UNIVERSITY OF CATANIA
DEPARTMENT OF ELECTRICAL, ELECTRONICS
AND COMPUTER ENGINEERING
PH.D. COURSE IN SYSTEMS, ENERGY, COMPUTER AND
TELECOMMUNICATION ENGINEERING

RADIO RESOURCE MANAGEMENT IN 5G CELLULAR NETWORKS

Doctoral Dissertation of:
Salvatore Riolo

Tutor:
Prof. Daniela Panno

Coordinator:
Prof. Paolo Arena

XXXIII Cycle

You've achieved success in your field
when you don't know whether
what you're doing is work or play.
— Warren Beatty

Abstract

THE fifth generation (5G) of cellular networks aims at providing connectivity for a large number of applications. To achieve this goal, 5G has been designed considering three generic services with vastly heterogeneous requirements: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC). To accommodate these wide range of services, a key role is played by scheduling and radio resource allocation, whose aim is allowing efficient sharing of the limited radio spectrum among different services.

The aim of this Dissertation is to investigate and to design Radio Resource Management (RRM) techniques and scheduling strategies that are suitable to meet the heterogeneous requirements of eMBB and mMTC usage scenarios. For this reason, the analysis provided considers different frequency band, like the mmWave transmissions and the sub-6GHz transmissions, different access techniques, like Time Division Multiple Access (TDMA), Orthogonal Frequency Division Multiple Access (OFDMA) and the novel Sparse Code Multiple Access (SCMA), the support of advanced radio access technologies, such as beamforming technique and device-to-device (D2D) communications, the definition of very simple random access procedure to meet the requirements of low-complexity connected devices, and various network architectures, like Millimeter-Wave Mobile Broadband (MMB) and single-cell.

In the context of radio resource allocation, we present four research

activity. First, we provide an OFDMA-based Quality-of-Service (QoS) aware scheduling framework to allocate radio resources among Guaranteed Bit Rate (GBR) and non-GBR services. Second, we consider a D2D-enabled MMB and propose a TDMA-based centralized access control scheme which jointly manages D2D communications and transmissions in both the access and the backhaul networks. Third, we propose a new access control scheme tailored for mMTC scenarios, where radio resources are allocated in the Physical Uplink Shared Channel (PUSCH) by means of the SCMA technique to properly multiplex a large number of small-sized date. Fourth, in order to reduce jointly the transmission energy consumption and the signaling overhead from the prospective of the MTC devices, we present the strategy of transmitting tagged preambles in the Physical Random Access Channel (PRACH) and present a rigorous analytical model to analyze the correct detection of both the preamble and the tag at the receiver next Generation NodeB (gNB), considering the presence of interference, noise, and multi-path fading.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The objective of this Chapter is to provide a brief *excursus* of the main concepts investigated in this Dissertation, whose aim is to study and propose new radio resource allocation strategies in 5G networks characterized by services with very heterogeneous requirements.

1.1 Background and Motivation

In the near future, it is expected a fully mobile and connected society, characterized by a huge growth in connectivity and traffic volume. Some typical trends include explosive growth of data traffic, great increase of connected devices and continuous emergence of new services. Today's statistics show that over 1 billion mobile users around the globe are intensely using the social networking media, streaming, and gaming services on a daily basis. At this regards, 5G technology has to support the proliferating traffic demand, providing a wide range of connected devices and services. Unlike earlier generations, 5G networks are required to simultaneously provide a diversity of services with different requirements in their service levels. Specifically, there is a broad consensus today that categorizes these services. They are defined by the Interna-

tional Telecommunication Union Radiocommunication Sector (ITU-R) and are divided into three categories [3]:

- **eMBB - Enhanced Mobile Broadband**: as an extension to 4G broadband services, the aim of this category is to have an ultra high-speed connection for both indoors and outdoors.
- **mMTC - Massive Machine Type Communications**: 5G will need to suit a whole raft of connected devices with heterogeneous quality of service requirements. The objective of this category is to have a unifying connectivity fabric with enough flexibility to support the exponential increase in the density of connected devices.
- **URLLC - Ultra-Reliable and Low Latency Communication**: this use case is related to services that are delay sensitive, and thus require stringent requirements for latency and reliability to ensure increased reactivity

The service requirements for each of the above categories are remarkably distinct in terms of reliability, throughput, latency, among others. The first category is data rate hungry (e.g., ultra high-definition (8K) videos at 120 fps and virtual/augmented reality wireless streaming). On the other hand, mMTC envisions the massive Internet of Things (mIoT) paradigm, that requires low power consumption and very low throughput, while URLLC services need to be extremely reliable with a target latency below 1 ms. The requirements for radio access technologies are depicted in Fig. 1.1 and listed in Table 1.1.

According to these objectives, 5G must be able to support users with throughput and peak throughput of 10 and 20 times higher than what available in legacy 4G networks, respectively. The density of the maximum connection will be 10 times more and the energy saving is of importance. In order to cater this vision and to satisfy the proliferating traffic demand, 5G technology is envisioned to support 1000 times increase in capacity and 100 times more connected devices than today's 4G networks. In this regard, the radio resource management has gained popularity for achieving an efficient utilization of the available radio frequency spectrum. Resource allocation, which involves scheduling of spectrum and power resources, represents a crucial problem for the per-

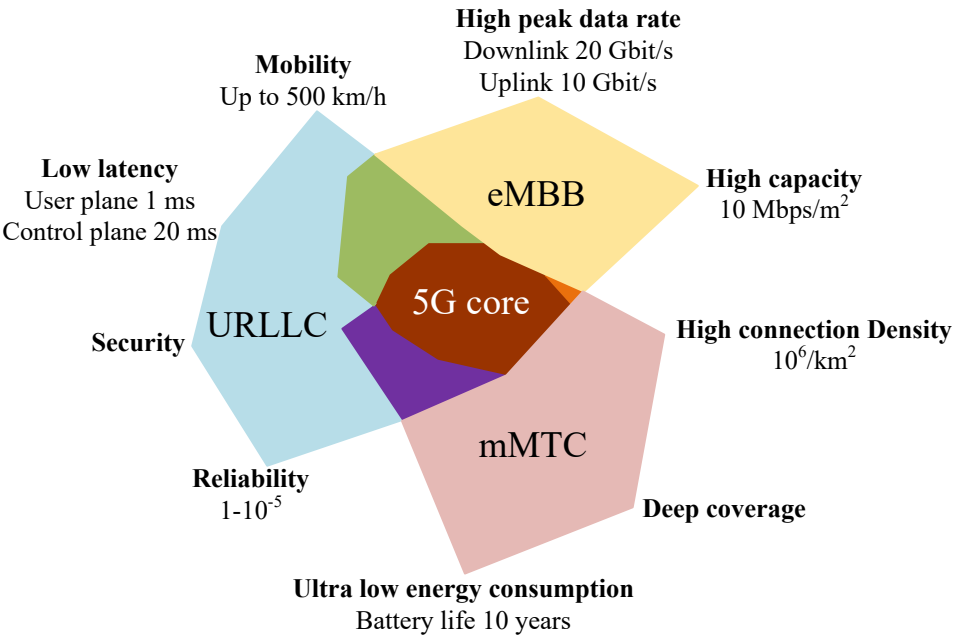


Figure 1.1: *5G usage scenarios*

Table 1.1: *Requirements for IMT-2020 5G*

Capability	5G requirement	Usage scenario
Downlink peak data rate	20 Gbit/s	eMBB
Uplink peak data rate	10 Gbit/s	eMBB
User experienced downlink data rate	100 Mbit/s	eMBB
User experienced uplink data rate	50 Mbit/s	eMBB
User plane Latency	$\leq$4 ms $\leq$1 ms	eMBB URLLC
Control plane Latency	$\leq$20 ms	eMBB/URLLC
Mobility	500 km/h	eMBB/URLLC
Connection density	10^6/km^2	mMTC
Energy efficiency	Equal to 4G	eMBB
Battery life	10 years	mMTC
Area traffic capacity	10 Mbps/m^2	eMBB
Peak downlink spectrum efficiency	30 bit/s/Hz	eMBB
Reliability	1-10^{-5}	URLLC

formance of 5G networks to achieve these heterogeneous service requirements. In this Dissertation, we investigate resource allocation problems in 5G network characterized by eMBB service and mMTC service by exploring a large variety of optimization techniques and available technologies.

