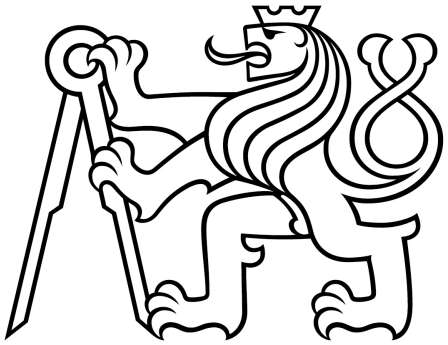
Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Telecommunication Engineering



Hierarchical Scheduling for Cloud Radio Access Networks

Doctoral Thesis

Mohammed Elfiky, MSc

Prague - Czech Republic, February 2024

Ph.D. programme: P2612 Electrical Engineering and Information Technology
Branch of study: 2601V013 Telecommunication Engineering

Supervisor: Prof. Ing. Zdeněk Bečvář, Ph.D.
Supervisor-Specialist: Ing. Pavel Mach, Ph.D.

Declaration

I declare that the work presented in this dissertation thesis is my work, and the achieved results are the findings of my research monitored and directed by my supervisors in parallel with the contributions of other co-authors in the corresponding published papers.

In Prague, February 2024 .....

Mohammed Elfiky

Acknowledgements

" Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning." *—Albert Einstein*

This quote has always resonated with me, but as I approach the end of my PhD, its profound meaning has taken on new significance in my life. As a matter of fact, completing a PhD is a monumental achievement, and I want to raise a toast to everyone who has been a part of this incredible journey with me.

First and foremost, I would like to thank my supervisors, Prof. Zdenek Becvar and Dr. Pavel Mach, for their guidance and unlimited support. Your exceptional academic expertise, coupled with your commitment to supporting early career researchers truly make you a role model in the often competitive realm of academia. Your guidance and encouragement have been invaluable.

In addition to my supervisors, I am indebted to my exceptional lab mates, whose support has been a constant source of motivation. Our collaborative writing sessions and informal chats, whether conducted via screens during lockdowns or in person whenever circumstances allowed, provided a lifeline during the most challenging times. I am proud to say that we became more than just lab partners, but good friends. My appreciation also goes out to all of my friends for their encouragement and support throughout this part of my life.

A special thanks goes out to Mr. Pavel Bubak for his invaluable support. You have been one of my biggest cheerleaders throughout the most crucial part of this journey. Your support and belief in me have been a source of strength and motivation.

This endeavour would not have been possible without my adorable family, who believe in my abilities and support. Your endless encouragement played an integral role in my accomplishments. To my lovely mom, Nariman; my dearest dad, Albahy; my beloved brothers Ahmad and Abdulrhman; and my sweetie sister, Asmaa. Thank you so much for everything. I love you all to the moon and back! Completing a PhD feels like the natural culmination of our shared journey, and I hope you are proud of the significant role you have played in shaping my path. Kindly let me dedicate this PhD thesis to all of you.

Last but not least, I want to express my deepest gratitude to my amazing wife, Barbora. Thank you for always being my rock, accompanying me through the highs and lows of this academic journey. My moon! Your support and belief in me have always been a boundless source of strength and motivation.

Abstract

This thesis delves into scheduling approaches for 5G networks based on Cloud Radio Access Network (C-RAN) architecture, addressing limitations in existing solutions through the introduction of a novel hierarchical scheduling framework. Motivated by the critical challenge of fronthaul delay in C-RAN, the proposed solution strategically distributes radio resource scheduling tasks among centralized and distributed units. The first part of the thesis focuses on mitigating the impact of fronthaul delay, resulting in a significant enhancement in network throughput of up to 26%. The second part introduces a dynamic hierarchical scheduler for individual User Equipments (UEs), effectively adjusting scheduling periods based on Channel State Information (CSI) and fronthaul delay. Comparative analysis reveals the dynamic schedulers outperform centralized and partially distributed counterparts, achieving a remarkable 30% and 27% increase in network throughput, respectively. The final part addresses resource scheduling for delay-sensitive services, proposing a dynamic resource pre-allocation framework within the hierarchical scheduling paradigm. Simulation results demonstrate a 39% increase in goodput while minimizing transport block loss rate and mean absolute percentage error in pre-allocated resources for Hybrid Automatic Repeat reQuest (HARQ) by 38% and 57%, respectively. This thesis contributes valuable insights into overcoming scheduling challenges in C-RAN-based 5G networks, paving the way for more adaptive, efficient, and self-optimized mobile networks.

Keywords: 5G, NR, C-RAN, RRH, BBU, resource allocation, ARIMA, function splitting, hierarchical scheduling, channel prediction, HARQ.

Abstrakt

Tato práce se zabývá principy pro plánovaní přenosu dat v sítích 5G založených na "cloudové" architektuře v rámci přístupové sítě (Cloud Radio Access Network, C-RAN). Práce konkrétně řeší problémy existujících řešení spojené se zpožděním na tzv. fronthaulu a navrhuje hierarchické řešení jak strategicky alokovat rádiové prostředky na centralizovaných a distribuovaných jednotkách. První část práce se zaměřuje na zmírnění negativního dopadu zpoždění na fronthaulu, což vede k významnému zvýšení propustnosti sítě až o 26%. Druhá část pak představuje dynamické hierarchické plánování přenosu dat pro jednotlivá uživatelská zařízení, které efektivně adaptuje plánování přenosu na základě informací o stavu kanálu (Channel State Information, CSI) a na konkrétním zpožděním na fronthaulu. Komparativní analýza ukazuje, že návrh je schopný zvýšit propustnost vůči plně centralizovanému a částečně distribuovanému řešení o 30%, respektive a 27%. Závěrečná část práce se věnuje plánování prostředků pro služby citlivé na zpoždění a navrhuje metodu pro dynamické alokování prostředků založené na hierarchickém plánovači přenosu dat. Výsledky simulací ukazují až 39% nárůst tzv. goodputu při minimalizaci ztrát transportních bloků (až o 38%) a minimalizaci průměrné absolutní procentuální chyby (až o 57%) při před-alokaci rádiových prostředků pro potřeby HARQ (Hybrid Automatic Repeat reQuest). Tato práce přináší cenné poznatky při řešení problému jak plánovat přenosy v sítích C-RAN založenýCH na 5G, a tím otevírající cestu k adaptivnějším, efektivnějším a samo-optimalizujícím se mobilním sítím.

Klíčová slova: 5G, NR, C-RAN, RRH, BBU, alokace zdrojů, ARIMA, rozdělení funkcí, hierarchické plánování, predikce kanálu, HARQ.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The fifth generation New Radio (5G NR) of mobile networks is expected to support various traffic patterns and unlock numerous applications for low latency and reliable communication [1] [2] [3] [4]. The various applications imply a broad range of requirements on radio resource management [5]. However, meeting such stringent requirements is challenging, as it requires efficient radio resource management. The radio resource management encompasses many functionalities with various complexity and operating at different timescales [6]. Some of these functionalities, such as resource allocation, can be centralized (see, e.g., [7] [8]). The centralization of the radio resource management functionalities can be efficiently accomplished through the Cloud-Radio Access Network (C-RAN) architecture.

The C-RAN is a novel concept of mobile network architecture enabling centralized baseband processing to meet 5G network multiple challenges, including a reduction in cost and energy consumption [9] [10] [11]. The typical C-RAN network architecture comprises a centralized baseband unit (BBU) and multiple distributed remote radio heads (RRHs). The BBU is interconnected with different RRHs via fronthaul links [9] [12] [13]. The C-RAN architecture brings low power consumption and cost reduction benefits. At the same time, C-RAN enables an elastic network structure and large-scale coordination [14]. However, C-RAN introduces several challenges related to delay [15] and splitting of management functions between the BBU and the RRHs (also referred to as a "functional split" [16]).

The functional split is suggested mainly to relax the requirements on the fronthaul, where the control functionalities (i.e., protocol layers) can be split among the BBU and RRHs as defined by the 3^{rd} Generation Partnership Project (3GPP), see [17] [18]. The functional split determines which functions are carried out locally at the RRHs and which are performed centrally at the BBU [16]. The local processing in RRH relaxes requirements on the fronthaul and reduces the delay. On the contrary, the centralized processing in the BBU facilitates the cost and energy consumption benefits of C-RAN. Hence, the more functions are carried out in the BBU, the higher the amount of data is transmitted over the fronthaul. Note that the more functions are carried out in the BBU, the higher the amount of data is transmitted over the fronthaul [19]. The functional split different options in the 5G NR are shown in Fig. 1.

Each split of the control functions is preferable for different network conditions and requirements of the supported services [17]. Thus, in [16], the RAN-as-a-Service (RANaaS) concept is proposed to allow centralized management and processing to adapt according to the actual service demands. Still, a flexible and fully dynamic

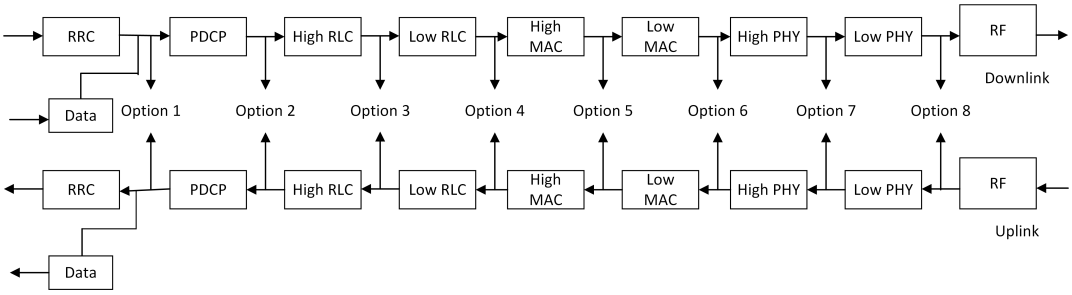


Figure 1. The functional split options in 5G NR [20]

change of the deployed functional split, according to the user's requirements and network status while reflecting capabilities and current load of the fronthaul, is an ongoing problem, as pointed out in [21]. With the control functions carried out in the centralized BBU, a low-quality fronthaul negatively affects the overall network performance and users' quality of service. Hence, the data transmission over the radio interface can be impaired as the radio channel changes over time. For example, the channel quality information (exploited, e.g., for scheduling of radio resources) might be outdated and might not correspond to the actual channel quality at the time of data transmission if the fronthaul with non-zero delay is deployed.

In this thesis, we focus on dynamic functional splitting in the context of 5G. Dynamic functional splitting typically refers to the ability to allocate and configure the processing functions between the BBU and RRHs dynamically. The functional split allows for flexibility and optimization in terms of resource utilization and network performance. In this regard, the objective of this thesis is to develop a solution that can handle the scheduling of radio resources efficiently in the C-RAN-based architecture.

In the rest of this chapter, we present the motivation for the research work carried out in the thesis and provide an outline of the thesis.

1.1 Motivation

The advent of 5G technology has ushered in a new era of connectivity, promising unprecedented speed, capacity, and versatility to support a wide array of applications. Cloud Radio Access Network (C-RAN) [9] has emerged as a cornerstone for 5G infrastructure, providing centralized processing capabilities [14]. However, as the demand for diverse services grows, the static nature of traditional functional splitting in C-RAN proves to be a limiting factor in optimizing resource utilization. This thesis is motivated by the exploration of resource allocation strategies within the context of Dynamic Functional Splitting (DFS) in C-RAN for 5G, with a specific focus on mitigating challenges associated with fronthaul delay.

Traditional C-RAN architectures often struggle to adapt to the dynamic nature of 5G applications due to static resource allocation strategies. The lack of flexibility hinders the efficient utilization of resources, especially in scenarios with varying traffic loads and diverse service requirements. In this context, dynamic functional splitting offers a promising solution by enabling the dynamic allocation of processing functions between the centralized BBU and RRH. However, the dynamic nature of this allocation introduces challenges, particularly in the form of fronthaul delay. DFS introduces a paradigm shift by allowing the dynamic allocation of processing functions based on real-time network conditions. This dynamic adaptability promises improved resource utilization, scalability, and responsiveness. The flexibility offered by DFS aligns with the dynamic requirements of 5G, making it a compelling avenue for research. This thesis aims to delve into the intersection of resource allocation strategies and dynamic functional splitting, mainly focusing on the challenges presented by fronthaul delay.

Fronthaul, the link connecting the centralized BBU to distributed RRHs, poses inherent latency challenges. The dynamic allocation of processing functions may introduce additional signaling overhead and data exchange, impacting fronthaul delay. This becomes particularly critical for applications with stringent latency requirements, such as autonomous vehicles, smart grids, and augmented reality. The varying traffic loads and synchronization constraints further compound the challenge of minimizing fronthaul delay while optimizing resource allocation.

The motivation for this thesis lies in addressing the critical need for efficient resource allocation within the dynamic framework of DFS in C-RAN for 5G networks. By investigating innovative resource allocation strategies, adaptive algorithms, and compression techniques, this research aims to strike a balance between dynamic functionality and minimal fronthaul delay. The outcomes are expected to not only contribute to the theoretical understanding of resource allocation in C-RAN but also provide practical insights for implementing responsive and resource-efficient 5G networks.

In conclusion, this thesis is motivated by the imperative to optimize resource allocation strategies within the evolving landscape of 5G networks, particularly within the context of dynamic functional splitting in C-RAN. By addressing the challenges associated with fronthaul delay, this research aims to contribute to the development of resource-efficient and responsive C-RAN architectures. The outcomes are anticipated to play a pivotal role in advancing the capabilities of 5G networks, ensuring they are well-equipped to handle the diverse and dynamic demands of the modern era.

1.2 Organization of the thesis

In this section, organization of the thesis is described.

- **Chapter 2 - State of the art:** provides a deep insight into the related topics to the subject of the thesis.
- **Chapter 3 - Thesis objectives:** defines thesis objectives based on the state of the art and the stated motivation.
- **Chapter 4 - Hierarchical Scheduler Framework:** outlines a general concept of the proposed hierarchical scheduler. This chapter illustrates how the proposed hierarchical scheduler concept enables flexibility of resource scheduling in the RRHs and the BBU.
- **Chapter 5 - Dynamic Approach of the Hierarchical Scheduler:** proposes a novel solution that enables dynamic adjustment of the scheduling period individually for each UE to maximize the sum throughput of the UEs.
- **Chapter 6 - HARQ Retransmission in Hierarchical Scheduler:** illustrates a comprehensive framework for the HARQ resource pre-allocation in the C-RAN, considering the hierarchical scheduling scheme.
- **Chapter 7 - Conclusions:** states the main thesis's contributions, summarizes all achieved results and outlines the future research directions.

Chapter 2

State of Art

This chapter overviews the recent up-to-date information about the fundamental principles and functionalities that interact with or influence the proposed thesis resource scheduler's design (i.e., the Hierarchical Scheduler). First, the chapter briefly explains the concept of C-RAN, the thesis proposed network architecture, in section 2.1. Since Functional Splitting introduces a flexible approach for overcoming the negative influence of the fronthaul delay in C-RAN by dynamically distributing processing functions between the BBU and RRHs in C-RAN, section 2.2 explains the functional splitting and its protocol layers in more detail. Section 2.3 comprehensively introduces wireless resource scheduling, including the scheduling algorithm and the main standard competitive resource schedulers, as the central theme of the thesis is resource scheduling. Part of the thesis's proposed solution is related to the Hierarchical Scheduler behavior with erroneous data retransmission in the context of Hybrid Automatic Repeat Request (HARQ), which is why section 2.4 explains the HARQ main challenges in C-RAN and different HARQ types.

2.1 Cloud-RAN in 5G

A Cloud Radio Access Network (C-RAN) is a network architecture that centralizes the processing of radio resources in a cloud or data center, rather than distributing them across traditional base stations (BTS) [9]. The C-RAN is an architecture that has been employed in the context of mobile communication networks, including 5G. C-RAN is a virtualized and centralized network architecture designed to enhance the efficiency and flexibility of radio access network operations. A typical network C-RAN architecture consists of a centralized BBU and distributed RRHs interconnected with a fronthaul, typically represented by a high-quality wired or wireless transport link as is shown in Fig. 2. While the RRHs perform digital signal processing, digital-to-analog conversion, power amplification, filtering, etc., the BBU centrally manages baseband processing for the RRHs. The BBUs' computation power is combined into a virtualized BBU pool that serves the RRHs.

The benefits provided by C-RAN are not limited just to the low energy consumption and the cost reduction, but C-RAN also enables elasticity of the network and large-scale coordination of network control and management [14]. For example, centralized control in BBU can efficiently facilitate interference mitigation techniques, e.g., coordinated multipoint (CoMP), as the BBU contains information about the whole network (or at least about a larger area covered by the BBU). On the other hand, moving some control functionalities from the RRHs to the BBU inevitably

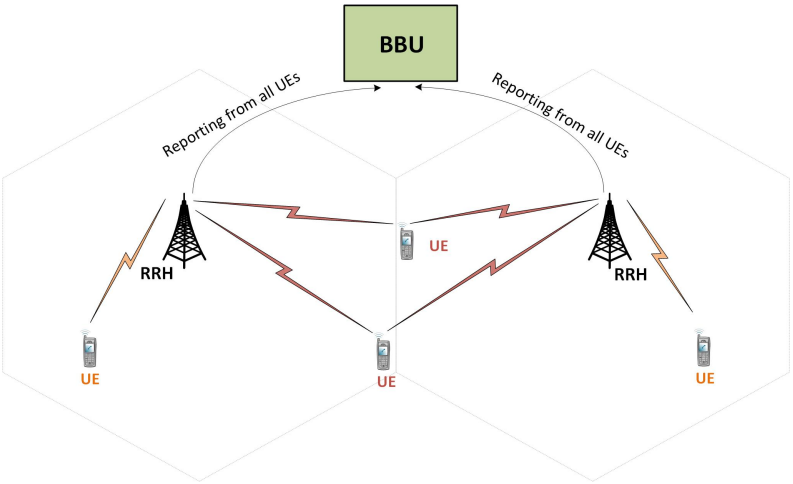


Figure 2. C-RAN architecture overview

introduces some challenges due to high requirements on the fronthaul capacity and latency to transport huge amounts of the baseband signals. The complexity of 5G C-RAN networks, coupled with the diverse requirements of different applications and services, demands intelligent and adaptive resource scheduling algorithms. Ongoing research and development continue to refine these strategies to address the challenges and opportunities presented by next-generation wireless networks. The main aspects and strategies that are taken into consideration in this thesis are listed as follows:

- **Dynamic Resource Allocation:** 5G C-RAN requires dynamic allocation of resources to adapt to varying traffic patterns and user demands.
- **Fronthaul Capacity Management:** Efficient scheduling of fronthaul resources is essential for ensuring low-latency communication between RRHs and BBU.
- **Interference Management:** Centralized coordination in C-RAN enables effective interference management.
- **Fault Tolerance and Redundancy:** Resource scheduling strategies need to account for fault tolerance (i.e., HARQ) and redundancy. Redundant resources can be allocated to handle failures, ensuring the continuity of services.

2.2 Functional Split in 5G based on C-RAN

In 5G, the functional split is often discussed in the context of the radio access network (RAN), which is responsible for connecting user devices (UEs) to the core network. Functional split in the context of 5G mobile networks refers to the division of network functions between different network elements or entities [7]. This functional splitting is a fundamental aspect of the architecture of 5G networks, enabling flexibility, scalability, and efficient resource utilization. Practically, the separation of the base station characterizes the functional split in 5G RAN functions into different components, typically known as the BBU and the RRH. The functional split between the BBU and the RRH allows for more distributed and centralized processing, offering advantages in terms of resource optimization, latency reduction, and scalability. The specific functions assigned to BBU and RRH may vary based on the 5G deployment scenario [7].

As a part of the study item for 5G New Radio (5G NR), 3GPP started studying different functional splits between BBU and RRHs [17]. They have presented roughly eight various split options, which are depicted in Fig.**??**. Basically, each functional split determines how many functions are carried out locally at the RRHs (with the benefit of relaxing requirements on the fronthaul and with lower latency) and how many functions are centralized at the BBU, thus exploiting the benefits offered by C-RAN. While the RRH performs digital processing, digital to analog conversion, power amplification, filtering, etc., the BBU centrally manages all baseband processing for the RRHs. The BBUs' computation power is combined together into a virtualized BBU pool that is able to serve tens, hundreds, or even thousands of RRHs. Several options for how to split functions between the BBU and RRHs depending on latency and capacity requirement of fronthaul links are defined (for more details see, e.g., [19] and [22]). An example of the individual options currently investigated within the 3rd Generation Partnership Project (3GPP) is depicted in Fig. 1 [17].

In general, the more functions are carried out in the BBU, the higher the amount of data is transmitted over the fronthaul. Consequently, the highest requirements on fronthaul capacity are observed for Option 8, where all protocol layers are moved from the RRHs into the BBU. On the other hand, in the case of Option 1, only radio resource control (RRC) layer functions are carried out in the BBU, while the rest of the functions are still performed in the RRH as in conventional mobile networks. To better understand each split option, we should start by explaining the protocols layer in 5G [17]. The 5G wireless communication system is designed with a layered architecture similar to its predecessors in cellular networks. The architecture consists of several protocol layers, each responsible for specific functions. The protocol layers are presented as follows [17]:

- **RRC layer:** The Radio Resource Control (RRC) layer is a crucial component in the control plane of cellular communication systems, including 3G, 4G LTE, and 5G networks. The RRC layer's primary responsibility is to manage the establishment, maintenance, and release of radio connections between a user device (such as a mobile phone or IoT device, i.e., UE) and the cellular network infrastructure (base station or eNodeB in LTE/5G networks, i.e., RRH). RRC is involved in managing the radio link and controlling the transition between different states (e.g., idle, connected).

- **PDCP layer:** The Packet Data Convergence Protocol (PDCP) layer is responsible for the compression and decompression of IP packets and the handling of packet-related functionalities. PDCP is involved in functions such as header compression, ciphering/deciphering, and reordering of packets. It is particularly important for optimizing the transmission of IP packets over the radio interface.

- **RLC layer:** The Radio Link Control (RLC) layer is responsible for tasks related to the segmentation, reassembly, and error correction of data packets. It manages the reliable and efficient transfer of data between the transmitting and receiving ends. RLC provides functionalities such as segmentation and concatenation of data packets and error detection and correction.

- **MAC layer:** The Medium Access Control (MAC) layer is responsible for managing access to the communication channel, especially in shared and wireless environments. It controls how devices contend for and access the shared medium. MAC layer multiplexes data from different higher-layer connections into frames for transmission and demultiplexes received frames to the appropriate higher-layer connections. The MAC layer maps logical channels, defined by higher-layer protocols, to the physical channels of the radio interface. This involves associating different types of traffic (voice, data, control signaling) with appropriate channels.

  MAC layer creates and parses headers in data frames. These headers contain information necessary for channel access, frame identification, addressing, and other control information. In some cases, the MAC layer can concatenate smaller frames into larger ones for more efficient transmission. Conversely, it can segment larger frames into smaller ones if needed. The MAC layer manages random access procedures and contention resolution mechanisms in scenarios where multiple devices compete for access to the channel. This is particularly relevant in wireless environments with shared communication channels.

  In 4G LTE and 5G networks, the MAC layer often includes Hybrid Automatic Repeat reQuest (HARQ) mechanisms. HARQ is used to enhance reliability by combining error detection and correction with retransmission strategies. The MAC layer supports QoS management, ensuring that different types of traffic receive appropriate levels of service based on their requirements (e.g., latency, reliability). In advanced wireless networks like LTE and 5G, the MAC layer participates in scheduling resources, determining when and how devices can access the channel based on factors like priority and resource availability. The MAC layer may be involved in power control mechanisms, optimizing the transmission power levels of devices to conserve energy and improve overall network efficiency.

- **PHY layer:** The Physical Layer (PHY) is responsible for modulating digital data into analog signals for transmission and demodulating received analog signals back into digital data at the receiver. Common modulation techniques include QPSK (Quadrature Phase Shift Keying), QAM (Quadrature Amplitude Modulation), and OFDM (Orthogonal Frequency Division Multiplexing).

