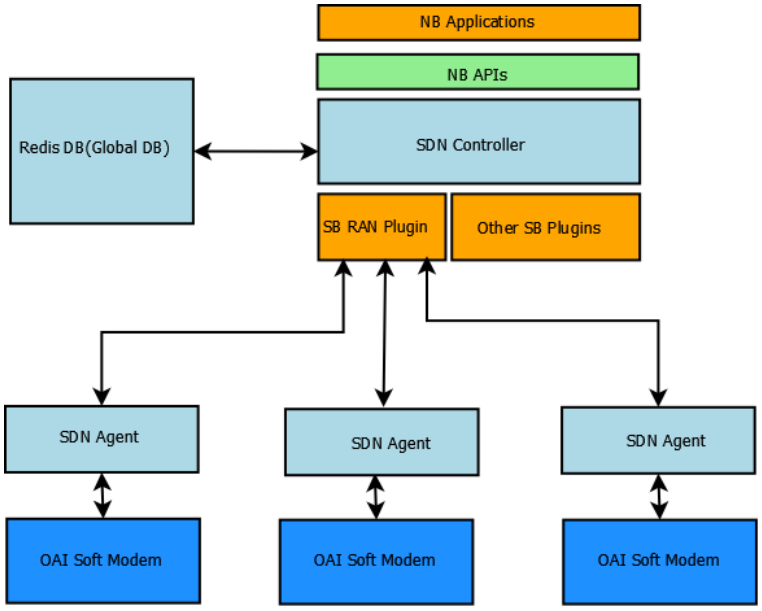


Figure 3.6 – The architecture of the C-RAN prototype including several BBUs controlled via SDN. The BBUs are OAI eNBs (a.k.a. OAI SoftModem) to which SDN SB agents have been added. The controller uses a Redis no-SQL database for storing measurement and control data.

data allows higher latency and generates less traffic than the user-plane data that needs to be shared for PHY multi-cell processing. Inside the BBU-pool, eNBs can exchange data using internal communication mechanisms without coordination required from the controller. Selected RAN state measurements from PHY and MAC layer are extracted from a local database by a dedicated SDN Agent and forwarded to the SB interface of the controller.

3.2.3.1 Southbound connection

We have implemented the SB protocol in the SDN controller for the abstraction of eNBs using Stream Control Transmission Protocol (SCTP) protocol due to its numerous advantages. It ensures message orientation and reliable message delivery by providing stronger (32 bits) end-to-end checksum compared to TCP and UDP. Also, it is the legacy transport protocol used in X2 interface defined for message transfer between distributed BSs in mobile networks [48, 79].

The SB protocol in the controller is a key element in bringing programmability to the RAN by

- collecting necessary measurements, configuration, and neighbor information of each eNB, and
- modifying/re-programming the functionality of each eNB by sending back new configuration parameters obtained from NB applications.

In order to support the deployment of user coordination algorithm, the SDN controller has been extended both in Service Abstraction Layer and also in NB API (see Figure 3.5) since the controller supports only fixed network control by default. Since the YANG model based XML data storage [80] in the controller is not persistent to maintain long term network history, we introduced No-SQL database (Mongo DB, Redis, etc) connectivity to the SDN controller to bring persistency.

Each eNodeB is equipped with an SDN Agent to establish communication with the centralized controller. The SDN Agent module is built using the SCTP client libraries similar to the SB protocol that is built using SCTP server libraries. Since SCTP supports point to multi-point communication, any number of eNBs can be served by single SDN controller in parallel. Moreover, the SDN Agent collects real-time measurements to be forwarded to the controller and also applies for example scheduling parameters received through the SB interface.

3.2.3.2 Controller adaptation and Northbound interface

The main motivation behind the application of SDN approach to RAN is to abstract the entire multi-cell network intelligence to the BBUs that can use coordination algorithms as NB applications to benefit from joint processing possibilities offered by C-RAN in order to enhance user experience. In our architecture, we have extended the NB Application Programming Interfaces (APIs) of the SDN controller to have the interfaces necessary for NB applications (coordination algorithms) operation, i.e., to provide measurements as well as to re-configure the RAN functions using updated parameters. In our architecture, NB applications interact with the external DB unit of the controller where set of tables have been created by the controller (e.g., measurements per user, scheduling information). During the operation, RAN service modules in the controller using the SB plug-in updates the RAN inventory on a per-subframe basis. NB APIs are built according to the requirements of the applications to accelerate their operation. The implementation of the SDN controller was based on the open-source Open Network Operating System (ONOS) distribution, however, it has been modified to ensure fast operation. The modules related to the management of fixed networks have not been included in the deployment. In a more advanced prototype where processing is split in microservices, a part of their control should be realized by additional modules dedicated to RAN management.

3.2.3.3 Northbound application

Since the processing dedicated to each cell is realized in a separate process, multi-cell coordination needs to be performed by a function external to the BBUs. In

the case of uplink JR, we need to synchronize the scheduling of cell-edge users in order to allocate the same PRB to the ones that have to be jointly detected. To summarize the operation of the application where the multi-cell coordination algorithm is implemented, we provide a description of the processing steps to be realized. The number of each step on Figure 3.7 represents the main functional block related to it.

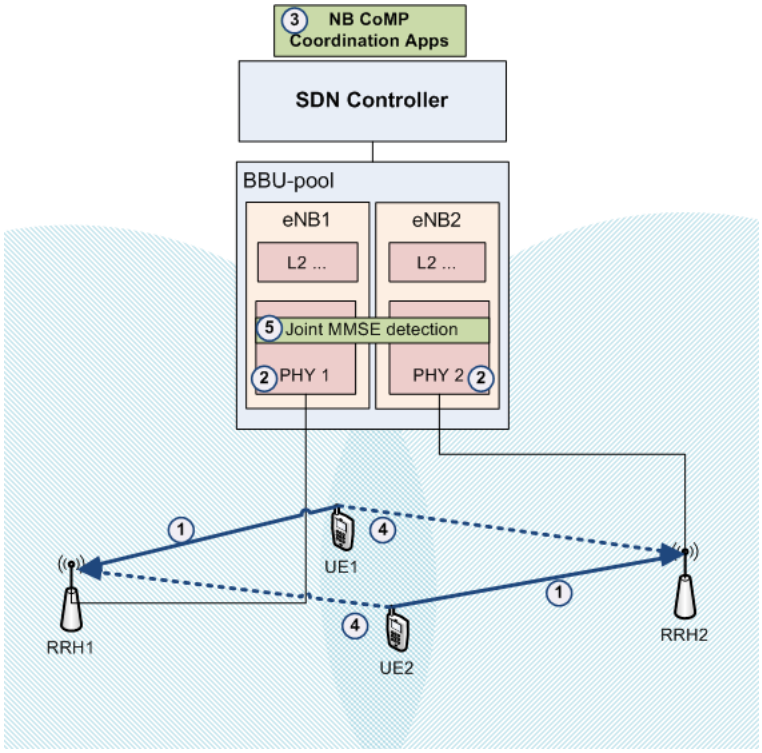


Figure 3.7 – Coordination process for multi-cell JR realized by the NB application

① Each UE transmits towards the RRH of the cell to which it is associated and single-user detection is performed in the BBU.

② After PHY processing, error rates and received signal characteristics (modulation order, received power) are sent to the coordinator.

③ Coordination algorithm detects high ICI and enables multi-user detection. It sends to the eNBs scheduling constraints for users involved in JD and instructions to activate joint detection functions (i.e., multi-user channel estimation, MMSE matrix computation and MMSE equalization, see details in Subsection 3.2.4).

④ UEs transmit according to new parameters. (They are intentionally scheduled on the same wireless resource, thus create inter-cell interference.)

⑤ Multi-cell joint MMSE detection is realized. Error rate is decreased w.r.t. previous transmission with single-user detection and effective throughput is improved.

We assume in this process that one user in each cell is involved in the joint reception for a given PRB, as intra-cell interference is avoided by the cell scheduler, this assumption does not constrain the operation of multi-cell coordination. If several groups of users transmit on different PRBs, scheduling decisions are sent by the NB application accordingly.

3.2.4 Multi-cell PHY processing

We provide in this subsection some insights about the implementation of multi-user multi-cell reception in C-RAN. We have added multi-user MMSE receiver function in the OAI BBUs and executed it in a multi-cell environment with the BBUs deployed in a C-RAN configuration along with the coordination application installed at the NB of the SDN controller. The main challenges of this implementation are described in the following.

3.2.4.1 Data sharing

To make available the received data needed for UL JR to several BBUs in the same CO an efficient data sharing solution is required. The fastest way of communicating between physical layer signal processing functions executed simultaneously is to store the data that need to be used by several streams (i.e., threads or processes) in a memory segment that can be accessed by all of them. This strategy allows to limit communication latency to the synchronization delay needed when a process has to wait before reading a data that the other processes finish to write it. One of the main advantages of C-RAN is to enable this very low latency communication between BBUs in order to efficiently execute multi-cell joint processing.

3.2.4.2 Channel estimation

Multi-user channel estimation is enabled in LTE by adapting the reference signal sent by the different users. For UL data transmissions, DMRS are used for channel estimation, the reference signals sent by the users are defined by shifting a base sequence with a cyclic shift angle α_i different for each user [31]. The eNB attributes the cyclic shift values to the UEs, in C-RAN its selection needs to be coordinated between cell performing joint reception. Applying different phase shift to the reference signal allows to minimize the interference between the DMRS at the receiver and estimate the channel accurately. The n^{th} reference

signal sequence of the UE i is given by

$$s_{\mathrm{UE}_i}(n) = e^{j\alpha_i n}\overline{s}(n), \tag{3.1}$$

where $\overline{s}(n)$ is the base sequence defined depending on the size of the reference signal, i.e., the number of PRBs allocated. For smaller sequence lengths $\overline{s}(n) = e^{j\phi(n)\pi/4}$ where the values of $\phi(n)$ allows to get a *constant amplitude zero autocorrelation* sequence (see Tables 5.5.1.2-1 and 5.5.1.2-2 in [31]). When at least 3 PRBs are allocated to the same user, a base sequence without autocorrelation is determined by the roots of the Zadoff-Chu sequence [81] which has for length the largest prime number lower than the size of the reference signal.

The value of cyclic shift phase $\alpha_i = 2\pi n_{cs,i}/12$ includes two deterministic quantities ($n^{(1)}_{\mathrm{DMRS}}$, $n^{(2)}_{\mathrm{DMRS}}$) and a pseudo-random sequence ($n_{\mathrm{PN}}(n)$) as follows

$$n_{cs,i} = (n^{(1)}_{\mathrm{DMRS}} + n^{(2)}_{\mathrm{DMRS}} + n_{\mathrm{PN}}(n)) \mod 12 \tag{3.2}$$

The index of the cyclic shift represented over 3 bits is sent by the BBUs to the UEs which then select the corresponding values of $n^{(1)}_{\mathrm{DMRS}}$ and $n^{(2)}_{\mathrm{DMRS}}$ in pre-defined tables (see Tables 5.5.2.1.1-1 and 5.5.2.1.1-2 in [31]). The successive index values ensure a maximal shift between the reference signals, so the index values have to be allocated in order. Note that to enable to allocate 12 different pilot signals, cyclic shift is combined with Orthogonal Cover Code (OCC) that adds a mask making successive reference signal sequences orthogonal to each other.

As each BBU has allocated the same cyclic shift values, they are known at the receiver in charge of performing channel estimation. For multi-cell reception, the channel estimation step is similar to the one needed in a single-cell multi-user MIMO configuration, since in each cell, pilot symbols from every user are received. Generally, ZF or MMSE channel estimation is used. For multi-cell joint reception, channel estimates have to be shared, thus the channel estimator function writes its output in a shared memory segment.

3.2.4.3 Multi-cell receiver

The physical layer processing of single-user MIMO receivers has usually a parallel structure which bears separate streams for the processing of the data received by each antenna, and combines them before the decoding. As we have discussed in Section 2.2.1 , MMSE receiver allows good performance with reasonable computational complexity and can be used in a MU-MIMO case for exploiting interference. We show in Figure 3.8 the architecture of multi-user MMSE receiver in C-RAN.

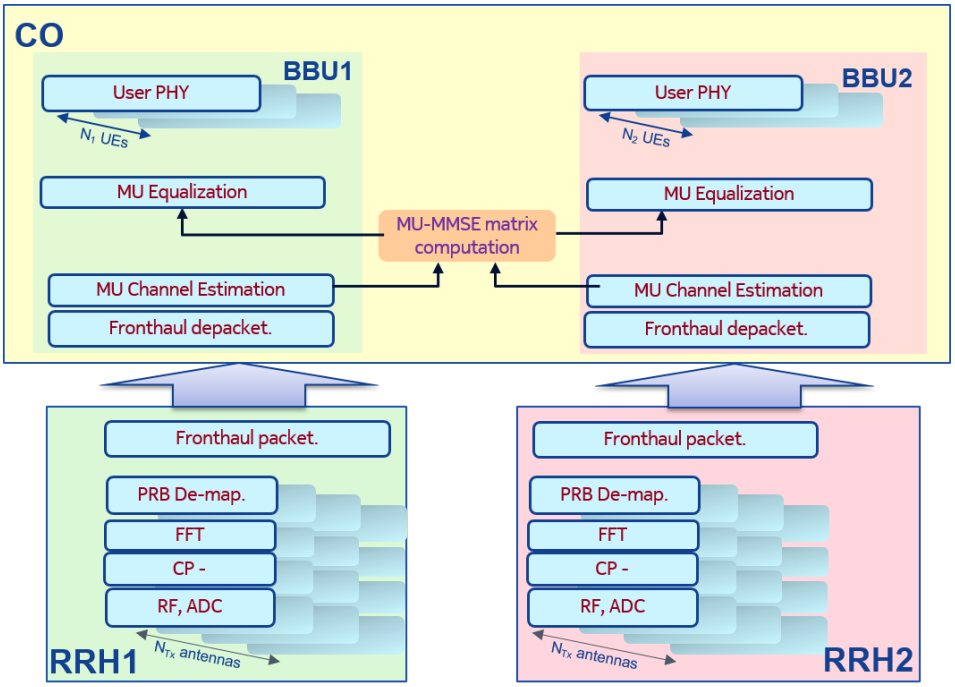


Figure 3.8 – Structure of PHY multi-user MMSE receiver in C-RAN. I/Q symbols of users selected for JR are forwarded through the fronthaul to the CO where multi-user (MU) channel estimation takes place, the MU-MMSE matrix is computed than distributed MU channel equalization in the BBUs isolates user signals.

In the case of a multi-user detection with as many users as antennas, for each PRB an MMSE matrix is computed based on the channel estimates of every user at every antenna. Multi-user MMSE detection allows to isolate the received signal of each user from the data streams coming from the whole set of receive antennas. After having performed this multi-user detection (i.e. channel equalization), each BBU can follow up with the user-PHY signal processing functions for the UEs located in its cell.

3.2.4.4 Number of jointly processed users

Despite the fact that linear MMSE receiver has significantly lower complexity compared to maximum likelihood detection, it still requires a matrix inversion that has a complexity of $O(N^3)$ with N co-channel users. Several approximations of the MMSE detector have been proposed (e.g., [82, 83]). However, their complexity is still around $O(N^2)$. Furthermore, since the design of the low-complexity algorithms exploits the properties of tall MIMO channel matrices, with many users, the performance of the receiver decreases and would result in much higher block error rates. To ensure high performance and still enable interference exploitation, in practical C-RAN deployments we can allocate to the

same PRB only a few users compared to the total number of available antennas located at the RRHs of the cell implied in multi-cell processing. Not only processing time, but synchronization delay is also reduced if multi-user reception is realized for several user groups of lower size, since less processing chains need to wait for each other. For further explanations see Section 4.2.

Part 4

Fronthaul allocation optimization for UL joint reception

In this chapter we aim to improve signal processing realized in future C-RAN where the previously described architecture accommodates multi-cell multi-user processing. User-selective fronthaul split and centralized SDN-control, detailed in the previous chapters, are the technological enablers for improving QoS of cell-edge users by JR in the BBU-pool. To further enhance performance and efficiency of UL transmissions, we investigate how to associate users involved in NOMA and how to quantize received signals when the cost of fronthaul usage is considered. We evaluate fronthaul allocation optimization first in the ideal case of perfect channel estimation, then, in a case when only channel covariances are known and based on them we can approximate the average throughput.

4.1 System model for multi-cell NOMA in C-RAN

Aiming to achieve higher spectral efficiency and satisfy 5G requirements, NOMA techniques are considered both for the downlink [84] and the uplink [85, 86]. NOMA is also helpful in a multi-cell environment to manage inter-cell interference [87]. It is similar to the uplink joint reception for cell-edge users that we have described previously in this thesis, its implementation in C-RAN requires to implement the same architectural and control solutions as the ones detailed in Sections 2.3.3 and 3.1.2. Indeed, exploiting interference that affects UL users can significantly improve their QoS and spectral efficiency. In C-RAN, multi-cell NOMA allows such interference exploitation. In the following analysis and numerical evaluations, we focus on the configuration where UEs are in the region covered by several RRHs, since we expect more gains by using multi-cell NOMA

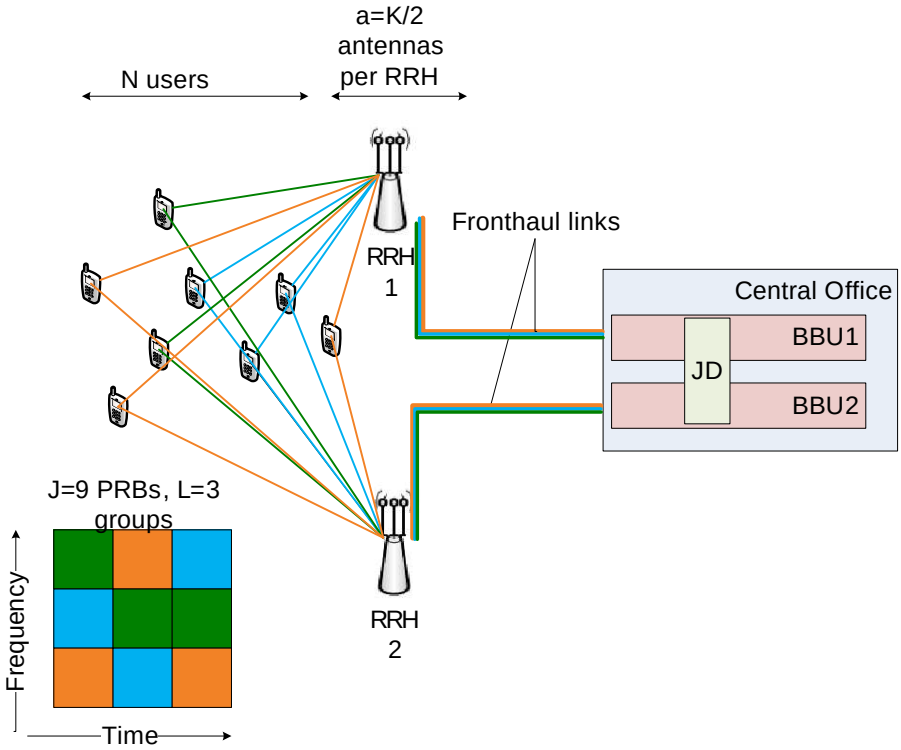


Figure 4.1 – System model of uplink NOMA in C-RAN

in this particular case. As cell-center UEs transmit usually with less power, we can assume that the inter-cell interference generated by them can be processed at the receiver together with the channel noise.