1.1.1 Enhanced mobile broadband usage scenario

The term enhanced mobile broadband (eMBB) unifies network capacity and peak data rates in the always-changing area of consumer needs and wants. eMBB has to cover not only a diverse present but also a future that will follow consumers wherever they may lead. This family represents a variety of services, with different QoS attributes.

In this context, 3GPP developed a new Radio Access Technology (RAT) termed 5G New Radio (NR) [4] for the 5G mobile network. It was designed to be the global standard for the air interface of 5G networks. Like in 4G wireless systems, for 5G NR Orthogonal Frequency Division Multiple Access (OFDMA) has been selected as the physical layer technology, but a flexible numerology has been introduced in order to efficiently satisfy the different requirements of 5G services. Since eMBB represents a great variety of services, in order to differentiate the QoS attributes, un the Next Generation Core Network a flow-based QoS concept is adopted. The 5G QoS model [1] supports both QoS Flows which require a Guaranteed Flow Bit Rate, termed Guaranteed Bit Rate (GBR) QoS Flows, and QoS Flows which do not require a GFBR, termed non-GBR QoS Flows. In the 5G Radio Access Network (RAN), the QoS Flows for these services are mapped into GBR Data Radio Bearers (DRBs) and non-GBR DRBs, respectively [5].

It is clear that, in order to satisfy the QoS requirements of DRBs, radio resources should be properly allocated among User Equipments (UEs), i.e., the Radio Resource Management (RRM), which is responsible for packet scheduling, plays a key role in the system performance. The basic packet scheduling in 5G is already standardized [6], but the optimization is still open to research and can be adapted for one or more usage scenarios [7]. In order to efficiently utilize radio resources in 5G NR, rational scheduling algorithms need to be designed, since they largely impact on the quality and the system performance, the guarantee of QoS parameters, as well as the system throughput, and the fairness in throughput among DRBs.

In parallel, to meet the stringent requirements of the eMBB services, especially in terms of peak and user experienced data rate, several new technologies and network paradigms built into 5G have emerged [8]. Some of them are the following.

- **mmWave radio transmission**. Bands in this millimeter wave range are characterized by higher available bandwidth than bands in the sub-6GHz band, which means high-speed and high-capacity data links.
- **Massive MIMO and beamforming technique**. Multiple Input, Multiple Output refers to techniques that increase cellular coverage and capacity through the use of large numbers of antennas.
- **Unlicensed spectrum sharing**. The use of unlicensed spectrum permits to unlock more spectrum and to extend the 5G network. In fact, 5G is designed to support all current spectrum types, with the flexibility to use sharing paradigms.
- **Device-to-device (D2D) communications**. They enable direct communications between devices in cellular networks, thus improving the spectrum utilization, enhancing the overall throughput, and increasing energy efficiency. D2D communication has the potential to enable a large number of services among nearby users, such as video sharing, gaming, proximity-aware services, popular content downloading [9].

These technologies are characterized by some issues. As regards the mmWave radio transmission, they suffer from high path loss, and susceptibility to blockage from physical barriers. Due to the high path loss, the coverage area is very reduced. In fact, most of the early works available in the literature consider as typical mmWave scenario a wireless personal area network (WPAN) limited to the coverage area of a single piconet coordinator (PNC), typically 15m [10–12]. As for the beamforming technique, one of the numerous difficulties is to fit a large number of antennas inside a mobile device, especially at low frequencies. The main issue of the unlicensed spectrum is the uncontrollable interferences, in the case of a massive use of Wi-Fi and Bluetooth transmissions in the same frequency range. The D2D communications need the management of interferences, not only with traditional cellular communications, but also among D2D communications themselves.

It is clear that the use of these technologies requires a thorough study of the related characteristics, important changes in cellular network ar-

chitecture and, accordingly, the radio resource allocation must be strongly revised.

1.1.2 Massive Machine Type Communications usage scenario

The mMTC network consists of a dense deployment of low-power and low-cost MTC devices, which transmit small packets sporadically with relaxed delay requirements. Connected MTC devices include, for instance, smart meters, smoke detectors, and consumer electronic devices, whose number is attended to be 10^6 per km^2, as reported in Table 1.1. In this context, the problem of radio resource allocation becomes of primary importance to manage an enormous quantity of access requests and to multiplex a large number of small-sized data transmissions since the scarce radio spectrum need to be allocated in a more efficient way.

As regards the multiplexing of the data in the Physical Uplink Shared Channel (PUSCH), one way to cope with this problem is to increase the transmission efficiency with the adoption of a Non-Orthogonal Multiple Access (NOMA) technique that can multiplex small-sized data in a more efficient way than the traditional Single Carrier Frequency Division Multiple Access (SC-FDMA). The adoption of the proper NOMA technique requires the study of these new innovative resource allocation techniques and the verification of their feasibility.

On the other hand, the problem of managing the random access of a huge number of MTC devices is of main importance. In fact, when several MTC devices simultaneously initiate their procedure to access to the network, a large number of access requests collide, the MTC devices will re-attempt their access, thus causing a severe congestion problem in the system. For this reason, the 3GPP proposed the adoption of Access Class Barring (ACB) schemes [13], i.e., congestion control schemes designed for limiting the number of simultaneous access attempts, thus reducing the number of Random Access (RA) failed attempts. In the conventional ACB scheme, the Base Station periodically broadcasts an ACB factor $p \in [0, 1]$ to the MTC devices. Then, each MTC device which has data to be transmitted draws a uniform random number $q \in [0, 1]$, and transmit only whether $q \leq p$. The main challenge of these ACB schemes is to adapt the ACB factor p according to the traffic load. Another way to manage the huge number of access request is to increase

the resources allocated to Physical Random Access Channel (PRACH). However, due to limited uplink resources, if the PRACH resources are increased, the amount of resources available for PUSCH decreases. Consequently, many MTC devices which have successfully complete RA procedure, could not find enough transmission resources in the PUSCH. This means that it is necessary to determine an optimal division of the uplink resources.

In spite of the adoption of an ACB scheme and/or a dynamic uplink radio resource dimensioning, in a massive scenario the collision probability in each RA attempt remains significantly high, thus the average access attempt number before success is high. This high number of access attempts not only increases delays, but also increases signaling transmissions and, therefore, energy consumption [14] which is a key point for the MTC devices, as shown in Table 1.1. Currently, in LTE / LTE-Advanced (LTE-A) uplink communication, a grant-based 4-step RA procedure is adopted. For each RA attempt, the device needs to carry out two signaling transmissions. In addition, if the 4-phase RA procedure was successful, further signaling messages have to be exchanged before starting the data packet transmission [14]. It is clear that the legacy RA procedure is highly inefficient for supporting the very short transmissions of MTC traffic. The 3GPP has already introduced, in Release 13, the enhanced Machine Type Communication and the NarrowBand IoT technologies [15], that are optimized for granting lower complexity, and providing longer battery life. Nevertheless, eMTC inherits both the 4-phase RA procedure and the data transmission from the conventional LTE/LTE-A, while NB-IoT inherits the RA procedure but allows the device to transmit the data packet immediately after the RA procedure. These technologies, although more suitable to support machine-type communications according to the IMT-Advanced requirements (connection density of 10^5 devices per km^2), are still inefficient for the 5G mMTC scenario. For this reason, in order to save energy, we necessitate to propose a new random access procedure, tailored for sporadic transmissions of small packets, enabled for reducing the number of transmissions per access attempt and the signaling overhead.