  The Physical Layer may include channel encoding to add redundancy to transmitted data, aiding in error detection and correction. At the receiver, decoding

is performed to recover the original data. PHY determines the bit rate and symbol rate of the transmitted signals. The bit rate is the rate at which bits are transmitted, while the symbol rate represents the rate at which symbols (modulated forms of bits) are transmitted. PHY manages multiplexing techniques to allow multiple signals to share the same channel efficiently. In wireless networks, multiple access schemes like FDMA (Frequency Division Multiple Access), TDMA (Time Division Multiple Access), CDMA (Code Division Multiple Access), and OFDMA (Orthogonal Frequency Division Multiple Access) are used. PHY adapts to the characteristics of the transmission medium, whether it is guided (e.g., copper or fibre-optic cables) or unguided (e.g., air interface in wireless communication).

In wireless networks, the PHY layer may include mechanisms for controlling the transmission power of devices to optimize coverage, minimize interference, and conserve energy. PHY layer performs basic error detection and, in some cases, error correction to ensure the integrity of transmitted data. The PHY layer may include mechanisms for estimating channel conditions and applying equalization techniques to mitigate the effects of channel impairments. PHY ensures synchronization between the transmitter and receiver, including timing synchronization and frequency synchronization, to facilitate reliable communication. PHY manages carrier frequencies, ensuring that different communication channels operate on distinct frequencies to avoid interference.

As a part of the 3GPP framework, the main features of each split option are briefly described as follows [17]:

- **Option 1 (RRC/PCDP split):** This split allows for flexibility in deploying and scaling RRC and PDCP functions independently. By separating RRC and PDCP functions, it is possible to optimize the allocation of resources, leading to more efficient network operation. The split can contribute to reducing latency in the network, especially in scenarios where specific functions need to be processed more quickly. RRC, being responsible for radio resource management, can more effectively control and adapt to changes in radio conditions. The separation of RRC and PDCP functions aligns with the principles of Network functions virtualization (NFV), making it easier to virtualize and deploy these functions on cloud infrastructure.

- **Option 2 (PDCP/RLC Split):** This split allows for the independent scaling and optimization of the PDCP and RLC functions based on specific predefined service requirements. This split enables the efficient use of network resources and adaptability to varying conditions. The separation of responsibilities between PDCP and RLC can contribute to reducing latency in data transmission. This split allows for specific optimizations in each layer, enhancing overall system performance. This split also provides flexibility in the deployment of PDCP and RLC functions, catering to different network architectures and deployment scenarios. The functional split aligns with the principles of network function virtualization (NFV), facilitating the virtualization and deployment of PDCP and RLC functions on cloud infrastructure.

- **Option 3 (High RLC/Low RLC split, Intra RLC split):** This option splits the RLC sublayer into High RLC and Low RLC sublayers such that for

RLC Acknowledge Mode operation, all RLC functions may be performed at the High RLC sublayer residing in the central unit (i.e., BBU), while the segmentation may be performed at the Low RLC sublayer residing in the distributed unit (i.e., RRH).

- **Option 4 (RLC-MAC split):** The split in this option allows for flexibility in deploying and scaling RLC and MAC functions independently based on specific network requirements and conditions. By separating RLC and MAC functionalities, it is possible to optimize the allocation of resources, leading to more efficient network operation. The split can contribute to reducing latency in the network, especially in scenarios where specific functions need to be processed more quickly. The separation aligns with the principles of NFV, making it easier to virtualize and deploy these functions on cloud infrastructure.

- **Option 5 (Intra MAC split):** By splitting the MAC layer in this split into 2 entities (e.g., High-MAC and Low-MAC), the services and functions provided by the MAC layer will be located in the BBU, in the RRH, or in both. In this split, an overall scheduler is centralized in the BBU, and a MAC sublayer is local in each RRH to handle time-critical processing. From this split and below, the time-critical procedures in the HARQ are performed locally in the RRH, as well as the functions where performance is proportional to latency.

  In split option 5, the BBU communicates with the RRH through scheduling commands and HARQ reports. The reduced delay requirements on the fronthaul interface ensure that the distance to the BBU can be longer. The latency requirements depend highly on the realization and interaction of the scheduling functions carried out locally and centrally.

  On the other hand, much of the processing must be performed locally, limiting the benefits of shared processing. The high MAC sublayer controls the low MAC sublayers and manages Inter-Cell Interference Coordination (ICIC) [23]. On the other hand, using this split might lead to fronthaul delays due to the centralized scheduling decisions, which will have limitations for the Coordinated Multi-Point transmission/reception (CoMP) scheme [24].

- **Option 6 (MAC-PHY split):** The MAC-PHY split allows for the centralization of certain functions in a network, enabling more efficient resource utilization and management. This split also facilitates the virtualization of network functions, allowing them to be run on cloud infrastructure. By decoupling the MAC and PHY layers, network operators can scale and upgrade each layer independently. This flexibility is particularly important as networks evolve and new technologies are introduced. The split typically involves well-defined and open interfaces between the MAC and PHY layers [25]. Open interfaces promote interoperability, allowing equipment from different vendors to work together seamlessly. The MAC-PHY split aligns with NFV principles, allowing network functions traditionally implemented in dedicated hardware to be virtualized and run as software on general-purpose hardware [25].

  In scenarios where certain MAC functions are centralized, processing can be done in a more centralized and powerful entity (e.g., cloud data center), leading to improved efficiency and coordination [25]. Software updates and upgrades for the MAC layer and PHY layer can be performed independently, simplifying the upgrade process and allowing for faster adoption of new features or

improvements. With a more flexible architecture, dynamic resource allocation becomes more feasible. Resources can be allocated based on real-time network conditions and demand. Centralized or coordinated control over both MAC and PHY layers allows for better optimization of resources, interference management, and overall network performance [25].

- **Option 7 (Intra PHY split):** In this split, the precoding and resource element mapper are included in the RRH. The fronthaul link is responsible for transporting subframe symbols. This results in a slightly lower fronthaul link bitrate but also a more complicated RRH and less shared processing in the BBU [27]. Because the Fast Fourier Transform (FFT), the main algorithm used in Orthogonal Frequency-Division Multiplexing, which is the primary method for combining signals on modern digital radio standards, and resource element mapper are included in the RRH, the fronthaul connection has a variable bitrate starting with this split and all splits below [26].

  As a result, they may be transferred utilizing a specialized transport interface. This split option provides CoMP functions without sacrificing performance [27]. Due to the large separation in the physical layer, an in-band protocol is required to facilitate Physical Resource Block (PRB) allocation in this split option.

- **Option 8 (PHY-RF split):** This option allows the separation of the RF and the PHY layer. This split permits the centralization of processes at all protocol layer levels, resulting in very tight coordination of the RAN. This allows efficient support of functions such as CoMP, MIMO, load balancing and mobility [27].

Each split of the control functions is preferable for different network conditions and requirements of the supported services [28]. The functional split in 5G RAN allows for more flexibility in deploying and managing network resources. The functional split facilitates the use of virtualization technologies and cloud-native architectures, making it easier to scale, upgrade, and adapt the network to different deployment scenarios and service requirements. The specific functional split configurations can vary based on the deployment model and the goals of the network operator [28].

2.3 Wireless Resource Scheduling

Resource scheduling in wireless networks is a crucial aspect of network management that involves allocating and managing resources such as frequency spectrum, time slots, and transmit power among competing users or devices. Effective resource scheduling is essential to optimize network performance, enhance user experience, and ensure fair resource access. In this subsection, we discuss the main scheduling algorithms and common competitive schedulers.

2.3.1 Scheduling Algorithms

Resource scheduling algorithms in wireless networks play a crucial role in efficiently managing the available resources such as frequency spectrum, time slots, and transmit power to meet the demands of various users and applications. Resource scheduling algorithms commonly used in wireless networks are expressed as follows [29]:

- **Proportional Fairness:** The Proportional Fairness (PF) scheduling algorithm is a widely used approach in wireless communication systems, particularly in cellular networks, to allocate radio resources among users. Its goal is to achieve a balance between maximizing system throughput and ensuring fairness among users. The Proportional Fairness scheduling algorithm strikes a balance between maximizing system throughput and ensuring fairness among users. Its adaptability to changing conditions makes it suitable for dynamic wireless environments, and it is commonly employed in cellular networks to allocate resources to users efficiently with varying data rate requirements.

- **Round Robin:** The Round Robin (RR) scheduling algorithm is a simple and widely used approach for process or task scheduling in operating systems, particularly in time-sharing systems. A preemptive scheduling algorithm allocates a fixed time slice or quantum to each process in a circular order, allowing each process to execute for a specified time before moving on to the next process in the queue. Round Robin is a straightforward scheduling algorithm suitable for systems where simplicity is prioritized. While it provides fairness, it may not be the most efficient choice for all scenarios, and the choice of the time quantum is crucial in balancing fairness and system performance.

- **Max C/I (Carrier-to-Interference):** The Max C/I (Carrier-to-Interference) scheduling algorithm is a resource allocation strategy commonly used in wireless communication systems, particularly in cellular networks. The primary goal of this algorithm is to maximize the Carrier-to-Interference ratio for each user or device, optimizing spectral efficiency and overall system performance. The Max C/I scheduling algorithm is a valuable strategy for optimizing spectral efficiency in wireless communication systems. It is commonly used in cellular networks to allocate resources based on the quality of communication links. However, it is important to consider the trade-offs and limitations, and the algorithm may be used in conjunction with other scheduling strategies to address diverse network conditions and requirements.

2.3.2 Schedulers in C-RAN

This subsection overviews key aspects of distributed, centralized, and partially distributed scheduling. Compared to the thesis proposed schedulers, these three schedulers are being represented as competitive schedulers.

Distributed Scheduling

Distributed scheduling is an analogy to traditional scheduling in 4G networks, where each base station schedules the radio resources for the users individually without interacting with other base stations. In C-RAN, the distributed scheduling corresponds to the case when the scheduling is carried out in the RRHs (i.e., the function split options 1-4) [17]. This solution benefits from a fast response to any changes in the user's channel quality and a fast retransmission of erroneous packets. However, the scheduling performed individually by each RRH without any coordination with neighbouring RRHs results in strong mutual interference, and the performance of the cell-edge users is degraded. To mitigate the interference, the RRHs can coordinate their transmission to the cell-edge users, for example, by means of ICIC or CoMP. Then, several neighbouring RRHs interact with each other and perform scheduling, considering the interference of other RRHs on the users. However, such a solution implies a mutual signaling exchange among the coordinated neighbouring RRHs and, consequently, a significant load of the direct connections among the RRHs and a high complexity of the network management. Thus, the coordination is not easy and significantly increases the network deployment cost.

Centralized Scheduling

The centralized scheduling is performed solely in the centralized unit. This case is represented by the function split options 5-8 [17], where the scheduling-related functionalities are located in the centralized BBU. On the one hand, such a solution can efficiently facilitate interference mitigation techniques, as the centralized unit contains information about the whole network or at least about the more extensive area covered by this unit. The centralized scheduling also preserves C-RAN benefits of a lower cost and energy consumption. On the other hand, centralized scheduling requires delivering the scheduling-related information (channel quality, buffer status, etc.) from the RRHs to the centralized unit. In such cases, the fronthaul quality plays a substantial role. The fronthaul with a high delay can heavily degrade the overall network performance and outweighs the gains introduced by the interference mitigation techniques because the scheduling is performed with outdated channel state information. The fronthaul delay also impairs and prolongs the retransmissions of erroneous packets.

Partially Distributed Scheduling

A scheduling combining both centralized and distributed approaches is introduced in [30], where the scheduling functions are split between the centralized BBU and partially distributed radio aggregation units (RAUs). The RAU is a new semi-distributed entity concentrating the control functions for several underlying RRHs. Hence, multiple RAUs are deployed in the network, and each RAU performs the scheduling for the several underlying RRHs so that each is under the control of just one RAU. In parallel to the scheduling in the RAUs, the BBU performs its own

scheduling for the whole network. The decision of whether the scheduling derived in the BBU or the RAUs is exploited depends on the fronthaul delay. Suppose the fronthaul delay prevents timely scheduling delivery from the BBU to the RAUs. In that case, the scheduling done by the RAUs is exploited. On the contrary, if the scheduling from the BBU arrives on time, the scheduling by the RAUs is overruled by the one from the BBU. From the function split perspective, the scheduling is not done in the RRHs; thus, the partially distributed scheduling corresponds to the centralized scheduling and the split options 5-8 [17].

2.4 Hybrid Automatic Repeat request (HARQ) in 5G based on C-RAN

HARQ is a communication protocol widely used in wireless communication systems and combines automatic repeat request (ARQ) and error correction coding to improve data transmission reliability. The idea behind the HARQ is to model a system that detects the received erroneous data transport block and then requests a retransmission of the erroneous blocks. HARQ is particularly important in the context of C-RAN, where centralized processing (i.e., BBU) is employed to enhance the efficiency of radio access networks. However, the fronthaul in C-RAN introduces challenges related to high goodput and low latency requirements to ensure a swift exchange of the baseband signals over the fronthaul links between the BBU and the RRHs [9]. The high latency at the fronthaul would negatively impact the data transmissions and time-critical radio resource management protocols, such as error correction via HARQ [31]. The whole retransmission process should be accomplished in up to 8 Time Transmission Intervals (TTIs) [32]. Thus, the TTI duration, which is ranging from 62.5µs to 1 ms [33] [34] in 5G, imposes challenges on the HARQ process in C-RAN [34] [35].

There are several works addressing the problem of the HARQ targeting the aspect of low latency for ultra-reliable and low latency communications (URLLC) in 5G, see, e.g., [36] [37] [38] [39]. In [36], the authors present a semi-persistent scheduling of resources for the UEs' retransmissions. To this end, for any potential retransmissions, a pre-defined amount of resources is shared by a pre-defined group of UEs based on the Block Error Rate (BLER) of the UEs' first transmission. In [37] and its extension in [38], a periodic radio resource allocation is proposed for retransmissions of individual UEs to meet latency and reliability requirements. The solution is based on selecting an optimal modulation and coding scheme (MCS) and subsequent allocation of the required resources. The paper [39] exploits the queuing model to optimize the HARQ resource requirement in URLLC. However, none of the works presented in [36] [37] [38] [39] assume the C-RAN architecture with the realistic fronthaul with non-zero delay for the HARQ and resource allocation.

Since the main theme of this thesis is resource scheduling, many works in the literature investigate resource allocation in 5G based on C-RAN. For instance, the resource allocation for C-RAN, considering strict HARQ requirements, is assumed in [40]. The HARQ itself is, however, not optimized in any way. The optimization of HARQ tailored for C-RAN is assumed in [41], [42], [43], [44]. The authors in [41] propose a centralized low-complexity packet scheduling scheme to reduce communication delay. Nevertheless, the inter-cell interference (ICI) among the deployed UEs is neglected, and this work is limited only to URLLC traffic. In [42], the authors consider sharing computing resources among multiple RRHs for the uplink in the C-RAN architecture to improve the HARQ retransmission process. However, the work considers only a single-user scenario, and extension toward a practical multi-user scenario is not straightforward. In [43] and [44], the authors focus on a proactive HARQ, which transmits proactively redundancy versions until the receiver indicates correct reception with ACK. This leads to reduced latency of HARQ, but, at the same time, it also lowers spectral efficiency notably. To this end, the authors in [43] propose a feedback prediction scheme for C-RAN to reduce the redundancy in proactive HARQ. The paper is further extended in [44], where machine-learning-assisted HARQ prediction schemes for C-RAN are proposed in order to decrease the maxi-

mum transmission latency. Still, neither [43] nor [44] deals with the pre-allocation of resources for HARQ in C-RAN; thus, this thesis can be seen as complementary to these. In this thesis, we examine our proposed scheduler (i.e., the Hierarchical Scheduler) with different types of HARQ that are listed as follows:

2.4.1 Adaptive HARQ (Type I HARQ):

In adaptive HARQ, the transmission parameters, such as modulation and coding scheme, can be adjusted dynamically based on the channel conditions. If a transmission is unsuccessful, the receiver sends a negative acknowledgment (NACK) to the transmitter, indicating that the received data contains errors. The transmitter, upon receiving a NACK, can adapt its transmission parameters to improve the chances of successful retransmission [35]. This adaptation may involve using a different modulation scheme, changing the coding rate, or adjusting other parameters. Adaptive HARQ is more flexible and can adapt to varying channel conditions, providing better performance in dynamic and changing environments.

2.4.2 Non-adaptive HARQ (Type II HARQ):

In non-adaptive HARQ, the transmission parameters are fixed and do not change based on the channel conditions. If a transmission is unsuccessful, the receiver sends a NACK to the transmitter. The transmitter, upon receiving a NACK, retransmits the same data using the same transmission parameters without any adaptation. Non-adaptive HARQ is simpler and has lower signaling overhead compared to adaptive HARQ [35]. However, it may not perform as well in environments with varying channel conditions.

The choice between adaptive and non-adaptive HARQ depends on the specific requirements of the communication system and the characteristics of the channel. Adaptive HARQ is often used in wireless communication systems, such as cellular networks, where channel conditions can change rapidly. Non-adaptive HARQ may be sufficient for more stable communication channels or when simplicity and low overhead are crucial.

However, the HARQ Process can also be classified based on the timing relationship between the original transmission and its retransmissions. This classification results in two types: synchronous HARQ and asynchronous (non-synchronous) HARQ.

2.4.3 Synchronous HARQ:

In synchronous HARQ, the retransmissions are closely synchronized with the original transmission in terms of timing [35]. The retransmissions are sent in predefined time intervals or slots that are synchronized with the communication system's frame structure. This synchronization simplifies the receiver's task in managing and decoding the retransmitted data, as it knows when to expect the retransmissions.

2.4.4 Asynchronous (Non-synchronous) HARQ:

In asynchronous HARQ, there is no strict synchronization between the original transmission and its retransmissions [35]. The retransmissions can occur at arbitrary

time instances, not necessarily aligned with predefined time slots. Asynchronous HARQ is more flexible in handling variable delays and may be suitable for communication systems with less stringent timing requirements.

The choice between synchronous and asynchronous HARQ depends on the characteristics of the communication system and the nature of the transmission medium [35]. Synchronous HARQ is often used in systems with well-defined frame structures and timing, such as cellular networks with time division multiple access (TDMA) or orthogonal frequency-division multiple access (OFDMA) schemes. Asynchronous HARQ may be employed in scenarios where the timing is less predictable or where the flexibility of not adhering to a strict schedule is beneficial, such as in packet-switched networks [35].

Chapter 3

Thesis Objectives

Based on the motivation and state of the art mentioned earlier, we can specify many objectives to optimize resource allocation strategies and address the challenges posed by fronthaul delay in the quest for responsive and efficient 5G networks. The thesis objectives are defined as:

- **Objective 1:** Propose a novel approach for scheduling in C-RAN-based mobile networks, i.e., Hierarchical Scheduler. The new solution is based on splitting the MAC scheduler into a centralized scheduler (C-Sc) located at the BBU and a distributed scheduler (D-Sc) located in the RRHs to suppress the negative impact of the fronthaul delay on the network throughput.
- **Objective 2:** Enable a dynamic adjustment of the scheduling period individually for each UE in the Hierarchical Scheduler to maximize the sum throughput of the UEs.
- **Objective 3:** Propose a comprehensive framework for the HARQ resource pre-allocation in the C-RAN, considering the hierarchical scheduling to maximize the goodput of UEs via minimizing the transport block loss rate and maximizing the resource pre-allocation accuracy.

Chapter 4

Hierarchical Scheduler Framework

In this chapter, we outline the general concept of the proposed hierarchical scheduler. This chapter focuses on demonstrating how the proposed hierarchical scheduler suppresses the negative impact of the non-ideal fronthaul. Additionally, we elucidate the adaptability of the proposed hierarchical scheduler in allocating resources based on the varying availability of processing resources within both RRHs and BBU. Furthermore, we showcase the efficient management of HARQ retransmissions by the proposed hierarchical scheduler.

This chapter is organized as follows: a high-level overview of the proposed concept is shown in section 4.1 before the definition of a centralized scheduler (C-Sc) running in the BBU, and a distributed scheduler (D-Sc) located in the RRH is defined in section 4.2. The thesis's proposed way for UEs' classification is explained in Section 4.3, and the resource preallocation method for retransmission of erroneous packets is detailed in Section 4.4. The dynamic deployment of scheduling-related functionalities is presented in section 4.5, where section 4.6 shows the competitive scheduling algorithms. Section 4.7 outlines the chapter's main findings and outcomes. Finally, section 4.8 provides a brief Conclusions of the chapter.

4.1 High-level overview of the concept

In this thesis, we propose a new framework for hierarchical scheduling, based on C-RAN architecture, combining the benefits of both a centralized scheduling in the BBU and a distributed scheduling in the RRHs to suppress the negative impact of the fronthaul's latency. The centralized scheduler is responsible for long-term scheduling, especially for the cell edge UEs, as these can benefit from interference mitigation techniques. The distributed scheduler mitigates a negative impact of the fronthaul delay on the non-cell edge UEs and enables efficient retransmission of erroneous blocks. The proposed hierarchal scheduler is composed of two tiers: a centralized scheduler (C-Sc) running in the BBU and a distributed scheduler (D-Sc) located in the RRH (see Fig. 3).

The objective of the C-Sc is to provide high-level and long-term scheduling with awareness of the mutual interference among the cells. The D-Sc then adjusts scheduling in real time according to the local needs of the UEs belonging to individual

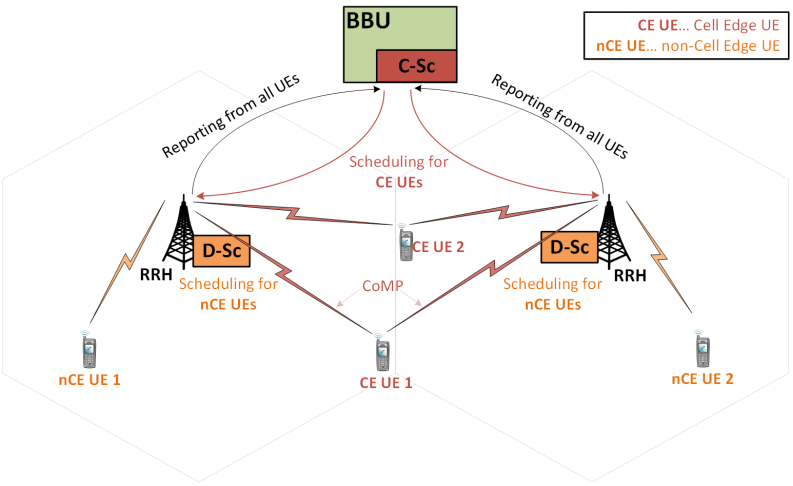


Figure 3. Overview of the Hierarchical Scheduler.

RRHs. As the interference among cells is critical mainly for the UEs at the cell edge, the C-Sc primarily schedules data transmission for these cell edge users (CE UEs), which suffer the most from interference due to proximity to neighbouring cells (in downlink) and to the UEs in adjacent cells (in uplink). The C-Sc can exploit knowledge of these mutual interferences; thus, the C-Sc can handle resource allocation for CoMP or any other advanced intercell interference cancellation technique and schedule the resources with awareness of the global interference relations. In contrast to the C-Sc, the D-Sc schedules data transmission only for non-cell edge UEs (nCE UEs), which are not influenced by interference from the other cells, and, thus, there is no need to keep awareness of the interferences.

4.2 Scheduling interval of C-Sc and D-Sc

As we mentioned earlier, The C-Sc in the BBU makes long-term scheduling for the CE UEs' data transmission while considering, at the same time, the needs of the nCE UEs. Long-term scheduling is understood as a scheduling decision not only for one-time transmission interval (TTI) but for N consecutive TTIs (i.e., NxTTI). In contrast, the D-Sc schedules resources for the nCE UEs on a short-term basis for each TTI (see Fig. 4). The long-term scheduling reduces requirements on the processing power of the C-Sc and lowers the amount of signaling overhead between the RRHs and the BBU as the resources are allocated for the CE UEs in a semi-persistent-like manner over NxTTI interval. Unfortunately, a high value of N may lead to potential performance degradation as the scheduling does not reflect actual radio conditions (i.e., channel information used for scheduling may not be valid anymore in later TTIs [50]).