We define here the system model of UL NOMA transmissions in C-RAN that we will use throughout this chapter. This system model is depicted in Figure 4.1. Since we are focusing on UL NOMA for cell-edge users, we show in this figure the particular case of 2 cells in a C-RAN deployment, with 1 RRH in each cell.

To provide analysis for a general case, we consider M RRHs located at the cell sites with a antennas each, thus the total number of antennas is $K = a \cdot M$. All of the RRHs are connected via digital communication links to the same CO which can perform signal processing. We have N single-antenna users and consider that each user is allowed to communicate with all the M RRHs. The number of users is constant in our analysis. We assume that the system architecture allows for all the N users to centralize the user-specific physical layer functions, i.e., the ones after demapping user signals from the attributed PRBs. This enables to perform multi-cell multi-user JR on the uplink (for details c.f. Section 2.3).

The RRHs, after receiving the uplink signals sent by the users, quantize them and forward through the fronthaul links to the CO. To send the signal received

by each antenna via the fronthaul link, a given quantization rate is used for the compression of radio signals to digital base-band symbols. The transmission rate over the fronthaul link depends on the quantization rate that will be considered in our optimization problem in Sections 4.5 and 4.6.

In the system, we have in total J orthogonal PRBs available commonly for all the RRHs at each channel use. Note that if the same data is sent by a user on several PRBs, the signal received by a given antenna will be combined in the RRH after PRB demapping and then forwarded to the CO.

Notations We denote random variables with uppercase letters with non-italic fonts, e.g., S, for scalars, bold and non-italic fonts, e.g., $\mathbf{V}$, for vectors, and bold and sans serif fonts, e.g., M, for matrices. Deterministic variables are denoted with italic letters, e.g., a scalar x or N (for quantities) , lowercase bold for a vector $\boldsymbol{v}$, and uppercase bold letters for a matrix $\boldsymbol{M}$. Logarithms denoted by log(.) are in base 2, the ones denoted by ln(.) mean the natural logarithm and superscript $(.)^H$ denotes the conjugate transpose of a vector or a matrix. $\mathbb{E}_{\mathrm{X}}[.]$ means the expectation w.r.t. the random variable X, $h(.)$ denotes the entropy of a random variable and $I(\mathrm{X},\mathrm{Y})$ the mutual information between X and Y.

The N users are uniformly distributed in the region covered by every RRH. The channel of each user n in the set $\mathcal{S}_N = \{1, ..., N\}$ towards all the antennas among the RRHs is denoted by $\boldsymbol{h}_n$, which is a K dimensional vector following the Gaussian distribution $\mathcal{N}(0, \boldsymbol{R}_n)$ with $\boldsymbol{R}_n \in \mathbb{C}^{K\times K}$. In the numerical evaluations, the channel covariances are computed following the one-ring scatterer model [88], thus they reflect the position of each User Equipment (UE) with respect to the RRHs. Note that one can consider other spatial distribution of the N users, it possibly has an impact on network performance, but the optimization methods considered in this thesis still remain feasible. The positions of 40 users in the cell-edge area are depicted in Figure 4.2, they are randomly selected from a uniform distribution. We can remark that some of them are almost co-located, while others can be isolated.

The UEs transmitting on the same PRB are said to form a group that has channel matrix given by the juxtaposition of the users' channel vectors. The number of user groups is denoted by L (with $L \leq J$). There are s_l users in the group denoted by Π_l, with $l \in \{1, ..., L\}$. The multi-user channel of the group Π_l, comprising the users with indexes $\pi_i^l \in \mathcal{S}_N$ with $i = 1, ..., s_l$, towards the K antennas is the $K \times s_l$ dimensional matrix $\boldsymbol{H}_l = \left[\boldsymbol{h}_{\pi_1^l}, ..., \boldsymbol{h}_{\pi_{s_l}^l}\right]$. The complete set of the groups $\{\Pi_1, ..., \Pi_L\}$ is a partition of the set of the users $\mathcal{S}_N$.

The Gaussian channel noise vector is denoted by $\boldsymbol{n}_l \sim \mathcal{N}(0, \sigma_z^2 \boldsymbol{I}_K)$. Power of the input signal is normalized, so that noise covariance is $\sigma_z^2 = \frac{1}{\mathsf{SNR}}$. The signal

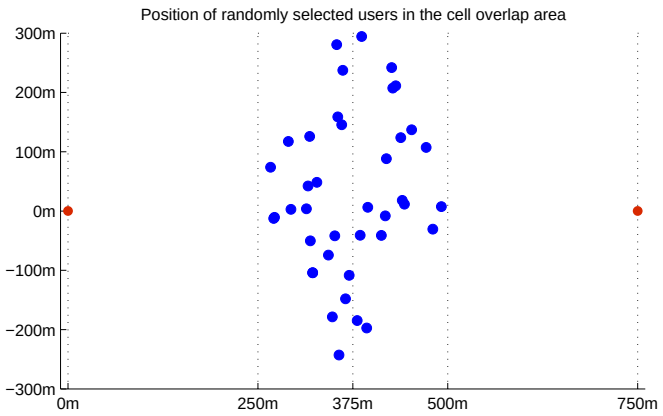


Figure 4.2 – Placement of randomly selected users in the cell-edge region. Red markers show the placement of the two RRHs and blue ones are the positions of the UEs.

received by the whole set of antennas for group Π_l is the K-dimensional vector $\boldsymbol{y}_l$ and it is given by the superposition of the signals sent by all of the s_l users in the group such that

$$\boldsymbol{y}_l = \sum_{i=1}^{s_l} \boldsymbol{h}_{\pi_i^l} x_{\pi_i^l} + \boldsymbol{n}_l. \tag{4.1}$$

After receiving $\boldsymbol{y}_l$ for each group $l \in \{1, ..., L\}$, the RRHs quantize this received signal and forward it to the CO. The fronthaul rate used for transmitting the quantized form of $\boldsymbol{y}_l$ is $\boldsymbol{c}^{(l)} = \{c_{l1}, ...c_{lK}\}$, where c_{lk} with $k \in \{1, ..., K\}$ denotes the fronthaul rate attributed to the forwarding of the component of $\boldsymbol{y}_l$ received by the antenna k, i.e., y_{lk}. The total fronthaul rate of group Π_l over the link between the RRH m and the CO is denoted by $c_m^{(l)}$ with $l \in \{1, ..., L\}$ and $m \in \{1, ..., M\}$. We can write $c_m^{(l)}$ using fronthaul rates $\{c_{lk}\}$ as follows:

$$c_m^{(l)} = \sum_{k=(m-1)\cdot a+1}^{m\cdot a} c_{lk}. \tag{4.2}$$

If the capacity of the fronthaul links is limited, we denote the maximum capacity available for the whole set of user groups on the link m by $\bar{c}_m$.

The signal received by the CO for group l is denoted by $\hat{\boldsymbol{y}}_l$. Following the analysis detailed in Section 2.2.2.2, it can be written as

$$\hat{\boldsymbol{y}}_l = \boldsymbol{\alpha}_l \left(\sum_{i=1}^{s_l} \boldsymbol{h}_{\pi_i^l} x_{\pi_i^l} + \boldsymbol{n}_l + \boldsymbol{d}_l \right), \tag{4.3}$$

where $\boldsymbol{\alpha}_l$ is the vector composed by the scaling factors α_{lk} for antennas $k \in \{1, ..., K\}$ and $\boldsymbol{d}_l$ is the K-dimensional vector of the Gaussian noise representing

the distortion created by the fronthaul quantization at the rates given by $\boldsymbol{c}_l$. The variance $\sigma^2_{d_{lk}}$ of the components of $\boldsymbol{d}_l$ is upper bounded as follows

$$\frac{\sigma^2_{d_{lk}}}{\alpha_{lk}} \leq \frac{\sigma^2_{\mathrm{Y}_{lk}|\mathbf{H}} 2^{-c_{lk}}}{1 - 2^{-c_{lk}}}. \tag{4.4}$$

In the current chapter we will discuss for the system defined above the following:

- the partial NOMA scheme where the overall set of UEs is split to several groups and each group transmits on its own PRB (Subsection 4.2),
- the possibility to associate users in groups in a deterministic manner (Section 4.3),
- the optimization of fronthaul rate allocation among the user groups when channel gains are known (Section 4.5),
- and the optimal fronthaul rate allocation based on channel statistics available in real C-RAN deployments (Section 4.6).

4.2 Partial NOMA

In this section, we describe the multi-cell NOMA scheme that is adapted to satisfy implementation constraints. We adopt this transmission strategy in C-RAN for the study of optimal fronthaul rate allocation maximizing the gain that operators get from transmissions in a practical deployment scenario.

4.2.1 Uplink NOMA

While with orthogonal multiple-access techniques we attribute to each UE a different time-frequency resource, in NOMA, users are intentionally scheduled in a way that they use the same PRBs. NOMA allows to increase the overall spectral efficiency if multi-user reception is applied for uplink, but requires this additional signal processing on the receiver side. Fortunately, in C-RAN, with multiple antennas, we can apply MIMO techniques, such as linear MMSE, for multi-user detection.

To increase the overall throughput while serving the whole set of users in order to ensure – some – fairness, the best scheduling strategy is to allow to as many users as possible to transmit on all of the available PRBs. We will call this strategy full NOMA in the coming discussions. Indeed, the sum rate of s_l users transmitting together over J PRBs is higher than the sum of their rates while each of them uses alone J/s_l PRBs. Furthermore, frequency diversity is improved by scheduling all of the UEs on all of the subcarriers. However,

regarding the implementation and the execution of receiver processing, it is not the optimal choice to apply full NOMA for many users.

4.2.2 Practical limitations at the receiver

As we have mentioned in Subsection 3.2.4, implementation of a multi-user receiver in C-RAN rises several challenges. To allow low-latency transmissions, processing time has to be short, thus applying low-complexity reception techniques is crucial. Often, it is faster to process less users in one function such as the computation of MMSE equalizer, and execute several instances of the function simultaneously to process the whole set of users. Furthermore, if less users are involved in multi-user reception, synchronization overhead is lower, since to start user-PHY processing, it is required to complete multi-user reception only for a smaller user group, not the whole set of users.

Evaluations in [82, 83] also point out the interest of limiting the number of jointly received UEs, they show that for MMSE receiver it is possible to implement low-complexity methods, however these ones result in a good performance where the number of users is significantly less compared to the number of receive antennas.

4.2.3 Partial NOMA groups

To deal with practical constraints such as receiver complexity, it has been proposed in [89] to perform the uplink multi-user detection needed in NOMA over several smaller subsets of users, while each subset transmits on a different channel. An ideal tradeoff between the throughput and the complexity of receive processing is to schedule users by groups and attribute different PRBs to each group in order to ensure orthogonality between them. We will call this strategy partial NOMA. It aims to achieve significantly higher overall throughput with respect to single user transmission. Its advantage compared to full NOMA is to require multi-user receive processing over a smaller number of users for each group, thus, receiver complexity remains low and the signals of different groups can be processed simultaneously. Figure 4.3 illustrates the scheduling in orthogonal, full and partial NOMA transmissions.

We define as a user group the UEs that are granted to transmit over the same set of PRBs. As we would like to illustrate different aspects of optimization with partial NOMA, we adopt a simple model regarding the structure of user groups. Therefore, we assume that each of the N users in our system is included in one group exactly (see conditions (4.5b)-(4.5c) below) and for the sake of fairness, each group contains the same number of users when possible or at most one user more than another group (see condition (4.5d)). Also, we attribute the same

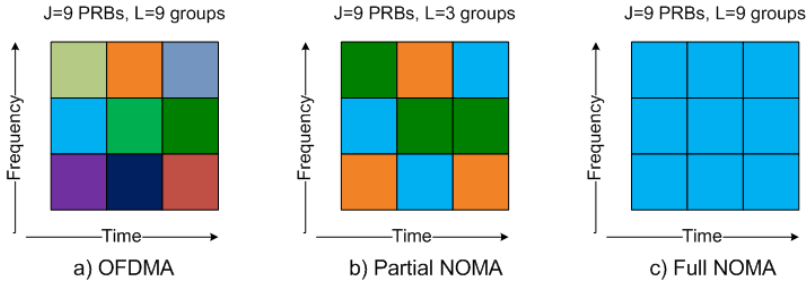


Figure 4.3 – Comparison of user scheduling strategies

number of PRBs to each group.

The following function $g(.)$ is used to denote the partitioning of the set of users $\mathcal{S}_N$ into L groups under the conditions (4.5b)-(4.5d):

$$g : \mathcal{S}_N \mapsto \{\Pi_1, ..., \Pi_L\}, \tag{4.5a}$$

$$\bigcup_l \Pi_l = \mathcal{S}_N, \tag{4.5b}$$

$$\Pi_i \cap \Pi_j = \emptyset \ \forall i, j \mid i \neq j, \tag{4.5c}$$

$$\mid s_i - s_j \mid \leq 1 \ \forall i, j \mid i \neq j. \tag{4.5d}$$

The proposed partial NOMA scheme allows to maintain the advantages of full NOMA, i.e., frequency and antenna diversity, at a reduced computational cost on the receiver side. In principle, the throughput of partial NOMA increases linearly with the number of users in each group, while the complexity of the MMSE receiver used for the detection in each group will increase proportionally to the square of the number of users. This would make MMSE receive processing inefficient and add significant computational delay when handling many users. As in 5G systems, low transmission round trip time is targeted [90] and retransmission protocols (such as Hybrid Automatic Repeat reQuest) require low receive processing latency, full NOMA strategy is not practical for 5G, for example to support enhanced Mobile BroadBand (eMBB) service. With partial NOMA, we can achieve high throughput compared to single-user scheduling and keep computational complexity and delay low.

4.3 Association of cell-edge users in groups

For cells with the same CO, we apply the partial NOMA scheme defined in Subsection 4.2.3 to multiple cells and turn inter-cell interference to useful signal by exchanging received data and CSI inside the CO. As a result, we schedule on the same PRB users on edge of neighbor cells to enable the reception of a

user's signal by several RRHs located in adjacent cell sites. Low-latency data exchange enabled by C-RAN allows to apply multi-user reception techniques and support it by scheduling coordination so that we assign the same PRB to the users who can cooperate more efficiently, thus improve the throughput and spectral efficiency with respect to random selection of cooperating users.

Several papers focus on opportunistic user association for JR, it was also studied for LTE in [91], where methods choosing user pairs giving maximal gain on the direct links of the resulting MIMO channel and minimal gain on the interference links. To improve fairness compared to these methods and realized computation without the knowledge of theCSI, double proportional fair algorithm was proposed in [92]. User pairing performed jointly with scheduling is studied in [93], where the proposed algorithm achieves significant improvement, but has high complexity. In [94] authors consider user pairing between neighboring cells for the whole coverage area, they compare various methods and evaluate their performance following several metrics. Sum-rate maximization-based pairing using bipartite graphs is described in [95]. This method ensures low complexity, but similarly to most of related works, it allows to process only two users per group.

The C-RAN configuration is also more favorable for NOMA reception than a single AP with K antennas [49], thanks to better channel diversity created by the placement of the receive antennas at several spatially distributed RRHs. Users can have channel gain mainly contained in different matrix subspaces, i.e., the correlation between their channels can be very low in some cases. Scheduling such users in the same group results in higher spectral efficiency, since the interference between them is then low too.

Although, in a practical system, it would introduce very high signaling overhead to realize channel estimation for all of the users at every PRBs. Finding by exhaustive search the partitioning which gives the maximal throughput has also very high complexity, and it requires the knowledge of CSI between each user and each antenna, on all of the frequency resources. Furthermore, the problem of joint user grouping and scheduling optimization of N users into L groups scheduled on J PRBs would have an extremely high complexity, especially for a system with many users. To avoid the overhead introduced by the pilot sequences used for channel estimation we can use the second order channel statistics for finding the right users to associate, since they do not change with the carrier frequency. Assuming that some CSI is available for each user, the covariance can be easily computed, however it characterizes less precisely the channel compared to real-time CSI. Based on the statistics, we can approximate the group sum rates and determine the user groups that are likely to achieve higher throughput, even without the knowledge of concrete channel realizations. This method

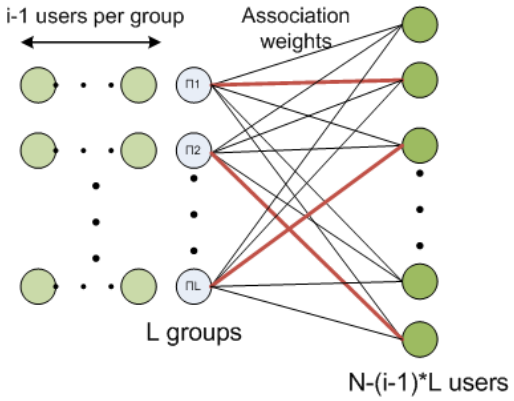


Figure 4.4 – Association between groups and users using bipartite graph (i^{th} round). The weight associated to each edge of the graph is the value of the function $\hat{f}$ for the group and the user that are the nodes linked by the given edge. We select exactly one edge for each user group in a way that the sum of the weights is maximal.

also reduces the complexity, since the grouping can be realized separately from the scheduling. The achievable sum rate of the user group Π_l is given by

$$\mathbb{E}[\log \det(\boldsymbol{I}_K + \mathsf{SNR} \cdot \mathbf{H}_l \mathbf{H}_l^H)]. \tag{4.6}$$

Using Jensen's inequality, we get the following function that is an upper bound of the sum rate depending only on the channel covariance $\boldsymbol{R}_l = \mathbb{E}[\mathbf{H}_l \mathbf{H}_l^H]$:

$$\hat{f}(\Pi_l) = \log \det(\boldsymbol{I}_K + \mathsf{SNR} \cdot \boldsymbol{R}_l). \tag{4.7}$$

We can compute the value $\hat{f}(.)$ based on the CSI available in a practical system, though, it only approximates the sum rate. Optimizing $\hat{f}(.)$ cannot guarantee that the actual sum rate is maximal, but it is likely to increase it.

A user pairing method maximizing the total sum rate was proposed in [95], it optimizes the user association by solving the assignment problem for a bipartite graph where users are denoted as the nodes and the value associated to the edge between two nodes is the sum rate that they can achieve together. Since the problem of finding the best user grouping with more than two users in each group is very complex to evaluate by exhaustive search, we can use an iterative algorithm built on the maximum sum assignment in bipartite graphs [96]. One round of the user association using a bipartite graph is represented in Figure 4.4. In our proposed scheme, we will solve the assignment problem N/L times to find a user grouping scheme which is expected to achieve high sum rate. The details of the proposed user grouping method are described in Algorithm 1. Obviously, the proposed user grouping method is not optimal, since in addition

Input: Sets of channel vectors $\{\boldsymbol{h}_1, ... \boldsymbol{h}_N\}$ enabling to compute $\boldsymbol{R}_l$
Result: User grouping solution $g(\mathcal{S}_N)$
Initialize $\Pi_l \leftarrow \emptyset \ \forall l \in \{1, ..., L\}$;
for $i = 1 : \lceil \frac{N}{L} \rceil$ **do**
  Set $W = \{w_{k,m}\}_{\substack{1 \leq k \leq L \\ 1 \leq m \leq N-(i-1)L}} = \{0\}_{L \times (N-(i-1)L)}$;
  for $l = 1 : L$ **do**
    for each user n not included in any group **do**
      $\mathbf{H}_l^{(n)} \leftarrow [\mathbf{H}_l, \boldsymbol{h}_n]$;
      $\boldsymbol{R}_l^{(n)} \leftarrow \mathbb{E}[\mathbf{H}_l^{(n)} {\mathbf{H}_l^{(n)}}^H]$;
      $w_{ln} \leftarrow \hat{f}(\Pi_l \cup \{n\})$;
    end
  end
  Find $\{n_1^*, ..., n_L^*\} \leftarrow \underset{1 \leq n_l \leq s_l}{\operatorname{argmax}} \sum_{l=1}^{L} w_{ln_l}$ using the Hungarian algorithm [96];
  for $l = 1 : L$ **do**
    $\Pi_l \leftarrow \Pi_l \cup \{n_l^*\}$;
    $\mathbf{H}_l \leftarrow [\mathbf{H}_l, \boldsymbol{h}_{n_l^*}]$;
  end
end

Algorithm 1: User grouping algorithm for sum rate improvement

to the approximation of the sum rate, we assign the users to the groups in several iterations and keep the decisions of the previous iterations unchanged.