1.2 Research Contributions and Thesis outline

Motivated by the aforementioned technical challenges, the general objective of these Ph.D. research activities is to develop efficient radio resource allocation and interference management algorithms for 5G cellular networks and beyond.

For each research activity, we carried out several studies and a critical analysis of the state of art. The objectives to be achieved had been clearly reported and mathematically formalized. When the analytical solution of the formulated problem is computationally high, we provided heuristic proposals for the given problem. The goodness of the proposed solutions has been verified through a large number of simulations in comparison with other works available in the literature.

Through mathematical analysis, we provide a proper radio network planning for a proposed mmWave architecture, and methodologies based on the probability theory for analyzing the performance of the access requests in mMTC scenarios. Furthermore, we present an accurate model for analyzing the detection probability distribution of the signal processed by the gNB receiver, considering a realistic radio channel, in mMTC scenarios.

The main contributions of this Ph.D. Dissertation are summarized as follows.

- The first research activity focuses on the radio resource allocation in sub-6GHz 5G Networks for achieving the heterogeneous QoS requirements of different services. The results have been published in the conference work [16] and in the Wireless Networks journal [17], and are described in Chapter 2. Therein, we present a new radio resource scheduling scheme designed to manage a heterogeneous traffic in Downlink OFDMA-based 5G NR network. We provide a heuristic QoS-aware scheduling framework, with two different channel aware scheduling strategies, at the aim of choosing a different trade-off between the goals of maximizing the system throughput and reaching the fairness among DRBs. Moreover, the proposed scheme is designed to work well in realistic scenarios where non-ideal Connection Admission Controls (CACs) are adopted.

- The second research activity is described in Chapter 3. We address the problem of radio resource allocation in a D2D-enabled mmWave Mobile Broadband (MBB) scenario, consisting of several Access Points (APs) interconnected among themselves through a wireless backhaul network. The results have been published in the conference works [18, 19] and in the IEEE Access journal [20]. We propose a slotted Time Division Multiple Access (TDMA) based radio access scheme with concurrent transmission support, where time slots are organized into variable-length frames and the access request is carried out by means of a polling technique. We propose a data flow management strategy and a multi-criteria scheduling algorithm based on greedy graph vertex-coloring techniques to jointly manage D2D communications and transmissions in the access and the backhaul network. We aim to maximize the system throughput, to minimize the end-to-end delay, and to improve the fairness among users. In addition, we provide a proper radio network planning, consisting of both coverage and capacity planning. We derive a mathematical analysis specific for the peculiarities of mmWave technology and the adopted Centralized Access Control scheme, taking into account the transmission delays, the physical layer control delays, and the log-normal fading.

- The third research activity addresses the radio resource allocation problem in a 5G mMTC scenario. The results have been published in the conference works [21, 22], in IEEE Internet of Things Journal [23], and in IEEE Communications Letters [24]. A detailed description is reported in Chapter 4. Therein, we propose to increase the transmission efficiency in the PUSCH by applying the Sparse Code Multiple Access (SCMA) technique. We analyze also the feasibility of our SCMA-based resource allocation by considering both the complexity and the signaling overhead points of view. Furthermore, in order to support a large number of access requests, we propose a load-aware Dynamic Uplink Resource Allocation (DURD) scheme to properly allocate the uplink radio resources between the PRACH and the PUSCH. Then, in order to further increase the number of succeeded communications even in the presence of a very high traffic load, we present the innovative idea to exploit also

the unused PUSCH resources to serve some MTC devices that have failed their access attempt. Finally, to reduce the energy consumption of the MTC device, we propose an optimized grant-based connectionless 2-step Random Access (RA) procedure, based on the transmission of a tagged preamble sequence. This technique reduces the number of signaling transmissions per access attempt, the overall number of steps in the access procedure, and, accordingly, the energy consumption per device.

- In light of the promising results achieved by means of the proposed 2-step RA access procedure, in the last research activity we mathematically analyze the strategy of transmitting tagged preambles in a realistic radio channel. The results have been accepted for publication in IEEE Transactions on Wireless Communications [25], and presented in Chapter 5. Despite the undoubted benefits introduced by the transmission of tagged preamble sequences, the main disadvantage is the complexity of the detection of these sequences by the gNB receiver, especially in a realistic radio channel. In literature, the performances of these advanced detection procedures were obtained only by running a large number of simulations that are typically highly time-consuming. For this reason, we present the first analytical model to analyze the correct detection of both the preamble and the tag transmitted by each MTC device in the presence of interference, due to other preambles and tags, of noise and multi-path fading. The high accuracy of the proposed model is verified through simulations. In addition, we show how our analytical study can be a good tool to investigate and derive innovative detection strategies.

CHAPTER 2

A QoS-aware radio resource allocation in 5G NR sub-6GHz

In this chapter, we analyze the radio resource allocation problem for an eMBB usage scenario in 5G NR FR1 (sub 6-GHz frequency bands), where services require different QoS attributes. At this aim, in the Next Generation Core Network a flow-based QoS concept is adopted. A QoS Flow corresponds to the user plane traffic that receives the same QoS treatment, within a PDU session. The 5G QoS model [1] supports both QoS flows which require a Guaranteed Flow Bit Rate (GFBR), termed Guaranteed Bit Rate (GBR) QoS Flows, and QoS Flows which do not require a GFBR, termed non-GBR QoS Flows. A GBR QoS Flow may be further qualified as delay-critical GBR when the end-to-end latency requirements are very tight. So, any QoS Flow is characterized by a QoS profile, that defines the QoS parameters applied to the QoS Flow, which includes resource type (i.e., GBR, delay-critical GBR, and non-GBR), GFBR attribute for GBR QoS Flows, priority, packet delay budget, packet error rate, and so on. The Service Data Adaptation Protocol (SDAP) is responsible for the QoS Flow handling across the 5G air interface. In particular, SDAP maps a specific QoS Flow within a PDU Session to a corresponding Data Radio Bearer (DRB) in the Radio Ac-

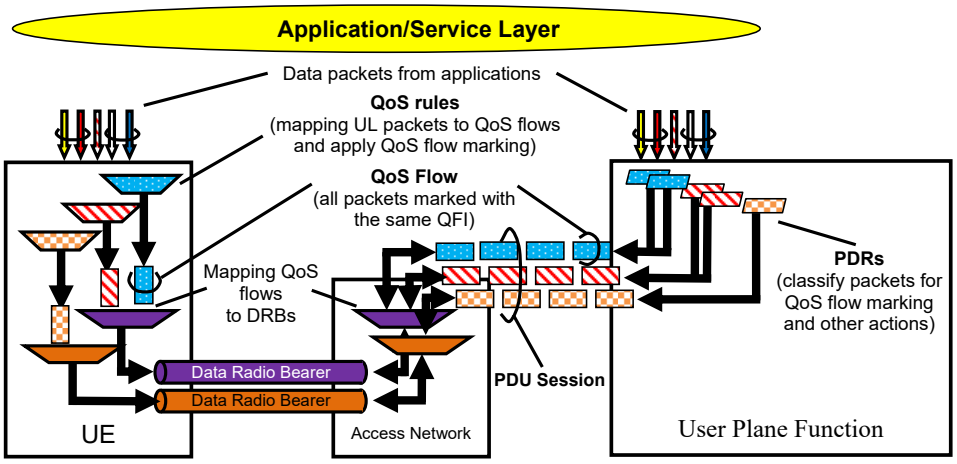
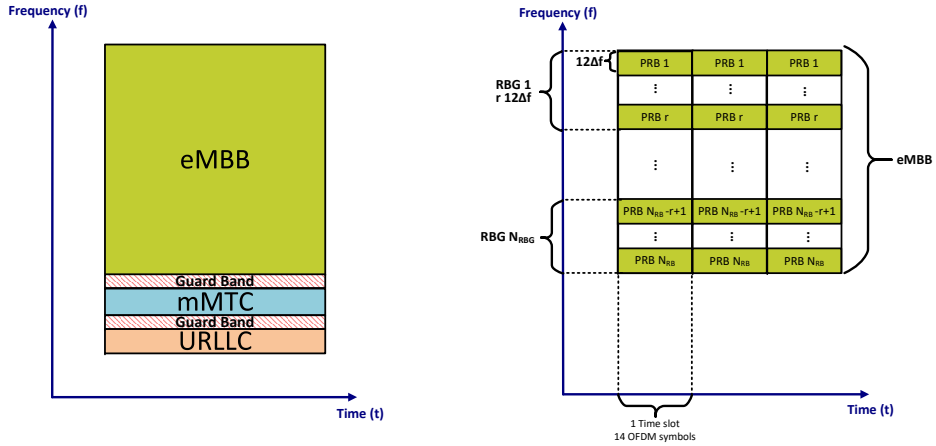