To cope with the potential performance degradation of the CE UEs due to the high N, we suggest scheduling the transmissions to the CE UEs in the early TTIs followed just after the centralized scheduling decision is made, as shown in Fig. 4. In an extreme case, all TTIs at the beginning of N can be dedicated solely to the CE UEs, while the nCE UEs are not scheduled whatsoever at these. Then, progressively, more resources are allocated to the nCE UEs. Since the RRHs are able to dynamically adapt long-term scheduling for the nCE UEs in each TTI according to the actual

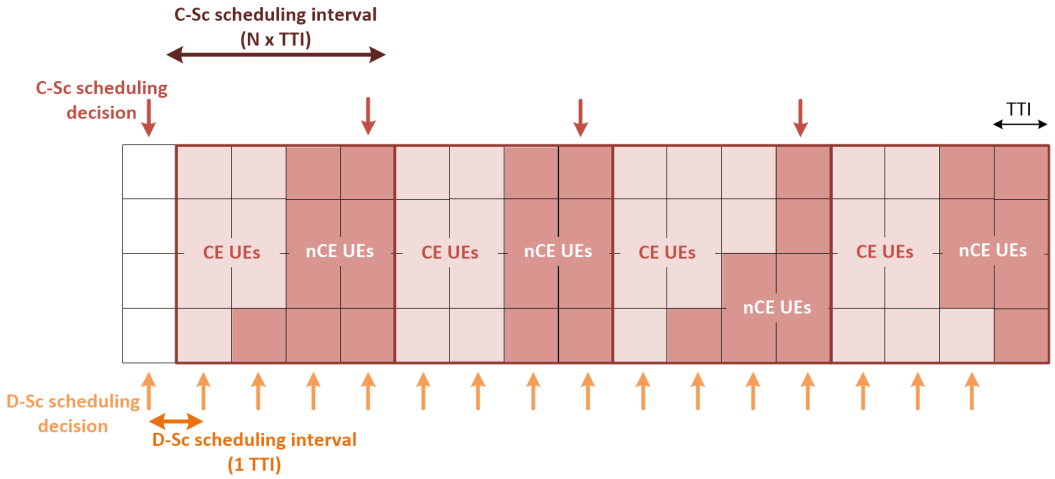


Figure 4. Example of different scheduling intervals for C-Sc (scheduling is done for four consecutive TTIs) and D-Sc (conventional single TTI).

channel information, there is no negative impact on the nCE UEs. Of course, the ratio between the resources for the CE and nCE UEs in each TTI should consider standard quality of service requirements, such as packet delay or priority.

Please note that the hierarchical scheduling enables to adapt N for individual cell-edge users according to their channel statistics. More specifically, the actual value of N is individually and dynamically set for each user according to the magnitude and frequency of the users' channel fluctuation over time. Thus, the negative effect of the outdated channel quality information is suppressed via a low N if the channel fluctuates significantly. On the contrary, the amount of signaling overhead is reduced by setting a larger N if the channel is stable.

4.3 Classification of UEs

Classification of the UEs to the CE and nCE UEs is crucial for the performance of the hierarchical scheduler. The classification process is done solely in the BBU, which has the global information needed to perform the classification efficiently. The UE's classification to the CE and nCE UEs can be done according to several criteria such as signal-to-interference plus noise ratio (SINR) (UEs that experience SINR below some threshold are considered to be CE UEs while the UEs with SINR above this threshold are classified as nCE UEs) or CoMP gain (UEs for which the CoMP improves the performance are assumed to be the CE UEs while the rest of them are considered to be the nCE UEs).

The classification of the users to the cell-edge and non-cell-edge can also be done based on the channel quality, interference level, total number of occupied resources, etc. As the hierarchical scheduling supports the cooperative interference mitigation techniques, the classification based on the benefit of the RRHs' cooperation for the interference mitigation is suggested. Hence, the user is classified as the cell-edge if the cooperation of at least two RRHs on the transmission to this user reduces the number of resources required to serve this user. Otherwise, the user is labeled as the non-cell-edge. The (re-)classification process should be updated either proactively (on a regular or irregular basis) or reactively if the network's performance is degraded, and

a re-classification of the UEs would improve network performance.

Hence, we further classify the UEs into K_{CE} CE UEs and K_{nCE} nCE UEs so that $K = K_{CE} + K_{nCE}$. This classification is based on the experienced SINR via individual UEs. Since, intuitively, the CE UEs experience more substantial inter-cell interference from adjacent RRHs than the nCE UEs, an Inter-Cell Interference Coordination (ICIC) technique [51] is adopted to mitigate such interference.

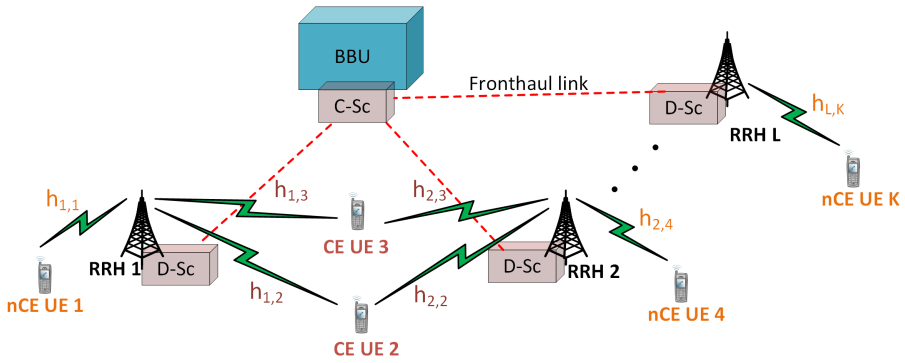


Figure 5. The ICIC in the hierarchical scheduler.

In the ICIC (see Fig. 5), a set of RRHs (i.e., $ICIC$ set, L_k^{ICIC}) cooperates together to mitigate the ICI of individual CE UEs. More specifically, each CE UE is served via orthogonal resource blocks (RBs) with respect to the transmission of other RRHs in the same ICIC set. The ICIC set encompasses L_k^{ICIC} RRHs that are involved in the cooperation to improve the k-th CE UE's SINR. The new RRH is added to the ICIC set for the k-th CE UE (i.e., to the L_k^{ICIC}) if and only if the CE UE communication goodput is improved by adding such RRH(s). Hence, the RRH is included in the UE's ICIC set if the RRH satisfies the following condition:

$$\Gamma_k < \alpha_C \frac{n_k^S}{n_k^{ICIC}} \tag{1}$$

where Γ_k defines the number of RRHs in L_k^{ICIC} (i.e, Γ_k ¿ 1), n_k^S represents the number of RBs required for the transmission of data to the k-th CE UE without ICIC (i.e., single RRH), n_k^{ICIC} corresponds to the number of RBs allocated for the data transmission to the k-th CE UE from each RRH in the ICIC set, and α_C is the ICIC threshold (i.e., α_C>1). The parameter α_C indicates the transmission efficiency with the ICIC utilization so that the RRH is added to the ICIC set if the ratio of the number of required RBs without the ICIC to the amount of the required RBs with the ICIC is α_C times higher than the number of RRHs in the ICIC set. This way, we can guarantee that the ICIC is exploited if and only if the ICIC increases the communication goodput of the CE UE. Considering the ICIC is used, the CE UE's SINR between the l-th RRH and the k-th CE UE ($\gamma_{k,l}^{CE}$) is calculated as:

$$\gamma_{l,k}^{CE} = \frac{p_l . h_{l,k}}{\sigma_n + \sum_{i \notin L_k^{ICIC}} p_i g_{i,k}} \tag{2}$$

where p_l is the transmission power of l-th RRH, $h_{l,k}$ is the channel gain between the l-th RRH and the k-th CE UE, σ_n is the noise power, and the term $\sum_{i \notin L_k^{ICIC}} p_i . h_{i,k}$ represents the inter-cell interference from all the RRHs except those included in

L_k^{ICIC}. Likewise, since the UEs' classification depends on individual UEs' SINR, the nCE UE's SINR between the l-th RRH and the k-th nCE UE is calculated as:

$$\gamma_{l,k}^{nCE} = \frac{p_l.h_{l,k}}{\sigma_n + \sum_{i=1,i\neq l}^{i=L} p_i h_{i,k}} \tag{3}$$

where the term $\sum_{i=1,i\neq l}^{i=L} p_i h_{i,k}$ represents the inter-cell interference from all but the serving RRHs.

4.4 Retransmission of Erroneous Packets

The hierarchical scheduling enables a fast and efficient handling of erroneous packets' retransmissions. The hybrid automatic repeat request procedure provides the fast error correction that combines retransmission of the erroneous packets with forward error correction, see [67]. For the non-cell-edge users, the retransmission process follows a standard hybrid automatic repeat request procedure in the mobile networks, i.e., the distributed unit schedules the resources for the retransmission if the packet is received with errors by the user (indicated by a request for retransmission). For the retransmissions of erroneous data, the distributed unit selects the most suitable resources, which are not dedicated to the cell-edge users at the scheduling moment.

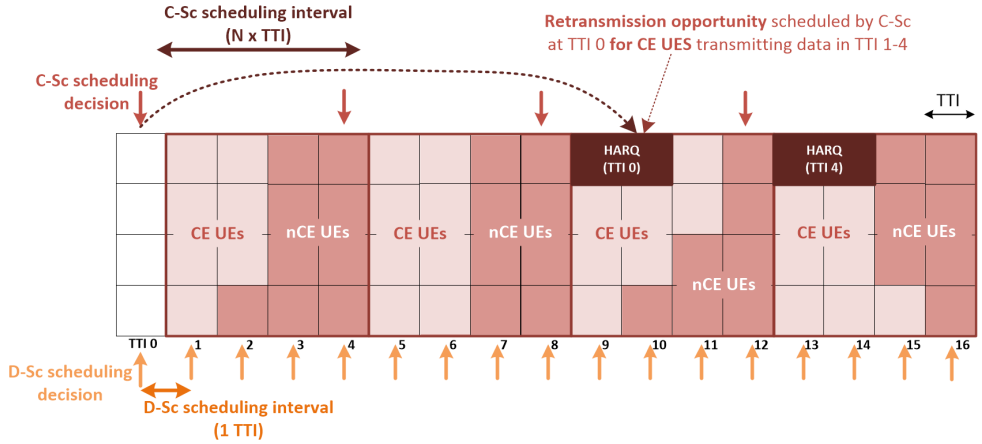


Figure 6. Allocation of resources for potential retransmission (HARQ) to CE UEs scheduled by C-Sc; nCE UEs retransmission is scheduled by D-Sc on per TTI bases according to ACK/NACK.

The retransmission of erroneous data is more complicated for cell-edge users. Handling the retransmissions by the centralized unit would lead to a high packet delay as the requests for retransmission should be delivered to the centralized unit, and then, a new scheduling decision for the retransmission should be sent back to the distributed unit. To avoid this long-lasting process, the centralized unit pre-allocates specific radio resources for potential retransmissions of the erroneous packets for the cell-edge users during each scheduling period N (see Fig. 6 with two retransmission opportunities for the scheduling done in TTI 0 and TTI 4).

The retransmission for the cell-edge users is handled solely by the distributed units at the pre-allocated radio resources with no intervention from the centralized unit. Thus, the retransmission process is shortened, and the fronthaul delay does not

affect the retransmissions at all. To avoid wasting the radio resources pre-allocated for the retransmission, the distributed unit arbitrarily assigns unused pre-allocated resources to any non-cell-edge user(s) since the interference from other neighboring cells is low.

4.5 Dynamic Deployment of Scheduling Related Functionalities

A deployment of the hierarchical scheduler functionalities is critical as it directly impacts the network performance. Except the scheduling for the nCE UEs (done by the D-Sc), all other functionalities are supposed to be performed by the C-Sc.

In general, all the functionalities located in the BBU can also be carried out (computed) by any other entity (such as MEC server or RAU in [30]) with a sufficient computing power, if the C-Sc would not be able to perform the computing, e.g., due to a heavy computing load of the BBU. Examples of such entities are mobile edge computing servers or neighboring RRHs with sufficient computing capabilities. This assumes that the communication link between the entity performing the computation and the BBU is of a high quality and close-to-zero delay in order to avoid the throughput degradation.

The computing load of the C-Sc and the D-Scs related to the scheduling is balanced, if needed, by a dynamic adjustment of the ratio of the CE and nCE UEs and by an appropriate setting of the centralized scheduling period N. Consequently, we can dynamically control and balance the scheduling-related computing load of the D-Scs and improve the quality of service offered to the UEs depending on the current network status including fronthaul quality, radio channel quality, etc.

The dynamic split that changes allocation of the scheduling functions over time is implemented by a continuous and periodic re-classification of the UEs to the CE and nCE UEs. This re-classification process considers the network load, which influences the fronthaul delay. With a high fronthaul delay, more UEs are scheduled directly by the D-Sc to overcome the low quality fronthaul and the ratio of nCE UEs is increased. Contrary, with a low fronthaul delay, more UEs are scheduled by the C-Sc. Hence, the ratio of nCE UEs is lowered.

4.6 The Hierarchical scheduler Versus Competitive Scheduling algorithms

This section overviews the pros and cons of each competitive scheduling algorithm to identify the challenges for scheduling in C-RAN. To distinguish the difference among the competitive scheduling approaches in C-RAN, Fig. 7 presents a brief overview of the proposed HS scheduler (green color) in comparison with the competitive scheduling approaches. The competitive scheduling approach includes distributed scheduling (orange color), centralized scheduling (red color), and partially distributed scheduling (violet color). The main features of each competitive scheduling approach are explained as follows:

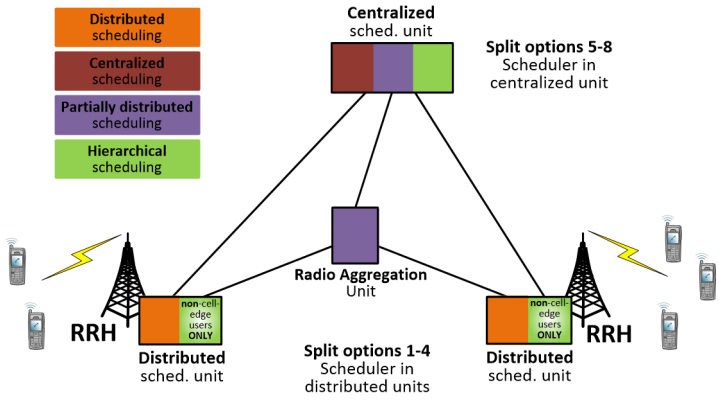


Figure 7. Overview of various scheduling options for C-RAN. Individual colors indicate nodes involved in the scheduling for each scheduling type.

4.6.1 Distributed scheduling

In C-RAN, the distributed scheduling corresponds to the case when the scheduling is carried out in the RRHs (i.e., the function split options 1-4) as indicated in Fig. 7 (orange color).

Summary of the *distributed scheduling*:

- *Pros*: low packet delay and easy management of the erroneous packets retransmission as the fronthaul does not impair scheduling.
- *Cons*: complex implementation resulting in a high cost (more features included in the RRHs); high amount of signaling for the coordinated interference mitigation.

4.6.2 Centralized scheduling

The centralized scheduling is performed solely in the centralized unit as shown in Fig. 7 (red color). This case is represented by the function split options 5-8, where the scheduling-related functionalities are located in the centralized BBU.

Summary of the *centralized scheduling*:

- *Pros*: low cost; easy coordination of scheduling among RRHs to suppress interference.
- *Cons*: performance potentially degraded due to the fronthaul delay; complicated management of erroneous packet retransmission leading to a high delay.

4.6.3 Partially distributed scheduling

A scheduling combining both centralized and distributed approaches is introduced in [30], where the scheduling functions are split between the centralized BBU and partially distributed radio aggregation units (RAUs), see Fig. 7 (violet color). The RAU is a new semi-distributed entity concentrating the control functions for several underlying RRHs. Hence, multiple RAUs are deployed in the network and each RAU performs the scheduling for the several underlying RRHs so that each is under the

control of just one RAU. In parallel to the scheduling in the RAUs, the BBU performs own scheduling for the whole network.

Summary of the *partially distributed scheduling*:

- *Pros*: reduced impact of the fronthaul delay comparing to the centralized approach; possible coordination among RRHs to suppress interference.
- *Cons*: higher cost than the centralized approach, as every scheduling is performed twice (in the BBU and the RAU) and the new entity is required; performance still limited by the actual status of the fronthaul and its parts (RRH to RAU and RAU to BBU).

4.7 Performance Evaluation

In this section, the proposed hierarchical scheduler performance is evaluated in MATLAB system-level simulator. To this end, simulation scenarios, parameters, and competitive algorithms are presented in the next subsection. Then, the simulation results are described and discussed.

4.7.1 Simulation scenario and models

We consider a square area with a size of 1x1 km encompassing a single BBU with C-Sc located in the middle of the area. Furthermore, 100 RRHs with D-Scs and 200 UEs are deployed randomly with the uniform distribution. Each UE is associated with the RRH providing the highest received signal strength. The channel between the UE and the RRH follows Urban Microcell model [62] with Rayleigh and Rician fading to model shadowing and fast fading, respectively, according to [63]. The performance is analyzed for the fronthaul delay from 0 ms (ideal fronthaul) to 30 ms, which encompasses ideal, near-ideal, sub-ideal, and non-ideal fronthaul delay ranges, as described in [64].

The modulation and coding scheme for transmission is determined according to reported channel quality indicator (CQI) in line with [67]. We consider a dynamic ICIC scheme based on [51] handled in the BBU. Note that the proposed concept is suitable also for any other techniques, such as CoMP. We demonstrate performance for ICIC as it implies less strict requirements on implementation in real networks comparing to CoMP. The HARQ retransmission process is implemented in line with HARQ defined by 3GPP in [67].

The performance of the proposed *hierarchical scheduler* is compared with: i) *centralized scheduler* deployed in the BBU (split options 6-8 according to 3GPP [17]), and ii) *partially distributed scheduler* proposed in [30]. Since [30] does not specify deployment of the RAUs, we consider a realistic case, where the RAUs are collocated with the underlying RRH that is closest to the center of the cluster of all underlying RRHs to avoid additional cost related to the RAUs' deployment. The BBU communicates directly with all RAUs and each RAU is connected to all underlying RRHs. As the performance of partially distributed scheduler depends on the number of RAUs and setting of the threshold, we depicts several realistic and well-performing options. All compared solutions (centralized, partially distributed, and hierarchical schedulers) exploit the same implementation of ICIC based on [51] for a fair comparison. Moreover, all solutions consider a conventional proportional fair scheduling, where the sum of logarithmic average of the UEs' throughput is maximized.

4.7.2 Simulation results

In this section, we present and discuss the simulation results to demonstrate the benefits of the proposed dynamic scheduling period with respect to other state-of-the-art solutions. Fig. 8 investigates the impact of the fronthaul delay on network throughput, defined as a sum throughput of all UEs per second. Regardless of the type of scheduler, the network throughput decreases with increasing fronthaul delay. This behavior is unsurprising as the fronthaul delay leads to a delay of information required for scheduling (channel quality reports) and delivery of the scheduling decision to the RRH and its application to transmitted data. However, the fronthaul delay impairs both centralized and partially distributed solutions significantly more than the proposed hierarchical scheduler.

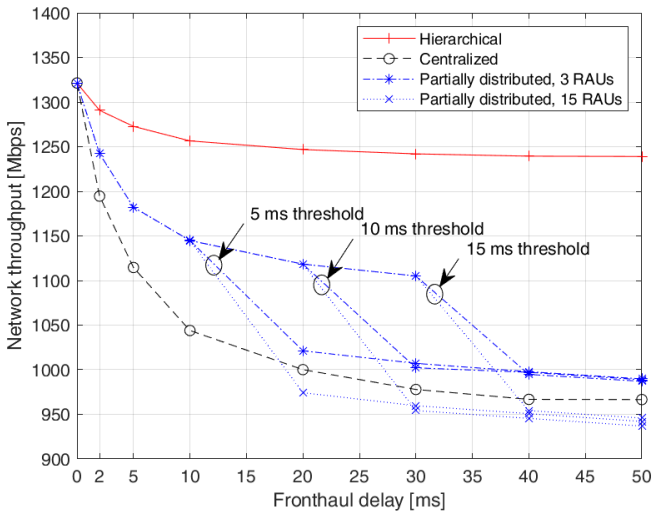


Figure 8. Impact of fronthaul delay on the network throughput for centralized (with N=1), partially distributed and hierarchical schedulers.

The proposed hierarchical scheduler improves the network throughput, especially for a lower quality fronthaul (i.e., with higher delay). Thus, for a 50 ms fronthaul delay, the hierarchical scheduler outperforms the centralized and partially distributed solutions by 19% and 17%, respectively. The gain of the hierarchical scheduler is achieved via an efficient suppression of the negative impact of fronthaul delay, as the centralized scheduling can be adjusted in the RRHs for the nCE UEs if the fronthaul delay is considerably significant. At the same time, the CE UEs can still benefit from well-coordinated CoMP as these are scheduled solely from the BBU. Note that the rapid drops in the network throughput of the partially distributed scheduler correspond to the threshold setting, determining whether the scheduling is done centrally in the BBU or in a distributed way in the RAUs. If the fronthaul delay exceeds this threshold, the centralized unit takes over the whole scheduling, and intermediate schedulers in the RAUs are not exploited. The drop appears at double the threshold value as the fronthaul delay applies twice (in uplink and downlink).

Furthermore, we investigate the impact of the prolonged scheduling period in the C-Sc in the BBU. The longer the scheduling period is, the less computation power is needed for the scheduling, and less signaling is required in the network. In Fig. 9, we can see that the network throughput is decreasing with N because the channel

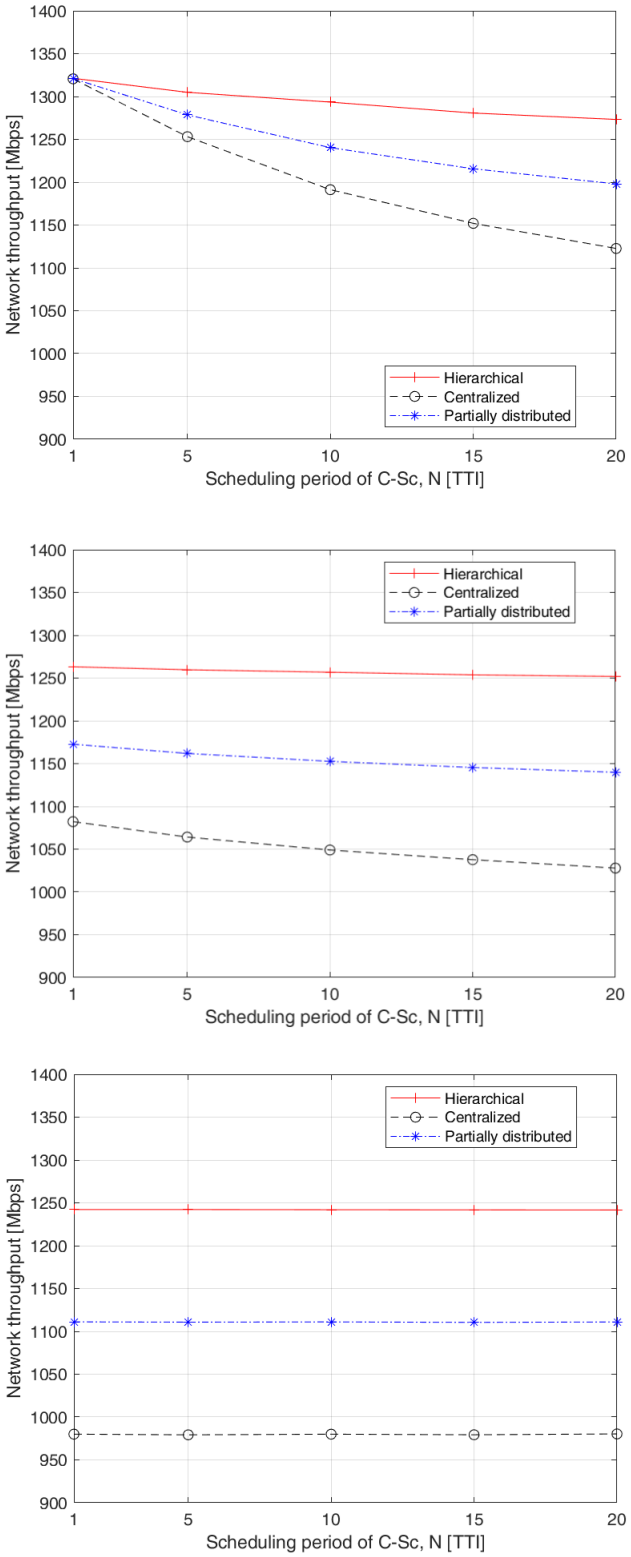


Figure 9. Impact of the scheduling period of the centralized scheduler C-Sc, N, on network throughput for fronhaul delay of 0 ms subplot (upper), 5 ms (middle), and 20 ms (bottom).

quality can change during the longer scheduling period, and the scheduling might not be accurate. However, as the hierarchical scheduler allows an adjustment of the scheduling in the RRHs for the nCE UEs. Thus, the network throughput degradation is suppressed (below 6% if N changes from 1 to 20) with respect to the conventional centralized scheduling, which leads to a drop of 15% (if N changes from 1 to 20).