We have evaluated the sum rate of the grouping scheme obtained using Algorithm 1 over a set of $N = 12$ users forming $L = 4$ groups and transmitting towards $M = 2$ RRHs with $a = 4$ antennas at each of them. We have compared the performance of the solution that we propose to exhaustive search, as well as to the average sum rate and the worst case grouping that we get by creating groups randomly. Table 4.1 shows the comparison of sum rates and also the complexity of each method.

We can see from Table 4.1 that the sum rate of the grouping scheme obtained by our proposed iterative algorithm that has the advantage of exploiting input data available in real radio access network deployments, is able to ensure a throughput only 1.2% lower than the globally best solution. Although the throughput achieved in average with random grouping is only slightly lower than the one with our solution, by the proposed deterministic grouping we can avoid a drop of the throughput by 5% in some cases. We can see that our algorithm, having the advantage of exploiting input data that is available without large signaling overhead and in reasonable computational time, is able to ensure higher throughput than random grouping. In fact, the gap between the worst grouping scheme that we can obtain randomly and the best one found by exhaustive search is 4 times larger than the gap between our solution and the best one. In use cases where throughput requirements are tight, it can be interesting

Grouping strategy	Single-user	Random grouping (worst case)	Random grouping (average)	Iterative algorithm (proposed)	Exhaustive search
Average throughput (bits per channel use)	157.27	441.38	455.66	458.88	464.45
Processing complexity	1	1	1	$O(\frac{3N^4}{4L}) \sim 4000$	$O(\frac{L^N}{L!}) \sim 700000$

Table 4.1 – Comparison of different user grouping methods in terms of throughput and computational complexity. Evaluation is done for $N = 12$ users with random location and $J = 12$ PRBs available. The partial NOMA scheme is realized with $L = 4$ groups of 3 users, each of them transmitting over 3 PRBs towards $K = 8$ receive antennas equally distributed between the $M = 2$ RRHs.

to use the proposed algorithm despite it requires more computation than random grouping. We can remark that the result of any practically implementable method exploiting realistic input data similarly to the one described above can only serve to fill the gap between random grouping and the optimal scheme.

4.4 Cost of fronthaul usage

We have selected the partial NOMA strategy for the study of fronthaul data transfer, since it accommodates practical limitations in C-RAN while improving the throughput compared to single-user transmission assisted by multi-cell scheduling. In Part 2, we have considered the most significant constraint in C-RAN deployments, i.e. the limited fronthaul from the point of view of the architecture. Though we can significantly decrease fronthaul usage by forwarding I/Q symbols only for the users implied in the JR, even for these ones there is a possibility to allocate the available fronthaul in an optimal way.

The net benefit of UL transmissions in C-RAN can be characterized by the actual gain related to the throughput minus the cost of using a fronthaul infrastructure at a given rate. In addition, as we have already mentioned in Section 2.2.2.2 the end-to-end performance depends on the fronthaul quantization rate, that results in a tradeoff between transmission benefit and fronthaul cost. In the current section we provide models of fronthaul usage cost in various scenarios that we will use in Sections 4.5 and 4.6 where metrics for characterizing throughput with partial NOMA are introduced.

To enable data processing in the CO, it has to be connected to the RRHs through high capacity and reliable communication links. The most commonly

used technology for fronthaul links is based on various deployments of optical fibers. Network operators have several options to connect their cell-sites to the CO [97], we describe in the following the modeling that we propose for the different fronthaul deployments.

Scenario 1: Fronthaul leasing A given capacity of fiber Ethernet can be leased from its owner who provides it either through a point-to-point link in some cases or through a switched network infrastructure. This scenario is modeled by limited fronthaul link capacity and a constant cost-per-bit $\lambda_k^{(1)}$. Assuming that the leased network capacity is accurately dimensioned, the cost-per-bit of the fronthaul transmissions in this model reflects the part of the total leasing cost for a unit rate. This linear model allows to dispatch the overall investment between all the transmissions over the leased link proportionally to the fronthaul rate that they use. As in our system model the antennas are located at several RRHs, due to differences between the fronthaul connections at each RRH, the cost can vary in function of the location of the antennas (we attribute a cost to an antenna k instead of an RRH in order to keep the framework independent of the number of antennas).Note that it is also possible to attribute different cost coefficients $\lambda_k^{(1)}$ to each group, for example if there is a priority ordering between the user groups. For simplicity, we do not consider this case, but our method and results remain valid as long as all the cost coefficients are positive. Then the total cost of the fronthaul transmission between the K antennas and the CO can be written as

$$q_1(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} \lambda_k^{(1)} c_{lk}. \tag{4.8}$$

Scenario 2: Owned point-to-point links The network operator can install its own point-to-point fiber link fully dedicated to the communication between a given RRH and the CO, thus the transmission cost is the consequence of the investment realized for the deployment and the operational costs such as energy consumption. In this case, the fronthaul capacity is limited and the cost-per-bit can be modeled as in (4.9). In this formulation, the first term decreases with the rate used, its role is to represent a portion of deployment costs. The more fronthaul is used, the lower is the cost-per-bit, since the constant deployment cost is distributed over a higher total rate. The second term $\lambda_k^{(2)}$ accounts for constant operational costs such as energy consumption.

$$q_2(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} \left(\frac{\mu_k^{(2)}}{c_{lk}} + \lambda_k^{(2)} c_{lk} \right) \tag{4.9}$$

with $\mu_k^{(2)}$ a real-valued constant. We use $\frac{\mu_k^{(2)}}{c_{lk}}$ as a decreasing function in our model, since the more straightforward is to distribute investment costs over the occurring transmissions proportionally to the rate used. However, note that any other positive decreasing function can be applied to represent a different way of distributing link deployment costs.

Scenario 3: Owned shared infrastructure Some operators own large fiber network infrastructures that are shared by various services and multiple sites. We expect that the cost model suitable to this scenario is when per-link fronthaul capacity is unlimited and the cost-per-bit of the fronthaul usage includes a penalty for preventing other services to use the given amount of rate. The first term of the cost then increases with the allocated rate and the second one represent the constant price. In this model, the more fronthaul rate is used, the higher is the price factor, since the allocated rate becomes unavailable for other services that may generate additional revenue. Here, we model the penalty pricing linearly with respect to the fronthaul rate (resulting in a quadratic term in the total cost), but other positive increasing functions can be also suitable. Note that the second term captures operational costs in the constant $\lambda_k^{(3)}$. The following equation describes this model where $\mu_k^{(3)}$ is a real-valued constant:

$$q_3(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} (\mu_k^{(3)} c_{lk} + \lambda_k^{(3)}) \cdot c_{lk}. \tag{4.10}$$

Generic cost function The cost functions (4.8)-(4.10) have various considerations of the fronthaul capacity, pricing, investment cost, and sharing costs. They can be generalized as follows:

$$q(\boldsymbol{c}^{(l)}) = \sum_{k=1}^{K} (\mu_k c_{lk}^n + \lambda_k c_{lk}) \tag{4.11}$$

where n is a real-valued exponent, μ_k and λ_k are non-negative coefficients. Note that although (4.11) is a general polynomial function, the optimization problem considering the net benefit can be solved as usual for any non-negative convex cost function.

When $n = 1$, $q(\boldsymbol{c}^{(l)}) = q_1(\boldsymbol{c}^{(l)})$ with $\lambda_k^{(1)} = \lambda_k + \mu_k$; $n = -1$ gives $q(\boldsymbol{c}^{(l)}) = q_2(\boldsymbol{c}^{(l)})$ with $\mu_k = \mu_k^{(2)}$ and $\lambda_k = \lambda_k^{(2)}$; while $n = 2$ covers Scenario 3 with $\mu_k = \mu_k^{(3)}$ and $\lambda_k = \lambda_k^{(3)}$.

4.5 Fronthaul allocation optimization with perfect CSI

We describe in this section the fronthaul allocation scheme allowing to maximize the net benefit of uplink partial NOMA transmission in C-RAN. We consider here that the channel is perfectly known at the receiver. Though, channel estimators that are generally used result in an estimation error, its impact on our fronthaul allocation framework is negligible.

As we adapt the quantization rate of the received signal after a given scheduling decision, we can use estimated channel gains within a channel coherence period. Regarding the delay of acquiring channel estimates, we assume that in the block fading channel model depicted in Figure 4.5, the coherence time T is long enough to consider adapting the fronthaul rate following uplink channel realizations. In terms of user mobility, e.g., for pedestrian users moving at 5 km/h, the channel changes approximately every 720 ms, in this time window we can update network parameters such as scheduling or quantization rate. Optimization is performed in the CO and results are fed back to the RRHs that can apply them on the mobile radio interface and the fronthaul interface respectively.

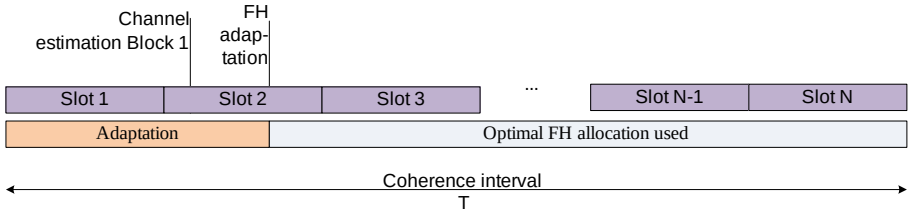


Figure 4.5 – Block fading model with coherence interval of length T

4.5.1 Net benefit of uplink NOMA transmissions

The gain of uplink transmissions increases with the rate, since high throughput allows mobile users to realize more data traffic that generates (direct or indirect) incomes for the mobile network operator. At the same time, the transmission generates operational costs among which we consider here the one related to fronthaul usage. To provide a practical evaluation of the net gain that we can get from a transmission we should realize a multi-objective optimization [98]. A way that has been identified to formulate such a problem is to convert it to a single-objective problem by including in the same goal function several aspects between which we need to find a tradeoff. It is also similar to the method used in Lagrangian optimization. For the case of uplink partial NOMA transmissions, we will use the objective function characterizing the net benefit of a transmission that is defined as the instantaneous sum rate minus the fronthaul cost.

Proposition 1. *The partial NOMA scheme in C-RAN architecture with fronthaul quantization the achievable sum rate of a group Π_l with a known channel realization $\boldsymbol{H}_l$ is given by*

$$\sum_{i=1}^{s_l} r_i \geq log\ det\left(I_{s_l} + \boldsymbol{H}_l^H \boldsymbol{V}_{s_l}^{-1} \boldsymbol{H}_l\right) \tag{4.12}$$

with $\boldsymbol{V}_{s_l}$ the equivalent noise covariance:

$$\boldsymbol{V}_{s_l} = \sigma_z^2 \boldsymbol{I}_K + \underset{k=\{1,..,K\}}{diag} \left(\frac{\sigma^2_{y_{lk}|\boldsymbol{H}_l} 2^{-c_{lk}}}{1 - 2^{-c_{lk}}} \right) \tag{4.13}$$

where $\sigma^2_{y_{lk}|\boldsymbol{H}_l}$ is the covariance of the signal received by the RRH at antenna k and c_{lk} is the number of bits that we use over the fronthaul link to forward this signal, i.e., the number of quantization bits.

We obtain the expression of the equivalent noise variance including fronthaul quantization rate from Equations (4.3) and (4.4) and inject it in the generic expression of achievable sum rate $\log \det(I_N + \mathsf{SNR} \cdot \boldsymbol{H}\boldsymbol{H}^H)$, that is in fact the minimum achievable sum rate since the channel and quantization noises are assumed to be Gaussian [99].

We use this formulation (4.12) of uplink transmissions in multi-user C-RAN in the objective function allowing to maximize the end-to-end benefit of the user group Π_l with s_l users towards the M RRHs with a antennas each. The parameters of this function are the following:

- The variance σ_z^2 of the Gaussian channel noise.
- The average received signal power from group l at antenna k given the channel estimate of the group: $\sigma^2_{y_{lk}|\mathbf{H}_l}$.
- The fronthaul rate c_{lk} with $k \in \{1, ..., K\}$ used to forward to the CO digital base-band I/Q symbols of group l received by antenna k. We denote by $\boldsymbol{c}^{(l)} = (c_{l1}, ..., c_{lK})^T$ the vector of rate values for each antenna and for group l.

The following function characterizes the net benefit of the transmission of group Π_l towards the whole set of receive antennas when the fronthaul rate allocated to the group is $\boldsymbol{c}^{(l)}$:

Given the parameters $\sigma_z^2, \sigma^2_{y_{lk}|\mathbf{H}_l}, \forall k \in \{1, ..., K\}$,

$$b_1(\boldsymbol{H}_l, \boldsymbol{c}^{(l)}) = \log \det \left(I_{s_l} + \boldsymbol{H}_l^H \boldsymbol{V}_{s_l}^{-1} \boldsymbol{H}_l\right) - q(\boldsymbol{c}^{(l)}) \tag{4.14}$$

where $\boldsymbol{V}_{s_l}$ is defined in (4.13). The first term of the function $b_1(.)$ in (4.14) gives the instantaneous sum rate during the coherence time block where the

channel matrix $\boldsymbol{H}_l$ holds and the second term is the total cost of the fronthaul transmission for group Π_l over the whole set of fronthaul links connecting the RRHs to the CO, denoted by the cost function $q(.)$ defined in equation (4.11).

4.5.2 Fronthaul optimization with limited rate available

The objective function defined in (4.14), characterizing the net benefit of the uplink transmission of a single user group. We use it to write the optimization problem allowing to allocate the available fronthaul capacity between the groups transmitting simultaneously on different PRBs. The fronthaul allocation scheme that maximizes the net benefit for the whole set of L users groups is the one that gives the highest sum of the metric (4.14) over all groups. In Scenario 1 and Scenario 2, where the available fronthaul rate on each fronthaul link is limited, we have a non-linear constraint giving the maximal fronthaul rate that can be used between one RRH and the CO to forward the received signals from every user group at every antenna belonging to this particular RRH. Then, to find the global fronthaul rate allocation resulting in the highest net benefit, we need to solve the following problem:

$$\begin{aligned} &\text{Find } \{\boldsymbol{c}^{(1)*}, ..., \boldsymbol{c}^{(L)*}\} = \underset{\{\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(L)}\}}{\operatorname{argmax}} \sum_{l=1}^{L} b_1(\boldsymbol{H}_l, \boldsymbol{c}^{(l)}) \\ &\text{subject to } \sum_{l=1}^{L} c_m^{(l)} \leq \bar{c}_m, \ \forall m \in \{1, ..., M\}. \end{aligned} \tag{4.15}$$

Let us recall that $c_m^{(l)} = \sum_{k=(m-1)\cdot a+1}^{m\cdot a} c_{lk}$, so the above constraint can also be written as:

$$\sum_{l=1}^{L} \sum_{k=(m-1)\cdot a+1}^{m\cdot a} c_{lk} \leq \bar{c}_m, \ \forall m \in \{1, ..., M\}. \tag{4.16}$$

Proposition 2. *The problem (4.15) is concave, thus admits a unique solution that gives the optimal capacity allocation scheme.*

Proof. The constraint in (4.15) is a linear function for each m, thus can be considered as concave. The subtracted cost function $q(.)$ is assumed to be convex, thus the concavity of the first term of $f(.)$ is sufficient to show that the problem is concave.

The function logdet($\boldsymbol{A}$) is concave if and only if the matrix $\boldsymbol{A}$ is non-negative definite. The sum of two non-negative definite matrices is also non-negative definite. Since the identity matrix satisfies this condition, we only need to show that the second term of the argument of the log det(.) in (4.14) is non-negative

definite.
The equivalent noise covariance matrix $\boldsymbol{V}_{s_l}$ is diagonal with positive elements which are its eigenvalues, thus it is positive definite. This property stands also for its inverse.
If a positive definite matrix $\boldsymbol{M}$ is multiplied by another matrix and its Hermitian as $\boldsymbol{B}^H\boldsymbol{M}\boldsymbol{B}$, the result is also positive definite if $\boldsymbol{B}$ is full rank. This is true for $\boldsymbol{H}_l^H\boldsymbol{V}_{s_l}^{-1}\boldsymbol{H}_l$ since the columns of $\boldsymbol{H}$ are independent, thus $\text{rank}(\boldsymbol{H}) = s_l$. Consequently, the matrix being the argument of log det(.) is positive definite and also non-negative definite, thus the first term of (4.14) which implies with the above reasons that (4.15) is concave. □

4.5.3 Fronthaul optimization without per-link rate constraint

In the fronthaul deployment described in Scenario 3, the connection between the RRHs and the CO is ensured by a converged Ethernet network owned by the mobile network operator and shared with other services. In this case, the available fronthaul capacity for connecting each RRH is considered to be unlimited, since the overall capacity of the network is much larger than the rate required for the fronthaul transmissions. However, this large capacity needs to be shared with other services, thus the cost function includes a term related to the cost of sharing the fronthaul network. We model this term of the price factor increasing proportionally to the fronthaul rate used, in order to limit the allocation of more rate than what is needed for accurate fronthaul transmissions.

In order to find the optimal fronthaul allocation scheme, we need to solve the following unconstrained optimization problem, which also aims to maximize the the sum of (4.14) over the whole set of user groups, but without having per-link rate constraints to be satisfied:

$$\begin{aligned} &\text{Find } \{\boldsymbol{c}^{(1)*}, ..., \boldsymbol{c}^{(L)*}\} = \underset{\{\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(L)}\}}{\text{argmax}} \sum_{l=1}^{L} b_1(\boldsymbol{H}_l, \boldsymbol{c}^{(l)}) \\ &\text{with } 0 < c_{lk}^* \ \forall k \in \{1, ..., K\}, \ \forall l \in \{1, ..., L\}. \end{aligned} \tag{4.17}$$

Note that fronthaul capacity values are specified to be positive, since the objective function for our specific configuration is meaningful only for this interval. However, it does not constrain the problem in practice. We have the following statement extended from Proposition 2, the proof is very similar and thus omitted.

Proposition 3. *The unconstrained optimization problem (4.17) is concave and admits a unique solution, as it has the same objective function as problem (4.15).*

We can remark that it is possible to solve the optimization problem (4.17)

using convex optimization algorithms adapted for unconstrained problems, or we can compute the gradient of the objective function with respect to the variable $(\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(L)})$ characterizing the fronthaul rate, and solving the equation where the all partial derivatives are equal to zero allows also to find the fronthaul rate allocation resulting in maximal net benefit for the given channel realization. Since the two solutions are equivalent, in order to use similar techniques in all of the three scenarios of deployment cost, we have optimized the objective function directly.