Figure 2.1: *User Plane marking for QoS Flows and mapping to Data Radio Bearers [1].*

cess Network (RAN), which has been established with the appropriate level of QoS. In addition, SDAP marks the transmitted packets with the correct QFI (QoS Flow ID), ensuring that each packet receives the correct forwarding treatment as it traverses the Core Network.

Fig. 2.1 shows the User Plane marking for QoS Flows in the Core Network and the mapping to DRBs in the RAN. Let us analyze the downlink user plane. Incoming data packets from Application / Service Layer are classified by the User Plane Function (UPF) based on the Packet Filter Sets of the DL Packet Detection Rules (PDRs) in the order of their precedence. Then, all packets are marked with a QFI and grouped into QoS flows, each one containing packets with the same QFI. The QFIs and the related QoS profiles are also provided to the RAN, which binds QoS Flows to DRBs. Unlike 4G systems where there is a strict 1:1 relation between EPC Bearers and Radio Bearers, one or more QoS flows could be mapped into one Data Radio Bearer (DRB) or vice versa [5]. It is clear that non-GBR QoS flows are mapped into non-GBR DRBs, as well as GBR QoS flows into GBR DRBs, characterized by a Guaranteed Bit Rate (GBR) value. Because the GFBR requirement of each GBR QoS flow is related to the end-to-end service, the RAN will assign to the related DRB a GBR value which is more stringent than the GFBR of the QoS flow.

In this chapter, we focus on the RAN taking into account GBR and non-GBR DRBs for eMBB usage scenarios. In order to satisfy the QoS



(a) *Frequency division for different usage scenarios.*

(b) *5G NR RA Type 0.*

Figure 2.2: *5G NR Downlink Frame Structure.*

requirements of DRBs in terms of GBR, radio resources should be properly allocated among User Equipments (UEs).

The radio resource allocation problem can be divided into two layers: one for time domain and one for frequency domain. In the frequency domain, different portions of spectrum should be allocated to different usage scenarios with a proper numerology (see Fig. 2.2a) [26, 27], and for each of them, in the time domain, the available radio resources should be assigned to DRBs with a packet scheduling every Time Transmission Interval (TTI). The Radio Resource Management (RRM) entity is responsible for packet scheduling and is located at the Next Generation NodeB (gNB). The basic packet scheduling in 5G is already standardized [6], but the optimization is still open to research and can be adapted for one or more usage scenarios [7]. In order to efficiently utilize radio resources in 5G NR, rational scheduling algorithms need to be designed, since they largely impact on the quality and the system performance.

An optimal scheduling scheme should maximize the system throughput and improve the fairness among DRBs while guaranteeing QoS parameters. A fundamental problem is that the above goals are typically in opposition to each other. On the one hand, adopting a scheduling strategy which aims to maximize the overall system throughput means serving, in a strongly unfair manner, only the UEs who have the best channel conditions (e.g., UEs near the base station), at the expense of those who have

worse channel condition (e.g., UEs at the cell edge). On the other hand, adopting a scheduling approach with the goal of maximizing the fairness in throughput among DRBs can result in a significant reduction of the system throughput. It is clear that the problem is not easy to solve, and an analytical optimization solution will take significantly long computation time, which is unacceptable for a scheduling procedure which should be made every Transmission Time Interval (TTI). Therefore, a heuristic algorithm which does not take a long computation time is needed to obtain near-optimal solutions.

At this regard, a considerable number of works have been done in resource allocation for OFDMA-based systems [28]. Scheduling techniques generally can be categorized as channel unaware schedulers (e.g., Round Robin [29]), channel aware schedulers (e.g., BestCQI, Proportional Fair [29]), and channel and QoS aware schedulers (e.g., QoS Guaranteed Resource Block Allocation [30]). Channel unaware schedulers assign radio resource units to UEs without taking into account the channel conditions, while channel aware ones take into account the channel conditions by means of periodic reporting of Channel State Information (CSI). Finally, channel and QoS aware schedulers also aim at satisfying the QoS requirements of DRBs.

However, the above scheduling algorithms, described in detail in Section 3.1, exhibit poor performance in achieving the aforementioned target. In this chapter, we present a new QoS aware scheduling scheme which jointly supports non-GBR and GBR DRBs, guaranteeing for the latter ones the minimum data rate required. In this framework, any chosen scheduling strategy can be implemented. In particular, we propose two different channel aware scheduling strategies, at the aim of being able to choose a different trade-off between the goals of maximizing the system throughput and reaching the fairness among DRBs. More specifically, the first one, called **BestCQI Highest Deviation (BestCQI_HD)**, is proposed to improve the system throughput at the expense of a limited loss in terms of fairness, while the second strategy, called **BestCQI Lowest Second (BestCQI_LS)**, aims at improving the fairness among DRBs at the expense of a limited loss in terms of system throughput.

The main contributions of this chapter can be summarized as follows.

1. We formulate the optimal joint scheduling problem for GBR and

non-GBR services, considering a variable balance between the goals of maximizing the system throughput and achieving the fairness.

2. We propose an enhanced Joint Scheduling scheme designed to manage heterogeneous traffic for eMBB scenarios: GBR DRBs with different minimum guaranteed bit rate value requirements and non-GBR DRBs.

3. The BestCQI_HD and BestCQI_LS scheduling strategies allow to select radio resources to be allocated, by taking full advantage of the complete view of the channel conditions perceived by all UEs. In fact, before allocating one resource unit to a DRB, they verify whether this choice does severely penalize any other DRB. Moreover, with respect to [16], the proposed scheduling strategies are improved. More specifically, when certain exceptions arise (see Sections 6.2 and 6.3), new selection criteria are adopted to allocate radio resources more efficiently rather than making purely random non-optimal choices.

4. The eJS scheme is extended to work well in realistic scenarios, where non-ideal Connection Admission Controls (CACs) are adopted.

5. Unlike the traditional scheduling algorithms available in the literature, here we consider the 3GPP specifications for 5G NR Release 15 [6] relating to both the CSI reporting methods and the radio resource allocation types. More specifically, we adopt the high-layer configured sub-band CSI reporting method, with the minimum granularity of a sub-band, and radio Resource Allocation Type 0, i.e., radio resources are allocated to DRBs by means of Resource Block Groups (RBGs).

6. Through a large number of simulations in the MATLAB environment, under different amount of UEs, type of services, and the GBR requirements, the performance of the proposed scheduling strategies have been compared with traditional OFDMA scheduling algorithms and the QoS Guaranteed Resource Block Allocation (QGRBA) [30]. The results obtained show that the proposed scheduling scheme reaches the best trade-off between system through-

put and fairness in throughput, while guaranteeing a larger number of GBR DRBs.