Note that the hierarchical scheduler does not allow scheduling adjustments for CE UEs as these are served by multiple RRHs, and uncoordinated scheduling updates by these RRHs would lead to the elimination of coordinated multipoint gain. Comparing subplots in Fig.9, representing the impact of N for different fronthaul delays, we can see that the impact of N is less notable if the fronthaul delay increases. This is because N becomes overwhelmed with the fronthaul delay as the fronthaul delay increases and, the scheduling is anyway done too much in advance (fronthaul delay plus N), and channel state information for scheduling is too outdated.

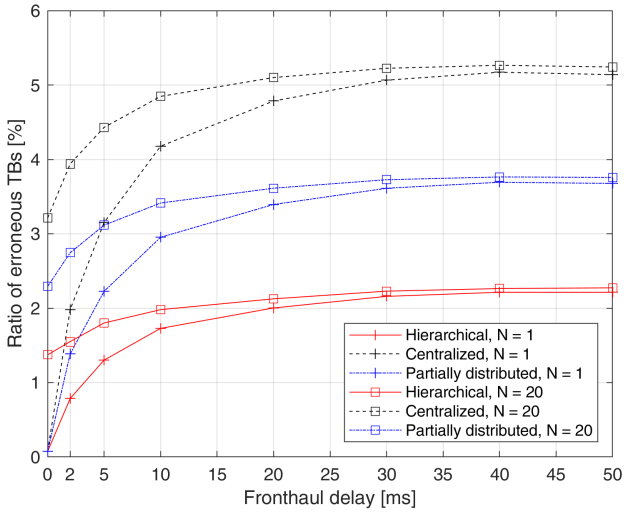


Figure 10. Ratio of transport blocks received with error and retransmitted by means of HARQ.

Fig. 10 shows the ratio of transport blocks received with error at the physical layer, i.e., the ratio of transport blocks that require retransmission by means of HARQ. The transport block error rate of all schedulers increases with the fronthaul delay and saturates when the fronthaul delay reaches about 30 ms. The saturation takes place because channel variation for this fronhaul delay is so significant that any impact of fast fading is random, and the difference between the real and reported value of CQI is statistically the same for any fronthaul delay above 30 ms. Also, the higher number of RRHs leads to a lower error rate as more opportunities for CoMP appear with a higher density of the RRHs. The proposed hierarchical scheduler reduces the error rate by about 60% and 46% compared to the centralized and partially distributed schedulers, respectively, for all investigated densities of RRHs. We can also see that a significant improvement is present even if we increase the scheduling period of the C-Sc in the BBU to N=20. The superiority of the hierarchical scheduler comes from its ability to suppress the negative impact of a fronthaul delay by scheduling a part of the UEs (nCEUEs) at the RRH.

4.8 Conclusions

In this chapter, we have outlined a new framework for the hierarchical scheduling in networks based on C-RAN architecture. The hierarchical scheduler encompasses distributed schedulers in the RRHs and centralized scheduler in the BBU. While the BBU performs long-term scheduling especially for the cell edge UEs that can benefit from the interference mitigation techniques, the distributed schedulers in the RRHs eliminates a negative impact of the fronthaul delay on the non-cell edge UEs and enables efficient handling of HARQ retransmissions. We also show that the hierarchical scheduler improves the network throughput as a result of suppression of the fronthaul delay. Also, error rate are reduced comparing to existing solutions of the scheduler.

Chapter 5

Dynamic Approach in the Hierarchical Scheduler

This chapter presents the dynamic approach of hierarchical scheduling in mobile networks with C-RAN architecture in vivid detail. Practically, to address the problem of the fronthaul delay, the hierarchical scheduler is proposed in chapter 4. In the mentioned hierarchical scheduler proposed version, the length of the scheduling period is set to a fixed value for all UEs. While the hierarchical scheduler is able to suppress the problem of a non-ideal fronthaul (i.e., the fronthaul with a non-zero delay) for scheduling in C-RAN, it also introduces some challenges. One of these challenges is the dynamicity setting of the scheduling period, N, for each UE to improve the individual UEs' throughput.

Thus, in this chapter, we propose a novel solution that enables an individual adjustment of the scheduling period for each UE to maximize the sum throughput of the UEs. We propose two approaches to derive N for individual UEs based on the UE's channel state information (CSI). The first approach derives the value of N for each UE solely from the UEs' CSI observed in the past. The second approach predicts a future evolution of the UE's CSI and, then, estimates N based on the predicted CSI.

This chapter is organized as follows: the system model is shown in section 5.1 before the chapter's problem is formulated in section 5.2. Section 5.3 presents the chapter's proposed solution, and section 5.4 summarizes the chapter's findings and primary outcomes. Finally, section 5.5 provides a brief Conclusions of the chapter.

5.1 System model

In this section, we describe the system model. We consider an area with a single BBU, L RRHs, and K UEs. We focus on the downlink, where the RRHs transmit data to the UEs. The RRHs are connected to the BBU via the fronthaul. Each UE is associated with the RRH, which provides the highest received signal to interference and noise ratio (SINR). Basically, the SINR between the l-th RRH and the k-th UE is calculated as:

$$\gamma_{l,k} = \frac{p_l \cdot g_{l,k}}{W_n + \sum_{i=1, i \neq l}^{i=L} p_i \cdot g_{i,k}} \tag{4}$$

where p_l is the transmission power of the l-th RRH, $g_{l,k}$ is the channel gain between

the l-th RRH and the k-th UE, W_n is the noise power, and $\sum_{i=1,i\neq l}^{i=L} p_i \cdot g_{i,k}$ represents the inter-cell interference from all but serving RRHs.

The CE UEs usually experience a strong inter-cell interference from adjacent RRHs. Thus, we adopt the Inter-Cell Interference Coordination (ICIC) [51] to mitigate the interference from other RRHs. In the ICIC, a set of RRHs cooperates together to mitigate the inter-cell interference for individual CE UEs. To suppress the negative effect of the inter-cell interference, the CE UEs are served via orthogonal Physical Resource Blocks (PRBs) with respect to the transmission of other RRHs in the same ICIC set. The ICIC set encompasses L_K RRHs that involved in the cooperation towards improving the k-th UE's SINR. To this end, a certain number of PRBs is allocated in each RRH from the ICIC set to serve the individual CE UEs. If the ICIC is exploited, the SINR of the CE UE is calculated as:

$$\gamma^*_{l,k} = \frac{p_l \cdot g_{l,k}}{W_n + \sum_{i=1,i\notin R_k}^{i=L} p_i \cdot g_{i,k}} \tag{5}$$

where R_k is the ICIC set of the k-th UE and the R_k includes the k-the UE's serving RRH plus all other RRHs coordinating their transmission to the k-th UE.

In our paper, we classify the UEs to the CE UEs and the nCE UEs based on the R_k. The UE is classified as the CE UE if $R_k > 1$ and as the nCE UEs if $R_k = 1$. The new RRH is added to the ICIC set for each UE if and only if the UE communication throughput is improved by adding the given RRH. Otherwise, the UE is classified as nCE UE (i.e., $R_k = 1$). In other words, the RRH is added to the ICIC set if the ratio of the number of required PRBs without the ICIC to the amount of the required PRBs with the ICIC is higher than the number of RRHs in the ICIC set. This way we guarantee that the ICIC is exploited only if the communication throughput of the UE is increased by the ICIC. Note that the UE SINR is calculated via (4) in case the UE is classified as the nCE UE and via (5) for the CE UE.

Now, let's define the amount of overhead consumed for scheduling. The channel is split into Transmission Time Intervals (TTI). In 3GPP-based mobile networks, the Physical Downlink Control Channel (PDCCH) and the Physical Downlink Shared Channel (PDSCH) are carried in each TTI. The PDCCH carries the control and scheduling information while the PDSCH carries the UEs' data. The amount of signaling overhead in PDCCH is related to the network configuration including the following factors: i) the number of scheduled UEs, ii) the CSI of individual UEs, and iii) the frequency of an update of the scheduling decision (i.e., scheduling period) [52]. Note that, in general, the amount of signaling overhead increases with an increasing number of UEs and with a decreasing UE's channel quality [52] [53]. At the same time, the amount of the signaling overhead decreases with an increasing scheduling period [54].

The signaling overhead is quantified as a number of symbols occupied by the PDCCH in the downlink subframe. The PDCCH can occupy from 1 to 3 symbols [55]. There is no standard format in 3GPP to deduce the exact PDCCH size based on the channel quality indicator (CQI) index, as there is a range of possible PDCCH sizes to be selected for any given CQI index [56]. We derive the PDCCH size and, thus, the amount of the signaling overhead, based on the CQI index. More specifically, the amount of PDCCH symbols, S_k, required at a specific CQI index is determined as:

$$S_k = \begin{cases} 2 \quad if \quad 1 \leq \text{CQI index} \leq 6, & (a) \\ 1 \quad if \quad 7 \leq \text{CQI index} \leq 15 & (b) \end{cases} \tag{6}$$

If the UE's scheduled data is transmitted over more than one TTI (i.e. $N_k > 1$), the PDCCH is not included at every TTI (see [57]). The PDCCH indicates the allocation of the PRBs for individual UEs' for the whole scheduling period. It means that for the k-th UE with a scheduling period N_k, the first TTI carries the UE's PDCCH. This PDCCH includes the signaling and control information not only for the first TTI, but for the whole scheduling period of the k-th UE (i.e., N_k). Thus, from the 2^{nd} TTI till the N_k-th TTI, the PDCCH size for the k-th UE is equal to zero.

The network throughput calculation depends on the number of scheduled PRBs and the selected modulation and coding scheme (MCS) of each UE, as detailed in 3GPP [67]. The number of scheduled PRBs is allocated to the UE during one transmission time interval (TTI). The MCS is defined by channel quality indicator (CQI) based on UE signal to noise and interference ratio (SINR). The mapping of UE CQI to UE MCS and UE MCS to UE TBS are both defined and derived in line with 3GPP [67].

5.2 Problem Formulation

The objective of this chapter is to maximize the throughput of individual UEs by enabling a dynamic adjustment of the scheduling period, N, for each CE UE. To this end, we determine the scheduling period N_k of the k-th CE UE maximizing the UE throughput C_k and, thus, the problem is formulated as:

$$N_k = \underset{N}{\operatorname{argmax}}\, C_{k,N,t_f} \;\forall\; k \in \{1, 2, \ldots, K\} \tag{7}$$

$$\begin{gathered}\text{s.t. } N \in \mathbb{Z}^+ \text{ (a)} \\ N_{min} \leq N \leq N_{max} \text{ (b)} \\ C_{k,N,t_f} \geq C_{min} \text{ (c)}\end{gathered}$$

where $C_{k,N}$ is the throughput of the k-th UE when the scheduling period length is N and the fronthual delay is t_f, the constraint (a) defines that N is a set of all positive integers, (b) defines the lower limit and the upper limits of N, N_{min} and N_{max}, respectively, and (c) defines that there is no CE UEs with zero PRB. The C_{min} equivelent to the C_{k,N,t_f} when the number of PRB is equal to one. Note that the setting of N_k is done independently for each CE UE to maximize the individual UEs throughput. It means that the sum throughput of UEs is also maximized due to the independence of C_{k,N,t_f} as a function of N among UEs.

5.3 Proposed dynamic setting of the scheduling period

This section explains two approaches for determining the UE's scheduling period. The motivation behind the dynamic scheduling period is to cope with the delay between the time when the channel quality (i.e., CSI) is reported and the time when individual UEs receive the data. Hence, the length of the scheduling period, N, should depend on 1) individual UEs' radio channel dynamicity and 2) the fronthaul delay. The length of N is determined by the C-Sc at the BBU. Therefore, the scheduling requires less computation (processing) power and the amount of signaling overhead between the BBU and the RRHs is reduced. Since the radio channel dynamicity is different for individual UEs, the value of N for each k-th UE (i.e. N_k) is also set individually. Generally, for highly dynamic radio channels with frequent and significant changes in the channel quality over time, the N_k should be set to lower values to keep the inputs to the scheduling decisions up to date. On the contrary, for the radio channels with a low dynamicity, i.e., with only rare and less significant changes in the channel quality, the value of N_k can be set to higher values as the channel characteristics remain almost the same for longer time intervals. The value of N_k is set for a future period of time denoted as T_f^*, which is expressed as:

$$T_f^* = t_f + N_{max} \tag{8}$$

In this chapter, we propose and compare two approaches for the setting of N_k. The first one, denoted as *history-based setting of N*, derives N_k solely from the individual UEs' CSI observed in the past. The second one, denoted as *prediction-based setting of N*, predicts a future evolution of the individual UEs' CSI and, then, estimates N_k based on the predicted future UE's CSI.

5.3.1 History-based setting of N

In the history-based solution, the actual value of N_k is derived solely from the k-th UEs' past CSI. The UEs' past CSI is understood as the CSI values in the time interval that is just prior to the time of the estimation (t_n), and is of a length equal to T (see Fig. 11).

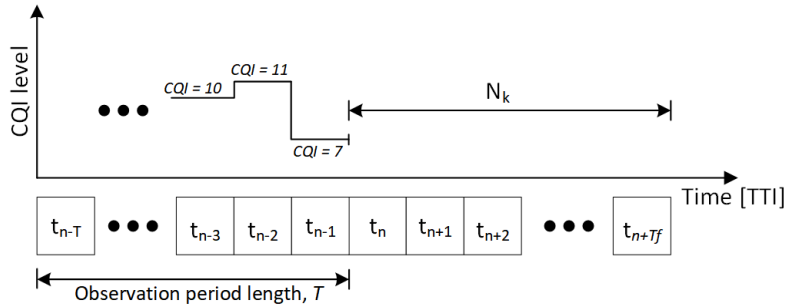


Figure 11. Evaluation of the number of changes in CQI for estimation of N.

We define the UE radio channel dynamicity via a number of CQI changes (U) resulting from the UE's channel fluctuation within an observation period. The observation period is defined as T consecutive TTIs. The U is estimated by calculating the sum of absolute values of the differences in the CQI between every two consecutive *TTI* within the observation period. To calculate U, the CQI difference for two consecutive TTIs is estimated, which is expressed as:

$$u_{k,t} = |CQI_{t-1} - CQI_t|, \ \forall \ t \in \{2, 3, \ldots, T\} \tag{9}$$

where $u_{k,t}$ is the CQI difference for the k-th UE for two consecutive TTIs within the observation period $[t_{n-T}\ t_{n-1}]$ (see Fig. 11). Thus, the U is calculated independently for individual UEs as follows:

$$U_{k,t} = \sum_{T} u_{k,t}, \ \forall \ k \in \{1, 2, \ldots, K\} \tag{10}$$

Then, for a future period of time, T_f^*, we rescale the N_k to the range of allowed values from N_{min} to N_{max} by the following equation:

$$N_k = \frac{N_{min} + [U_k - U_{k,min}]}{[U_{k,max} - U_{k,min}]}(N_{max} - N_{min}) \tag{11}$$

where N_{min} is set, by default, to 1 representing the lower limit of the UE's radio channel dynamicity, N_{max} is set up dynamically by the network operator and it depends on the overall average UEs' radio channel dynamicity. Note that the average UEs' channel dynamicity is updated regularly depending on the UEs' environmental conditions. The $U_{k,max}$ and $U_{k,min}$ are the maximum and minimum numbers of U_k, respectively and these are defined as:

$$U_{k,max} = \max\{U_{k,t}, U_{k,t+1}, \ldots, U_{k,T_s}\} \tag{12}$$

$$U_{k,min} = \min\{U_{k,t}, U_{k,t+1}, \ldots, U_{k,T_s}\} \tag{13}$$

where T_s is the communication session period. In real world applications, the $U_{k,max}$ and $U_{k,min}$ can be set based on the experienced U_k values over a long period of time.

5.3.2 Prediction-based setting of N

The second approach for determination of the scheduling period is based on prediction of the number of CQI changes in the future. In this approach, the value of N_k for a future period of time, T_f^*, is selected based on the predicted future CSI of the UE. First, we predict the future UE CSI for a period of time equal to T_f^* based on the UE CSI history. The UE CSI history is defined as a set of the CSI values in the time interval that is just prior to the time of the prediction (t_s), and is of a length equal to T_p (see Fig. 12).

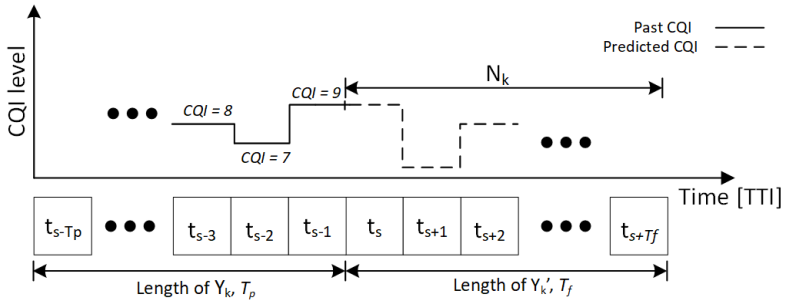


Figure 12. Evaluation of CQI prediction for estimation of N.

The CSI history is then exploited for the channel prediction. For the prediction, we choose the Auto-regressive Integrated Moving Average (ARIMA) model [58], as it

makes the prediction process flexible and reliable more than other statistical models such as the exponential smoothing model or the moving average algorithm [59]. The ARIMA model is defined by a combination of coefficients p, d, and q representing the order of the autoregressive model, the degree of differencing, and the order of the moving-average model, respectively. The ARIMA model of the input signal of the k-th UE, $Y_{k,t}$, is defined as:

$$(1 - \sum_{i=1}^{p} \varphi_i B^i)(1 - B)^d Y_{k,t} = (1 + \sum_{j=1}^{q} \theta_j B^j)\varepsilon_t \tag{14}$$

where φ_i is the coefficient of the auto-regressive component of ARIMA, θ_j is the coefficient of the Moving Average (MA) component of ARIMA, ε_t is the white noise, i is the index for lags of Y (up to p lags), j is the index for lagged forecast errors (up to q lags), and B is the lag operator. The combination of ARIMA coefficients is set individually per UE based on extensive experiments. We define the Bayesian information criterion (BIC) [60] for assessing the combinations of coefficients p_k, d_k, and q_k for the k-th UE. The combination of coefficients achieving the lowest BIC is considered for the UE's CSI prediction process. The lowest BIC is selected as the BIC contains the Maximum Likelihood Estimation, which penalizes free parameters to combat overfitting. Hence, the ARIMA-based prediction for the input Y_k targets:

$$Y'_{k,t} = \arg\min_{p,d,q} BIC(ARIMA(Y_{k,t})) \tag{15}$$

where Y'_k is the k-th UE predicted CSI (see Fig. 12). Note that the UE CSI is presented in Fig.12 by UE CQI levels for sake of simplicity as the UE CQI is one of the major components of the UE CSI. After the ARIMA predicts the UE's future CSI, the N_k is determined based on Y'_k. The N_k calculation is done by examining the impact of each possible scheduling period from N_{min} to N_{max} on the UE's throughput. Intuitively, the resulting throughput of individual UEs for each option of N is different based on individual UEs' predicted radio channel dynamicity (i.e, UEs' CSI). Hence, the value of N yielding the maximum throughput is selected as N_k, i.e.,:

$$N_k = \underset{N \in \{N_{min}, \ldots, N_{max}\}}{\operatorname{argmax}} C_k(Y'_{k,t}, N, t_f) \;\forall\; k \in \{1, 2, \ldots, K\} \tag{16}$$

where C_k is the k-th UE's throughput when the scheduling period is equal to N and the fronthual delay is equal to t_f.

5.4 Performance evaluation

In this section, the proposed dynamic hierarchical scheduler is evaluated in MATLAB system-level simulator. To this end, simulation scenarios, parameters, and competitive algorithms are presented in the next subsection. Then, the simulation results are described and discussed.

5.4.1 Simulation scenario and models

In our proposal, we implement the 3GPP LTE-A compliant model, which is based on LTE-A physical layer as described in [61]. In the downlink transmission, we assume the orthogonal frequency division multiple access. The propagation model between the RRHs and the UEs follows the Urban Microcell model [62] with Rayleigh and Rician fadings (see [63]) to model shadowing and fast fading, respectively. We consider the fronthaul delay to be within the range from 0 ms to 30 ms, which encompasses ideal, near-ideal, sub-ideal, and non-ideal fronthaul delay ranges, as described in [64].

We adopt the proportional fair scheduler [65] as the basis for the resource scheduling among the UEs. This scheduler provides a good trade-off between scheduling fairness and network throughput [66]. Furthermore, we assume the full buffer traffic model to investigate the performance of our proposal under heavy load conditions. Transmission errors and subsequent retransmissions are handled by HARQ process, as specified in 3GPP [67]. The simulation scenarios and parameters are listed in Table 5.1.

Table 5.1. Parameters and sitting for the chapter simulation.

Parameters and Assumptions	***Value***
Simulation case	3GPP Urban Micro Scenario [62]
Carrier frequency	2 GHz
System bandwidth	20 MHz (100 PRBs)
Number of BBU	1
Number of RRH	up to 100
Number of RAUs	up to 15
RRH transmit power	27 dBm
Fading model	Rayleigh and Rician
Lognormal shadowing std. dev.	4 dB (LOS), 7.82 dB (NLOS) [62]
Scheduler	Proportional fair
Traffic model	Full buffer
Fronthaul delay	0; 2; 5; 10; 20; 30 ms

5.4.2 Competitive algorithms and performance metrics

Our proposed history-based (denoted as *proposal-history*) and prediction-based (denoted as *proposal-prediction*) determination of the scheduling period N_k are compared with the following three competitive approaches: i) static hierarchical scheduler (denoted as *static hierarchical*), where the N is set to a fixed value for all UEs regardless of the individual UEs' radio channel state, ii) static centralized scheduler (denoted as *static centralized*), where the scheduling process takes place only at the BBU, iii) partially distributed static scheduler (denoted as *partially distributed*), where the scheduling process is performed at either the BBU or the RAU depending on the individual RRHs' fronthaul delay, as described in [30].

We also depict maximum throughput achieved under a perfect prediction of the channel for determination of N_k. This upper bound is denoted as *proposal-optimum* in the figures. The optimal N is determined based on a perfect prediction of individual UEs' CSI, i.e., the optimal N corresponds to selection of N_k for each UE assuming perfect knowledge of future UE CSI. Even if the optimum cannot be determined in real world networks, it illustrates the upper bound of the throughput and allows us to demonstrate an efficiency of both proposed solutions.

The performance of the proposed algorithms and all competitive solutions are assessed by means of two performance metrics. The first metric is the network throughput which is understood as the sum throughputs of all UEs. The second metric is the mean absolute percentage error (ϵ), which is a measure of the prediction accuracy of *ARIMA* (see, e.g., [58]). The purpose of the prediction error analysis is to evaluate the UEs' CSI prediction accuracy and to assess the hierarchical dynamic prediction performance in comparison to other counterparts. The mean absolute percentage error is expressed as:

$$\epsilon = \frac{1}{N_{max}} \sum_{t=1}^{N_{max}} \left| \frac{OP_t - Y'_t}{OP_t} \right| \tag{17}$$

where OP_t and Y'_t are the optimum value of UE CSI and predicted value of UE CSI at time t, respectively.