4.5.4 Performance evaluation

After having formalized the optimization problems (4.15) and (4.17) allowing to find the fronthaul allocation schemes that maximize the net benefit for the UL partial NOMA transmission, we have simulated its performance and compared it to uniform allocation of the whole amount of fronthaul rate available. We aim to highlight by these numerical results the benefit of cost-aware fronthaul allocation and show the feasibility of effective UL transmissions in C-RAN. We have evaluated the results of the fronthaul allocation optimization in the 3 deployment scenarios described in Subsection 4.4 with $N = 40$ users transmitting towards $M = 2$ RRHs located at 750 meters from each other. We have $K = 8$ antennas equally distributed between the RRHs. Note that the optimization problem can be solved efficiently using standard convex programming [100]. Channel gain is modeled using independent one-ring scatterer model for each user [88].

To provide an accurate comparison of the scenarios that we have defined in Section 4.4, we assume the following relation among the price factors $\lambda_k^{(i)}$ in order to illustrate the characteristics of the fronthaul architectures and commercial models respectively

$$\lambda_k^{(2)} \leq \lambda_k^{(3)} \leq \lambda_k^{(1)}. \tag{4.18}$$

4.5.4.1 Limited fronthaul rate available

In Scenarios 1 and 2, the fronthaul rate that we can allocate per link is constrained. The gain provided by optimal fronthaul allocation strategy depends on the amount of available fronthaul rate and the fronthaul cost.

We show in Figure 4.6 the benefit realized by uplink multi-user partial NOMA transmissions with $L = 10$ user groups following the metric defined in (4.14). For Scenario 2, we used $\mu_k = \frac{\lambda_k^{(2)}}{2}$ for this evaluation in order to model the fact that constant operational costs are higher than the cost related to the investment which is shared among all the transmissions occurring. We compare the optimal fronthaul allocation scheme to uniform fronthaul allocation for different amounts of available fronthaul capacity. In the case of uniform fronthaul allo-

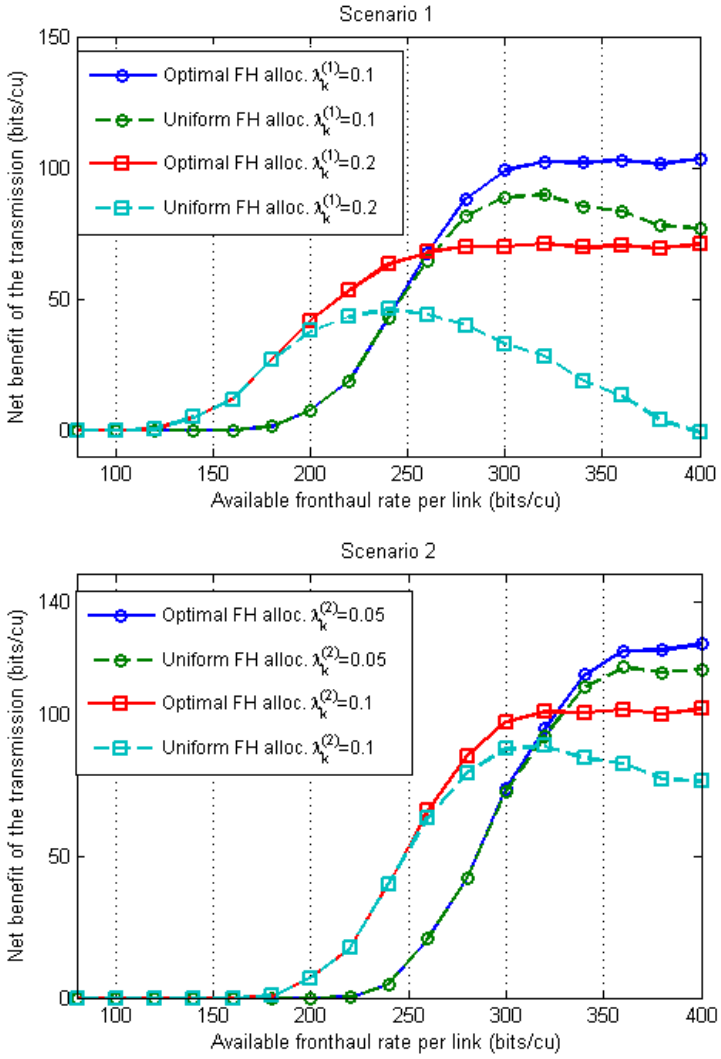


Figure 4.6 – Net benefit of uplink transmission with constrained fronthaul. In the uniform fronthaul allocation scheme we distribute all the available fronthaul rate uniformly between the user groups and the antennas.

cation, available fronthaul capacity is equally distributed to all groups and all antennas.

When the available fronthaul rate is low, both uniform and optimized allocation result in similar efficiency, since the constraint does not allow to achieve higher sum rate. With sufficient fronthaul rate, optimized fronthaul allocation allows to achieve higher transmission benefit, since the sum rate of each group can be improved by allocating more fronthaul to the received signals with higher powers. In other words, fronthaul allocation is adapted to the variations of channel gains for different users and antennas. In Scenario 2, when a point-to-point fronthaul link is owned by the network operator, we can see that the benefit of the transmission is higher when more fronthaul rate is available, since

the investment cost term is reduced thanks to higher sum rate, therefore higher gain.

4.5.4.2 Fronthaul allocation with different cost values

We have evaluated the maximal net benefit that we get by optimizing the fronthaul rate allocation for different cost coefficients. For Scenario 2, we have used $\mu_k^{(2)} = \lambda_k^{(2)}/2$ and for Scenario 3 $\mu_k^{(3)} = \lambda_k^{(3)}/4$.

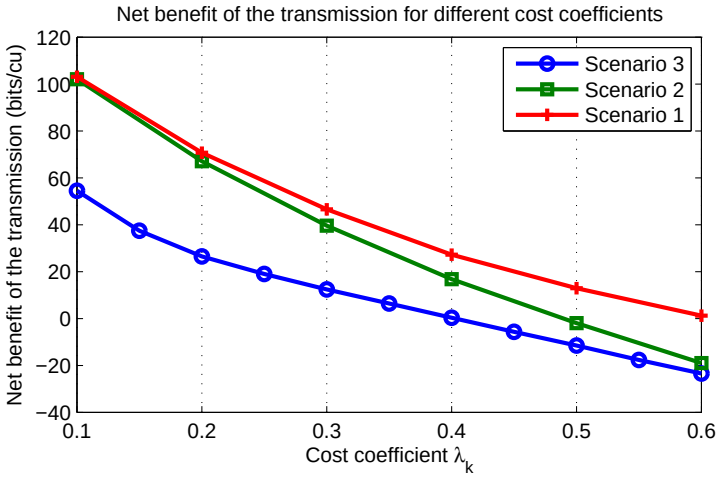


Figure 4.7 – Net benefit of uplink transmission for different cost coefficient values

We can observe in Figure 4.7 that the benefit in Scenario 2 is close to the one of Scenario 1 when exploitation costs are low, and for high cost it approaches the (lower) benefit in Scenario 3. Also, the benefit of the transmission decreases quickly when the cost increases. The benefit of the transmission can even happen to be negative despite optimization, whereas the cost of fronthaul usage can be higher than the total sum rate. Obviously, in this case it is better not to transmit or change the system parameters, e.g., the size of the NOMA groups.

4.5.4.3 Optimization for various group sizes

As we have detailed in Section 4.2, for practical considerations, the best choice is not necessarily to schedule as many users as possible on the same PRBs. However, regarding the fronthaul, one can expect that by reducing the number of user groups, we use less fronthaul and get higher benefit from the transmission. To quantify this benefit, we have evaluated the result of optimal fronthaul allocation for various group sizes. The number of PRBs is fixed to $J = 20$ and the $N = 40$ users are partitioned in groups varying from $L = 20$ to $L = 5$.

We can see in Figure 4.8 that thanks to fronthaul optimization, with groups of 4 users, we already get around 70% of the gain that we can get with 8 users

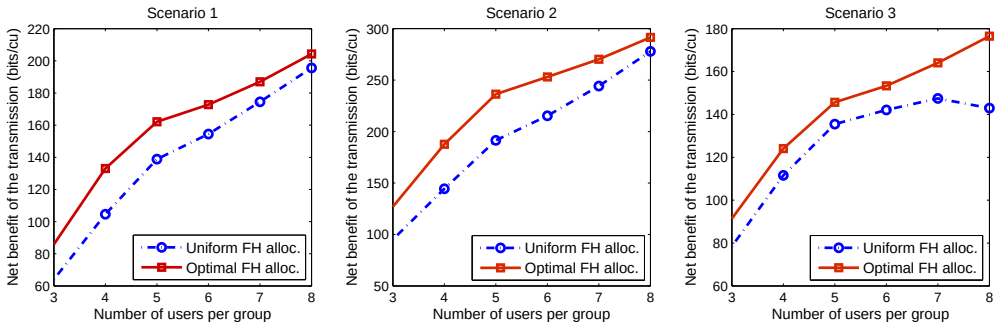


Figure 4.8 – Net benefit of uplink partial NOMA transmission for different group sizes. For Scenario 1 $\lambda_k^{(1)} = 0.2$, $\forall k \in \{1, ..., K\}$ and the available fronthaul rate is 400 $bits/cu/link$. For Scenario 2, $\lambda_k^{(2)} = 0.1$ and $\mu_k^{(2)} = 0.05$, $\forall k \in \{1, ..., K\}$ and the available fronthaul rate is also 400 $bits/cu/link$. For Scenario 3, $\lambda_k^{(3)} = 0.12$ and $\mu_k^{(2)} = 0.03$, $\forall k \in \{1, ..., K\}$. These values are set to follow the differences between deployment and operational cost as described in (4.18).

per group (note that the latter requires much higher computational cost). In Scenarios 1 and 2, we have a gap of about 10% of net benefit between optimized and uniform fronthaul allocation. In these scenarios, since the dominant cost term is the one with $\lambda_k^{(i)}$ which models exploitation costs, fronthaul allocation improves more the transmission gain for medium group size than for large group size. We can achieve a given value of net benefit for a partial NOMA transmission with less users when fronthaul allocation is optimized.

In Scenario 3, when more fronthaul is allocated per group, the cost term with $\mu_k^{(3)}$ that aims equity between the various services sharing the same fronthaul infrastructure, becomes dominant for large user group. Consequently, optimizing the fronthaul allocation gives more improvement compared to smaller group size. However, with optimal fronthaul allocation and a group size of 5 users, we can achieve higher net benefit than for any group size with uniform fronthaul allocation.

We compare in Figure 4.9 the efficiency of fronthaul usage, i.e., the ratio of the net benefit of the transmission and the total fronthaul rate used, in the different deployment scenarios. We can see again that optimizing fronthaul allocation improves the performance of transmissions for any group size in all of the 3 scenarios. With the different cost models used, the fronthaul is exploited with the highest efficiency in Scenario 3. In comparison, Scenario 2 can result in higher efficiency than Scenario 1, in both the uniform and optimized fronthaul allocation cases. We can observe in Figure 4.9 that by optimizing the fronthaul allocation in Scenarios 1 and 2, the efficiency of the fronthaul usage becomes close to the one that we get from uniform allocation under Scenarios 2 and 3, respectively.

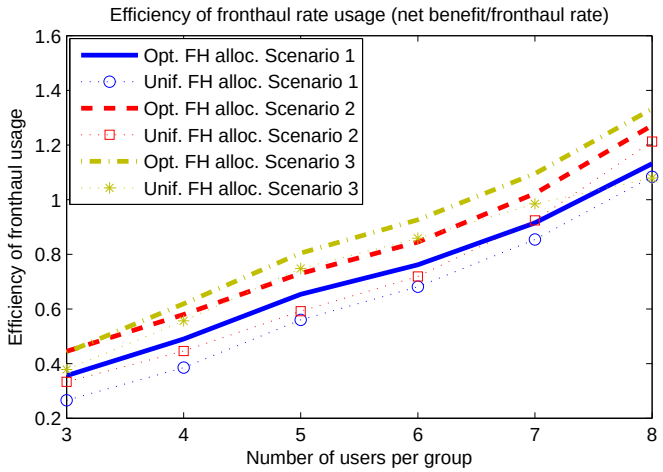


Figure 4.9 – The efficiency of fronthaul usage by uplink partial NOMA transmission for different group sizes in the 3 deployment scenarios defined in Subsection 4.4

4.5.4.4 Observations

Through these numerical evaluations, we have found that the more fronthaul rate is available on the link between the RRHs and the CO, the more fronthaul allocation optimization improves the net benefit with respect to uniform allocation. The comparison of deployment scenarios for different exploitation cost coefficients has shown that leasing fronthaul infrastructures can be the most beneficial except when the costs of deploying new links is negligible (e.g., very long term investments). We can also confirm that independently of the cost model, fronthaul allocation is useful for any group size and can compensate the loss of benefit due to partial NOMA instead of using full NOMA. By optimizing fronthaul rates, we can improve the efficiency of exploiting fronthaul links, for example optimal allocation in simpler models (Scenarios 1 and 2) can be as efficient as the more evolved one (Scenario 3) with uniform allocation. These show that by combining partial multi-cell NOMA on the C-RAN radio interface with cost-aware fronthaul allocation on the RRH-CO interconnection, we are able to ensure high spectral efficiency and throughput, by using affordable and practically implementable multi-user receiver and maximizing the benefit that operators get despite the fronthaul usage cost that is considered as the main limitation of Cloud RAN.

4.6 Fronthaul allocation optimization using channel statistics

To evaluate the fronthaul allocation strategy optimizing the net benefit of uplink partial NOMA in more realistic conditions of network operation, we consider in this section the channel estimation error made by the receiver and its impact on the overall performance. Furthermore, as we can formulate only the probability distribution of this error with a given channel estimation method, in a practical system, it is not possible to know how accurate a given channel estimate is. Thus, we use an approximation based on channel statistics that are more stable in time for a user in a given position and independent of the subcarrier frequency. This allows to provide the fronthaul allocation scheme maximizing this approximate value of the net benefit in advance, simultaneously with the scheduling of user groups, without having a delay due to the channel estimation. Consequently, when the CO selects the cell-edge users participating in the partial NOMA transmission, it creates users groups randomly or using a heuristic (see Section 4.3), then scheduling is realized according to the decided grouping scheme and the optimal fronthaul allocation scheme can be also computed from the users' channel statistics. Then, along with signaling containing UL configuration parameters, the CO can send to the RRHs the fronthaul quantization rate to be used on the uplink for each antenna and user group.

4.6.1 Characterization of channel estimates

Channels between each UE and each antenna are estimated at the receiver based on pilot sequences. We assume that co-channel users transmit orthogonal pilots, thus, there is no correlation introduced by the channel estimator. We consider that channel estimation takes place in the RRHs. Though the allocation of orthogonal pilot sequences following the scheduling decisions happens in the CO, then it forwards to the RRHs information about the pilot sequences allocated to every user so that each RRH can estimate the channel of every user toward each of its antennas on the PRBs allocated to each of them. Offloading this processing to the RRHs allows to reduce computation latency in the CO, but increases slightly control-plane traffic over the fronthaul links. Note that quantization of channel estimates can also have an effect on the performance of the receiver, however, as the overall amount of data corresponding to channel estimates is lower (thanks to extrapolation in time and frequency and antenna combining), we can normally allow high quantization rates to minimize distortion. Joint compression of data and channel estimates exploiting the correlation between them [101] can improve the efficiency of fronthaul usage.

The estimate of channel realization $\mathbf{H}_l$ is denoted by $\hat{\mathbf{H}}_l$ and the matrix of channel estimation error is $\tilde{\mathbf{H}}_l$. We denote by $\hat{\boldsymbol{h}}_{\pi_i^l}$ and $\tilde{\boldsymbol{h}}_{\pi_i^l}$ the columns of $\hat{\mathbf{H}}_l$ and $\tilde{\mathbf{H}}_l$ representing for each user in the group Π_l the channel estimates and the estimation errors respectively. They can be characterized by $\hat{\boldsymbol{h}}_{\pi_i^l} \sim \mathcal{N}(0, \hat{\boldsymbol{R}}_n)$ and $\tilde{\boldsymbol{h}}_{\pi_i^l} \sim \mathcal{N}(0, \tilde{\boldsymbol{R}}_n)$ where $\hat{\boldsymbol{R}}_n$ and $\tilde{\boldsymbol{R}}_n$ are complex valued matrices. Consequently, we have $\mathbf{H}_l \sim \mathcal{N}(0, \sum_{i=1}^{s_l} \boldsymbol{R}_i)$, $\hat{\mathbf{H}}_l \sim \mathcal{N}(0, \sum_{i=1}^{s_l} \hat{\boldsymbol{R}}_i)$ and $\tilde{\mathbf{H}}_l \sim \mathcal{N}(0, \sum_{i=1}^{s_l} \tilde{\boldsymbol{R}}_i)$. For a given user n we have $\boldsymbol{h}_n = \hat{\boldsymbol{h}}_n + \tilde{\boldsymbol{h}}_n$ and similarly for the group Π_l: $\mathbf{H}_l = \hat{\mathbf{H}}_l + \tilde{\mathbf{H}}_l$.

The relation between the covariance matrices can be written as $\boldsymbol{R}_n = \hat{\boldsymbol{R}}_n + \tilde{\boldsymbol{R}}_n$. Using Minimum Mean Square Error (MMSE) channel estimation, we have the following relation between $\boldsymbol{R}_n$ and $\hat{\boldsymbol{R}}_n$ [102]:

$$\hat{\boldsymbol{R}}_n = \boldsymbol{R}_n \cdot (\sigma_z^2 \boldsymbol{I}_K + \boldsymbol{R}_n)^{-1} \cdot \boldsymbol{R}_n. \tag{4.19}$$

4.6.2 Ergodic sum rate with imperfect CSI

Using the actual channel estimate matrix $\hat{\mathbf{H}}_l$, the receiver in the CO performs multi-user detection of the signal

$$\hat{\boldsymbol{y}}_l = \boldsymbol{\alpha}_l \left(\frac{1}{\sqrt{K}} \sum_{i=1}^{s_l} (\hat{\boldsymbol{h}}_{\pi_i^l} + \tilde{\boldsymbol{h}}_{\pi_i^l}) x_{\pi_i^l} + \boldsymbol{n}_l \right) \tag{4.20}$$

where the Gaussian noise vector $\boldsymbol{n}_l$ includes both channel and quantization noises. Note that the normalization factor $\frac{1}{\sqrt{K}}$ accounts for the fact that the signal power is split over the whole set of antennas. In the statistical description of the received signal we need to consider this averaging to derive results independently from the number of antennas used.