2.1 5G NR Downlink Transmission Model

2.1.1 Frame Structure

At the physical layer, 5G NR allows flexible bandwidth with a maximum channel bandwidth per carrier of 400MHz [4]. OFDMA is used for Downlink transmission in 5G NR, where users are dynamically multiplexed on a time-frequency grid. The basic unit is the Resource Element (RE), which consists of one OFDMA symbol and one sub-carrier. To fulfill the very different requirements of 5G usage scenarios, multiple OFDM numerologies are supported. They are defined by sub-carrier spacing (Δf), ranging from 15 kHz to 480 kHz, and Cyclic Prefix (CP) overhead. The proper numerology should be selected on basis of the usage scenario (eMBB, URLLC, and mMTC), and link type (uplink or downlink) [31]. Each symbol carries two, four, six or eight physical channel bits, depending on the modulation scheme (QPSK, 16-QAM, 64-QAM, or 256-QAM, respectively). The symbols are grouped into time slots of 14 OFDM symbols with normal CP. Resource Elements are also grouped into Physical Resource Blocks (PRBs), each of which consists of one time slot and 12 sub-carriers, as shown in Fig. 2.2b.

2.1.2 Channel State Information Reporting methods

To perform an efficient radio resource scheduling, radio resource should be allocated in accordance with the user's channel condition. At this aim, we consider procedures followed by the UE for reporting Channel State Information (CSI) provided by the standard [6]. CSI consists of channel information, such as Channel Quality Indicator (CQI), precoding matrix indicator, and so on. CQI is a four-bit quantity, which indicates the maximum data rate that the mobile can handle with a block error ratio of 10% or below. The CQI mainly depends on the received signal to interference plus noise ratio, but it also depends on the implementation of the mobile receiver, because an advanced receiver can successfully process the incoming data at a lower SINR than a more basic one. The gNB can configure the UE to report the CSI in the Physical Uplink Control

Channel (PUCCH) in three different ways. First, *Wideband reporting method* covers the whole of the downlink band with a single wideband CQI value. Second, a UE configured with the *UE selected sub-band reporting* method selects the sub-bands with the best channel quality. Then, it reports their locations, one CQI value that spans them, and a separate wideband CQI value. Third, by adopting the *higher layer configured sub-band reporting* method, the gNB divides the downlink band into equally sized sub-bands, and the UE reports one CQI value for each sub-band. This sub-band is defined as a fixed number of 4, 8, 16 or 32 contiguous PRBs. In this chapter, we adopt the third reporting method, which is the most accurate and allows to achieve high transmission efficiency, which is one of the stringent requirements for 5G networks [32].

2.1.3 Downlink Transmission

Given the spectrum portion available for a usage scenario, radio resources should be distributed among UEs every Transmission Time Interval (TTI), corresponding to a mini-slot, one slot or several slots. The gNB, on basis of the CQI values received by a UE, begins the downlink transmission procedure by sending to it a scheduling command on the Physical Downlink Control Channel (PDCCH) using the Downlink Control Information (DCI). It alerts the UE to a new data transmission related to a DRB and states how it will be sent. More specifically, it communicates, inter alia, the Resource Block Allocation Type, the Resource Block Assignment, and the Modulation and Code Scheme (MCS) Index, which is a number composed of 5 bits. Then, the gNB transmits data on the Physical Downlink Shared Channel (PDSCH) in the way defined by the scheduling command.

As regards Resource Allocation (RA) Type, the 3GPP in [6] provides two RA Types: Type 0 and Type 1. 5G RA Type 0 is similar to LTE RA Type 0, where PRBs are assigned in frequency domain by means of a bitmap indicating the RBGs that are assigned to the DRB. So, the gNB can allocate only multiples of RBG, where each RBG is a set of $r = 2, 4, 8, 16$ consecutive PRBs, as depicted in Fig. 2.2b. It is not required for the RBGs allocated to the DRB to be consecutive. As regards 5G RA Type 1, it is similar to LTE RA Type 2, where one or more consecutive PRBs could be allocated to each UE. Despite RA Type 1 allowing to

allocate the desired amount of Physical Resource Blocks (rather than multiples of RBG), in this chapter we adopt the RA Type 0 to overcome the constraint of allocating exclusively consecutive PRBs. In addition, RA Type 1 is the most straightforward choice to take full advantage of the adopted CSI reporting method.

2.2 System Model

In this chapter, we consider a downlink transmission scenario of a 5G NR system with a single antenna (i.e., 1 transmission layer). In the frequency domain, the available bandwidth has been divided into different portions of spectrum allocated to different usage scenarios, as shown in Fig. 2.2a. We consider a single portion of spectrum, corresponding to N_{RB} PRBs, dedicated to the eMBB usage scenario, with a fixed numerology and TTI duration. We assume that N_{UE} UEs are served by a gNB. Each UE could be associated one or more DRBs, in case the UE requests services which have different QoS requirements. We assume that the scheduler properly allocates radio resources to N_{DRB} DRBs, where $N_{DRB} \geq N_{UE}$. Each DRB i is a GBR or non-GBR DRB. $N_{DRB} = N_{GBR} + N_{BE}$, where N_{GBR} is the amount of GBR DRBs requiring a GBR, whereas N_{BE} is the amount of non-GBR DRBs, related to best effort services. Let $\mathbf{U_{GBR}} = [1, \ldots, N_{GBR}]^T$ be the vector of GBR DRBs, $\mathbf{U_{BE}} = [N_{GBR} + 1, \ldots, N_{DRB}]^T$ the vector of non-GBR DRBs, $\mathbf{U} = \left[\mathbf{U}_{\mathbf{GBR}}^T, \mathbf{U}_{\mathbf{BE}}^T\right]^T$ the vector of all DRBs, and $\mathbf{R} = [R_1, \ldots, R_{N_{DRB}}]^T$ the vector of the minimum data rate values, where $R_i > 0 \quad \forall i \in \mathbf{U_{GBR}}$, and $R_i = 0 \quad \forall i \in \mathbf{U_{BE}}$.

Every TTI, the scheduler allocates RBGs to DRBs. Let $\mathbf{G}$ be the vector of all available RBGs, where $N_{RBG} = \left\lfloor \frac{N_{RB}}{r} \right\rfloor$. UEs have been configured by the gNB to report CSI values in higher layer configured sub-band way, where each downlink sub-band is equal to $r \cdot 12 \cdot \Delta f$ kHz (i.e., the bandwidth of a RBG). Let $\mathbf{c}_i = [CQI_{i,1}, \ldots, CQI_{i,N_{RBG}}]$ be the row vector containing the CQI values associated to the ith DRB. Each $CQI_{i,k}$ is the CQI value measured in the kth sub-band and reported from the UE related to the ith DRB. $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{N_{DRB}}]^T$ is the matrix of size $N_{DRB} \times N_{RBG}$ containing all CQI values. It provides the complete view of the channel conditions perceived by each UE for each sub-band.

Table 2.1: *Relationship between CQI values and MCS indexes based on [2]*

CQI value	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MCS index	0	0	1	3	5	7	10	11	14	16	18	20	22	24	26	26

Let **A** be the Allocation Matrix of size $N_{DRB} \times N_{RBG}$, whose binary elements $a_{i,k}$ are indicators of the radio resources assigned to DRBs: $a_{i,k} = 1$ if DRB i is assigned RBG k, otherwise $a_{i,k} = 0$. The output of a scheduling algorithm are the $a_{i,k}$ values. CQI values of the RBGs allocated DRB i are reported in row vector $\mathbf{c}'_i$, that is, a subset of vector $\mathbf{c}_i$. It is obtained by extracting the elements of $\mathbf{c}_i$ whose indices are in the set $\{k|k \in \{1, \ldots, N_{RBG}\} \wedge a_{i,k} = 1\}$. As reported in Section 2.1, the MCS index specified by the gNB in the DCI depends on the CQI values in $\mathbf{c}'_i$, for each DRB i. Let us note that the MCS field is a five-bit index, while CQI is a four-bit quantity, therefore to estimate the throughput it is necessary to define a proper mapping between CQI and MCS index. An example based on [2] is shown in Table 2.1.