5.4.3 Simulation results

In this section, we present and discuss the simulation results to demonstrate the benefits of the proposed dynamic scheduling period with respect to other state-of-the-art solutions. The impact of the fronthaul delay on the network throughput is shown in Fig. 13. In general, the network throughput is decreasing with the increasing of the fronthaul delay for all investigated schedulers. This performance degradation with the fronthaul delay occurs, as the fronthaul delay leads to an outdated CSI, which directly negatively impacts the scheduling decision and, thus, a decrease in the network throughput is observed. If the fronthaul delay is increased from 0 to 30 ms, the network throughput of *static centralized* scheduler and *partially distributed* scheduler are decreased in comparison with the *proposal-prediction* scheduler by roughly up to 30% and 27%, respectively.

Both proposed dynamic approaches (i.e., *proposal-history* and *proposal-prediction*) suppress the negative impact of the outdated CSI. Thus, the network throughput is declined only by around 9%, if the fronthaul delay is increased from 0 ms to 30 ms. The reason for this behavior is that the scheduling at the BBU can be set in RRHs for

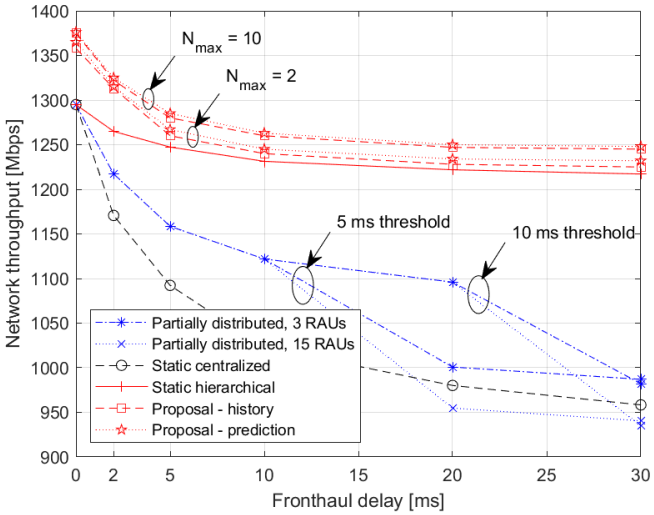


Figure 13. Impact of the fronthaul delay on the network throughput for individual scheduling approaches.

the nCE UEs if the fronthaul delay leads to a notable change in the channel quality. However, CE UEs are scheduled solely in the BBU and still benefit from the ICIC gain, as the RRHs are not allowed to change the scheduling decision to protect the ICIC gains. With respect to the *static hierarchical* scheduler, both *proposal-history* and *proposal-prediction* approaches improve the network throughput by around 6%. This gain is due to the fact as the scheduling period is dynamically set based on individual UE CSI to minimize the negative impact of the outdated CSI.

Another dimension of the performance is interpreted in Fig. 14, where the impact of the scheduling period (i.e., N for the static scheduling period and N_{max} for the dynamic scheduling period) on the network throughput is investigated. Fig. 14 shows that the network throughput initially increases with N for all presented solutions. The reason behind this increase is that the scheduling control information is sent only every N-th TTI instead of each TTI as in the case with $N = 1$. Then, the network throughput starts gradually decreasing for $N > 2$, with different slopes for individual scheduling options. With respect to the *static centralized* and *partially distributed* solutions, the *static hierarchical*, *proposal-history* and *proposal-prediction* approaches suppress the network throughput degradation for a longer duration of N. The hierarchical scheduler allows an adjustment of the scheduling in both BBU and RRH, and the benefit of the reduced overhead is preserved.

Generally, the dynamic scheduling approaches (i.e., *proposal-history*, *proposal-prediction*) outperform the static scheduler solutions (i.e., *static hierarchical*, *static centralized* and *partially distributed*). For instance, the *proposal-prediction* approach outperforms the *static hierarchical*, the *static centralized*, and the *partially distributed* solutions by approximately 6%, 17%, and 11%, respectively. The proposed solutions' outstanding performance is because the actual value of N for each UE is dynamically set according to the channel dynamicity over time reflecting the stability of the communication channel. Furthermore, the *proposal-prediction* option outperforms the *proposal-history* approach by around 2%. This small gain is at a cost of an additional computational complexity. The superior behavior of the *proposal-prediction* approach is not surprising as the *proposal-prediction* approach estimates the values

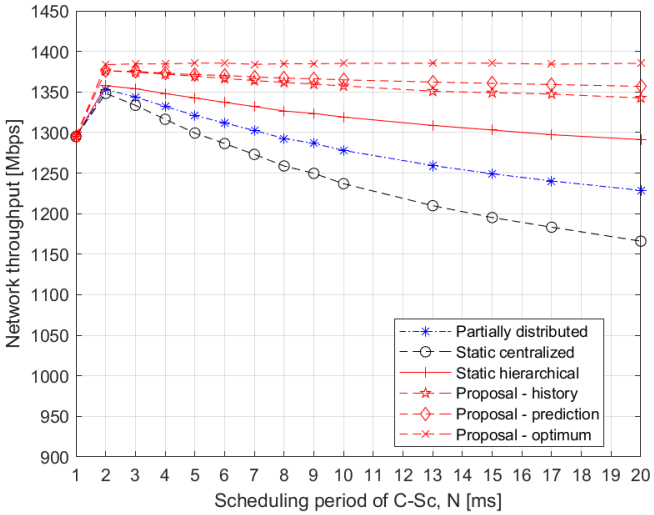


Figure 14. Impact of the scheduling period of centralized scheduler on the network throughput for different scheduling options.

of N for individual UE based on the predicted future of UE CSI instead of the history of UE CSI as in the *proposal-history* solution. The maximum network throughput achieved by the *proposal-optimum* solution is roughly only 2.3% and 3.2% above the maximum achieved by *proposal-prediction* approach and *proposal-history* approach for the ideal fronthaul with zero delay, respectively.

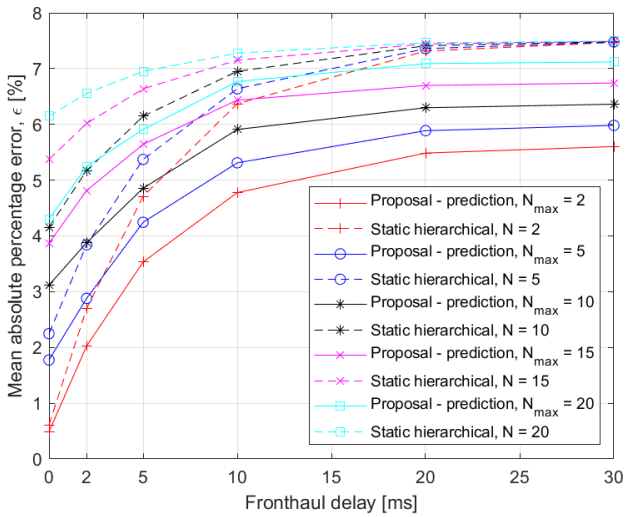


Figure 15. Impact of the fronthaul delay on the mean absolute percentage error, ϵ, of the proposal-prediction scheduler and the static hierarchical scheduler for different values of the scheduling period.

Fig. 15 illustrates the mean absolute percentage error, ϵ, of selection of N over the fronthaul delay. In general, the ϵ for all presented scenarios is below 8% and starts increasing with the fronthaul delay, because of the outdated CSI negative impact. The *proposal-prediction* approaches reach a lower values of the ϵ by up to around 24% in

comparison with the *static hierarchical* solutions. The reason is that the actual value of N is adapted based on the UE CSI instead of keeping it static for all deployed UEs. Furthermore, the values of ϵ go up with the increasing of the scheduling period (i.e., N, N_{max}) for all presented approaches as the scheduling information is not up to date for the later TTIs within the scheduling period. The values of ϵ of individual *static hierarchical* solutions are saturated after approximately 20 ms fronthaul delay. This saturation takes place as the channel variations at this fronthaul delay ($\sim$20 ms) are so large in such that the impact of the fast-fading is random and the probability of the channel information that utilized for scheduling being outdated is increased considerably.

5.5 Conclusions

In this chapter, we have presented a dynamic hierarchical scheduling in mobile networks with C-RAN architecture. The dynamic hierarchical scheduler estimates the UE's scheduling period via two proposed approaches: i) history-based and ii) prediction-based. The simulations confirm that the proposed dynamic schedulers increases the network throughput and notably outperforms the centralized scheduler and the partially distributed scheduler by around 30% and 27%, respectively.

In the next chapter, this work will be complemented with a profound resources allocation analysis of the proposed scheduler solution considering a dynamic resource allocation of erroneous packets retransmission. Furthermore, the proposed hierarchical scheduler will be evaluated in further realistic scenarios with a variety of mobility patterns.

Chapter 6

HARQ Retransmission in the Hierarchical Scheduler

This chapter proposes a framework for the HARQ resource pre-allocation in the hierarchical scheduler based on the C-RAN. One of the key challenges in the proposed hierarchical scheduler in the previous chapters (i.e., chapter 4 and chapter 5) is determining the amount of resources to be pre-allocated for the HARQ process. Since we allocate the resources for the HARQ in advance for multiple upcoming TTIs (i.e., in relatively long-term compared to traditional scheduling), we use the term "pre-allocation" in this chapter.

For the proposed hierarchical scheduler in the previous two chapters, the number of pre-allocated RBs for the HARQ process is set to a fixed amount regardless of the actual required retransmissions. Unfortunately, this fixed resource pre-allocation degrades the performance of the proposed hierarchical scheduler. Therefore, this chapter presents a dynamic resource pre-allocation scheme to cope with HARQ retransmissions. The chapter's objective, in other words, is to determine the required amount of resources to be pre-allocated for HARQ retransmissions depending on the individual UEs' actual needs. Our contributions in this chapter are summarized as follows:

- We propose a comprehensive framework for the HARQ resource pre-allocation in the C-RAN, considering the hierarchical scheduling to maximize the goodput of UEs via minimizing the transport block loss rate and maximizing the resource pre-allocation accuracy.
- We develop two distinct approaches to determine the required amount of pre-allocated resources for the HARQ and optimize them jointly to improve the resource pre-allocation accuracy.
- We also show that the idea of resource pre-allocation is not limited to hierarchical scheduling only, but it is extended to be applicable also to other existing types of schedulers, such as centralized and partially distributed schedulers.

This chapter is organized as follows: the system model is shown in section 6.1 before the chapter's problem is formulated in section 6.2. Section 6.3 presents the chapter's proposed solution, and section 6.4 summarizes the chapter's findings and primary outcomes. Finally, section 6.5 provides a brief Conclusions of the chapter.

6.1 System model

This section describes the system model based on the C-RAN architecture, the background on the hierarchical scheduler, and the HARQ process. Each part is described in the following subsections.

6.1.1 C-RAN based architecture

We assume a single BBU interconnected with L RRHs via the fronthaul. In our study, we consider the fronthaul from the perspective of fronthaul latency for data retransmissions, and we do not expect any errors to be originated at the fronthaul. Furthermore, K UEs are deployed randomly over an area covered by the RRHs. The UEs are individually associated with the RRH, providing the highest Signal to Interference and Noise Ratio (SINR). We further classify the UEs into K_{CE} CE UEs and K_{nCE} nCE UEs so that $K = K_{CE} + K_{nCE}$. This classification is based on the experienced SINR via individual UEs. Since, intuitively, the CE UEs experience more substantial inter-cell interference from adjacent RRHs than the nCE UEs, an Inter-Cell Interference Coordination (ICIC) technique [51] is adopted to mitigate such interference as been explained in chapter 2.

6.1.2 The hierarchical scheduler

The hierarchical scheduler, the basis of this work, splits the scheduling process into two tiers: a centralized scheduler (C-Sc) and a distributed scheduler (D-Sc), as described in chapter 4. The C-Sc runs in the BBU and D-Sc in the RRH. The D-Sc handles data transmission and manages the allocation of resources for the nCE UEs, as these UEs suffer less from inter-cell interference. However, the C-Sc schedules data transmission for the CE UEs, enabling a high level and long-term scheduling, reinforced by an awareness of the mutual interference among individual RRHs. The long-term scheduling is understood as a scheduling decision not only for a single TTI but for N consecutive TTI (i.e., $NxTTI$). The N is set individually for each k-th CE UE (i.e., N_k) as considered in our previous work [90] to maximize the sum goodput of the UEs. The value of the scheduling period, N_k, is set based on: i) the individual predicted future CSI of the individual CE UEs' radio channel dynamicity and ii) the fronthaul delay. Using one channel prediction tool, we predict the future UE CSI based on the UE CSI history record. The UE CSI history record is understood as CSI values in the time interval just before the time of the prediction and is presented as the input of the channel prediction tool, as explained in [90].

6.1.3 HARQ process

The idea behind the HARQ is to model a system that detects the received erroneous data transport block and then requests the needed retransmissions in case of the erroneous data transport block. As we explained in chapter 2, the retransmissions can be classified as adaptive and non-adaptive, where we have two ways to implement HARQ for downlink: synchronous HARQ and asynchronous HARQ [68]. In the adaptive HARQ, the MCS and other transmission attributes (such as the redundancy version and sub-carrier) have the option to be updated for each retransmission, where the transmission attributes are fixed or pre-defined in the non-adaptive context.

In practice, the HARQ course of action for the CE UEs and nCE UEs in the hierarchical scheduler are quite different. For the nCE UEs, the HARQ process is handled in a standard way at the RRH, as the nCE UEs are scheduled directly by the D-Sc in the RRH, and the fronthaul does not have any direct negative impact on the HARQ process. The standard way is understood so that the resources for the HARQ retransmissions are scheduled directly by the D-Sc at the RRH based on ACKs/NACKs received from individual nCE UEs served by the given RRH. In the case of the NACK, the D-Sc allocates any available RB(s) that are not dedicated to CE UEs at the moment of the retransmission (see [90] for more details). The HARQ process is being more complicated for the CE UEs due to the fronthaul delay intervention. The HARQ process would be significantly prolonged due to the transmissions taking place over the fronthaul in case the HARQ would be processed in the BBU.

To understand our proposal for pre-allocating resources for CE UEs' HARQ, described in the following sections, let us first define $R^*_{k,n}$ and $R_{k,n}$ as the estimated and actual numbers of the required RBs for the k-th CE UE's potential retransmissions in the n-th TTI within N_k, respectively. Based on that, we introduce a new performance evaluation parameter; the resource pre-allocation efficiency of individual CE UEs, ζ_k. This parameter is defined as the mean absolute percentage error (MAPE) of the amount of pre-allocated resources for the HARQ to estimate how far the $R^*_{k,n}$ value from $R_{k,n}$ value and is expressed as:

$$\zeta_k = \frac{1}{N_k} \sum\nolimits_{n=1}^{N_k} \frac{|R^*_{k,n} - R_{k,n}|}{R_{k,n}} * 100 \tag{18}$$

Then, considering also retransmissions due to HARQ, we can define the goodput experienced by the k-th UE as:

$$G_k = N_{Sym} N_{SC} \left(\sum\nolimits_{l=1}^{RB_k} CR_{l,k} log_2 M_{l,k} - \sum\nolimits_{r=1}^{R_k} CR_{r,k} log_2 M_{r,k} \right) (1 - OH_k) \tag{19}$$

where N_{Sym} represents the number of OFDM symbols per one RB, N_{SC} stands for the number of subcarriers per RB, $CR_{l,k}$ is the coding rate applied at the l-th RB allocated to the k-th UE, $M_{l,k}$ corresponds to the number of possible modulation states based on the modulation used for data transmission at the l-th RB allocated for the k-th UE, RB_k is the number of all RBs allocated to the k-th UE per second, R_k represents the number of RBs allocated only for retransmissions of the k-th UE per one second, and finally OH_k stands for the overhead due to various signaling and control messages to serve the k-the UE (expressed as a ratio between the amount of resources allocated for signaling to the resources allocated for data transmissions).

6.2 Problem formulation

This chapter aims to maximize the sum goodput of the CE UEs in the hierarchical scheduler architecture based on the C-RAN. This objective can be attained by optimizing the resource scheduling efficiency for the CE UEs' potential retransmissions. The main challenge in such optimization is estimating the required scheduling resources for any potential CE UEs' data retransmissions. One way to address this challenge is to perform a dynamic adjustment of the pre-allocated amount of RBs instead of pre-allocating a fixed amount of resources for the CE UEs' HARQ requirements, which is suggested in our earlier works in chapter 4 and chapter 5.

Basically, the individual CE UEs' HARQ resource requirements depend on many factors, including *i*) the CE UE radio channel condition, *ii*) the fronthaul delay, and *iii*) the scheduling period, N_k. The motivation behind presenting this work is to fulfill the varied resources needed to be pre-allocated for individual CE UEs' HARQ. Thus, the problem is formally formulated as:

$$\begin{aligned} \max_{R_k^*} \quad & \sum_{k=1}^{K_{CE}} G_k \\ \text{s.t.} \quad & \text{a) } 0 \leq R_k^* \leq R_{max} \\ & \text{b) } 0 \leq N_k \leq N_{max} \end{aligned} \tag{20}$$

where R_k^* is the amount of RBs pre-allocated for the HARQ of the k-th CE UE, R_{max} represents the maximum affordable RBs depending on the system bandwidth and the number of served UEs, and N_{max} stands for the maximum length of centralized period set up dynamically by the network operator based on the overall average CE UEs' radio channel dynamicity. The constraint a) in (20) limits the possible values of R_k^* to be pre-allocated to each k-th CE UE while constraint b) gives the lower and upper limits on the scheduling period N_k. Note that the calculation of the R_k^* is made independently for each CE UE to maximize the individual CE UEs' goodput and, hence, also to maximize the sum goodput of all UEs. The dynamic setting of the amount of the HARQ pre-allocated RBs also minimizes the unexploited pre-allocated resource in case of free-error delivery data.

The formulated problem can be classified as a non-linear integer programming problem. The reason is that the dependence of goodput on the amount of pre-allocated resources is non-linear with respect to the channel quality experienced by individual UEs. Moreover, both the objective function and the constraints are integer (discreet) variables as: *i*) the goodput in objective function strictly depends on selected coding rate and modulation and is limited to several discreet values (see (19)), and *ii*) constraints on the scheduling period N and the number of pre-allocated resource blocks for HARQ are also integer variables.

In general, the resource allocation formulated as non-linear integer problem is usually solved by various deterministic algorithms [69] [70] [71] [72] [73] or evolutionary algorithms [34] [74] [75] [76]. The main limitation of the deterministic algorithms is an enormous complexity since finding the optimal solution in an ample search space is infeasible while limiting the search to only a subset of the search space results in a poor and far from optimal solution [77]. Moreover, the uncertainty in the channel quality in the future (several TTIs for which the pre-allocation of resources for HARQ is done) adds another dimension to the complexity. Hence, utilizing such deterministic algorithms would make our proposal computationally complex, time-consuming, and impractical, especially for large-scale problems with multiple UEs. Therefore, the deterministic algorithms would not be a good fit for our problem, which demands swift and instantaneous pre-allocation scheduling decisions for the available resources in a horizon of milliseconds. Along similar lines, the evolutionary algorithms are not suitable for our problem as these are known to suffer from slow convergence [78].

In contrast to these traditional tools, the heuristic algorithms can be designed to be fast because they do not require a complete search in the search space [77]. Thus, the heuristic algorithms are practical, serving as fast and feasible solutions for planning and scheduling problems (i.e., see [77]) as targeted in our work (i.e., finding the R_k^*). Therefore, we adopt the heuristic approach to solve the defined optimization problem. The proposal is described in detail in the following section.

6.3 Dynamic resource pre-allocation for the HARQ

This section describes the proposed approach for the HARQ resource pre-allocation in the C-RAN based on the hierarchical scheduler. We tackle in this work the resource pre-allocation problem precisely for the CE UEs, since the retransmissions for the nCE UEs are handled directly by the RRHs, as explained in chapter 4 and chapter 5. Furthermore, the resource scheduling decision for the nCE UEs is not negatively impacted by the fronthaul delay. We first outline a high-level principle of the resource pre-allocation for the CE UEs' HARQ retransmissions. Then, we describe two proposed approaches for determining the number of pre-allocated RBs for the CE UEs' HARQ retransmissions.

6.3.1 High-level principle

Let us first illustrate the possible scenarios that can occur during the pre-allocation of resources for the CE UEs' HARQ and discuss the motivation behind the proposed work. To cope with the fronthaul delay affecting the HARQ process of the CE UEs, a part of RBs is pre-allocated in the BBU for any retransmission needs of all CE UEs transmitting data at any given TTI.

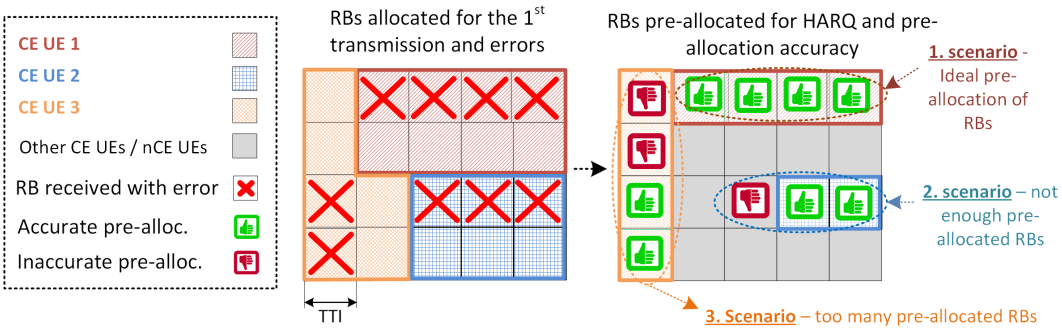


Figure 16. High-level overview of the HARQ resource pre-allocation when the number of pre-allocated RBs is equal to, lower than, and larger than the actual number of RBs required by the HARQ process.

Based on the amount of pre-allocated RBs in comparison with the actual required RBs for the HARQ retransmissions, we can distinguish three scenarios (see Fig.16):

- *1. scenario*: The pre-allocated amount of RBs is precisely equal to the amount of actually required RBs. Thus, there are no further actions to be taken since the number of pre-allocated RBs exactly matches the required RBs.
- *2. scenario*: The C-Sc at the BBU pre-allocates an insufficient number of RBs for retransmitting all erroneous data transport blocks. Consequently, some of the retransmitted data is delivered with an additional delay due to the fronthaul. In this scenario, the HARQ process is performed in the C-Sc at the BBU instead of the D-Sc in the RRHs. This additional delay postpones the retransmission process to later TTIs and increases the overall delay. The situation is getting even more critical for delay-sensitive services, such as URLLC in 5G mobile networks [79]. For such services, the retransmitted data transport blocks might even get rejected once the data retransmission deadline is expired due to the introduced additional delay in the HARQ process.

- *3. scenario*: The C-Sc pre-allocates too many RBs for CE UEs' HARQ retransmissions. This scenario alleviates the bottleneck of additional HARQ delay and reduces the probability that some retransmitted data transport blocks are not delivered in time. However, relatively, a lower number of RBs remains available for the new data transmission of UEs (both CE UEs and nCE UEs) due to the over-booking of scheduling resources for the CE UEs' HARQ retransmissions.

In order to rectify the problem of the HARQ resource pre-allocation and estimate a proper number of RBs in the C-RAN with the hierarchical scheduler, we propose a flexible and dynamic resource pre-allocation approach based on the estimated retransmission requirements of individual CE UEs. The following subsection illustrates the proposed framework and details the resource pre-allocation principle.

6.3.2 The proposed pre-allocation of resources for HARQ retransmissions

This section describes the proposed solution for estimating the number of pre-allocated RBs for any possible retransmissions of the k-th CE UE, R_k^*. The value of R_k^* is set independently for each CE UE based on the individual CE UEs' retransmission needs. The number of retransmissions is set up to a pre-defined maximum limit, δ_{max}. Note that the value of R_k^* is not set only for a single TTI but for N_k consecutive TTIs (i.e., for the whole scheduling period of the k-th CE UE).

In order to estimate the R_k^*, part of our proposal is to predict the evolution of individual CE UEs' CSI. This prediction is exploited via the Auto-regressive Integrated Moving Average model (ARIMA) [58]. Compared with other statistical models, such as the exponential smoothing model and the moving average algorithm, the ARIMA makes the prediction process more reliable and flexible [59]. Fundamentally, the ARIMA model is defined by a combination of coefficients p, d, and q representing the order of the autoregressive model, the degree of differencing, and the order of the moving-average model, respectively. This combination of ARIMA coefficients is adjusted individually for each CE UE based on extensive experiments on CE UE's CSI (see [90] for more details).