Thus, the achievable sum rate for the user group l is given by

$$\sum_{i=1}^{s_l} r_i \geq I(\mathbf{X}; \hat{\mathbf{Y}} \mid \hat{\mathbf{H}}) = h(\mathbf{X}) - h(\mathbf{X} \mid \hat{\boldsymbol{y}}, \hat{\mathbf{H}}) \tag{4.21}$$

As the realization of the channel estimation error can not be known at each transmission slot, we can only characterize the transmission in average, i.e, by the ergodic achievable sum rate that we obtain after multi-user receive processing at the CO.

$$\sum_{i=1}^{s_l} r_i \geq \mathbb{E}_{\hat{\boldsymbol{H}}_l} \left[\log \det \left(\boldsymbol{I}_{s_l} + \frac{1}{K} \hat{\boldsymbol{H}}_l^H \boldsymbol{V}_{e,s_l}^{-1} \hat{\boldsymbol{H}}_l \right) \right] \tag{4.22}$$

with $\boldsymbol{V}_{e,s_l}$ the equivalent noise covariance:

$$\boldsymbol{V}_{e,s_l} = \sum_{i=1}^{s_l} \tilde{\boldsymbol{R}}_i + \sigma_z^2 \boldsymbol{I}_K + \operatorname*{diag}_{k=\{1,\ldots,K\}} \left(\frac{\sigma^2_{y_{lk}|\hat{\mathbf{H}}_l} 2^{-c_{lk}}}{1 - 2^{-c_{lk}}} \right) \tag{4.23}$$

where the variance of the signal received at the antenna k is $\sigma^2_{y_{lk}|\hat{\mathbf{H}}_l} = \{\boldsymbol{R}_l\}_{kk} + \sigma_z^2$ where we denote by $\{\boldsymbol{R}_l\}_{kk}$ the k^{th} diagonal element of $\boldsymbol{R}_l$.

Computing this expression would require channel estimates over a long time period, thus a significant delay for performing the optimization. Furthermore, the computational complexity of the fronthaul optimization based on the ergodic achievable sum rate would be very complex compared to the case with perfect CSI. As an exact closed-form expression is not available, it is necessary to find an approximation of this quantity.

4.6.3 Approximation of the sum rate by deterministic equivalent

It has been shown in [103] that in large-scale systems the channel capacity converges towards a deterministic formulation which does not depend on real-time channel estimates, only on the channel covariance. In this section we investigate the possible usage of such an approximation to our use case, since the numerical evaluations in [104] indicate that for finite number of antennas it is already accurate. Furthermore we can benefit from the C-RAN architecture to apply it for the whole set of antennas available at the cooperating RRHs, thus the total number of antennas can be relatively high even if there are only a few of them at each RRH, only the number of users in each group is limited to keep low receiver complexity.

To be able to approximate the average sum rate, in equation (4.22), we separate the channel-related and the noise-related terms.

$$\begin{aligned}&\mathbb{E}_{\hat{\boldsymbol{H}}_l}\left[\log\det\left(\boldsymbol{I}_{s_l}+\frac{1}{K}\hat{\boldsymbol{H}}_l^H\boldsymbol{V}_{e,s_l}^{-1}\hat{\boldsymbol{H}}_l\right)\right]=\\&\mathbb{E}_{\hat{\boldsymbol{H}}_l}\left[\log\det\left(\boldsymbol{V}_{e,s_l}+\frac{1}{K}\hat{\boldsymbol{H}}_l\hat{\boldsymbol{H}}_l^H\right)\right]-\text{logdet}(\boldsymbol{V}_{e,s_l})=\end{aligned} \tag{4.24a}$$

$$\begin{aligned}&\mathbb{E}_{\hat{\boldsymbol{H}}_l}\left[\log\det\left(\boldsymbol{I}_K+\frac{1}{\sigma_z^2}\left(\sum_{i=1}^{s_l}\tilde{\boldsymbol{R}}_i+\underset{k=\{1,\dots,K\}}{\text{diag}}\left(\frac{\sigma^2_{y_{lk}|\hat{\mathbf{H}}_l}2^{-c_{lk}}}{1-2^{-c_{lk}}}\right)+\frac{1}{K}\hat{\boldsymbol{H}}_l\hat{\boldsymbol{H}}_l^H\right)\right)\right]\\&\quad-\log\det(\boldsymbol{V}_{e,s_l})+\log\det(\sigma_z^2\boldsymbol{I}_K)\end{aligned} \tag{4.24b}$$

Following the result in [105, Theorem 14.iv], we can approximate the average of the first term of (4.24) over the antennas, benefiting from the knowledge of the covariance of the channel estimates $\hat{\boldsymbol{R}}_l$.

$$\frac{1}{K}\mathbb{E}_{\hat{\boldsymbol{H}}_l}[\mathbf{S}_K]\approx\boldsymbol{Q}_K \tag{4.25}$$

where

$$\mathsf{S}_K = \log\det\left(\boldsymbol{I}_K + \frac{1}{\sigma_z^2}\left(\boldsymbol{B}_K(\sigma_z^2, \boldsymbol{c}^{(l)}) + \frac{1}{K}\hat{\boldsymbol{H}}_l\hat{\boldsymbol{H}}_l^H\right)\right) \tag{4.26}$$

and

$$\begin{aligned}\boldsymbol{Q}_K = \frac{1}{K}\log\det&\left(\boldsymbol{I}_K + \frac{1}{\sigma_z^2}\boldsymbol{B}_K(\sigma_z^2, \boldsymbol{c}^{(l)}) + \frac{1}{\sigma_z^2}\frac{1}{K}\sum_{i=1}^{s_l}\frac{\hat{\boldsymbol{R}}_i}{1+\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}\right)\\ &+ \frac{1}{K}\sum_{i=1}^{s_l}\log(1+\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})) - \frac{1}{K}\sum_{i=1}^{s_l}\frac{\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}{1+\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}\end{aligned} \tag{4.27}$$

with

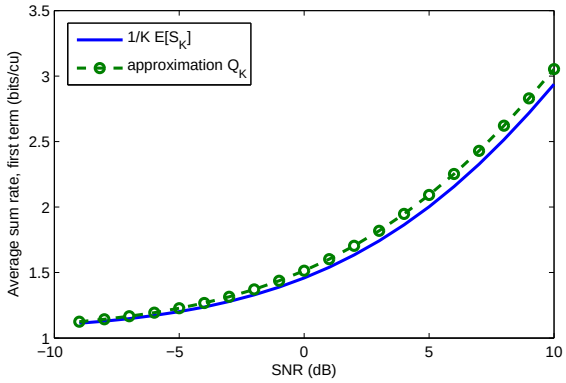
$$\boldsymbol{B}_K(\sigma_z^2, \boldsymbol{c}^{(l)}) = \sum_{i=1}^{s_l}\tilde{\boldsymbol{R}}_i + \underset{k=\{1,\ldots,K\}}{\text{diag}}\left(\frac{\sigma^2_{y_{lk}|\hat{\mathsf{H}}_l}2^{-c_{lk}}}{1-2^{-c_{lk}}}\right) \tag{4.28}$$

One can observe, that the expression $\boldsymbol{Q}_K$ contains the deterministic terms of the random matrix S_K without any change, and represents the random part related to the matrix $\hat{\boldsymbol{H}}_l$ using its statistical characterization. They include the Stieltjes transform $\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})$ describing the probability distribution of the eigenvalues of S_K. We can compute for every $i \in \{1, ..., s_l\}$ the Stieltjes transform adapted to the expression in (4.24b) iteratively, as the solution of the following fixed point equation:

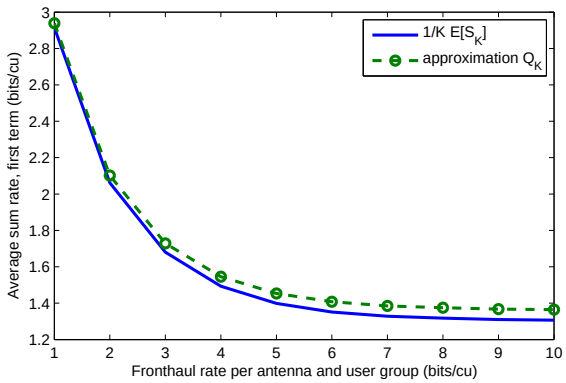
$$\delta_i(x) = \frac{1}{K}\text{tr}\left[\hat{\boldsymbol{R}}_i\left(\frac{1}{K}\sum_{n=1}^{s_l}\frac{\hat{\boldsymbol{R}}_n}{1+\delta_n(x)} + \boldsymbol{B}_K - x\boldsymbol{I}_K\right)^{-1}\right] \tag{4.29}$$

We can verify if for our simulation configuration with $K = 8$ antennas and $s_l = 4$ users the approximation is accurate. We have evaluated with respect to the noise variance the value of the first term of the average sum rate $\frac{1}{K}\mathbb{E}_{\hat{\boldsymbol{H}}_l}[\mathsf{S}_K]$ and the deterministic equivalent $\boldsymbol{Q}_K$.

We can observe in Figure 4.10 the variation of the first term of (4.24b) averaged over the number of antennas and its approximation by deterministic equivalent for different values of SNR with the fronthaul rate equal to 8 bits/channel use for each group and at each antenna so that it is enough high to minimize the influence of fronthaul quantization on the sum rate. As we can see, for low SNR values the approximation fits well the original expression, for higher SNR a small gap appears, however the approximation still follows the variation of the simulated values. For the fixed value of $\mathsf{SNR} = 5\ dB$, the evaluation of the approximation when the fronthaul rate allocated to the user group at 1 antenna varies is also shown in Figure 4.10. In this case, which is of more interest for the fronthaul optimization problem, we can also observe a good fit between the approximation and the simulation. Note that the first term of (4.24b) decreases



(a) Approximation of the sum rate vs. SNR.



(b) Approximation of the sum rate vs. fronthaul rate usage

Figure 4.10 – Comparison between the simulated ergodic sum rate and its approximation by deterministic equivalent

when more fronthaul rate allocated since its third term which is deterministic covers the increase of the total sum rate.

4.6.4 Fronthaul allocation using statistical approximation

4.6.4.1 With per-link fronthaul limitation

The approximation of the average sum rate by deterministic equivalent allows to optimize the fronthaul allocation scheme based only on the knowledge of the covariance of the channel estimate matrix. To find the optimal fronthaul allocation for every user group and receive antenna in terms of net benefit of the transmission using uplink partial NOMA, we can solve the following problem in the deployment scenarios where the fronthaul rate per link is limited (Scenario 1 and 2).

$$\begin{aligned} &\text{Find } \{\boldsymbol{c}^{(1)*}, ..., \boldsymbol{c}^{(L)*}\} = \underset{\{\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(L)}\}}{\text{argmax}} \sum_{l=1}^{L} b_2(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)}) \\ &\text{subject to } \sum_{l=1}^{L} c_m^{(l)} \leq \overline{c}_m, \ \forall m \in \{1, ..., M\}. \end{aligned} \tag{4.30}$$

where

$$b_2(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)}) = K \cdot \boldsymbol{Q}_K(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)}) - \log \det(\boldsymbol{V}_{e,s_l}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})) + \log \det(\sigma_z^2 \boldsymbol{I}_K) - q(\boldsymbol{c}^{(l)}) \tag{4.31}$$

Even though the approximation of the sum rate is a closed-form expression, it is difficult to determine its concavity with respect to the fronthaul rate, due to the iterative formulation of the Stieltjes transform. We can verify numerically the variation of the derivative of (4.31), in order to obtain at least an experimental confirmation of the applicability of the function $b_2(.)$ in the optimization problem (4.30).

The partial derivative of $b_2(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})$ with respect to $c_{lk}, \ \forall k \in \{1, ..., K\}$ is given by

$$\begin{aligned} \frac{\partial b_2(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})}{\partial c_{lk}} &= K \cdot \frac{\partial \boldsymbol{Q}_K(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})}{\partial c_{lk}} - \frac{1}{\ln 2} \text{tr} \left[\boldsymbol{V}_{e,s_l}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})^{-1} \frac{\partial \boldsymbol{V}_{e,s_l}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})}{\partial c_{lk}} \right] \\ &\quad - \frac{\partial q(\boldsymbol{c}^{(l)})}{\partial c_{lk}} \end{aligned} \tag{4.32}$$

where

$$\begin{aligned} \frac{\partial \boldsymbol{Q}_K(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})}{\partial c_{lk}} &= \frac{1}{K} \frac{1}{\ln 2} \text{tr} \left[\boldsymbol{A}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})^{-1} \frac{\partial \boldsymbol{A}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})}{\partial c_{lk}} \right] - \frac{1}{K} \frac{1}{\ln 2} \frac{\delta'_{ik}(-\sigma_z^2, \boldsymbol{c}^{(l)})}{1 + \delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})} \\ &\quad - \frac{1}{K} \frac{\hat{\boldsymbol{R}}_i \delta'_{ik}(-\sigma_z^2, \boldsymbol{c}^{(l)})}{(1 + \delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)}))^2}, \end{aligned} \tag{4.33}$$

with

$$\boldsymbol{A}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)}) = \boldsymbol{I}_K + \frac{1}{\sigma_z^2}\boldsymbol{B}_K(\sigma_z^2, \boldsymbol{c}^{(l)}) + \frac{1}{\sigma_z^2}\frac{1}{K}\sum_{i=1}^{s_l}\frac{\hat{\boldsymbol{R}}_i}{1+\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}$$

$$\text{and } \frac{\partial \boldsymbol{A}(\hat{\boldsymbol{R}}_l, \boldsymbol{c}^{(l)})}{\partial c_{lk}} = \frac{1}{\sigma_z^2}\left(\underset{j\in\{1,\dots,K\}}{\text{diag}}\{d_{jj}\} - \frac{1}{K}\sum_{i=1}^{s_l}\frac{\hat{\boldsymbol{R}}_i \delta'_{ik}(-\sigma_z^2, \boldsymbol{c}^{(l)})}{(1+\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)}))^2}\right),$$

$$\text{where } d_{jj} = \begin{cases} 0 & \text{if } j \neq k \\ (\{\boldsymbol{R}_l\}_{kk} + \sigma_z^2)\frac{-\ln 2 \cdot 2^{-c_{lk}}}{1-2^{-c_{lk}}} & \text{if } j = k \end{cases}$$

$$\text{and } \delta'_{ik}(-\sigma_z^2, \boldsymbol{c}^{(l)}) = \frac{\partial \delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}}$$

The partial derivative $\delta'_i(-\sigma_z^2, \boldsymbol{c}^{(l)})$ is the i^{th} element of the vector $\boldsymbol{\delta}'(-\sigma_z^2, \boldsymbol{c}^{(l)})$ which is obtained through the following matrix formulation

$$\begin{aligned}
&\boldsymbol{\delta}'(-\sigma_z^2, \boldsymbol{c}^{(l)}) = (\boldsymbol{I}_{s_l} - \boldsymbol{J}_{s_l})^{-1}\boldsymbol{v}_{s_l} \\
&\text{where } \boldsymbol{J}_{s_l} \in \mathbb{C}^{s_l \times s_l} \text{ and } \boldsymbol{v}_{s_l} \in \mathbb{C}^{s_l \times s_l} \text{ are defined by} \\
&\{\boldsymbol{J}_{s_l}\}_{mi} = \frac{\text{tr}\,[\hat{\boldsymbol{R}}_m \boldsymbol{T}_K^{-1}(\boldsymbol{c}^{(l)})\hat{\boldsymbol{R}}_i \boldsymbol{T}_K^{-1}(\boldsymbol{c}^{(l)})]}{K^2(1+\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)}))^2} \ \forall m, i \in \{1, ..., s_l\} \\
&\{\boldsymbol{v}_{s_l}\}_m = \text{tr}\,[\hat{\boldsymbol{R}}_m \boldsymbol{T}_K^{-1}(\boldsymbol{c}^{(l)}) \underset{j\in\{1,\dots,K\}}{\text{diag}}\{d_{jj}\}\boldsymbol{T}_K^{-1}(\boldsymbol{c}^{(l)})] \ \forall m \in \{1, ..., s_l\} \\
&\text{with } \boldsymbol{T}_K = \frac{1}{K}\sum_{n=1}^{s_l}\frac{\hat{\boldsymbol{R}}_n}{1+\delta_n(-\sigma_z^2, \boldsymbol{c}^{(l)})} + \boldsymbol{B}_K - \sigma_z^2 \boldsymbol{I}_k.
\end{aligned} \tag{4.34}$$

The derivation allowing to get the expression of $\boldsymbol{\delta}'(-\sigma_z^2, \boldsymbol{c}^{(l)})$ is provided in Appendix 1.

We have evaluated the value of the partial derivative (4.32) with the cost function $q(\boldsymbol{c}^{(l)}) = q_1(\boldsymbol{c}^{(l)})$ with price factors $\lambda_k^{(1)} = 0.01$ and $\lambda_k^{(1)} = 0.05$ for $\mathsf{SNR} = 5dB$. Note that since the approximation by deterministic equivalent is more accurate for transmissions at low SNR when the size of the system is not large, in order to keep the same system size as for numerical evaluations with perfect CSI assumed (see Section 4.5), we use lower values for the SNR and the cost coefficients. We show in Figure 4.11 the variation of the partial derivative of our objective function (4.31) with respect to the fronthaul rate allocated to the signal received from one user group at one antenna (i.e., c_{lk}). We can observe that for lower capacity values the derivative is positive, while for higher fronthaul capacity it turns to negative, meaning that the objective function has a maximum on the interval evaluated. Even though this maximum can be a local one in the whole definition domain of the function, for the usual range of fronthaul rate values, it appears that it is unique. Consequently, the objective function based on the approximation of the ergodic group sum rate

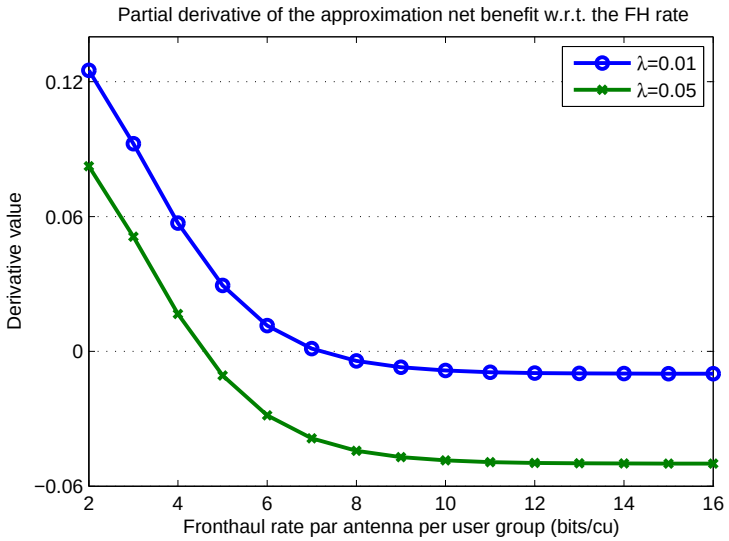


Figure 4.11 – Partial derivative of the objective function w.r.t. one component of the fronthaul rate based on approximated sum rate

by its deterministic equivalent can be used in a convex optimization problem to find the fronthaul allocation scheme that maximizes the average net benefit.

4.6.4.2 Optimal fronthaul allocation without per-link limitation

We use here the same fronthaul cost models as in Section 4.5 (for detailed description of the cost models see Section 4.4), thus we consider in Scenarios 1 and 2 limited available fronthaul rate for each link connecting a RRH to the CO, while in Scenario 3, unlimited fronthaul rate is available, and a cost is accounted for preventing other services from using that same transmission. As the fronthaul cost is part of the objective function instead of being a constraint, we can find the optimal fronthaul allocation by solving the system of equations where all the partial derivatives (4.32) are equal to zero. Solvers for non-linear systems of equations can efficiently find the optimal fronthaul rate values that we can then inject in the objective function to get the value of the net benefit achieved. This method allows to find the fronthaul quantization rates maximizing the overall net benefit of the transmission for a given grouping scheme.