2.3 Problem Formulation

In this section, we formulate the optimal joint scheduling problem for GBR and non-GBR services Given the data to be transmitted in Downlink for each DRB, the first goal of the optimal scheduling algorithm is to maximize the transmission efficiency, by maximizing the system throughput. It can be expressed as follows:

$$\max \sum_{i=1}^{N_{DRB}} T_i, \tag{2.1}$$

where T_i is the transmission rate achievable by each DRB i. The transmission rate achievable by each DRB in a simulation time t_{SIM} is equal to:

$$T_i = \frac{1}{t_{SIM}} \sum_{h=1}^{\lfloor \frac{t_{SIM}}{TTI} \rfloor} TBS_{i,h}, \tag{2.2}$$

where $TBS_{i,h}$ is the Transport Block Size, that is, the number of information bits transmitted in the hth TTI for DRB i and is estimated on basis of the MCS adopted for it in each RBG allocated to it.

The second goal of the optimal scheduling algorithm is to achieve the fairness in throughput among DRBs. At this aim, we consider the Jain's index [33], a parameter which quantitatively measures the degree of fairness offered by a system allocating resources to N_{DRB} contending requests, as follows:

$$J_{index} = \frac{\left(\sum_{i=1}^{N_{DRB}} x_i\right)^2}{N_{DRB} \sum_{i=1}^{N_{DRB}} x_i^2} \tag{2.3}$$

where x_i is the amount of resource associated to contender i. J_{index} ranges from $\frac{1}{N_{DRB}}$ (i.e., all resources allocated to a single DRB) to 1 (i.e., each DRB receives the same amount of resources). In the case considered, which supports heterogeneous traffic, the fairness condition needs to be evaluated considering that each DRB i requires a service with a minimum guaranteed data rate R_i, which is generally different from each other. For this reason, we evaluate the fairness on the basis of the Throughput Gain of each DRB, which is defined as the difference between the transmission rate achieved by the ith DRB and the guaranteed data-rate required. Therefore, the goal of improving the fairness among DRBs can be expressed as:

$$\max J_{index} \text{ in (3.25)}, \quad \text{where } x_i = T_i - R_i, \forall i \in \mathbf{U} \tag{2.4}$$

Now, we analyze the system constraints. First, the transmission rate achieved by each GBR DRB should be greater or equal to the minimum data rate required to satisfy the QoS. It can be formulated as follows:

$$T_i \geq R_i, \forall i \in \mathbf{U_{GBR}} \tag{2.5}$$

Second, since OFDMA is an Orthogonal Multiple Access (OMA), each RBG must be allocated to one DRB at most. It can be formulated as follows:

$$\sum_{i=1}^{N_{DRB}} a_{i,k} \leq 1, \forall k \in \mathbf{G}, \forall \text{ TTI}. \tag{2.6}$$

A fundamental problem is that goals (2.1) and (2.4) are typically in opposition to each other. On the one hand, adopting a scheduling strategy which aims to maximize the overall system throughput means serving, in a strongly unfair manner, only the DRBs related to UEs who have the

best channel conditions (e.g., UEs near the base station), at the expense of those related to UEs who have worse channel condition (e.g., UEs at the cell edge), such as by adopting BestCQI. On the other hand, adopting a scheduling approach with the goal of maximizing the fairness among DRBs can result in a significant reduction in the system throughput.

For the above reason, we need to define an optimal trade-off between the above goals. The scheduler should minimize the weighted difference between the system throughput normalized with respect to the maximum achievable throughput and the Jain's index normalized with respect to the maximum achievable Jain's index. It can be formulated as follows:

$$\min \left\{ \left| \alpha \frac{\sum_{i=1}^{N_{DRB}} T_i}{T_{max}} - (1-\alpha) \frac{J_{index}}{J_{max}} \right| \right\}, \tag{2.7}$$

where J_{index} is calculated by using (3.25) with $x_i = T_i - R_i$; T_{max} and J_{max} are the maximum throughput and the maximum Jain's index obtainable in the considered system, respectively; and $\alpha \in [0, 1]$ is a weight which considers the balance between the goals of maximizing the system throughput and improving the fairness. Obviously, in the extreme cases (i.e., $\alpha = 1$ or $\alpha = 0$), the optimal scheduler maximizes the system throughput without taking into account the fairness or vice versa, while for $\alpha = 0.5$ the two goals are perfectly balanced. In conclusion, given an α value, the problem of optimal scheduling is to obtain Matrix $\mathbf{A}$ so that (2.7) is achieved, subject to constraints (2.5) and (2.6).

This radio resource scheduling analysis is an optimization problem with high complexity, which increases with the number of constraints and variables (N_{DRB} and N_{RBG}). For this reason, this analytical solution will take significantly long computation time, which is unacceptable for a scheduling procedure which should be made every TTI. Therefore, a heuristic solution which does not take a long computation time is needed to obtain near-optimal solutions.

More heuristic algorithms are possible, which aim at incrementing the system throughput more than the fairness among DRBs, or vice versa, the strategy depends on the network operator. For this reason, in Section 6, we will propose a pseudo-optimal heuristic solution. It is a joint radio resource allocation scheme for GBR and non-GBR services that guarantees the minimum data rate required by GBR DRBs, with two possible

strategies to achieve a different trade-off between the above goals.

2.4 Reference Work

There are several scheduling algorithms known in the literature, which generally are categorized as channel unaware schedulers, channel aware schedulers, and channel and QoS aware schedulers.

Channel unaware schedulers assign radio resource units to UEs without taking into account the channel conditions (i.e., the CQI values of Matrix **C**) and the minimum data rates required by DRBs (i.e., vector **R**). One of the most popular channel unaware schedulers is the Round Robin (RR). This scheduler, for each TTI, assigns radio resources cyclically to DRBs. This is a very simple procedure giving the best fairness among DRBs in terms of number of PRBs allocated. Adopting the 5G RA Type 0, for each $i \in \{1, 2, ..., N_{DRB}\}$, it follows:

$$a_{i,k^*} = \begin{cases} 1, & \text{if } k^* = i + mN_{DRB}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

where $m = 0, 1, 2... \left[\frac{N_{RBG}-i}{N_{DRB}}\right]$. However, this solution offers poor performance both in terms of cell throughput and fairness in throughput. In fact, the same number of RBGs is assigned to DRBs related to different UEs which could have very different channel conditions.

The BestCQI and PF schedulers, unlike RR, take into account the channel condition reporting of UEs, but data rate requirements are still not considered. The first scheduler allocates, for each TTI, each radio resource unit to the DRB with the highest CQI value in it (i.e., the one which can receive data at the highest data rates). According to our system model, for each $k \in \{1, 2, ...N_{RBG}\}$, it follows:

$$a_{i^*,k} = \begin{cases} 1, & \text{if } i^* = \arg\max_{i \in \mathbf{U}} \{c_{i,k}\} \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

This solution maximizes the system throughput of the cell but is very unfair, as far UEs with lower channel condition may not get a chance to receive any data.

The PF scheduler tries to find over time (i.e., a large number of TTIs) a good trade-off between distributing radio resources equally to all DRBs (i.e., RR) and assigning radio resources only to DRBs with the best channel quality (i.e., BestCQI). Every TTI, all radio resources are assigned to a single DRB i' with the highest ratio between the estimated instantaneous data rate at time t, $T_i(t)$ and the average data rate achieved until time t ($\overline{T_i}(t)$), according to equation:

$$i' = \arg\max_{i \in \mathbf{U}} \frac{T_i(t)}{\overline{T_i}(t)} \tag{2.10}$$

where $\overline{T_i}(t) = \beta \overline{T_i}(t-1) + (1-\beta) T_i(t)$, $0 \leq \beta \leq 1$ and t is the discrete time index for the scheduling interval [34]. The value $\overline{T_i}(t)$ is calculated recursively over a defined period which is typically in the order of a second [35]. According to our system model, this means that, for each TTI, it follows:

$$a_{i,k} = \begin{cases} 1, \forall k & \text{if } i \text{ fullfills condition (2.10)} \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

This solution allows UEs near the gNB to be served with a large number of RBGs, whereas far UEs would still have some of the radio resources. However, PF algorithm does not guarantee the QoS requirements of DRBs and is not suitable for real-time services.