To assess the combinations of coefficients for individual CE UEs (i.e., p_k, d_k, and q_k), we exploit the Bayesian information criterion (BIC) [60]. The combination of coefficients achieving the lowest BIC is selected for the CE UE's CSI prediction process [60]. Note that the selected BIC (lowest BIC) contains the maximum likelihood estimation, which penalizes free parameters to combat overfitting. After individual CE UEs' CSI is predicted, the block error rate (BLER) and then the required resources for any retransmission at any given TTI (i.e., R_k^*) can be estimated.

The value of the R_k^* depends on three distinct parameters: *i*) the length of the scheduling period of the k-th CE UE, N_k, *ii*) the number of required retransmissions for the k-th CE UEs, δ_k, where $0 \leq \delta_k \leq \delta_{max}$, and *iii*) the number of pre-allocated RBs for each retransmission (i.e., τ retransmission) at the n-th TTI, $R_{k,n}^{\tau}$, where $R_{k,n}^* = \sum_{\tau=1}^{\delta_k} R_{k,n}^{\tau} \; \forall \; \tau \in (1, 2, \ldots, \delta_k)$.

Let us investigate these three parameters in more detail. First is the scheduling period's length N_k, which is determined according to the individual CE UEs' channel quality information, as introduced in our prior work [90]. Second, the number of required retransmissions, δ_k, which depends on the loss rate of the data transport blocks. The loss rate is calculated based on the CE UE's MCS, which is set to keep

the BLER below a certain threshold. Third and lastly, the number of pre-allocated RBs for each retransmission, $R^{\tau}_{k,n}$. The $R^{\tau}_{k,n}$ can be variant for each retransmission since the adaptive HARQ retransmission is assumed in our proposal, as explained in the system model.

Based on the parameters mentioned above for estimating the R^{*}_{k} (i.e., N_k, δ_k, and $R^{\tau}_{k,n}$), we can categorize our proposed solution into two distinct aspects: 1) the CE UE's error rate, and 2) the CE UE's scheduling period. Both aspects are described vividly in the following sub-sections. Then, both aspects are combined to make the HARQ resource pre-allocation estimation more precise.

The error rate aspect

In this subsection, we determine the amount of the pre-allocated resource for the HARQ based on the individual CE UEs' error rate. The individual CE UEs' error rate is one of the primary data transmission metrics influencing the average number of retransmissions for a successful data transport block delivery. Each transmission/retransmission is defined by a delivery state S so that $S = 1$ for the data packet received correctly and $S = 0$ for delivery with error(s). Hence, the resulting amount of pre-allocated resources in this aspect, R^{er}_{k}, can be derived via the Bernoulli random variable and the Poisson Binomial Distribution [80]. Since the received data transport block of the k-th CE UE has a delivery state at every TTI along the scheduling period N_k, it can be written as a vector V_k in such that: $V_k = \{S_1, ..., S_{N_k}\}$. We define all possible combinations of delivery states (i.e., vectors) of the k-th CE UE with the scheduling period N_k as a sample space (SS_{N_k}) in such that:

$$SS_{N_k} = \{V^1_k, ..., V^e_k, ..., V^{Q_k}_k\} = \begin{pmatrix} S^1_1 & S^1_2 & \cdots & S^1_{N_k} \\ S^2_1 & S^2_2 & \cdots & S^2_{N_k} \\ \vdots & \vdots & \ddots & \vdots \\ S^{Q_k}_1 & S^{Q_k}_2 & \cdots & S^{Q_k}_{N_k} \end{pmatrix} \tag{21}$$

where $e \in \{1, 2, ..., Q_k\}$ is defined as a single vector outcome out of $Q_k = 2^{N_k}$ possible vectors for N_k TTIs. Note that the selected vector of the delivery states of the k-th CE UE, V^s_k, over other outcome vectors in the SS_{N_k} depends on the predicted BLER of individual CE UEs at every TTI along the scheduling period, N_k. Hence, the amount of HARQ pre-allocated resources for the selected V^s_k is estimated as:

$$R[V^s_k] = \sum\nolimits_{n=1}^{N_k} V^s_{k,n} R^{\tau}_{k,n} \tag{22}$$

where $R[V^s_k]$ represents the number of pre-allocated RBs for HARQ along N_k TTI in case part, or all of data transport blocks of the k-th CE UE are received with an error. Note that the $R^{\tau}_{k,n}$ is estimated based on the predicted future evolution of the individual CE UEs' CSI at each TTI along the scheduling period, N_k.

Let us discuss the way to estimate R^{er}_{k}. Since the probability of error of individual CE UEs' is randomly distributed along the scheduling period N_k, we exploit the Poisson Binomial Distribution (PBD) to estimate the error probability distribution of the vector V^s_k for the k-th CE UE. Fundamentally in the PBD, two subsets of vectors are defined from the SS_{N_k}; a_k and a^c_k. The subset a_k is understood as a collection of vectors that occur for the k-th CE UE with the scheduling period equal to N_k. The subset a^c_k (i.e., $a^c_k = \{SS_{N_k} - a_k\}$) is complementary to a_k and includes vectors that are not occurring for the k-th CE UE with the same scheduling period,

N_k. The occurred and not occurred vectors in the respective subsets a_k and a_k^c depend on the error rate for each vector in the SS_{N_k}, $\eta_{k,e}$, and a pre-defined error rate threshold, ψ.

The value of the $\eta_{k,e}$ is expressed as $\eta_{k,e} = \prod_{n=1}^{N_k} \varphi_{k,e,n}$, where $\varphi_{k,e,n}$ is the predicted BLER for the vector e at every TTI along the scheduling period N_k of the k-th CE UE. Notice that each vector in the SS_{N_k} is indicated by an index: e. Therefore, we can define subset A_k and subset A_k^c as they correspond to the vectors' indices in the subset a_k and the subset a_k^c, respectively. Based on that, the classification of the vectors' indices either belonging to subset A_k or subset A_k^c subset as:

$$\eta_{k,e} = \begin{cases} \eta_{k,i} & \text{if } \eta_{k,e} \geq \psi_{N_k} \qquad \forall\ i \in A_k \\ \eta_{k,j} & \text{if } \eta_{k,e} < \psi_{N_k} \qquad \forall\ j \in A_k^c \end{cases} \tag{23}$$

where the i, and j refer to the individual vectors' indices within the subset A_k and subset A_k^c, respectively. The value of the ψ_{N_k} for the scheduling period N_k is determined based on the average of the experienced vectors' error rate over a long period of time so that:

$$\psi_{N_k} = \overline{\eta_{k,e}}(T_s|N_k) \tag{24}$$

where $T_s|N_k$ is the communication session period when the scheduling period length is equal to N_k. The communication session period, i.e., T_s, is defined as the period of time (in seconds) for a series of interactions between two communication endpoints (i.e., UE and RRH/BBU) that occur during the span of a single connection. Note that the ψ_{N_k} is calculated independently for each scheduling period, and then it varies depending on the N_k. Hence, the error probability distribution of the vector V_k^s for the k-th CE UE is written as follows:

$$P_k[V == V_k^s] = \sum\nolimits_{A_k} \prod\nolimits_{i \in A_k} \eta_{k,i} \prod\nolimits_{j \in A_k^c} (1 - \eta_{k,j}) \tag{25}$$

Then, the number of pre-allocated scheduling resources for the HARQ over N_k TTI is calculated as:

$$R_k^{er} = P_k[V_k^s]\, R[V_k^s] \tag{26}$$

The proposed pre-allocation based on the error rate aspect is summarized in Algorithm 1. Note that the algorithm is illustrated for any k-th CE UE. The algorithm starts with definition of SS_{N_k} matrix in line with (21) giving all possible combinations of delivery states (i.e., vectors) of the k-th CE UE with the scheduling period N_k (see line No. 1 in Algorithm 1). Based on SS_{N_k} matrix, the amount of HARQ pre-allocated resources for the selected vector V_k^s (i.e., $R[V_k^s]$) is calculated according to (22) (line No. 2). Then, $\eta_{k,e}$ (i.e., error rate for V_k^e) and ψ_{N_k} (i.e., pre-defined error rate threshold of the scheduling period N_k) is estimated according to (23) and (24), respectively (line No. 3). In the next step, the error probability distribution for any e-th combination out of Q_k is estimated via the Poisson Binomial Distribution (lines No. 4-9). Finally, the number of pre-allocated scheduling resources for the HARQ, R_k^{er}, is calculated via (26) (line No. 12).

Algorithm 1 The Error Rate Aspect

Define matrix SS_{N_k} in line with (21)
Calculate $R[V_k^s]$ according to (22)
Estimate $\eta_{k,e}$ and ψ_{N_k} via (23) and (24), resp.
for $e = 1 : Q_k$ **do**
    if $\eta_{k,e} \geq \psi_{N_k}$ **then**
        $e \in A_k$
    else if $\eta_{k,e} < \psi_{N_k}$ **then**
        $e \in A_k^c$
    end if
end for
Estimate $P_k[V_k^s]$ according to (25)
Calculate R_k^{er} according to (26)

The scheduling period aspect

Now, let us turn our attention to the scheduling period aspect for estimating the required pre-allocated resources for HARQ retransmission(s), R_k^{sp}. The importance of this aspect comes from the fact that the CE UE's scheduling period reflects two factors in its estimation: 1) the CE UE's radio channel dynamicity and 2) the fronthaul delay (see [90] for more details). Both factors are essential in the way for achieving network reliability and fulfilling retransmission requirements.

To estimate the number of pre-allocated resources for the HARQ retransmission of CE UEs, we reformulate the sample space in (21) into the number of error states, ρ, for each vector. In other words, the error state ρ indicates the number of TTIs in which the errors occur for the individual CE UEs along N_k TTIs. It means the sample space of error states, ES, is expressed as: $ES = \{0, 1, ..., \rho, ..., N_k\}$. Hence, the case $\rho = 0$ indicates an error-free transmission event(s) over the N_k, and the case $\rho = N_k$ refers to the event(s) with an error in each TTI within the scheduling period N_k. Then, the probability mass function of the error for each ρ in the ES with N_k scheduling period of the k-th CE UE is:

$$P_{k,N_k}(\rho) = \prod\nolimits_{\vartheta_\rho} P(\eta_{k,e}) \tag{27}$$

where ϑ_ρ represents the group of vectors that have the same number of errors ρ in such that $\vartheta_\rho = SS_{N_k}[ES == \rho]$, and $P(\eta_{k,e})$ is the probability of the vector's error rate, which is calculated as:

$$P(\eta_{k,e}) = \prod\nolimits_{n=1}^{N_k} P(\varphi_{k,e,n}) \tag{28}$$

where $P(\varphi_{k,e,n})$ is the probability of BLER at every n TTI within N_k for the vector e. The size of the ϑ_ρ is indicated by Υ_ρ, and represents the number of vectors that have the same number of erroneous TTI, i.e., ρ.

Each vector in the ϑ_ρ is indicated by V_{k,ϑ_ρ}^q, where q is the vector index in the ϑ_ρ in such that: $\vartheta_\rho = \{V_{k,\vartheta_\rho}^1, ..., V_{k,\vartheta_\rho}^q, ..., V_{k,\vartheta_\rho}^{\Upsilon_\rho}\}$. Each V_{k,ϑ_ρ}^q requires a number of pre-allocated RBs for the HARQ retransmissions as estimated in (**??**). Hence, the number of pre-allocated RBs required for the HARQ of the k-th CE UEs with the scheduling period, N_k and ρ erroneous TTI is written as follows:

$$R_{k,\Upsilon_\rho} = \overline{R}\{V_{k,\vartheta_\rho}^1, ..., V_{k,\vartheta_\rho}^q, ..., V_{k,\vartheta_\rho}^{\Upsilon_\rho}\} \tag{29}$$

where R_{k,Υ_ρ} is the average number of pre-allocated RBs for every V^q_{k,ϑ_ρ} in ϑ_ρ. Based on the selected number of erroneous TTI, ρ^s, the number of pre-allocated RBs is estimated. The selected ρ of the k-th CE UE, ρ^s_k, for a received data transport block depends on the predicted BLER of individual CE UEs at every TTI along the scheduling period, N_k. Then, the number of pre-allocated RBs for the HARQ over N_k TTI with ρ^s_k errors is calculated as:

$$R^{sp}_k = R_{k,\Upsilon_{\rho^s_k}} P_{k,N_k}(\rho^s_k) \tag{30}$$

The proposed solution for the scheduling period aspect is managed by Algorithm 2 as follows. First, the sample space SS_{N_k} defined in (21) is reformulated into the number of error states ρ (line No. 1 in Algorithm 2). Then, the probability mass function of the error for each ρ (i.e., $P_{k,N_k}(\rho)$) and the probability of the vector's error rate ($P(\eta_{k,e})$) is estimated in (27) and (28), respectively (lines No. 2-3). Based on that, the number of pre-allocated RBs required for the HARQ of the k-th CE UEs with all possibilities of errors R_{k,Υ_ρ}) is estimated according to (29). Finally, R^{sp}_k is calculated in line with (30).

Algorithm 2 The Scheduling Period Aspect

1: Reformulate the sample space in (21) into ρ
2: Calculate $P_{k,N_k}(\rho)$ according to (27)
3: Estimate $P(\eta_{k,e})$ according to (28)
4: Estimate R_{k,Υ_ρ} according to (29)
5: Calculate R^{sp}_k from (30)

Joint optimization of the aspects

This subsection describes the combination of both HARQ resource's pre-allocation approaches presented in previous subsections (i.e., the error rate aspect and the scheduling period aspect) in order to make the estimation more precise. Following three cases of the HARQ resource pre-allocation can take place: i) $R^{er}_k = R^{sp}_k$, ii) $R^{er}_k > R^{sp}_k$, and iii) $R^{er}_k < R^{sp}_k$.

The first case explains when pre-allocation estimation outcomes are identical in both aspects (i.e., $R^{er}_k = R^{sp}_k$); therefore, no further action is needed since the number of pre-allocated RBs is validated by both aspects. For the second case (i.e., $R^{er}_k > R^{sp}_k$), we pre-allocate the larger of both values (i.e., R^{er}_k) since the amount and placement of erroneous TTI are estimated in advance (i.e., the selected vector, V^s_k). The solution for the third case is quite different since the larger value, i.e., R^{sp}_k, shows only the number of erroneous TTIs and does not contain information on which TTI the erroneous data is placed (i.e., the selected number of erroneous TTI, ρ^s). Therefore, we initially pre-allocate the number of resources indicated in R^{er}_k. Then, in addition, the difference in the number of pre-allocated RBs between R^{er}_k and R^{sp}_k, which is denoted as R^d_k: $R^d_k = R^{sp}_k - R^{er}_k$, is also considered for the HARQ needs. In other words, the number of resources R^d_k is pre-allocated as shared resources for any k-th CE UE retransmission needs. This way, we can fulfill all CE UEs' retransmissions needs and, simultaneously, improve the scheduling resource utilization since the shared pre-allocated resources (i.e., R^d_k) can be fully re-scheduled in case of CE UEs' error-free data delivery.

Thus, the number of pre-allocated RBs for the HARQ in the t-th TTI is determined as follows:

$$R^*_{k,t} = \max_{n=N_k} \{R^{er}_k, R^{sp}_k\} \tag{31}$$

Finally, the total amount of pre-allocated resources for the HARQ retransmissions of all CE UEs over a communication session period, T_S, is:

$$R^* = \sum_{t=1}^{T_s} \sum_{k=1}^{K_{CE}} R^*_{k,t} \tag{32}$$

The integration of both aspects of the proposal is managed by Algorithm 3. At the beginning, the initialization of the algorithm is done by setting N_{max}, R_{max}, δ_{max}, and K_{CE} representing the maximum length of the scheduling period, the maximum number of resources available for pre-allocation, the maximum number of possible retransmissions, and the total number of CE UEs, respectively (see line No. 1 in Algorithm 3). After that, the centralized scheduling period N_k is estimated for all CE UEs (line No. 2). Then, Algorithm 1 and Algorithm 2 are executed to obtain R^{er}_k and R^{sp}_k, respectively (lines No. 4-5). In the sequel, the following two cases of the HARQ resource pre-allocation can take place: i) $R^{er}_k \geq R^{sp}_k$ or ii) $R^{er}_k < R^{sp}_k$. Based on this, R^*_k is calculated for the k-the CE UE (lines No. 6-10). The steps in lines No. 3-11 are repeated for each k-th CE UE. Finally, the overall number of pre-allocated resources R^* are calculated based on (32) (line No. 12).

Algorithm 3 Joint Optimization of Aspects

1: Initialization: N_{max}, R_{max}, δ_{max}, K_{CE}
2: Estimate $N_k \; \forall \; k$
3: **for** $k = 1 : K_{CE}$ **do**
4: Execute Algorithm 1 for k-th CE UE (obtain R^{er}_k)
5: Execute Algorithm 2 for k-th CE UE (obtain R^{sp}_k)
6: **if** $R^{er}_k \geq R^{sp}_k$ **then**
7: $R^*_{k,t} \Longleftarrow R^{er}_k$
8: **else if** $R^{er}_k < R^{sp}_k$ **then**
9: $R^*_{k,t} \Longleftarrow R^{sp}_k$
10: **end if**
11: **end for**
12: Calculate R^* according to (32)

6.3.3 Discussion on suitability of various types of HARQ

In this section, we discuss a suitability of our proposal for various HARQ types and we outline any potential modifications that need to be done. In general, the HARQ types can be classified according to several criteria:

- *Synchronous vs. asynchronous HARQ* - In the synchronous HARQ, each HARQ process occurs at predefined times relative to the initial transmission. Thus, signaling of the HARQ process number is unnecessary and can be inferred from transmission timing. In the asynchronous HARQ, the retransmissions can occur at any time. Thus, the HARQ process number is necessary to correctly associate each retransmission with the corresponding initial transmission. In other words, the main difference for both HARQs is the retransmission timing. In our work, we adopt the asynchronous HARQ, since it is used in 5G networks [35]. Still, our proposal can be easily adapted also for the synchronous HARQ and only time of individual retransmissions may need to be changed for the synchronous HARQ while the number of pre-allocated resources is unaffected.

- *Adaptive vs. non-adaptive HARQ* – The adaptive HARQ allows to change modulation, coding rate, or number of resource blocks for retransmissions while the non-adaptive HARQ keeps these parameters the same as for the first transmission. In our work, we assume adaptive HARQ process, since it is used in 5G (please see [35]). As the result, the number of pre-allocated resource blocks is modified with respect to initial transmission depending on the current channel quality. In principle, even the non-adaptive HARQ process can be utilized for our proposal. In this case, however, the number of pre-allocated resource blocks for individual retransmissions should be the same as in case of the initial transmission of data.

- *HARQ type I-III* – In HARQ type I (chase combining), the same information and parity bits are retransmitted each time. In HARQ Type II (incremental redundancy), multiple different sets of code bits are generated for the same information bits used in each transmission. The HARQ type III is based on HARQ type II, but each retransmitted packet is self-decodable. In our work, we do not specify HARQ type, as these rather relates to the physical layer and, thus, are not relevant to the proposed pre-allocation of resource targeting higher layers. Hence, all three HARQ types can be used for our proposal.

6.4 Performance evaluation

The performance is evaluated in the MATLAB system-level simulator. To this end, the simulation setup, competitive algorithms, and performance metrics are introduced in the following subsections. Then, the simulation findings are presented and comprehensively discussed.

6.4.1 Simulation scenario

We assume a square area of 1000x1000 m encompassing a single BBU located in the middle, up to 100 RRHs, and 200 UEs deployed randomly with uniform distribution. Each UE is associated to the RRH, providing the highest SINR. In this work, we implement the 3GPP 5G-compliant model described in [61]. The orthogonal frequency division multiple access is assumed for the downlink transmission. The channel between any UE and RRH, including shadowing and fast fading, is modeled according to the Urban Micro-cell model [62] with mixed Line-Of-Sight (LOS) and Non-LOS communication (see [81]). We adopt ICIC for interference management, as explained in [51].

We assume a realistic fronthaul with latencies between 0 ms to 30 ms in line with the Small Cell Forum model [64], which is widely adopted by researchers. The proportional fair scheduler [65] is adopted as a basis for resource scheduling among the UEs, as this scheduler provides an adequate trade-off between network goodput and fairness [66]. For the traffic model, we select the full buffer model to examine the performance of our proposal under heavy-load network conditions.

The BLER calculation is based on the Cyclic Redundancy Check (CRC) evaluation, which is attached to transport blocks to detect the error at the receiver side (i.e., UE). The incremental redundancy (IR) HARQ with a 1/3 turbo encoder is considered. The reason behind adopting IR-HARQ is its higher coding gain compared to the chase combining HARQ [82]. The retransmitted data transport blocks are sent with an initial coding rate of 1/2 or 3/4, and the maximum number of simultaneous downlink HARQ processes is limited to 8 [61].

The HARQ adopts the N-channel stop-and-wait protocol, offering low buffering requirements and low acknowledgment (ACK) / negative acknowledgment (NACK) feedback overhead. In particular, the data packet must be delivered with a packet error rate (PER) of less than 10^{-5}, either with or without retransmission(s), as detailed in [81].

The HARQ RTT is scaled by the TTI length, which is assumed as the default time unit in this work. The TTI length depends on the number of OFDMA symbols and the subcarrier spacing of the OFDMA modulation is $t_{TTI} = N_S(1/\Delta_f + t_{CP})$, where N_S is the number of OFDMA symbols per TTI, Δ_f represents subcarrier spacing, and t_{CP} stands for the duration of a cyclic prefix. We adopt a standard system configuration with the carrier spacing equal to 15kHz and the normal duration of t_{CP} equal to approximately 4.7 us. Hence, considering 14 OFDMA symbols per one TTI, the length of each TTI is equal to 1 ms. Note that the proposal can be adapted for any sub-carrier spacing and any TTI defined for 5G networks (see Table I in [83] with 5G numerologies).

Our proposed resource pre-allocation approach focuses on one part of RTT: decreasing the retransmission delay part. Since we propose HARQ retransmission at the RRH instead of the BBU, our proposal shortens the retransmission delay by at least double fronthaul. Note that the retransmission delay is understood as an ad-

Table 6.1. Parameters and sitting for the chapter simulation

Parameters	***Values***
Number of BBU, RRH, UEs	1, up to 100, 200
α_C	1.2
Number of retransmissions	up to 3 attempts
TTI length	1 ms
HARQ RTT	8 TTI
Antenna configuration	Single input single output
Centralized sched. period	1; 5; 10; 15; 20 ms

ditional delay caused by data transport blocks needing retransmission(s). It means the other HARQ RTT components are considered negligible for this purpose. Because the 5G networks have scalable TTIs, we assume 1 ms as a TTI length in this work. Since our proposal pre-allocates part of the scheduling resources for individual CE UEs' HARQ needs at RRH(s) for immediate retransmission(s) without the BBU intervention, any retransmission(s) is admitted and scheduled during individual CE UEs' scheduling period (i.e., N_k). Otherwise, the data packet is assumed to be lost. Some simulation parameters and settings for this chapter are listed earlier in table 5.1, the rest are presented in table 6.1.

6.4.2 Competitive algorithms and performance metrics

To show the gain of the proposal, we compare the results of the proposed scheduling with related competitive approaches. The proposed scheduling settings comprise: *i*) the scheduling period selection, as explained in [89] [90], and *ii*) the dynamic pre-allocation of resources for HARQ retransmission, as proposed in this work. The following approaches are compared:

1) *Centralized scheduler (CS)*: The conventional scheduling process is done only at the BBU for all UEs without any functional split (i.e., split options 6-8 acc. to 3GPP [17]).

2) *Partially–Distributed Scheduler (PDS)*: The conventional scheduling is performed at either the BBU or the partially distributed radio aggregation units (RAU) depending on the individual UEs' fronthaul delay, as proposed in [30]. Since the authors in [30] do not specify any deployment scheme of the RAUs, a realistic case with the RAUs are collocated with the underlying RRH closest to the cluster's center of all underlying RRHs is assumed.

3) *Hierarchical scheduler (HS)*: The conventional hierarchical scheduler based on our previous works [89] [90], where only fixed pre-allocation of resources for the HARQ process is done. This way, we demonstrate the impact of the proposed dynamic pre-allocation.

4) *CS - Proposal*: The conventional CS implemented with our proposed scheduling settings (i.e., dynamic pre-allocation of resources for HARQ retransmissions).