We can also use unconstrained optimization by solving a simple system of equations to determine the optimal per-link fronthaul usage for a given set of system characteristics $\{K, s_l, L, \sigma_z^2\}$ that are relatively stable in time, and various random association of channel covariance measurements at possible user positions. By this offline optimization we should accurately predict the total amount of fronthaul rate to be used during the period when the parameters hold. If the difference of the optimal fronthaul usage for the various grouping possibilities is

not significant, ideal fronthaul rate dimensioning should be realized in advance to the actual radio transmissions.

4.6.5 Numerical evaluations

We have carried out numerical simulations aiming to evaluate the performance of fronthaul optimization exploiting only channel statistics. Similarly to the configuration used in Section 4.5, we have fixed the number of RRHs to $M = 2$ and the number of antennas at each RRH to $a = 4$, except where otherwise indicated. The distance between the RRHs is set to 750 meters. The overall number of users is $N = 40$, they are placed in the overlapping area of the two neighbor cells and they perform uplink transmission following the partial NOMA scheme. Users are randomly associated in $L = 10$ groups, if other value is not specified. The channel covariances between each user and the whole set of receive antennas are computed using the one-ring scatterers model [88]. To compute ergodic sum-rates, we have generated for each user i independent realizations of the channel estimate vectors following the Gaussian distribution with zero mean and covariance $\hat{\boldsymbol{R}}_i$.

According to the definition of the various cost functions in the three deployment scenarios, we evaluate them with $\lambda_k^{(1)} = 0.02,\ \lambda_k^{(2)} = 0.01,\ \mu_k^{(2)} = \lambda_k^{(2)}/2$ and $\lambda_k^{(3)} = 0.015,\ \mu_k^{(3)} = \lambda_k^{(3)}/4$. Note that we use lower values for these parameters in order to perform efficiently the optimization even with low SNR, where the approximation by deterministic equivalent is more precise.

4.6.5.1 Fronthaul allocation with per-link fronthaul constraint

First, we have evaluated the covariance-based fronthaul allocation with limited fronthaul rate available on each link in Scenarios 1 and 2. Numerical results are shown in Figure 4.12. Note that the slight difference between the ergodic and the approximated values comes from the small gap that we can see also in Figure 4.10, and as we approximate the average sum rate of each user group, a more significant difference appears in the obtained values. However, it does not affect the optimization, since the variation of the functions is the same, thus they reach their maxima for the same fronthaul capacity values.

We can observe, that by optimizing of the objective function that includes the cost of fronthaul rate usage, we can avoid to operate in a regime where too much fronthaul can be allocated for the spectral efficiency that we are able to achieve. In other words, when more fronthaul rate is available than needed for accurate transmission between the RRHs and the CO, we use only as much as we need to maximize the overall benefit. Though, as we evaluate the fronthaul allocation that is optimal for the net benefit achieved in average, optimal allocation does

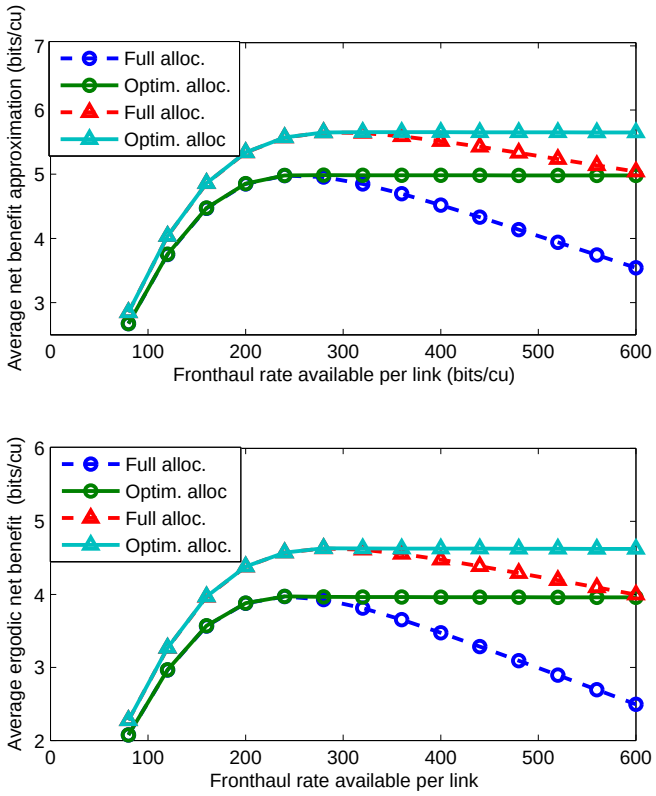


Figure 4.12 – Comparison of fully allocated and optimized fronthaul rate for approximated and ergodic average net benefit for $\sigma_z^2 = 5\ dB$.

not result in an opportunistic gain is it did for a given channel realization that is known.

4.6.5.2 Fronthaul dimensioning using static parameters

We can observe in Figure 4.12 that when the limitation of available fronthaul rate is selected such as the cost of fronthaul cost is not predominant over the realized sum rate, the gap between allocating completely the available fronthaul and applying the optimal scheme is negligible, thus it can be interesting to only compute the fronthaul limitation where the net benefit is maximal. Furthermore, as our objective function depends only on parameters that we can set constant for a given time period when the number of users that need to be involved in the partial NOMA transmission does not change, we can compute the typical value of the optimal fronthaul rate limitation. For this computation, we need to exploit per-user channel covariance measurements from the considered area, they can be known either from previous transmissions or by realizing offline channel measurements, as channel states are relatively constant in time.

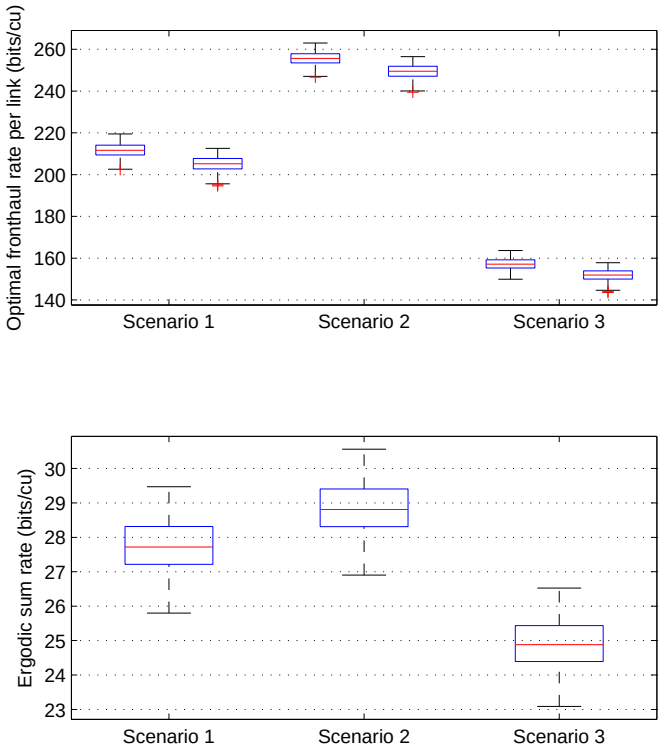


Figure 4.13 – Variation of the optimal per-link fronthaul rate and net benefit for different realizations of user grouping in the three deployment scenarios.

We show in Figure 4.13 the optimal fronthaul rate limitation and the corresponding sum rates for different random user grouping schemes. One can see that the values obtained by optimization are close to each other, thus dimensioning the fronthaul rate available per link would not decrease significantly the net benefit of the transmission compared to the maximal value. As in our system model the users are located in the same area, they have similar channel covariances, thus changing the grouping scheme does not affect much the result of the optimization.

The evaluation of per-link fronthaul rate and the corresponding sum rate for different SNR values in the three scenarios is shown in Figure 4.14. The aim of this simulation was to investigate the effect of channel noise or different power allocation that can happen for example when cell size is larger on the fronthaul rate requirements. We can see that similar throughput can be realized with optimal fronthaul allocation in all of the deployment scenarios. However, in the case of converged fronthaul infrastructure (Scenario 3), fronthaul rate is significantly lower than in the other scenarios, and this difference increases with the SNR. Note that even if fronthaul rate usage is the highest in Scenario

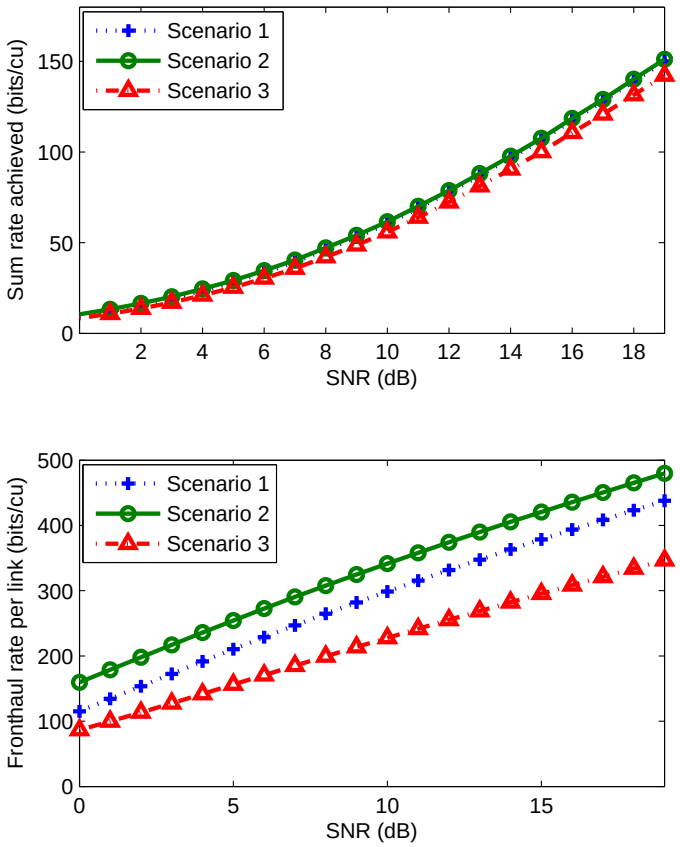


Figure 4.14 – Sum rate and per-link fronthaul rate with optimal allocation w.r.t. the SNR. We compare the throughput and the used fronthaul rate obtained by optimizing the objective functions with the cost formulations of the three scenarios.

2, as we use a dedicated point-to-point link which probably needs to be over-dimensioned to accommodate traffic variations, the higher fronthaul rate impacts less the fronthaul cost. However, these results suggest that for new deployments it is more efficient to use fronthaul deployments of Scenario 1 (leasing) or 3 (converged network).

4.6.5.3 Ideal number of antennas with limited fronthaul

As the fronthaul rate usage is related to the number of antennas, we have evaluated the benefit of the transmission and the fronthaul usage in the deployment scenarios where per-link fronthaul rate is limited (Scenarios 1 and 2). The results depicted in Figure 4.15 indicate, as expected, that with more antennas we can achieve higher performance while using less fronthaul. We can also observe, that the difference between the two scenarios is less significant with 8 anten-

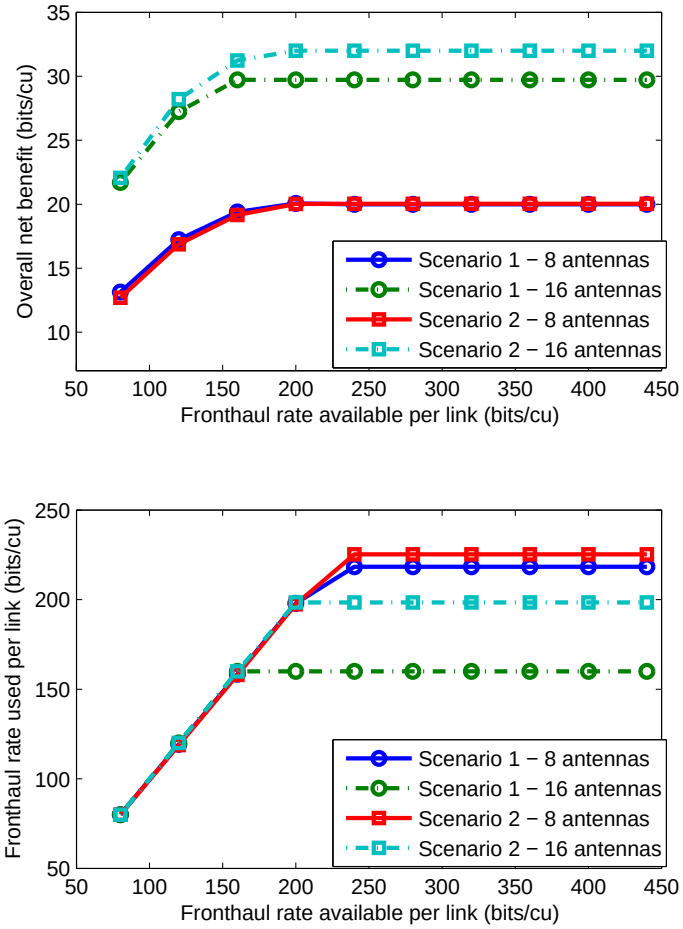


Figure 4.15 – Optimal net benefit and fronthaul rate usage with 4 and 8 antennas at each of the two RRHs vs. the fronthaul rate available per link in Scenarios 1 and 2.

nas in total than with 16. In Scenario 1, it can be more interesting to deploy more antennas since using the optimization of the net benefit we obtain better result with less fronthaul rate used, while in Scenario 2, more antennas require also more fronthaul rate to achieve higher performance. It worth noting that when low fronthaul rate is available only, we can almost double the net benefit with two times more antennas and the same total fronthaul rate, but when more fronthaul is used the improvement obtained by adding antennas is slightly lower.

4.6.5.4 Observations

Through the various numerical evaluations that we have carried out, we have shown that it is possible to optimize the net benefit of uplink multi-user trans-

missions based on channel statistics is useful to avoid allocating superfluous fronthaul rate that does would decrease the overall gain of the mobile network operator. We have also illustrated that the formulation of the objective function that we propose allows to find by offline computation the ideal amount of the fronthaul rate to be used per link, which can be useful for dimensioning fronthaul links previously to actual transmissions. By comparing the net benefit for different antenna numbers we could see in which cases it gives more gain to have more receive antennas. These results provide a complementary viewpoint to the ones that we have shown for the case when we have assumed the knowledge of perfect real-time CSI (see Section 4.5).

Part 5

Conclusion and outlook

5.1 Concluding remarks

We have studied in this thesis various aspects of multi-user communications in the Cloud RAN architecture and provided an end-to-end framework for practical deployments. As the main limitation of C-RAN is the fronthaul connectivity between the RRHs and the CO, we have proposed solutions for functional architecture and optimized operation aiming to satisfy future network requirements. We have also investigated and shown in a proof-of-concept demonstration how low-latency coordination can be realized among base-band processes hosted in the same C-RAN server. Hopefully, our results are useful to improve the performance of RAN in 5G and beyond, where denser deployment of radio access points is foreseen.

In Part 2, generalities about Cloud RAN, current mobile networks and physical layer processing and current tendencies for 5G were described. We could see that according to related works, a major part of the benefits of C-RAN is related to its capability to perform multi-cell processing, however, under the 5G requirements their practical realization can be challenging. As the main technological limitation in C-RAN is the capacity of the fronthaul links, we have evaluated the amount of information that multi-cell processing requires to be transferred over these links and we have proposed a dynamic placement of physical layer functions that minimizes the fronthaul rate usage and also enables multi-cell cooperation both on the downlink and the uplink.

Then, we have focused on coordination and implementation aspects of future C-RAN deployments in Part 3. The current research about realizing RAN processing by software processes running on generic computing units and also bringing methods and architectures used in software products into applications in telecommunications has been described. Fitting in this background, we have presented the architecture that integrates SDN to C-RAN deployments. Details

about the realization of our prototype platform demonstrating how SDN enables low-latency multi-cell coordination were also provided in this part.

Finally, we have studied from a theoretical perspective, but still based on practical considerations, how to use fronthaul links efficiently for multi-user transmissions in the C-RAN architecture that we have defined. As the benefit of high over-the-air throughput can be compromised by the cost of using high fronthaul rate, we have optimized fronthaul allocation following a metric that characterizes the overall net benefit of uplink transmissions. In a first approach, we have assumed that channel realizations are known by the receiver, thus it can benefit from an opportunistic gain using fronthaul allocation adapted to the current channel gain. A second approach consisted in finding the best fronthaul allocation in average, based only on channel covariance values and system parameters. We could point out the utility of this optimization for longer time periods in dimensioning fronthaul links offline, that already allow efficient operation from an end-to-end perspective without computation overhead.

To provide a general conclusion of the thesis, we can note the different dimensions of practical realization of multi-cell multi-user transmissions that we have studied show that using a careful end-to-end design of C-RAN architecture and processing, we can benefit from its advantages. As for many new technologies, implementation challenges need to be solved by improving several elements of the whole system. Often simple solutions with low complexity can provide almost as much gain as complex processing, and they are more likely to be adopted in real products. Even though any of the design elements alone does not enable the exploitation of inter-cell interference in C-RAN, their joint usage solves the major challenges of this implementation.

5.2 Future research axes

5.2.1 Next steps

For numerical evaluations of the various proposed solutions, we have realized Monte-Carlo simulations with a channel model that computes the channel gain independently for each user. It would be very interesting to integrate together all of the elements, i.e., the user-selective split, the SDN-based control, the real-time fronthaul allocation and the covariance-based dimensioning of the fronthaul links in an end-to-end network simulator with several cells and where the fronthaul infrastructure follows real-word cases instead of mathematical modeling. Furthermore a spatially consistent channel model where randomly placed scatterers are also correlated from one user to another - not only the deterministic part based on the position of the user - would make the performance results

more realistic. We should also tune the parameters of the cost functions based on observation of real-word data, particularly regarding energy consumption in the different deployment scenarios.

Regarding the prototype platform, we should implement the dynamic fronthaul split, perform the control of C-RAN and the control of the converged Ethernet fronthaul network jointly. In order to better show the benefits of software defined centralized network control regarding scalability and flexibility, we should extend the prototype with more BBUs and UEs.

5.2.2 New perspectives

Several open research topics related to the C-RAN architecture should be investigated based on the work realized in this thesis. One of them is to study multi-cell coordination for transmissions in the millimeter-wave (mmwave) band that is planned to be used for a part of 5G communications. Indeed, large unexploited frequency bands are available in that part of the wireless spectrum, however, propagation conditions are very different compared to the sub-6 GHz bans currently used for mobile communications. Despite more important fading, on short transmission ranges high throughput can be achieved in directional transmissions, i.e. using beamforming. When distributed access points connected to a central unit ensure the coverage of an area, serving UEs in a coordinated manner may allow to improve the overall network performance. With directional transmissions interference is not really significant, however managing user association to realize load balancing between the access points can increase spectral efficiency. Furthermore, when high capacity line-of-sight connection is available between the RRHs and the CO, by efficient coordination transmissions over the RAN air interface and in-band mmwave fronthauling should be possible, making the exploitation of the spectrum even more efficient.