As regards channel and QoS aware schedulers, we focus on the QoS Guaranteed Resource Block Allocation (QGRBA) algorithm [30], which takes into account the GRB value of GBR DRBs and the channel condition reporting of UEs. The goal of this scheduling algorithm is to maximize the overall system throughput under the constraints of guaranteeing the required data rates. More specifically, it is proposed a suboptimal radio resource allocation algorithm that comprises two steps. In the first one, for each GBR DRB, the scheduler calculates the ratio between its GBR value and the average value of the CQIs reported by it. Then, on basis of the values obtained, the scheduler estimates the portion of the available bandwidth to be associated to each GBR DRB, in terms of number of PRBs. In the second step, the scheduler defines a resource allocation priority: the GBR DRB with the highest average CQI value is selected first (in the event of a tie, the scheduler selects the GBR DRB

with the lowest GBR value). It allocates to this DRB the amount of PRBs estimated in step 1, by selecting the RBs in which this DRB is experiencing the best channel quality. Then, the scheduler verifies whether the GBR DRB is satisfied. If so, it selects the next GBR DRB in order of priority. Otherwise, an additional PRB is associated recursively to the selected DRB, until it has been satisfied. The overall operation is performed until no PRB is available or all GBR DBRs have been satisfied. In the latter case, in order to improve the system throughput, all the remaining PRBs are assigned only to the DRB with the highest priority.

The main drawbacks of this solution are the following. First, it maximizes the system throughput in a strongly unfair manner, giving any "extra" available radio resource to only one DRB. Second, the minimum number of PRBs allocated to each DRB is established on basis of the average CQI value, thus this amount of PRBs could be overestimated. In addition, QGRBA does not fully exploit the global view of the channel conditions perceived by each UE for each RBG, in the terms that are explained in the following Section. As regards the resource allocation procedure, PRBs are allocated to DRBs without taking into account any RA Type. Also, it is assumed that the scheduler knows the CQI values of each PRB, which is in contrast to the CSI reporting methods provided by the standard.

2.5 Enhanced Joint Scheduling Scheme

Our joint scheduling scheme to support GBR and non-GBR services is summarized in pseudo-code 1 and is based on the following key elements. In order to improve the fairness among DRBs, the scheduler cyclically allocates to DRBs one RBG at a time. The basic principle of the proposed scheme is to give priority to GBR DRBs, by first allocating RBGs only to GBR DRBs until their minimum requirement (R_i) is met, as reported in Lines 1 to 12. Once all GBR DRBs have been satisfied, the scheduler assigns one RBG (if any) to each non-GBR DRB, because they have not received any radio resources yet (Lines 13-16). Finally, non-GBR and GBR DRBs are treated the same way, by allocating RBGs cyclically until there are no more channels (i.e., RBGs) available (see Lines 17-20). Let us underline that in Lines 4, 15, and 19 RBG alloca-

tion is properly done by means of one out the two scheduling strategies that are proposed and explained in following subsections.

It is clear that this procedure can work well only whether the available resources (i.e., the amount of RBGs) are enough to satisfy the requirement of all GBR DRBs. In other words, when the network becomes congested, it is required an ideal Connection Admission Control (CAC) that automatically reduces the amount of accommodated GBR DRBs, so that each GBR requirement can be satisfied [36]. For the sake of simplicity, the procedure in pseudo-code 1 has been described under the assumption of an ideal CAC. In addition, the complete description of our eJS scheme, that takes into account a realistic scenario with a non-ideal CAC, is provided in subsection 2.5.4.

Pseudo-code 1 QoS aware Joint Scheduling Scheme

Definitions:

- $\mathbf{U} = [1, 2, \ldots, N_{DRB}]^T$: vector of DRBs
- $\mathbf{U_{GBR}} = [1, 2, ..., N_{GBR}]^T$: vector of GBR DRBs
- $\mathbf{U_{BE}} = [N_{GBR} + 1, N_{GBR} + 2, ..., N_{DRB}]^T$: vector of non-GBR DRBs
- $\mathbf{U_C}$: column vector containing the DRBs to be scheduled
- $\mathbf{G} = [1, 2, ..., , N_{RBG}]$: vector of all RBGs
- **G'**: vector of all available RBGs
- $\mathbf{R} = [R_1, ..., R_{N_{DRB}}]^T$: vector of the guaranteed data rates
- **C**: CQI matrix whose elements are $c_{i,k}$

Iteration:

1: $\mathbf{U_C} = \mathbf{U_{GBR}}$; $\mathbf{G'} = \mathbf{G}$ {RBGs will be allocated only to GBR DRBs}
2: **while** $\mathbf{U_C} \neq \emptyset \wedge \mathbf{G'} \neq \emptyset$ **do**
3: Give $\tilde{\mathbf{C}}$ as a sub-matrix of **C** containing elements $c_{i,k}$ so that $i \in \mathbf{U_C}$ and $k \in \mathbf{G'}$
4: Assign one RBG to each DRB $i \in \mathbf{U_C}$, by means of the selected Scheduling Strategy (see Pseudo-code 2).
5: Remove the allocated RBGs from vector **G'**.
6: **for** each DRB $i \in \mathbf{U_C}$ **do**
7: Estimate T_i.
8: **if** $T_i \geq R_i$ **then**
9: Remove i from vector $\mathbf{U_C}$.
10: **end if**
11: **end for**
12: **end while**
13: **if** $\mathbf{G'} \neq \emptyset$ {All GBR DRBs have been satisfied and there are still available RBGs} **then**
14: $\mathbf{U_C} = \mathbf{U_{BE}}$ {One RBG will be allocated only to each non-GBR DRB}
15: Give $\tilde{\mathbf{C}}$ and assign one RBG to each DRB $i \in \mathbf{U_C}$, by means of the selected Scheduling Strategy (see Pseudo-code 2). Remove from vector **G'** the allocated RBGs.
16: **end if**
17: **while** $\mathbf{G'} \neq \emptyset$ **do**
18: $\mathbf{U_C} = \mathbf{U}$ {The remaining RBGs will be allocated one by one to GBR and non-GBR DRBs}
19: Give $\tilde{\mathbf{C}}$ and assign one RBG to each DRB $i \in \mathbf{U_C}$, by means of the selected Scheduling Strategy (see Pseudo-code 2). Remove from vector **G'** the allocated RBGs.
20: **end while**

2.5.1 Overview of the Proposed Scheduling Strategies

As regards our scheduling strategies, unlike BestCQI and QGBRA, in order to achieve the established goals, we propose to allocate radio resources, by fully exploiting the complete view of the channel conditions perceived by each UE for each RBG.

In the considered eMBB scenario, the ideal solution is that each DRB is allocated the RBG in which the best channel condition is experienced (i.e., the highest value of CQI), so that the highest possible MCS index will be adopted in the PDSCH. Therefore, data will be received with the maximum value of throughput achievable for it. However, if two or more DRBs experience the best channel quality in the same RBG k, the scheduler must choice to which DRB the sub-band k should be allocated, penalizing the other DRBs.

For this reason, unlike the reference schedulers, both the proposed scheduling strategies, before allocating one RBG to a DRB, estimate how much this allocation could severally penalize the system throughput and/or the fairness in throughput among DRBs.

The first strategy (BestCQI Highest Deviation, BestCQI_HD) is proposed to improve the system throughput at the expense of a limited loss in terms of fairness, while the second one (BestCQI Lowest Second, BestCQI_LS) aims at improving the fairness among DRBs at the expense of a limited loss in terms of the system throughput.

To illustrate the basic idea of each proposed scheduling strategy in comparison with the behavior of traditional BestCQI scheduler, we present a very simple example with 3 non-GBR DRBs and 3 RBGs.