5) *PDS - Proposal*: The conventional PDS implemented with our proposed scheduling settings.

6) *HS - Proposal*: The conventional HS implemented with our proposed scheduling settings.

7) *HS − Optimum*: We also show a theoretical upper bound of the hierarchical scheduler in terms of network goodput and the CE UEs goodput. The scheduling settings are dynamic and optimally adjusted for individual CE UEs. The HS-Optimum comprises two parts: *i*) an estimation of the optimal scheduling period length, N_{opt}, and *ii*) estimation of the optimal amount of pre-allocated resources for the HARQ process R^*_{opt}, while perfect prediction of the future signal characteristic for the monitored period (50 ms in our case) is assumed to estimate both. The first part, N_{opt}, is determined individually for each CE UE according to its "channel dynamicity". Channel dynamicity is understood to be significant in the changes in CQI per a monitored period of time. In general, the more significant the CQI changes within the monitored period, the lower N is set for the CE UE and vice versa. To find N_{opt}, we subsequently set the scheduling period N from 1 to N_{max} and N yielding the maximum goodput is selected as the optimal individually for each CE UE. Regarding the second part, the HS-Optimum always pre-allocates the exact amount of RBs needed for any retransmission (please see scenario No. 1 in Fig. 16 representing the ideal pre-allocation). Even though the *HS-Optimum* cannot be determined in real-world networks, it represents the upper bound performance and allows us to demonstrate the efficiency of the proposed solutions.

The performance of the competitive solutions and the proposed algorithms are assessed by four performance metrics:

1) *Network goodput*: Sum goodput over both CE UEs and nCE UEs. The calculation of network goodput is based on (19). Note that the signaling overhead (i.e., OH_k in (19)) is also taken into consideration when the goodput is estimated. Basically, the overhead size usually varies between 7% and 14% of the downlink subframe size.

2) *CE UEs goodput*: Sum goodput only over the CE UEs for whom the proposal is tailored specifically.

3) *Downlink transport block loss rate*: The ratio of data transport blocks not retransmitted to the CE UEs within 8 consecutive TTI due to a lack of scheduling resources divided by the total number of received transport blocks.

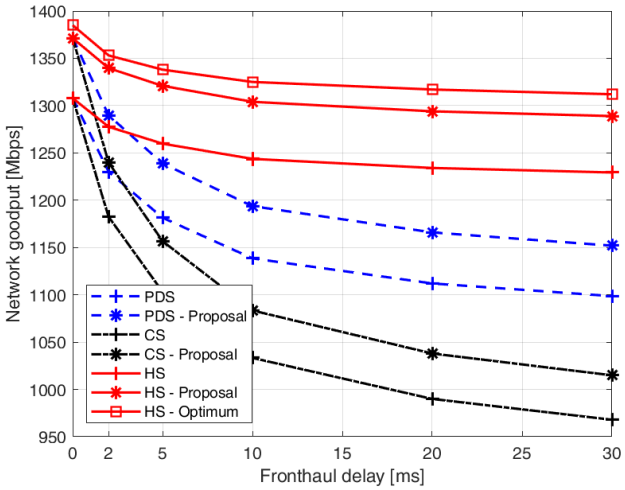
4) *MAPE*: The evaluation of the number of pre-allocated RBs accuracy of the thesis proposal compared to other counterparts (see (18) for the calculation of the MAPE).

6.4.3 Simulation results

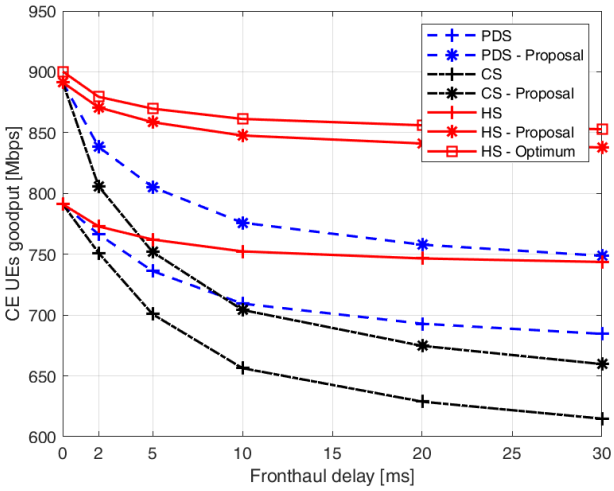
This subsection summarizes the chapter's findings and contributions made. Let us start with the impact of the fronthaul delay on the goodput of all UEs (i.e., the network goodput) as Fig. 17a, and also on the goodput of only CE UEs (see Fig. 17b). Disregarding the scheduler type, both the network goodput and the CE UEs goodput gradually decrease with the fronthaul delay increasing. This is because the fronthaul delay postpones the required scheduling information, i.e., channel quality reports and delivery of the scheduling decision to the RRHs. The fronthaul delay impairs the *PDS*, *CS*, and *HS* approaches' goodput (i.e., the CE UEs and the network) more significantly with respect to the proposed ones (*PDS - Proposal*, *CS - Proposal*, and *HS - Proposal*), especially for higher fronthaul delays. More specifically, for the *PDS*, *CS*, and *HS* approaches, the network goodput and the CE UEs goodput are notably decreased compared to the optimum hierarchical scheduler approach (i.e., *HS - Optimum*) by up to 39% if the fronthaul delay is increased from 0 to 30 ms (see Fig. 17). At the same time, the performance gap between *HS - Proposal* and *HS - Optimum* is only by up to 2%. The gain of the hierarchical scheduler is attained via an efficient suppression of the negative fronthaul delay impact as the centralized scheduling decision can be adjusted in the RRHs for the nCE UEs in case the fronthaul delay leads to a notable change in the channel quality. At the same time, the CE UEs can still benefit from ICIC gain, as these are scheduled solely by the BBU.

Another important finding is that the *HS - Proposal* outperforms the *HS* by up to 13% for both the network goodput and the CE UEs goodput (see Fig. 17). Besides, the added value of our proposal is that we can improve the performance of conventional approaches, i.e., *CS* and *PDS*, if they exploit our pre-allocation algorithm. In particular, the CE UEs goodput and the network goodput of *CS-Proposal*, and *PDS-Proposal* approaches outperform the *CS*, and *PDS* approaches' by up to around 11%. This improvement is because the dynamicity of the HARQ resource pre-allocation minimizes the probability of transport blocks getting lost due to the lack of pre-allocated resources. Moreover, the dynamicity of the HARQ resource pre-allocation decreases the amount of unexploited resources since the HARQ pre-allocated resources vary based on the individual CE UEs' actual needs. However, in the *CS*, *PDS*, and *HS* approaches, the amount of the HARQ pre-allocated resources is fixed. Based on that, the probability of data transport blocks getting lost due to a lack of resources is notably increased since there is no flexibility in the amount of HARQ pre-allocated resources.

The impact of the prolonged scheduling period in the C-Sc on the network goodput and the CE UEs goodput is investigated in Fig. 18a and Fig. 18b, respectively. Intuitively, the longer the scheduling period is, the less signaling overhead is required. In Fig. 18, we observe that both the network and the CE UEs goodput increase with the prolonging of N until the maximum goodput is reached at some point. Then, the goodput starts decreasing. In the first phase (i.e., the goodput raising phase from 1 ms to 2 ms scheduling period), the network goodput and the CE UEs goodput are increased as the signaling overhead related to the scheduling is notably reduced and more resources remain for new data transmissions. Moreover, the channel quality for low values of N is generally stable, and the loss in goodput due to a higher error rate resulting from potentially outdated channel state information for scheduling is negligible. In the second phase (i.e., the goodput declining phase from about 2 ms scheduling period and onward), the goodput starts decreasing gradually with N



(a)

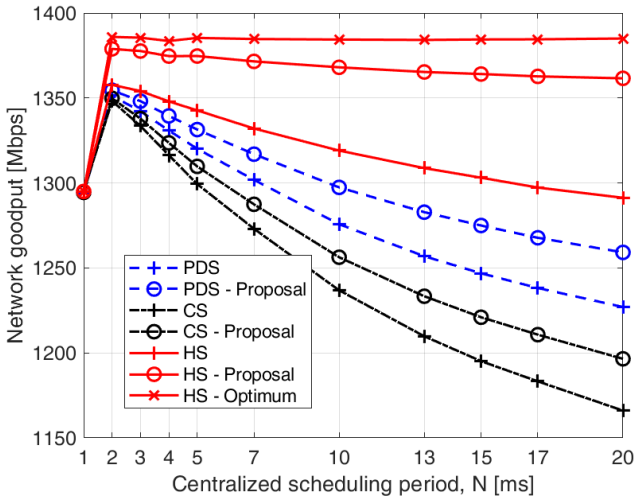


(b)

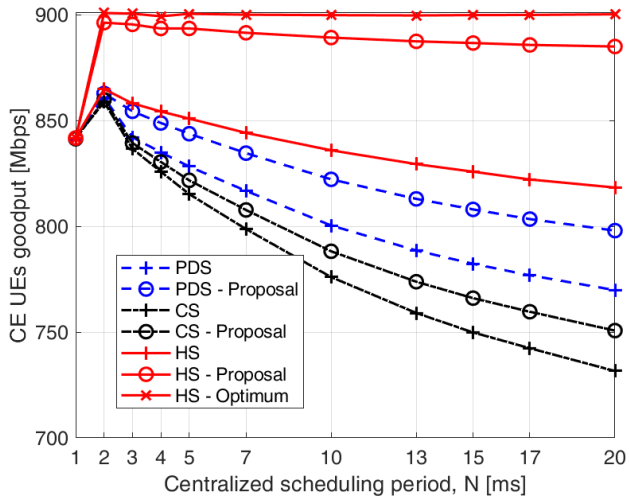
Figure 17. Impact of fronthaul delay on the network goodput (a) and on the CE UEs goodput (b) for centralized, partially distributed, and hierarchical schedulers (note that N is set dynamically up to 20 ms).

as the impact of outdated channel state information becomes more significant and dominates the gain introduced by the overhead saving.

Fig.18 also indicates that the degradation of goodput in case of *HS - Proposal* are suppressed to be below 5% and 2%, if N is set to 1 ms and 20 ms, respectively. This is because the scheduling period and the retransmission of pre-allocated resources are dynamically set based on individual CE UEs CSI and the HARQ resources need on the way to minimize the negative impact of the outdated CE UEs CSI. Note that the hierarchical scheduler does not allow adjusting the scheduling resources for CE UEs at the RRHs, as multiple RRHs serve these resources, and uncoordinated scheduling updates by these RRHs would lead to potentially strong interference. However, the goodput accomplished by *PDS* and *CS* severely decreases to around 15% for the same range of N.



(a)



(b)

Figure 18. Impact of scheduling period on the network goodput (a) and on the CE UEs goodput (b) for centralized, partially distributed, and hierarchical scheduler (fronthaul delay=0 ms).

Our results also cast a new light on the HARQ resource's pre-allocation options. Generally, *HS - Proposal*, *CS - Proposal*, *PDS - Proposal* outperform the conventional approaches (i.e., *HS*, *CS*, *PDS*) by nearly up to 9%. The proposal approaches outstanding performance is not surprising as the amount of the HARQ RBs for individual CE UEs is dynamically set according to their actual needs. Overall, the *HS-Proposal* approach is the one that obtained the most robust results in comparison with the *HS - Optimum*. The results show that the *HS-Proposal* is dropped only by around 2% compared to the *HS -Optimum*, while the *HS* is fallen by 10%.

Fig. 19 shows the probability of data transport blocks being lost due to insufficient pre-allocated resources for HARQ retransmissions (i.e., within up to 8 TTI). Fig. 19 reveals that the transport block loss probability increases with the fronthaul delay for all schedulers because the fronthaul delay negatively influences the scheduling

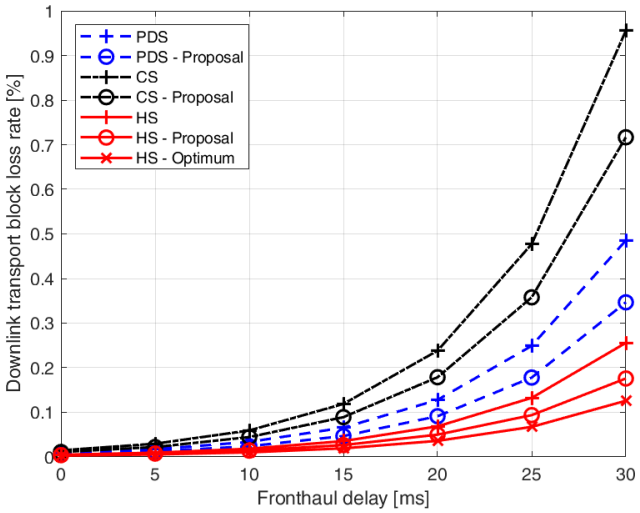


Figure 19. Impact of the fronthaul delay on the transport block loss rate.

decision. Intuitively, a higher error rate is observed for high fronthaul. However, the *HS - Proposal* significantly reduces the loss rate by up to 83% and 65%, compared to the *PDS* and *CS*, respectively, due to the dynamicity of the HARQ resource pre-allocation. Moreover, a notable improvement is observed even if we compare the *HS - Proposal* with the *HS*, where our proposal reduces the loss rate by roughly up to 30% for the longer fronthaul delay. The superiority of the dynamic HARQ approach for hierarchical scheduler comes not only from its ability to suppress the negative impact of a fronthaul delay by scheduling part of the UEs (i.e., nCE UEs) at RRHs but also from the dynamic amount of RBs that are assigned for the HARQ process based on individual CE UEs radio channel characteristics. However, even better results are achieved using our proposed scheduling setting on both *CS* and *PDS*, where the loss rate is remarkably reduced by up to 38% and 27%, respectively.

Fig. 20 shows the impact of the fronthaul delay and the scheduling period prolongation on the MAPE of the amount of the HARQ pre-allocated resources, ζ. The ζ for all shown scenarios starts increasing with the fronthaul delay due to the negative impact of outdated CSI. Our approach, *HS - Proposal*, reaches the ζ of up to 57% lower compared to the *HS*, regardless of the length of the scheduling period. This is because the actual value of retransmission pre-allocated resources is adjusted according to the CE UEs' HARQ actual needs rather than keeping it static for all CE UEs. Furthermore, the ζ values increase with the scheduling period (i.e., N) for all presented approaches as the scheduling information is not up to date for the later TTIs within the scheduling period. Still, the proposed dynamic pre-allocation of resources decreases ζ by roughly three times compared to static allocation.

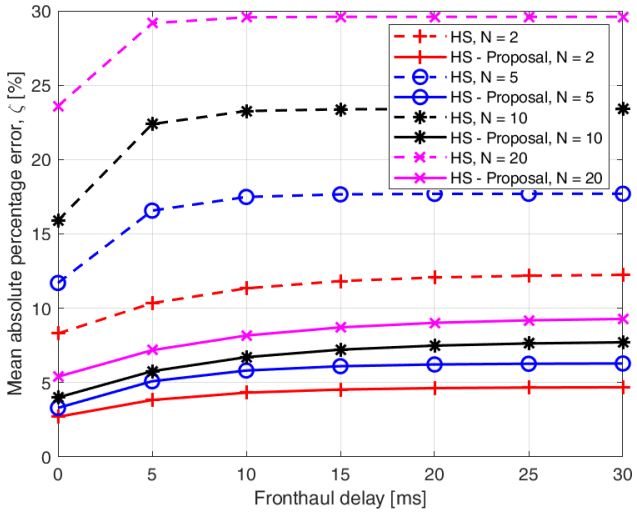


Figure 20. Impact of the fronthaul delay on the MAPE of the proposed hierarchical scheduler and the conventional hierarchical scheduler for different values of the scheduling period.

6.5 Conclusions

In this chapter, the dynamic resource pre-allocation framework for hierarchical scheduling in mobile networks with C-RAN architecture is presented. The dynamic resource pre-allocation calculates the pre-allocated resources for the HARQ by combining two distinct approaches: 1) the error rate and 2) the scheduling period. In both aspects, we derived analytical expressions for estimating the amount of resources needed. Our simulation results illustrate that the proposed dynamic pre-allocation scheduler increases the goodput with other presented schedulers by around 39% and, at the same time, minimizes the transport block loss rate and mean absolute percentage error of the amount of pre-allocated resources for the HARQ by 38% and 57%, respectively.

Chapter 7

Conclusions

This thesis has delved into the intricacies of resource allocation strategies within the dynamic landscape of dynamic functional splitting C-RAN for 5G, specifically focusing on addressing the challenge posed by fronthaul delay. By exploring various strategies and solutions, this study has contributed valuable insights into optimizing resource utilization in the context of C-RAN architectures. Therefore, this chapter presents a synopsis of the thesis, followed by an elucidation of the research contribution derived from the presented work. Bringing this chapter to a close, a direction for future research is outlined.

7.1 Thesis Summary

In this thesis, an overview of the scheduling approaches based on C-RAN for 5G has been presented. Motivated by the limitations of the existing scheduling solutions, a new framework for hierarchical scheduling in networks with C-RAN architecture has been proposed for real-time self-optimization of mobile networks. Several objectives are outlined in pursuit of this goal, and corresponding solutions have been furnished.

The first part of this thesis (i.e., chapter 4) describes a solution for resource scheduling in 5G based on C-RAN. One of the main limitations of the mobile networks based on C-RAN is fronthaul delay between BBU and RRHs. The consideration of fronthaul delay as a key challenge has added a layer of complexity to the resource allocation dynamics, emphasizing the critical need for an adaptive and efficient scheduling framework. Thus, we have proposed hierarchical scheduling as a solution to overcome the current limitation by encompassing resource scheduling into distributed units and centralized units. While the centralized unit performs the long-term scheduling (i.e., scheduling period, N), especially for the cell-edge users that can benefit from the interference mitigation techniques, the distributed units eliminate a negative impact of the fronthaul delay on the non-cell-edge users and enable efficient handling of the error correction. The hierarchical scheduling notably improves the network throughput (up to 26%) via suppressing the negative impact of the fronthaul delay. This gain can be further increased by several percent via a dynamic setting of the scheduling period of the centralized unit.

The second part of the thesis (i.e., chapter 5) focuses on maximizing the throughput of individual UEs via enabling a dynamic adjustment of the scheduling period, N, for each CE UE (i.e., the dynamic hierarchical scheduler). The motivation behind the dynamic scheduling period is to cope with the delay between the time when the channel quality (i.e., CSI) is reported and the time when individual UEs receive the data. We propose and compare two approaches for setting N_k in this solution. The first approach, denoted as the history-based setting of N, derives the value of N_k solely from the individual UEs' CSI observed in the past. The second one, denoted as the prediction-based setting of N, predicts a future evolution of the individual UEs' CSI and, then, estimates N_k based on the predicted future UE's CSI. The simulations confirm that the proposed dynamic schedulers increase the network throughput and notably outperform the centralized and partially distributed schedulers by around 30% and 27%, respectively.

Lastly, the third part of the thesis (i.e., chapter 6) presents a solution for resource scheduling based on C-RAN for delay-sensitive services such as HARQ in the thesis proposed hierarchical scheduling. This part of the thesis has proposed the dynamic resource pre-allocation framework for hierarchical scheduling in mobile networks with C-RAN architecture for tackling such a problem. The dynamic resource pre-allocation calculates the pre-allocated resources for the HARQ by combining two distinct approaches: 1) the error rate and 2) the scheduling period. The error rate approach determines the amount of the pre-allocated resource for the HARQ based on the individual CE UEs' error rate, while the scheduling period approach reflects two factors in its estimation: 1) the CE UE's radio channel dynamicity and 2) the fronthaul delay. Analytical expressions for estimating the amount of resources needed have been derived in both aspects. Simulation results illustrate that the proposed dynamic pre-allocation scheduler increases the goodput with other presented schedulers by around 39% and, at the same time, minimizes the transport block loss rate and mean absolute percentage error of the amount of pre-allocated resources for the HARQ by 38% and 57%, respectively.

7.2 Research contributions

We have presented the thesis proposed solutions in many published papers, appearing in various sorts, such as magazines, journals, conferences, and technical reports. The research contributions of these papers and the related objectives in each paper are presented as follows:

- A new framework for resource scheduling in 5G based on C-RAN has been proposed, combining the benefits of both centralized scheduling in the BBU and distributed scheduling in the RRHs to suppress the negative impact of the fronthaul delay. The proposed solution (i.e., hierarchical scheduler) notably improves the network throughput and reduces the received transport block error rate. This solution, described in chapter 4, published in [89] and fulfills the Objective No. 1.

- In Chapter 5, we have presented a dynamic hierarchical scheduling in mobile networks with C-RAN architecture. The dynamic hierarchical scheduler estimates the UE's scheduling period via two proposed approaches: i) history-based and ii) prediction-based. The simulations confirm that the proposed dynamic schedulers increase the network throughput compared to the proposed solution in chapter 4. This solution, vividly described in Chapter 5, and published in [90], fulfills Objective No. 2.

- We further extended previous work in chapter 4 and chapter 5 by proposing a dynamic resource pre-allocation framework for hierarchical scheduling in mobile networks with C-RAN architecture. The dynamic resource pre-allocation calculates the pre-allocated resources for the HARQ individually per UE by combining two distinct approaches: 1) the error rate and 2) the scheduling period. In both aspects, we derived analytical expressions for estimating the amount of resources needed for HARQ. Our simulation results illustrate that the proposed dynamic pre-allocation scheduler increases the goodput with other presented schedulers and, at the same time, minimizes the transport block loss rate and mean absolute percentage error of the amount of pre-allocated resources for the HARQ. This solution, described in chapter 6 and published in [91], fulfills Objective No. 3.

7.3 Future Research Directions

The exploration of resource allocation with dynamic functional splitting in C-RAN for 5G networks opens avenues for future research. As we delve into this dynamic landscape, several promising directions emerge, offering opportunities for further investigation and advancements. Practically, several future directions and trends can be anticipated for scheduling resources in 5G, considering the hierarchical scheduling in mobile networks with C-RAN architecture, which could include the following:

- **Machine Learning (ML) and artificial intelligence (AI) Integration:** Machine learning and artificial intelligence can be a promising integration on the hierarchical scheduling to be incorporated for self-optimization. The fact that these technologies can analyze vast amounts of data and, based on that, make real-time scheduling decisions is helping to enhance the overall hierarchical scheduling network performance in terms of efficiency, flexibility, and adaptability of resource management. These technologies can be applied to address the dynamic and complex nature of 5G networks, where diverse services with varying requirements coexist.

  For instance, ML algorithms can analyze historical data and real-time network conditions to predict traffic patterns and demand. This information can contribute to dynamically pre-allocating resources based on predicted future requirements. Furthermore, AI can facilitate cross-layer optimization by considering interactions between different network layers, enabling more holistic and efficient resource scheduling.

- **Energy-Efficient Scheduling:** An energy-efficient resource allocation model can be a perfect research extension for this work to obtain better bandwidth performance, minimize network operations' environmental impact, and improve the sustainability of 5G networks. In the proposed hierarchical scheduling, energy consumption can be adopted as a parameter in user resource scheduling decisions. Dynamically allocating network resources based on traffic demand and user requirements can help reduce unnecessary power consumption. Moreover, techniques such as load balancing and adaptive modulation and coding can be employed to optimize resource allocation dynamically.

- **Fronthaul Latency Reduction Techniques:** Investigate novel techniques to reduce fronthaul latency, such as using low-latency transport protocols, edge computing, and caching mechanisms, could be helpful as a future direction of the hierarchical scheduling. This could involve designing intelligent algorithms that dynamically adapt the functional split based on fronthaul delay measurements.

- **Network Slicing for QoS Differentiation:** Explore the concept of network slicing to create virtualized, customized slices of the network with specific quality of service (QoS) requirements. This could involve designing resource allocation mechanisms based on the hierarchical scheduling that considers the unique demands of different slices, especially with respect to fronthaul delay sensitivity.

- **Dynamic Functional Splitting in Heterogeneous Environments:** Extend research of the hierarchical scheduling to cover heterogeneous environments, including the integration of different radio access technologies and deployment scenarios. Develop adaptive algorithms based on hierarchical scheduling that can dynamically adjust functional splitting based on the specific characteristics of each deployment.
- **Security Considerations:** Address security challenges associated with hierarchical scheduling and dynamic functional splitting. Explore potential vulnerabilities and develop mechanisms to ensure the integrity and confidentiality of data transferred between BBU and RRH units in C-RAN.