A different technique that attracts currently a lot of attention is concerns full duplex transceivers. Though practical considerations regarding self-interference and implementation complexity prevents its generalized usage, it has been shown that for relay systems it can be useful. In a C-RAN deployment with wireless fronthaul, RRHs that can simultaneously receive and transmit would increase significantly the performance of the network. In addition, spatially separating uplink and downlink APs, enabled by C-RAN, can completely remove self-interference between transmit and receive antennas at the RRH. These new research topics fitting to the architecture with partially centralized base-band and SDN-based control would be challenging, but interesting extensions of the contributions presented in this thesis.

Appendix A

Derivative of the Stieltjes transform used in the approximation of ergodic sum rate

We provide here the derivation that allowed to obtain the formulation (4.34) of the derivative of the Stieltjes transform (4.29) that we have used in the approximation by deterministic equivalent of the group sum rate (4.27). Let us first define the partial derivative with respect to the fronthaul rate for one group and one antenna of the vector $\boldsymbol{\delta}(-\sigma_z^2, \boldsymbol{c}^{(l)})$ gathering the values of $\delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})$ obtained by solving (4.29) for every user $i \in \{1, ..., s_l\}$.

$$\frac{\partial \boldsymbol{\delta}(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}} = \begin{pmatrix} \frac{\partial \delta_1(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}} \\ \frac{\partial \delta_2(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}} \\ \vdots \\ \frac{\partial \delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}} \\ \vdots \\ \frac{\partial \delta_{s_l}(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}} \end{pmatrix} \tag{A.1}$$

Then, one element of this vector can be computed as

$$\begin{aligned} &\frac{\partial \delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})}{\partial c_{lk}} = \frac{1}{K} \mathrm{tr} \left[\hat{\boldsymbol{R}}_i \frac{\partial (\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1})}{\partial c_{lk}} \right] \\ &\text{with } \boldsymbol{T}_K(\boldsymbol{c}^{(l)}) = \frac{1}{K} \sum_{i=1}^{s_l} \frac{\hat{\boldsymbol{R}}_i}{1 + \delta_i(-\sigma_z^2, \boldsymbol{c}^{(l)})} + \boldsymbol{B}_K(\boldsymbol{c}^{(l)}) - \sigma_z^2 \boldsymbol{I}_k. \end{aligned} \tag{A.2}$$

With the derivative chain rule we get

$$\begin{aligned}\frac{\partial(\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1})}{\partial c_{lk}} &= -\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1}\frac{\partial \boldsymbol{T}_K(\boldsymbol{c}^{(l)}}{\partial c_{lk}}\boldsymbol{T}_K(\boldsymbol{c}^{(l)^{-1}} \\ \text{where } \frac{\partial \boldsymbol{T}_K(\boldsymbol{c}^{(l)})}{\partial c_{lk}} &= -\frac{1}{K}\sum_{n=1}^{s_l}\frac{\hat{\boldsymbol{R}}_n\delta_n' k(-\sigma_z^2,\boldsymbol{c}^{(l)})}{(1+\delta_n(-\sigma_z^2,\boldsymbol{c}^{(l)}))^2} + \underset{j\in\{1,\dots,K\}}{\text{diag}}\{d_{jj}\}.\end{aligned} \tag{A.3}$$

By decomposing the trace of the sum of matrices to the sum of their traces we obtain

$$\begin{aligned}\frac{\partial\delta_i(-\sigma_z^2,\boldsymbol{c}^{(l)})}{\partial c_{lk}} = \frac{1}{K^2}\sum_{n=1}^{s_l}\text{tr}\left[\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1}\hat{\boldsymbol{R}}_i\frac{\hat{\boldsymbol{R}}_n\delta_{nk}'(-\sigma_z^2,\boldsymbol{c}^{(l)})}{(1+\delta_n(-\sigma_z^2,\boldsymbol{c}^{(l)}))^2}\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1}\right] \\ + \frac{1}{K}\text{tr}\left[\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1}\hat{\boldsymbol{R}}_i\underset{j\in\{1,\dots,K\}}{\text{diag}}\{d_{jj}\}\boldsymbol{T}_K(\boldsymbol{c}^{(l)})^{-1}\right]\end{aligned} \tag{A.4}$$

that gives in a matrix formulation the expression in (4.34), that is similar to the derivative provided in [105], Theorem 21.

Appendix B

List of acronyms

3GPP	3rd Generation Partnership Project
ACK	acknowledgment message
ADC	Analog-Digital Conversion
AP	Access Point
API	Application Programming Interface
AWGN	Additive White Gaussian Noise
BB	base-band
BBU	Base-band Unit
BCJR	Bahl, Cocke, Jelinek, Raviv (its inventors)
BLER	Block Error Rate
BPSK	Binary Phase Shift Key
BS	Base Station
CAPEX	Capital Expenditure
CB	Coordinated Beamforming
CoMP	Coordinated Multi-Point
CO	Central Office
CPRI	Common Public Radio Interface
CP	Cyclic Prefix
C-RAN	Cloud Radio Access Network
CRC	Cyclic Redundancy Check
CS	Coordinated Scheduling
CSI	Channel State Information
CU	Central Unit
DAC	Digital-Analog Conversion
DCI	Downlink Control Indication
DFT	Discrete Fourier Transform
DL	Downlink
DMRS	De-Modulation Reference Signal

DPB	Dynamic Point Blanking
eICIC	Enhanced ICIC
embb	Enhanced Mobile BroadBand
eNB	eNodeB
EPC	Evolved Packet Core
ETSI	European Telecommunication Standards Institute
FDD	Frequency Division Duplex
FEU	Front-End Unit
FFT	Fast Fourier Transform
FPGA	Field-Programmable Gate Array
Gbps	Giga-bit per second
GPP	General Purpose Processor
GPU	Graphical Processing Unit
HARQ	Hybrid Automatic Repeat reQuest
ICI	Inter-cell Interference
ICIC	Inter-cell Interference Cancellation
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
I/Q	In-phase/Quadrature
IRC	Interference Rejection Combining
IT	Information Technology
JR	Joint Reception
JT	Joint Transmission
KPI	Key Performance Indicator
LTE	Long Term Evolution
LTE-A	LTE Advanced
LoS	Line-of-Sight
MAC	Medium Access Control Layer
Mbps	Mega-bit per second
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MMTC	Massive Machine Type Communications
MNO	Mobile Network Operator
MU-MIMO	Multi-User MIMO
MRC	Maximum Ratio Combining
NACK	non-acknowledgment message
NB	NorthBound
NFV	Network Function Virtualization

NGFI	Next Generation Fronthaul Interface
NLoS	Non-Line-of-Sight
NOMA	Non-Orthogonal Multiple Access
OAI	OpenAirInterface
OCC	Orthogonal Cover Code
OFDM	Orthogonal Frequency Division Multiplexing
OPEX	Operating Expenditure
ONOS	Open Network Operating System
OS	Operating System
PAPR	Peak-to-average power ratio
PHY	Physical Layer
PRB	Physical Resource Block
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RAN	Radio Access Network
RF	Radio Functions
RRH	Remote Radio Head
RTT	Round Trip Time
RU	Remote Unit
SB	SouthBound
SCN	Small Cell Network
SC-FDMA	Single-Carrier Frequency Division Multiple Access
SC-OFDM	Single-Carrier Orthogonal Frequency Division Multiplexing
SCTP	Stream Control Transmission Protocol
SDN	Software Defined Networking
SER	Symbol Error Rate
SIC	Successive Interference Cancellation
SISO	Single Input Single Output
SNR	Signal-to-noise-ratio
SQL	Stuctured Query Language
STBC	Space-Time Block Code
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TDD	Time Division Duplex
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
URLLC	Ultra-Reliable and Low-Latency Communications
USRP	Universal Software Radio Peripheral
VM	Virtual Machine

VNF	Virtual Network Function
WDM	Wavelength-Division Multiplexing
XML	Extensible Markup Language
ZF	Zero Forcing

Appendix C

Résumé en français

1 Evolution des réseaux mobiles

Les réseaux de communication radio pour terminaux mobiles sont en constante évolution pour répondre à la demande accrue de transmissions de données. De nouvelles applications basées sur une connexion réseau stable et de grande capacité apparaissent tous les jours, ainsi, les opérateurs de réseaux mobiles veulent attirer et fidéliser leurs clients en offrant un service fiable et de bonne qualité, mais souhaitent également minimiser les coûts d'installation et d'opération de leurs infrastructures. Pour leur offrir un compromis convenable entre ces deux aspects, de nouvelles solutions technologiques et fonctionnelles s'imposent.

1.1 De la technologie LTE vers la 5G

La quatrième génération des réseaux mobiles, dont la technologie principale, LTE (Long Term Evolution) est déployée par les opérateurs partout dans le monde a été conçue pour répondre aux besoins de connectivité des années 2010, notamment l'accés aux contenus mulitmédia en différé. Elle utilise une tehcnique de transmission radio multi-porteuse connue de longue date et répandue dans divers domaines des communications filaires ou sans fil, l'OFDM (Orthogonal Frequency Division Multiplexing). Cette technique permet une transmission simultanée sur des sous-porteuses indépendentes, chacune ayant une fréquence différente, ces sous-porteuses espacées de 15 kHz en LTE remplissent la largeur de bande disponible qui peut aller jusqu'à 20 MHz. Avec un multiplexage sur les différentes fréquences par transformée de Fourier, l'interférence entre les sous-porteuses est évitée, de même, la modulation quadratiques en amplitude (QAM) - pouvant aller jusqu'à un ordre de 256 dans certains cas avec des conditions de canal favorables - assure l'absence d'interférence entre les symboles. De plus, le codage turbo à rendement variable permet de rendre la transmission fiable sans pour autant sacrifier la performance. Tous ces élements intégrés dans la

couche physique LTE permettent d'atteindre un débit de quelques centaines de mégaoctets par chaîne de transmission, dont la normalisation permet d'avoir 8 au plus.

Malgré leur capacité importante, les réseaux 4G ne pourraient pas faire fonctionner des terminaux avec réalité augmentée ou réalitée virtuelle, nécessitant un débit encore plus importante, ni des véhicules connectés qui exigent une fiablilité extrême et une latence de transmission très réduite. Ces applications - entre autres - sont visées par les réseaux de cinquième génération (5G). Le spectrum des fréquences radio étant saturé, il n'est pas possible d'augmenter significativement la largeur de bande pour atteindre une meilleure capacité par rapport à LTE. Il est donc nécessaire de déployer des points d'accés radio avec une plus grande densitée géographique. De plus, l'utilisation des bandes de fréquences des ondes millimétriques ayant un évanouissment plus marqué va également dans le sens des déploiements denses.

1.2 Vers une architecture centralisée *Cloud*

Afin de rendre les réseaux plus efficaces, simplifier leur maintenance et leur évolution, les unités de calculs ont été séparées des éléments de transmission (antennes et convertisseurs analogique-numérique). Dans une seconde étape, ces unités de calcul dédiées aux traitements de signal en bande de base (BBU) ont été rassemblées dans un même emplacement, puis les traitements de signal ont été attribués à un même serveur composé de plusieurs éléments de calcul dédié à des fonctions spécifiques. Pour aller encore plus loin, dans une architecture *Cloud RAN* l'évolution des hardwares permettrait de réaliser ces calculs avec des logiciels en environnement virtualisé, exécutés sur un plateforme générique, donc moins coûteux.

Les avantages de *Cloud RAN*consistent non seulement à faire des économies dans l'installation et l'opération, mais aussi augmenter la performance des transmissions sans fil grâce au traitement conjoint des signaux de plusieurs cellules radio. En effet, le partage des unités de calcul permet de distribuer les donnés attribuées aux différentes cellules entre les fonctions de celles-ci. Ce traitement de signal conjoint peut être appliquée en voie déscendante et en voie montante afin d'améliorer les transmissions en particulier dans les régions des bords de cellule où l'interférence peut être plus importante. L'implémentation, dans le cas de la voie montante, nécessite de surmonter des difficultés technologiques, car il demande d'une part le traitement conjoint des signaux dans la couche physique et d'autre part la coordination de l'allocation des resources entre les cellules réalisés dans la couche de contrôle. De plus, l'optimisation de cette allocation et du transfer des données entre les sites radio et le serveur central

permet d'augmenter davantage l'efficacité bout-à-bout des transmissions. Ces points sont abordés dans les différentes contributions de cette thèse.

2 Réseaux définis par logiciel: clé pour techniques multi-cellulaires

Nouveaux paradigmes de design des logiciels de télécommunication Dans l'architecture de réseaux d'accés radio cloud (C-RAN), afin de rendre l'exécution des logiciels plus efficace, des techniques répandues dans l'industrie informatique sont adoptées pour les applications de communications radio. L'architecture software basée sur des micro-services isolés entre eux et dédiés à une fonction donnée, ainsi que le concept des réseaux définis par logiciel (Software Defined Networking - SDN) permettant une séparation entre le plan des données et le plan de contrôle et une abstraction des données de contrôle. Toutefois, l'application de SDN au réseau d'accès radio demande certaines adaptations par rapports aux réseaux fixes, dû aux caractéristiques différentes des transmissions sans fil.

Implémentation de la coordination entre cellules avec SDN Afin de montrer la faisabilité de l'utilisation d'un contrôlleur SDN pour la coordiantion des transmissions entre les cellules pour lesquelles les traitements en bande de base sont exécutés dans le même serveur C-RAN. Pour des techniques multi-cellulaires sur la voie montante et la voie déscendante, nous avons déployé des applications coordinant l'allocations des ressources radio entre les utilisateurs des cellules voisines. Ces applications ont été installées à l'interface nord du contrôlleur SDN, ainsi, nous avons pu montrer que le temps de confuguration dans cette architecture était compatible avec une utilisation à faible latence.

3 Amélioration des communications multi-utilisateurs dans Cloud RAN

Dans cette thèse, nous avons étudié différentes possibilités pour améliorer la performance des transmissions multi-utilisateurs dans C-RAN. D'une part, sur l'interface sans fil entre les terminaux mobiles et les têtes radio multi-antennes, on choisit un schéma pour l'allocation des ressources radio aux terminaux. D'autre part, sur l'interface dit de fronthaul entre les têtes radio et le serveur centralisé, nous allouons le débit des liens de fronthaul aux signaux reçus par chaque antenne, en modifiant le taux de quantisation appliqué.

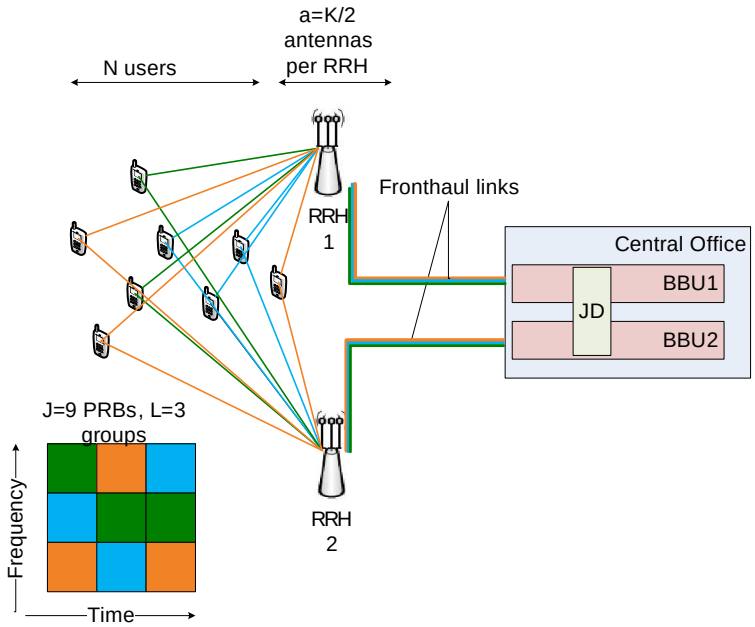


Figure C.1 – Modèle de système pour transmissions sur la voie montante dans l'architecture C-RAN

3.1 Multi-accès à orthogonalité partielle

Pour un meilleur débit réalisé par la transmission en air sur la voie montante, dans un scénario tenant compte des contraintes pratiques, notamment de la compléxité des traitements de signal à réaliser par le récepteur, nous avons adapté un schéma d'allocation de resource multi-utilisateur au contexte mutli-cellulaire en C-RAN. Nous proposons de répartir les utilisateurs en bordure de cellules - donc impliqués dans l'interférence inter-cellulaire - en groupes de quelques terminaux. Chaque groupe dispose de ses propres ressources radio, mais les utilisateurs dans un même groupe ont accès à la totalité des ressources du groupe et transmettent sur ces ressources simultanément. De cette maniére lorsqu'une réception multi-utilisateur est appliqée, l'efficacité spectrale augmente significativement. Pour les cas où il y a une forte contrainte concernant le débit, nous avons proposé une algorithme qui exploite les statistiques de canal afin de distribuer les utilisateurs dans les groupes itérativement en maximisant à chaque itération la somme d'une borne supérieur des efficacités spectrales des groupes.

3.2 Allocation de débit sur les liens fronthaul

Dans le modéle de système (voir Figure C.1) que nous avons défini pour les transmissions sur la voie montante dans l'architecture C-RAN selon la stratégie de multi-accès à orthogonalité partielle décrite ci-dessus, chaque antenne des têtes

radio reçoit le les signaux de tous les groupes d'utilisateurs indépendamment. Cela permet de choisir un niveau de quantisation différent pour chacun de ces signaux reçus pour le transfer fronthaul, permettant d'adapter le débit aux caractéristiques de la transmission. Nous avons étudié l'allocation du débit fronthaul par ce moyen dans deux cas: avec l'hypothèse que nous connaissons les réalisations de canal instantannées, et dans le cas où nous avons accès uniquement aux statistiques des estimations de canal et des erreurs d'estimation correspondantes. Dans les deux cas, nous avons utilisé pour l'optimisation une fonction objective qui inclut le bénéfice que la transmissions apporte à l'opérateur du réseau en terme de débit bout-à-bout et le coût de la transmission sur le lien fronthaul en fonction du taux de quantisation appliqué.

Optimisation avec connaissance de canal parfaite Lorsque nous considérons le cas idéal avec connaissance de canal parfaite, nous pouvons adapter le taux de quantisation de signaux reçus par chaque antenne et pour chaque group d'utilisateurs, afin de satisfaire une éventuelle contrainte sur la capacité de chacun des liens de fronthaul et maximiser le gain net instantanné de l'opérateur du réseau sur la transmission. En évaluant les résultats que nous obtenons par optimisation convexe, nous avons observé un amélioration du gain net d'environ 10 % par rapport à l'allocation uniforme du débit total disponible lorsque la capacité des liens de fronthaul est convenable. De plus, par l'optimisation nous pouvons éviter d'utiliser un débit trop élevé pour la qualité de canal - notamment en terme de rapport signal-à-bruit - à un instant donné.

Optimisation avec estimation de canal imparfaite et caractéristiques statistiques Suivant un scénario plus réaliste, nous n'avons pas de moyen pratique de connaître les réalisations de canal, mais seulement leur estimations qui peuvent comporter une part d'erreur, et celle-ci peut être caractérisée par ses statistiques seulement. Dans cette seconde approche, nous exploitons uniquement des données statistiques liées aux estimations de canal, ce qui signifie également que nous pouvons optimiser les paramètres de transmission sur un échelle de temps à plus long terme, en visant une meilleure performance en moyenne. Le problème d'optimisation que nous solvons est similaire à celui défini dans le cas avec connaissance de canal parfaite, mais intègre l'efficacitée spectrale moyenne. En réalisant cette optimisation, nous pouvons donc choisir sur des périodes de temps où le volume du trafic varie peu un taux de quantisation donnant performance très proche de l'optimum. De plus, selon le modèle de coût des transfers fronthaul, nous pouvons estimer en avance le bénéfice des transmissions dans différentes configurations, en particulier pour différents nombres d'antennes, afin de paramétrer les déploiements pour maximiser le gain net obtenu.

4 Conclusion

Les différentes contributions de cette thèse visaient à définir des améliorations pour une opération performante du réseau d'accès radio dans les déploiements centralisés *cloud*. Les aspects architecturaux et fonctionnels ont été analysés pour le plan de données et le plan de contrôle, dont la séparation est un concept de base pour les réseaux futurs. Ainsi, nous avons pu proposer et démontrer - soit par une réalisation pratique, soit par une évaluation par simulations suivant un modèle théorique - des solutions innovantes pour les segments clés des réseaux 5G couplés avec C-RAN.

Bibliography

[1] P. Marsch and G. P. Fettweis, *Coordinated Multi-Point in Mobile Communications: from theory to practice*. Cambridge University Press, 2011.

[2] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development*, vol. 54, pp. 4:1–4:12, January 2010.

[3] C. M. R. Institute, "C-RAN the road towards green RAN," October 2011.

[4] P. Marsch and G. Fettweis, "Uplink CoMP under a constrained backhaul and imperfect channel knowledge," *IEEE Transactions on Wireless Communications*, vol. 10, pp. 1730–1742, June 2011.

[5] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Communications Letters*, vol. 3, pp. 365–368, Aug 2014.

[6] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint base station selection and distributed compression for cloud radio access networks," in *2012 IEEE Globecom Workshops*, pp. 1134–1138, Dec 2012.

[7] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1295–1307, June 2014.

[8] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Performance evaluation of multiterminal backhaul compression for cloud radio access networks," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, March 2014.

[9] H. Guan, T. Kolding, and P. Merz, "Discovery of cloud-RAN," in *Cloud-RAN Workshop*, 2010.

[10] M. Weldon, *The Future X Network: A Bell Labs Perspective*. CRC Press LLC, 2016.

[11] D. Boviz and S. Yang, "Fronthaul allocation for uplink multi-user transmissions in cloud RAN with statistical CSI,"

[12] D. Boviz, C. S. Chen, and S. Yang, "Effective design of multi-user reception and fronthaul rate allocation in 5G cloud RAN," *Journal on Selected Areas of Communications, special issue on Deployment and Performance Challenges for 5G*, 2017.

[13] M. Artuso, D. Boviz, A. Checko et al., "Enhancing LTE with cloud-RAN and load-controlled parasitic antenna arrays," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 183–191, 2016.

[14] D. Boviz, C. S. Chen, and S. Yang, "Cost-aware fronthaul rate allocation to maximize benefit of multi-user reception in C-RAN," in *IEEE Wireless Communications and Networking Conference (WCNC)*, (San Francisco, USA), Mar. 2017.

[15] D. Boviz and Y. El Mghazli, "Fronthaul for 5G: low bit-rate design enabling joint transmission and reception," in *IEEE Global Telecommunications Conference (Globecom), 5G RAN Design Workshop*, December 2016.

[16] D. Boviz, N. Abbas, G. Aravinthan, C. S. Chen, and M. A. Dridi, "Multi-cell coordination in cloud RAN: architecture and optimization," in *International Conference on Wireless Networks and Mobile Communications (WINCOM)*, (Fez, Morocco), pp. 271–277, Oct. 2016.

[17] D. Boviz and S. Yang, "Optimal fronthaul capacity allocation and training for joint uplink receiver in C-RAN," in *European Wireless (EW2016)*, (Oulu, Finland), pp. 415–420, May 2016.

[18] D. Boviz, G. Aravinthan, C. S. Chen, and L. Roullet, "Physical layer split for user selective uplink joint reception in SDN enabled Cloud-RAN," in *2016 Australian Communications Theory Workshop (AusCTW)*, (Melbourne, Australia), pp. 83–88, Jan. 2016.

[19] D. Boviz, M. A. Dridi, N. Abbas, and G. Aravinthan, "Multi-cell coordination in cloud RAN enabled by SDN (demonstration)," in *IEEE Wireless Communications and Networking Conference - Student Outreach Program*, (San Francisco, USA), Mar. 2017.

[20] D. Boviz, G. Aravinthan, N. Trabelsi, and L. Roullet, "C-RAN fronthaul enhancements using software defined networking," in *Cloud- and fog-based PHY communications in 5G - GDR ISIS Workshop*, (Paris, France), Nov. 2015.

[21] J. Segel and M. Weldon, "Lightradio portfolio: White paper 1," 2011.

[22] S. Scholz and H. Grob-Lipski, "Reallocation strategies for user processing tasks in future cloud-RAN architectures," in *IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–6, Sept 2016.

[23] N. Alliance, "Further study on critical C-RAN technologies," 2015.

[24] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings," *Journal of Communications and Networks*, vol. 18, pp. 162–172, April 2016.

[25] K. Lee and J. Eidson, "IEEE-1588 standard for a precision clock synchronization protocol for networked measurement and control systems," in *In 34 th Annual Precise Time and Time Interval (PTTI) Meeting*, pp. 98–105, 2002.

[26] Nokia, "5G use cases and requirements," 2015.

[27] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[28] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009.

[29] 3GPP TS 36.212, "Multiplexing and channel coding,"

[30] Y. Léost, M. Abdi, R. Richter, and M. Jeschke, "Interference rejection combining in LTE networks," *Bell Labs Technical Journal*, vol. 17, pp. 25–49, June 2012.

[31] 3GPP TS 36.211, "Physical channels and modulation,"

[32] A. Paulraj and T. Kailath, "Increasing capacity in wireless broadcast systems using distributed transmission/directional reception (dtdr)," Sept. 6 1994. US Patent 5,345,599.

[33] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communications," in *Global Telecommunications Conference*, vol. 3, pp. 1809–1814, IEEE, 1996.

[34] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1451–1458, Oct 1998.

[35] G. D. Golden, C. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electronics Letters*, vol. 35, pp. 14–16, Jan 1999.

[36] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 742–758, Oct 2014.

[37] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Transactions on Information Theory*, vol. 60, pp. 7112–7139, Nov 2014.

[38] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, pp. 3590–3600, November 2010.

[39] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, pp. 74–80, February 2014.

[40] J. Vihriala, N. Ermolova, E. Lahetkangas, O. Tirkkonen, and K. Pajukoski, "On the waveforms for 5G mobile broadband communications," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2015.

[41] B. Farhang-Boroujeny and H. Moradi, "OFDM inspired waveforms for 5G," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 2474–2492, Fourthquarter 2016.

[42] X. Zhang, L. Chen, J. Qiu, and J. Abdoli, "On the waveform for 5G," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 74–80, 2016.

[43] C. Ibars, U. Kumar, H. Niu, H. Jung, and S. Pawar, "A comparison of waveform candidates for 5G millimeter wave systems," in *Signals, Systems and Computers, 2015 49th Asilomar Conference on*, pp. 1747–1751, IEEE, 2015.

[44] Nokia, Alcatel-Lucent Shanghai Bell, "Waveform proposal for carrier frequencies beyond 40 GHz," in *3GPP TSG-RAN WG1-86bis*, 2016.

[45] CMCC, "Discussion and evaluation of UL waveforms," in *3GPP TSG-RAN WG1-86bis*, 2016.

[46] AT&T, "Summary of link level analysis of candidate waveforms for NR," in *3GPP TSG-RAN WG1-86bis*, 2016.

[47] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, 2010.

[48] 3GPP TS 36.420, "X2 general aspects and principles,"

[49] L. Liu and R. Zhang, "Optimized uplink transmission in multi-antenna C-RAN with spatial compression and forward," *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5083–5095, 2015.

[50] K. Balachandran, J. H. Kang, M. K. Karakayali, and K. M. Rege, "Base station cooperation with non-ideal backhaul and non-full buffer traffic," in *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*, pp. 1–5, IEEE, 2014.

[51] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[52] S.-H. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 69–79, 2014.

[53] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE trans. signal process*, vol. 64, no. 16, pp. 4138–4151, 2016.

[54] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Technical Journal*, vol. 18, pp. 105–128, June 2013.

[55] China Mobile Research Institute, Alcatel-Lucent, Nokia Networks, ZTE Corporation, Broadcom Corporation, Intel China Research Center, "White paper of NGFI (Next Generation Fronthaul Interface)," october 2015.

[56] S. C. Forum, "Small cell virtualization functional splits and use cases," June 2015.

[57] S. Khalili and O. Simeone, "Uplink HARQ for distributed and cloud RAN via separation of control and data planes," *arXiv preprint arXiv:1508.06570*, 2015.

[58] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, pp. 2307–2359, May 2010.

[59] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, pp. 4232–4265, July 2014.

[60] N. Nikaein, "Processing radio access network functions in the Cloud: Critical issues and modeling," in *MCS 2015, 6th International Workshop on Mobile Cloud Computing and Services, in conjunction with MOBICOM*, (Paris, FRANCE), 09 2015.

[61] R. McDougall and J. Anderson, "Virtualization performance: Perspectives and challenges ahead," *SIGOPS Oper. Syst. Rev.*, vol. 44, pp. 40–56, Dec. 2010.

[62] A. Gopalasingham, D.-G. Herculea, C. S. Chen, and L. Roullet, "Virtualization of radio access network by virtual machine and Docker: Practice and performance analysis," in *IFIP/IEEE International Symposium on Integrated Network Management*, (Lisbon, Portugal), 05 2017.

[63] "What is docker?." `https://www.docker.com/whatisdocker/`. [Online].

[64] "Lxc – linux containers." `https://linuxcontainers.org/`. [Online].

[65] ETSI NFV ISG, "Network Function Virtualisation Use Cases," 2013.

[66] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network store: Exploring slicing in future 5G networks," in *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture*, MobiArch '15, pp. 8–13, 2015.

[67] "Docker swarm." `https://www.docker.com/products/docker-swarm`. [Online].

[68] "Kubernetes." `https://kubernetes.io/`. [Online].

[69] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. D. Girolamo, and G. Giuliani, "RAN as a service: Challenges

of designing a flexible ran architecture in a cloud-based heterogeneous mobile network," in *2013 Future Network Mobile Summit*, pp. 1–8, July 2013.

[70] A. Outtagarts, L. Roullet, B. Mongazon-Cazavet, and G. Aravinthan, "When IT meets telco: RAN as a service," in *IEEE/ACM UCC*, pp. 422–423, 2015.

[71] S. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On scalability of software-defined networking," *IEEE Communications Magazine*, vol. 51, pp. 136–141, Feb. 2013.

[72] M. Mendonca, K. Obraczka, and T. Turletti, "The case for software-defined networking in heterogeneous networked environments," in *ACM Conference on CoNEXT Student Workshop*, pp. 59–60, 2012.

[73] A. R. Curtis, J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, and S. Banerjee, "DevoFlow: Scaling flow management for high-performance networks," *SIGCOMM Comput. Commun. Rev.*, vol. 41, pp. 254–265, Aug. 2011.

[74] K.-K. Yap, M. Kobayashi, R. Sherwood, T.-Y. Huang, M. Chan, N. Handigol, and N. McKeown, "OpenRoads: Empowering research in mobile networks," *SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 125–126, Jan. 2010.

[75] M. Bansal, J. Mehlman, S. Katti, and P. Levis, "OpenRadio: A programmable wireless dataplane," in *First Workshop on Hot Topics in Software Defined Networks*, pp. 109–114, 2012.

[76] A. Gopalasingham, L. Roullet, N. Trabelsi, C. S. Chen, A. Hebbar, and E. Bizouarn, "Generalized software defined network platform for radio access networks," in *IEEE Consumer Communications and Networking Conference*, 2016.

[77] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A flexible platform for 5G research," *SIGCOMM Comput. Commun. Rev.*, vol. 44, pp. 33–38, Oct. 2014.

[78] "OpenAirInterface." `https://openairinterface.org/`. [Online].

[79] L. Ong and J. Yoakum, "RFC 3286, An introduction to the stream control transmission protocol (SCTP)." http://www.faqs.org/rfcs/rfc3286.html, 2002.

[80] A. Lindem, L. Berger, D. Bogdanovic, and C. Hopps, "Network Device YANG Organizational Models," Aug. 2016. Work in Progress.

[81] D. Chu, "Polyphase codes with good periodic correlation properties (corresp.)," *IEEE Transactions on Information Theory*, vol. 18, pp. 531–532, Jul 1972.

[82] S. Wang, Y. Li, and J. Wang, "Low-complexity multiuser detection for uplink large-scale MIMO," in *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 224–229, IEEE, 2014.

[83] L. Fang, L. Xu, and D. D. Huang, "Low complexity iterative MMSE-PIC detection for medium-size massive MIMO," *IEEE Wireless Communications Letters*, vol. 5, pp. 108–111, Feb 2016.

[84] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE 77th Vehicular Technology Conference (VTC-Spring)*, pp. 1–5, June 2013.

[85] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, pp. 74–81, September 2015.

[86] B. Kim, W. Chung, S. Lim, S. Suh, J. Kwun, S. Choi, and D. Hong, "Uplink NOMA with multi-antenna," in *IEEE 81st Vehicular Technology Conference (VTC-Spring)*, pp. 1–5, May 2015.

[87] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *arXiv preprint arXiv:1611.01607*, 2016.

[88] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, 2000.

[89] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *11th International Symposium on Wireless Communications Systems (ISWCS)*, pp. 781–785, Aug 2014.

[90] Nokia, "5G use cases and requirements," 2015.

[91] 3GPP, "UL virtual MIMO system level performance evaluation for E-UTRA," TSG-RAN1 R1.051422, 3rd Generation Partnership Project (3GPP), Nov. 2005.

[92] X. Chen, H. Hu, H. Wang, H. h. Chen, and M. Guizani, "Double proportional fair user pairing algorithm for uplink virtual MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 7, pp. 2425–2429, July 2008.

[93] J. Fan, D. Lee, G. Y. Li, and L. Li, "Multiuser scheduling and pairing with interference mitigation for LTE uplink cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 2, pp. 481–492, 2015.

[94] S. Dhakal and J. Kim, "Statistical analysis of user-pairing algorithms in virtual MIMO systems," in *Wireless Telecommunications Symposium (WTS), 2010*, pp. 1–5, April 2010.

[95] E. Viterbo and A. Hottinen, "Optimal user pairing for multiuser MIMO," in *IEEE 10th International Symposium on Spread Spectrum Techniques and Applications*, pp. 242–246, 2008.

[96] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[97] E. Wireless, "Economics of Backhaul." `http://www.exaltcom.com/Economics-of-Backhaul.aspx`, 2016. [Online].

[98] E. Bjornson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Processing Magazine*, vol. 31, pp. 14–23, Nov 2014.

[99] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Transactions on Information Theory*, vol. 49, pp. 951–963, April 2003.

[100] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[101] J.-K. Kang, O. Simeone, J. Kang, and S. Shamai, "Joint signal and channel state information compression for uplink network MIMO systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 875–878, IEEE, 2013.

[102] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 160–171, February 2013.

[103] S. Wagner, R. Couillet, M. Debbah, and D. T. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4509–4537, 2012.

[104] J. Hoydis, M. Kobayashi, and M. Debbah, "Optimal channel training in uplink network MIMO systems," *Signal Processing, IEEE Transactions on*, vol. 59, pp. 2824–2833, June 2011.

[105] J. Hoydis, *Random matrix theory for advanced communication systems.* Theses, Supélec, Apr. 2012.

Titre : Communications multi-utilisateurs dans les réseaux d'accès radio centralisés: architecture, coordination et optimisation

Mots clefs : multi-antennes, réseaux d'accès radio, traitement de signal, fronthaul, couche physique

Résumé : Dans les réseaux mobiles du future, un déploiement plus dense des points d'accés radio est prévu pour satisfaire la demande accrue de débit, mais les terminaux utilisateurs peuvent être affectés par une interférence inter-cellulaire plus forte. Par chance, la centralisation des traitements de signal en bande de base dans l'achitecture Cloud RAN (C-RAN) offre la possibilité de la coordination et du traitement conjoint de plusieurs cellules. Pour réellement permettre de déployer ces techniques, une étude bout-à-bout du C-RAN est nécessaire selon plusieurs aspects, notamment l'architecture fonctionnelle, la stratégie de coordination, l'implémentation du traitement de signal multi-utilisateur et les optimisations possibles pour un fonctionnement plus efficace.
Dans cette thèse, nous proposons en premier une architecture qui définit le placement des fonctions du traitement en bande de base entre les unités distribuées et le serveur central. Le but de ce design est de permettre la réalisation des fonctions multi-utilisateurs en transmettant avec la moins de débit possible sur les liens de fronthaul reliant les différentes entités. Dans un second temps, nous présentons comment il est possible de coordiner les différentes cellules servies par le C-RAN en utilisant le concept de réseaux définis par logiciels adapté pour les réseaux d'accès radio. Nous avons mis en place un prototype démontrant la faisabilité de la méthode de contrôle proposée. Finalement, nous étudions l'allocation adaptative du débit sur les liens de fronthaul transportant les symboles numériques quantifiés des utilisateurs en besoin de traitement multi-cellulaire sur la voie montante pour exploiter l'interférence entre eux. Nous proposons un modèle d'optimisation qui inclut le coût des transmissions fronthaul pour maximiser ainsi le gain obtenu par l'opérateur du réseau où la communication multi-utilisateur a lieu. Nous réalisons l'optimisation pour différents modèles de coût et en utilisants deux types de données: d'abord les estimations de canal supposées parfaites et disponibles en temps réel, puis seulement les statistiques du canal. Nous montrons que la méthode d'optimisation proposée permet d'exploiter plus efficacement les liens de fronthaul dans l'architecture précedemment définie.

Title : Multi-user Communication in Cloud Radio Access Network: Architecture, Coordination and Optimization

Keywords : Multiple input multiple output, radio access network, signal processing, fronthaul, physical layer

Abstract : In future mobile networks denser deployment of radio access points is planned to satisfy demand of higher throughput, but an increased number of mobile users can suffer from inter-cell interference. Fortunately, the centralization of base-band processing offered by Cloud Radio Access Network (C-RAN) architecture enables coordination and joint physical layer processing between cells. To make practical deployment of these techniques possible, we have to study C-RAN in an end-to-end view regarding several aspects: the functional architecture of a deployment, the multi-cell coordination strategy, the implementation of multi-user signal processing and possibilities for optimization to increase operational efficiency.
In this thesis, first, we propose an architecture defining the placement of base-band processing functions between the distributed remote units and the central processing unit. The aim of this design is to enable multi-cell processing both on the uplink and the downlink while requiring low data rate between the involved entities. Secondly, we study how low latency coordination can be realized inside the central unit using software defined networking adapted to radio access networks. Our demonstration through a real-time prototype deployment shows the feasibility of the proposed control framework. Finally, we investigate adaptive allocation of fronthaul rate that is used for transferring quantized base-band symbols for users participating in uplink multi-cell reception in order to exploit interference between them. We propose an optimization model that includes the cost of fronthaul tranmissions and aims to maximize the gain of network operators from multi-user transmissions in C-RAN. We solve the optimization problem for different fronthaul pricing models, in a scenario where real-time and accurate channel estimates are available and in another where only channel statistics are exploited. Using our method - fitting in the architecture that we have defined - cost efficiency of fronthaul usage can be significantly improved.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France