In conclusion, future research directions for resource allocation with dynamic functional splitting in C-RAN for 5G networks hold great promise. These future research directions aim to address the complexities of dynamic resource allocation, ensuring responsiveness, efficiency, and minimal fronthaul delay negative influence in the dynamic landscape of 5G.

References

[1] Shayea, Ibraheem, Mustafa Ergen, Marwan Hadri Azmi, Sultan Aldirmaz Çolak, Rosdiadee Nordin, and Yousef Ibrahim Daradkeh. "Key challenges, drivers and solutions for mobility management in 5G networks: A survey." IEEE access 8 (2020): 172534-172552.

[2] Gkonis, Panagiotis K., Panagiotis T. Trakadas, and Dimitra I. Kaklamani. "A comprehensive study on simulation techniques for 5g networks: State of the art results, analysis, and future challenges." Electronics 9, no. 3 (2020): 468.

[3] Erunkulu, Olaonipekun Oluwafemi, Adamu Murtala Zungeru, Caspar K. Lebekwe, Modisa Mosalaosi, and Joseph M. Chuma. "5G mobile communication applications: A survey and comparison of use cases." IEEE Access 9 (2021): 97251-97295.

[4] Chochliouros, Ioannis P., Michail-Alexandros Kourtis, Anastasia S. Spiliopoulou, Pavlos Lazaridis, Zaharias Zaharis, Charilaos Zarakovitis, and Anastasios Kourtis. "Energy efficiency concerns and trends in future 5G network infrastructures." Energies 14, no. 17 (2021): 5392.

[5] Xu, Dongzhu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu, Congkai An, Yiming Shi, Liang Liu, and Huadong Ma. "Understanding operational 5G: A first measurement study on its coverage, performance and energy consumption." In Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, pp. 479-494. 2020.

[6] Manap, Sulastri, Kaharudin Dimyati, Mhd Nour Hindia, Mohamad Sofian Abu Talip, and Rahim Tafazolli. "Survey of radio resource management in 5G heterogeneous networks." IEEE Access 8 (2020): 131202-131223.

[7] Matoussi, Salma, Ilhem Fajjari, Salvatore Costanzo, Nadjib Aitsaadi, and Rami Langar. "5G RAN: Functional split orchestration optimization." IEEE Journal on Selected Areas in Communications 38, no. 7 (2020): 1448-1463.

[8] Foukas, Xenofon, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K. Marina. "Network slicing in 5G: Survey and challenges." IEEE communications magazine 55, no. 5 (2017): 94-100.

[9] Checko, Aleksandra, Henrik L. Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S. Berger, and Lars Dittmann. "Cloud RAN for mobile networks—A technology overview." IEEE Communications surveys & tutorials 17, no. 1 (2014): 405-426.

[10] Sexton, Conor, Nicholas J. Kaminski, Johann M. Marquez-Barja, Nicola Marchetti, and Luiz A. DaSilva. "5G: Adaptable networks enabled by versatile radio access technologies." IEEE Communications Surveys & Tutorials 19, no. 2 (2017): 688-720.

[11] Agiwal, Mamta, Abhishek Roy, and Navrati Saxena. "Next generation 5G wireless networks: A comprehensive survey." IEEE communications surveys & tutorials 18, no. 3 (2016): 1617-1655.

[12] Lee, Hyun-Suk, Seokjae Moon, Do-Yup Kim, and Jang-Won Lee. "Packet-Based Fronthauling in 5G Networks: Network Slicing-Aware Packetization." IEEE Communications Standards Magazine 7, no. 2 (2023): 56-63.

[13] Vinnakota, Venu, Kirtan Gopal Panda, and Debarati Sen. "Fronthaul Design In Cloud Radio Access Networks: A Survey." Advanced Computing and Communications (2019).

[14] Chih-Lin, I., Jinri Huang, Ran Duan, Chunfeng Cui, Jesse Jiang, and Lei Li. "Recent progress on C-RAN centralization and cloudification." IEEE Access 2 (2014): 1030-1039.

[15] Agrawal, Rajeev, Anand Bedekar, Troels Kolding, and Vishnu Ram. "Cloud RAN challenges and solutions." Annals of Telecommunications 72, no. 7-8 (2017): 387-400.

[16] Rost, Peter, Carlos J. Bernardos, Antonio De Domenico, Marco Di Girolamo, Massinissa Lalam, Andreas Maeder, Dario Sabella, and Dirk Wübben. "Cloud technologies for flexible 5G radio access networks." IEEE Communications Magazine 52, no. 5 (2014): 68-76.

[17] 3GPP TR38.801, "Study on new radio access technology: Radio access architecture and interfaces," v.14.0.0, March 2017.

[18] Townend, Dave, Ryan Husbands, Stuart D. Walker, and Andy Sutton. "Challenges and Opportunities in Wireless Fronthaul." IEEE Access (2023).

[19] Dötsch, Uwe, Mark Doll, Hans-Peter Mayer, Frank Schaich, Jonathan Segel, and Philippe Sehier. "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE." Bell Labs Technical Journal 18, no. 1 (2013): 105-128.

[20] https://moniem-tech.com/2022/03/13/why-7-2x-split-is-the-best-split-option/

[21] Peng, Mugen, Chonggang Wang, Vincent Lau, and H. Vincent Poor. "Fronthaul-constrained cloud radio access networks: Insights and challenges." IEEE Wireless Communications 22, no. 2 (2015): 152-160.

[22] Maeder, Andreas, Massissa Lalam, Antonio De Domenico, Emmanouil Pateromichelakis, Dirk Wübben, Jens Bartelt, Richard Fritzsche, and Peter Rost. "Towards a flexible functional split for cloud-RAN networks." In 2014 European Conference on Networks and Communications (EuCNC), pp. 1-5. IEEE, 2014.

[23] Soret, Beatriz, Antonio De Domenico, Samer Bazzi, Nurul H. Mahmood, and Klaus I. Pedersen. "Interference coordination for 5G new radio." IEEE Wireless Communications 25, no. 3 (2017): 131-137.

[24] Sawahashi, Mamoru, Yoshihisa Kishiyama, Akihito Morimoto, Daisuke Nishikawa, and Motohiro Tanno. "Coordinated multipoint transmission/reception techniques for LTE-advanced [Coordinated and Distributed MIMO]." IEEE Wireless Communications 17, no. 3 (2010): 26-34.

[25] Mountaser, Ghizlane, Maria Lema Rosas, Toktam Mahmoodi, and Mischa Dohler. "On the feasibility of MAC and PHY split in cloud RAN." In 2017 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6. IEEE, 2017.

[26] Harutyunyan, Davit, and Roberto Riggio. "Flex5G: Flexible functional split in 5G networks." IEEE Transactions on Network and Service Management 15, no. 3 (2018): 961-975.

[27] Larsen, Line MP, Aleksandra Checko, and Henrik L. Christiansen. "A survey of the functional splits proposed for 5G mobile crosshaul networks." IEEE Communications Surveys & Tutorials 21, no. 1 (2018): 146-172.

[28] Alba, Alberto Martinez, and Wolfgang Kellerer. "A dynamic functional split in 5G radio access networks." In 2019 IEEE Global Communications Conference (GLOBECOM), pp. 1-6. IEEE, 2019.

[29] Jabbar, A. I. A., and Fawaz Y. Abdullah. "Long term evolution (LTE) scheduling algorithms in wireless sensor networks (WSN)." International Journal of Computer Applications 121, no. 10 (2015).

[30] Huang, Min, and Xu Zhang. "Distributed MAC Scheduling Scheme for C-RAN with Non-Ideal Fronthaul in 5G Networks." In 2017 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6. IEEE, 2017.

[31] Kadan, Fehmi Emre, and Ali Özgür Yılmaz. "A theoretical performance bound for joint beamformer design of wireless fronthaul and access links in downlink C-RAN." IEEE Transactions on Wireless Communications 21, no. 4 (2021): 2177-2192.

[32] Holma, Harri, Antti Toskala, and Takehiro Nakamura, eds. 5G technology: 3GPP new radio. *John Wiley & Sons*, 2020.

[33] Voigtländer, Florian, Ali Ramadan, Joseph Eichinger, Jürgen Grotepass, Karthikeyan Ganesan, Federico Diez Canseco, Dirk Pensky, and Alois Knoll. "5G for the factory of the future: Wireless communication in an industrial environment." arXiv preprint arXiv:1904.01476 (2019).

[34] Ejaz, Waleed, Shree K. Sharma, Salman Saadat, Muhammad Naeem, Alagan Anpalagan, and Naveed Ahmad Chughtai. "A comprehensive survey on resource allocation for CRAN in 5G and beyond networks." Journal of Network and Computer Applications 160 (2020): 102638.

[35] Ahmed, Ashfaq, Arafat Al-Dweik, Youssef Iraqi, Husameldin Mukhtar, Muhammad Naeem, and Ekram Hossain. "Hybrid automatic repeat request (HARQ) in wireless communications systems and standards: A contemporary survey." IEEE Communications Surveys & Tutorials 23, no. 4 (2021): 2711-2752.

[36] Abreu, Renato, Preben Mogensen, and Klaus I. Pedersen. "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications." In 2017 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-5. IEEE, 2017.

[37] Han, Yishu, Salah Eddine Elayoubi, Ana Galindo-Serrano, Vineeth S. Varma, and Malek Messai. "Periodic radio resource allocation to meet latency and reliability requirements in 5G networks." In 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), pp. 1-6. IEEE, 2018.

[38] Elayoubi, Salah Eddine, Patrick Brown, Matha Deghel, and Ana Galindo-Serrano. "Radio resource allocation and retransmission schemes for URLLC over 5G networks." IEEE Journal on Selected Areas in Communications 37, no. 4 (2019): 896-904.

[39] Jang, Hyeondeok, Junsung Kim, Wonsuk Yoo, and Jong-Moon Chung. "URLLC mode optimal resource allocation to support HARQ in 5G wireless networks." IEEE Access 8 (2020): 126797-126804.

[40] Sharara, Mahdi, Sahar Hoteit, Patrick Brown, and Véronique Vèque. "On Coordinated Scheduling of Radio and Computing Resources in Cloud-RAN." IEEE Transactions on Network and Service Management (2022).

[41] Karimi, Ali, Klaus I. Pedersen, and Preben Mogensen. "Low-complexity centralized multi-cell radio resource allocation for 5G URLLC." In 2020 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6. IEEE, 2020.

[42] Bassi, Francesca, and Hatem Ibn Khedher. "HARQ-aware allocation of computing resources in C-RAN." In 2020 IEEE Symposium on Computers and Communications (ISCC), pp. 1-6. IEEE, 2020.

[43] Göktepe, Barış, Cornelius Hellge, Thomas Schierl, and Slawomir Stanczak. "A hybrid HARQ feedback prediction approach for Single-and Cloud-RANs in the sub-THz regime." In GLOBECOM 2022-2022 IEEE Global Communications Conference, pp. 2309-2315. IEEE, 2022.

[44] Göktepe, Barış, Cornelius Hellge, Thomas Schierl, and Slawomir Stanczak. "Distributed Machine-Learning for Early HARQ Feedback Prediction in Cloud RANs." IEEE Transactions on Wireless Communications (2023).

[45] Saha, Rony Kumer, Shinobu Nanba, and Kosuke Nishimura. "A technique for cloud based clustering and spatial resource reuse and scheduling of 3D in-building small cells using CoMP for high capacity CRAN." IEEE Access 6 (2018): 71602-71621.

[46] Akhtar, Tafseer, Christos Tselios, and Ilias Politis. "Radio resource management: approaches and implementations from 4G to 5G and beyond." Wireless Networks 27 (2021): 693-734.

[47] Jiang, F., Gong, Z., Hao, K., Zhang, Y. (2020). Multiple Access Techniques. In: Shen, X.(., Lin, X., Zhang, K. (eds) Encyclopedia of Wireless Networks. Springer, Cham. https://doi.org/10.1007/978-3-319-78262-1 236

[48] https://academy.infinetwireless.com/en/online-education/wireless-networking-fundamentals/5

[49] Saadia, Rahat, and Noor M. Khan. "Single carrier-frequency division multiple access radar: Waveform design and analysis." IEEE Access 8 (2020): 35742-35751.

[50] Pontois, Nicolas, Megumi Kaneko, Thi Ha Ly Dinh, and Lila Boukhatem. "User pre-scheduling and beamforming with outdated CSI in 5G fog radio access networks." In 2018 IEEE Global Communications Conference (GLOBECOM), pp. 1-6. IEEE, 2018.

[51] Nardini, Giovanni, Giovanni Stea, Antonio Virdis, Dario Sabella, and Marco Caretti. "Practical large-scale coordinated scheduling in LTE-Advanced networks." Wireless Networks 22 (2016): 11-31.

[52] Ahmadzadeh, Amir M., Lucas Cuadra, Miguel A. del Arco-Vega, J. Antonio Portilla-Figueras, and Sancho Salcedo-Sanz. "Influence of overhead on LTE downlink performance: a comprehensive model." Telecommunication Systems 67 (2018): 485-517.

[53] Pedersen, Klaus I., Gilberto Berardinelli, Frank Frederiksen, Preben Mogensen, and Agnieszka Szufarska. "A flexible 5G frame structure design for frequency-division duplex cases." IEEE Communications Magazine 54, no. 3 (2016): 53-59.

[54] Jiang, Dajie, Haiming Wang, Esa Malkamaki, and Esa Tuomaala. "Principle and performance of semi-persistent scheduling for VoIP in LTE system." In 2007 International Conference on Wireless Communications, Networking and Mobile Computing, pp. 2861-2864. IEEE, 2007.

[55] Love, Robert, Ravi Kuchibhotla, Amitava Ghosh, Rapeepat Ratasuk, Weimin Xiao, Brian Classon, and Yufei Blankenship. "Downlink control channel design for 3GPP LTE." In 2008 IEEE Wireless Communications and Networking Conference, pp. 813-818. IEEE, 2008.

[56] Kucera, Stepan, and David Lopez-Perez. "Inter-cell interference coordination for control channels in LTE heterogeneous networks." IEEE/ACM Transactions on Networking 24, no. 5 (2015): 2872-2884.

[57] A.A. Zaidi, R. Baldemair, M. Andersson, S. Faxér, V. Molés-Cases, and Z. Wang, "Designing for the future: the 5G NR physical layer," *Ericsson Technology Review*, pp. 1-13, 2017.

[58] Sayeed, Zulfiquar, Ed Grinshpun, Dave Faucher, and Sameer Sharma. "Long-term application-level wireless link quality prediction." In 2015 36th IEEE Sarnoff Symposium, pp. 40-45. IEEE, 2015.

[59] Zheng, Yadan, Shubo Ren, Xiaoyan Xu, Ying Si, Mingke Dong, and Jianjun Wu. "A modified ARIMA model for CQI prediction in LTE-based mobile satellite communications." In 2012 IEEE International Conference on Information Science and Technology, pp. 822-826. IEEE, 2012.

[60] S. Watanabe, "A widely applicable Bayesian information criterion," *Journal of Machine Learning Research*, pp. 867-897, 2013.

[61] 3GPP, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures. TS 36.300 version 17.1.0 Release 17, 2022.

[62] 38.901, G.T., Study on channel model for frequencies from 0.5 to 100 GHz. 3GPP TR 38.901, 2022. 17.0.0.

[63] Sun, Wenbin, Qiyue Yu, Weixiao Meng, and Victor CM Leung. "Transmission mechanism and performance analysis of multiuser opportunistic beamforming in Rayleigh and Rician fading channels." IEEE Transactions on Vehicular Technology 67, no. 10 (2018): 9459-9473.

[64] Small Cell Release, SCF159, "Small cell virtualization functional splits and use cases," 2016.

[65] Kwan, Raymond, Cyril Leung, and Jie Zhang. "Proportional fair multiuser scheduling in LTE." IEEE Signal Processing Letters 16, no. 6 (2009): 461-464.

[66] Mosleh, Somayeh, Lingjia Liu, and Jianzhong Zhang. "Proportional-fair resource allocation for coordinated multi-point transmission in LTE-advanced." IEEE Transactions on Wireless Communications 15, no. 8 (2016): 5355-5367.

[67] 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," V. 16.5.0, 2021.

[68] Sesia, Stefania, Issam Toufik, and Matthew Baker. LTE-the UMTS long term evolution: from theory to practice. *John Wiley & Sons*, 2011.

[69] Wang, Zhaoying, Yifei Wei, F. Richard Yu, and Zhu Han. "Utility optimization for resource allocation in multi-access edge network slicing: A twin-actor deep deterministic policy gradient approach." IEEE Transactions on Wireless Communications 21, no. 8 (2022): 5842-5856.

[70] Ren, Yin, Aihuang Guo, Chunlin Song, and Yidan Xing. "Dynamic resource allocation scheme and deep deterministic policy gradient-based mobile edge computing slices system." IEEE Access 9 (2021): 86062-86073.

[71] Angelakis, Vangelis, Ioannis Avgouleas, Nikolaos Pappas, Emma Fitzgerald, and Di Yuan. "Allocation of heterogeneous resources of an IoT device to flexible services." IEEE Internet of Things Journal 3, no. 5 (2016): 691-700.

[72] Sharma, Nidhi, and Krishan Kumar. "Resource allocation trends for ultra dense networks in 5G and beyond networks: A classification and comprehensive survey." Physical Communication 48 (2021): 101415.

[73] Kim, Sungwook. "Asymptotic shapley value based resource allocation scheme for IoT services." Computer Networks 100 (2016): 55-63.

[74] Kim, Minhyeop, and In-Young Ko. "An efficient resource allocation approach based on a genetic algorithm for composite services in IoT environments." In 2015 IEEE international conference on web services, pp. 543-550. IEEE, 2015.

[75] Yin, Peng-Yeng, and Jing-Yu Wang. "A particle swarm optimization approach to the nonlinear resource allocation problem." Applied mathematics and computation 183, no. 1 (2006): 232-242.

[76] Mohamed, Ali Khater, Ali Wagdy Mohamed, Ehab Zaki Elfeky, and Mohamed Saleh. "Solving constrained non-linear integer and mixed-integer global optimization problems using enhanced directed differential evolution algorithm." Machine learning paradigms: Theory and application (2019): 327-349.

[77] Sangaiah, Arun Kumar, Ali Asghar Rahmani Hosseinabadi, Morteza Babazadeh Shareh, Seyed Yaser Bozorgi Rad, Atekeh Zolfagharian, and Naveen Chilamkurti. "IoT resource allocation and optimization based on heuristic algorithm." Sensors 20, no. 2 (2020): 539.

[78] Plachy, Jan, Zdenek Becvar, Pavel Mach, Radek Marik, and Michal Vondra. "Joint positioning of flying base stations and association of users: Evolutionary-based approach." IEEE Access 7 (2019): 11454-11463.

[79] Soldani, David, Y. Jay Guo, Bernard Barani, Preben Mogensen, I. Chih-Lin, and Sajal K. Das. "5G for ultra-reliable low-latency communications." Ieee Network 32, no. 2 (2018): 6-7.

[80] Wang, Yuan H. "On the number of successes in independent trials," *Statistica Sinica* (1993): 295-312.

[81] 3GPP TS 22.261 Service requirements for the 5G systems; Stage 1 (Release 17)," 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, Tech. Rep., 2022

[82] Frenger, Pal, Stefan Parkvall, and Erik Dahlman. "Performance comparison of HARQ with chase combining and incremental redundancy for HSDPA." In IEEE 54th Vehicular Technology Conference. VTC Fall 2001. Proceedings (Cat. No. 01CH37211), vol. 3, pp. 1829-1833. IEEE, 2001.

[83] Kihero, Abuu B., Muhammad Sohaib J. Solaija, and Hüseyin Arslan. "Inter-numerology interference for beyond 5G." IEEE Access 7 (2019): 146512-146523.

[84] Capozzi, Francesco, Giuseppe Piro, Luigi Alfredo Grieco, Gennaro Boggia, and Pietro Camarda. "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey." IEEE communications surveys & tutorials 15, no. 2 (2012): 678-700.

[85] Abu-Ali, Najah, Abd-Elhamid M. Taha, Mohamed Salah, and Hossam Hassanein. "Uplink scheduling in LTE and LTE-advanced: Tutorial, survey and evaluation framework." IEEE Communications surveys & tutorials 16, no. 3 (2013): 1239-1265.

[86] Yang, Hui, Jie Zhang, Yuefeng Ji, and Young Lee. "C-RoFN: Multi-stratum resources optimization for cloud-based radio over optical fiber networks." IEEE Communications Magazine 54, no. 8 (2016): 118-125.

[87] Capozzi, Francesco, Giuseppe Piro, Luigi Alfredo Grieco, Gennaro Boggia, and Pietro Camarda. "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey." IEEE communications surveys & tutorials 15, no. 2 (2012): 678-700.

[88] Becvar, Zdenek, Pavel Mach, Morten Høgdal, and Andrijana Popovska Avramova. "Hierarchical resource scheduling method of wireless communication system." U.S. Patent 10,624,105, issued April 14, 2020.

[89] Becvar, Zdenek, Pavel Mach, Mohammed Elfiky, and Mitsuo Sakamoto. "Hierarchical scheduling for suppression of fronthaul delay in C-RAN with dynamic functional split." IEEE Communications Magazine 59, no. 4 (2021): 95-101.

[90] Elfiky, Mohammed, Zdenek Becvar, and Pavel Mach. "Dynamic Adjustment of Scheduling Period in Mobile Networks Based on C-RAN." In 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), pp. 1-7. IEEE, 2021.

[91] Elfiky, Mohammed, Zdenek Becvar, and Pavel Mach. "Allocation of Resources for HARQ Retransmission in Mobile Networks Based on C-RAN," in IEEE Systems Journal, doi: 10.1109/JSYST.2023.3327823.

Chapter 8

List of Thesis-Related Publication and Research Works

- Elfiky, Mohammed, Zdenek Becvar, and Pavel Mach. "Allocation of Resources for HARQ Retransmission in Mobile Networks Based on C-RAN," in IEEE Systems Journal, doi: 10.1109/JSYST, 2023, 3327823.
- Elfiky, Mohammed, Zdenek Becvar, and Pavel Mach. "Dynamic Adjustment of Scheduling Period in Mobile Networks Based on C-RAN." In IEEE 94th Vehicular Technology Conference (VTC2021-Fall), pp. 1-7. IEEE, 2021, doi: 10.1109/VTC2021-Fall52928, 2021, 9625329.
- Becvar, Zdenek, Pavel Mach, Mohammed Elfiky, and Mitsuo Sakamoto. "Hierarchical scheduling for suppression of fronthaul delay in C-RAN with dynamic functional split." IEEE Communications Magazine 59, no. 4 (2021): 95-101, doi: 10.1109/MCOM.001.2000697.
- Becvar, Zdenek, Pavel Mach, and Mohammed Elfiky, "Preliminary design of split scheduler, Report on preliminary design of dynamic scheduler splitting," FOXCONN-CTU project, Mobile edge computing and functional splitting for scheduling of radio resources. Deliverable D3, Version 1.1.1, March 31, 2018.
- Becvar, Zdenek, Pavel Mach, and Mohammed Elfiky, "Preliminary design of split scheduler, High-level design and initial implementation of hierarchical scheduler in MATLAB," FOXCONN-CTU project, Mobile edge computing and functional splitting for scheduling of radio resources. Deliverable D2, Version 1, September 30, 2017.

Chapter 9

List of Research Projects

- Project number: SGS23/171/OHK3/3T/13, "Allocation of communication, control, and caching resource in 6G networks", funded by Czech Technical University in Prague, 01/2023 - 12.2025
- Project number: SGS20/169/OHK3/3T/13, "Resource management based on machine learning for 6G mobile networks", funded by Czech Technical University in Prague, 01/2020 - 09/2020.
- Project number: 13132/830/8301636C000−1NE, "Mobile Edge Computing and Functional Splitting for Scheduling of Radio Resources", funded by Foxconn (Taiwanese multinational electronics contract manufacturer), 10/2018 - 09/2019.
- Project number: SGS17/184/OHK3/3T/13, "Flexible radio access for future mobile communications", funded by Czech Technical University in Prague, 01/2017 - 12/2019.