Let us consider the following Matrix $\tilde{\mathbf{C}}$, containing all CQI values.

$$\tilde{\mathbf{C}} = \begin{bmatrix} 12 & 11 & 8 \\ 14 & 6 & 10 \\ 11 & 9 & 8 \end{bmatrix}. \tag{2.12}$$

By applying the BestCQI Algorithm the Allocation Matrix and the CQI values related to the RBGs assigned to each DRB ($\mathbf{c}'_i$ for each DRB i) are reported in Table 2.2.

In order to quantitatively evaluate goals (2.1) and (2.4), the T_i of each DRB i must be calculated. Let us remember that the throughput is a func-

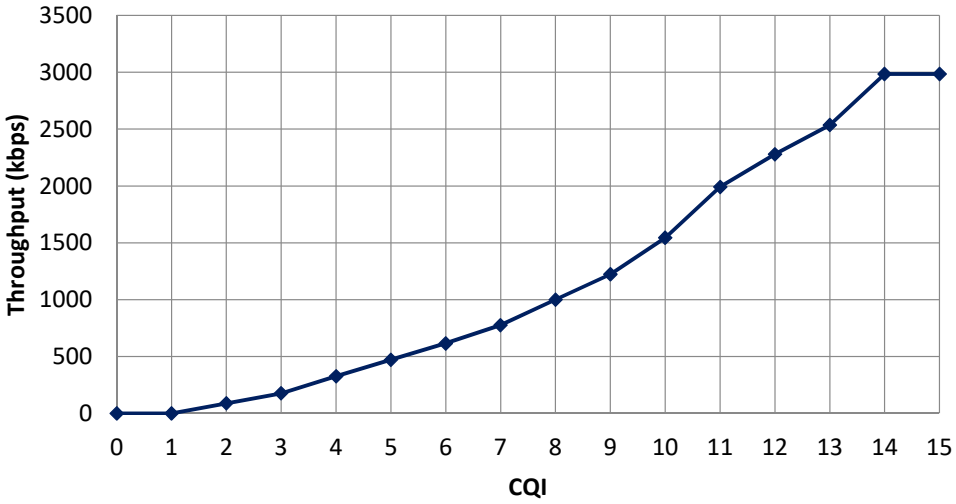


Figure 2.3: *Relationship between the CQI values and the throughput achieved with 1 RBG, when $r = 4$ and $\Delta f = 15kHz$.*

tion of the CQI values in $\mathbf{c}'_i$. By adopting Table 2.1 and the guidelines provided in [2, 6], in the way illustrated in Section 2.6.1, a possible relationship between each CQI value and the related throughput in case of $r = 4$ is depicted in Fig. 2.3. It shows that the throughput is a monotonically increasing function of CQI values in the sub-interval CQI $\in [1, 14]$. For this reason, a proper Throughput Indicator (i.e., an indirect estimate of the throughput) can be obtained as follows:

$$\tilde{T}_i = \sum_{k=1}^{N_{RBG}} a_{i,k} c_{i,k}. \tag{2.13}$$

Then, we define a System Throughput Indicator as:

$$\tilde{T} = \sum_{i=1}^{N_{DRB}} \tilde{T}_i. \tag{2.14}$$

Also, a Fairness Indicator ($\tilde{J}$) can be estimated by using (3.25), where $x_i = \tilde{T}_i$. Therefore, the output of the BestCQI algorithm applied in example (4.2) yields the maximum System Throughput Indicator value ($\tilde{T} = 35$), but this has been achieved in a strongly unfair manner. In fact, only two out of three DRBs are served, and $\tilde{J} = 0.58 \ll 1$ (see Table 2.2).

Now, we analyze how the BestCQI_HD works in the same use case (4.2). Because 3 DRBs are competing with the same RBG (i.e., RBG 1),

in order to take into account the impact of the choice which will be made, for each DRB i the BestCQI_HD scheduler evaluates a *deviation* value, which represents the difference between its maximum CQI value and its second maximum. Then, the scheduler assigns RBG 1 to the DRB with the greatest value of deviation, i.e., DRB 2 which presents the highest deviation from its maximum reachable throughput. Consequently, a different resource allocation for DRB 2 would lead to a greater deterioration of the system throughput.

Once RBG 1 is assigned to DRB 2, in matrix $\tilde{\mathbf{C}}$ row 2 and column 1 are deleted:

$$\tilde{\mathbf{C}} = \begin{bmatrix} - & 11 & 8 \\ - & - & - \\ - & 9 & 8 \end{bmatrix}. \tag{2.15}$$

Next, because 2 DRBs are competing with the same RBG (i.e., RBG 2), the above strategy is applied again. The scheduler assigns RBG 2 to DRB 1. Then, the remaining RBG is assigned to DRB 3.

Finally, the output of BestCQI_HD is reported in Table 2.2. We obtain that the system throughput indicator ($\tilde{T} = 33$) is slightly lower than BestCQI, whereas fairness is strongly improved. In fact, all DRBs are served, and $\tilde{J} = 0.94$.

As regards the second strategy (BestCQI_LS) applied to the same use case (4.2), let us remember that it aims to improve the fairness among DRBs, avoiding that a DRB experiences too low performance, even at the cost of reducing the overall system throughput. While BestCQI_HD assigns a RBG to the DRB which presents the largest deviation between its maximum CQI value and its second maximum, the approach of BestCQI Lowest Second takes into account the value of its second maximum CQI. More specifically, RBG 1 is assigned to DRB 3 which presents the lowest second maximum value. This strategy aims to prevent a DRB from being served with a too little CQI value, therefore with a too low transmission rate. Row 3 and column 1 are deleted from matrix $\tilde{\mathbf{C}}$.

$$\tilde{\mathbf{C}} = \begin{bmatrix} - & 11 & 8 \\ - & 6 & 10 \\ - & - & - \end{bmatrix}. \tag{2.16}$$

Table 2.2: *Output of Scheduling Strategies*

Scheduler	Allocation Matrix	CQI values of assigned RBGs	$\tilde{T}$	$\tilde{J}$
BestCQI	$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$	$\mathbf{c}'_1 = \{11\}$ $\mathbf{c}'_2 = \{14, 10\}$ $\mathbf{c}'_3 = \{\}$	35	0.58
BestCQI_HD	$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\mathbf{c}'_1 = \{11\}$ $\mathbf{c}'_2 = \{14\}$ $\mathbf{c}'_3 = \{8\}$	33	0.94
BestCQI_LS	$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$	$\mathbf{c}'_1 = \{11\}$ $\mathbf{c}'_2 = \{10\}$ $\mathbf{c}'_3 = \{11\}$	32	0.998

Now, all DRBs are experiencing their best channel conditions in a different RBG, therefore, the scheduler assigns RBG 2 to DRB 1, and RBG 3 to DRB 2.

The output of BestCQI_LS is reported in Table 2.2. We obtain that the system throughput ($\tilde{T} = 32$) is slightly lower than the one of BestCQI_HD, whereas fairness is improved ($\tilde{J} = 0.998$), very close to the ideal value.

Until now, we have described the basic approach of our strategies, as proposed in [16]. However, several exceptions could occur and the new solutions adopted by our strategies are described in detail in the following subsections.

2.5.2 BestCQI_Highest Deviation

In this subsection, we describe the improved BestCQI_HD scheduling strategy, which can be adopted in Lines 4, 15 and 19 of pseudo-code 1.

As reported in pseudo-code 2, the inputs of this scheduling strategy are: column vector $\mathbf{U_C}$ containing DRBs to be scheduled, row vector **G'** containing the available RBGs, and matrix $\tilde{\mathbf{C}}$, that is, a sub-matrix of $\mathbf{C}$ containing elements $c_{i,k}$ so that $i \in \mathbf{U_C}$ and $k \in$ **G'**. Each element of $\tilde{\mathbf{C}}$ is $\tilde{c}_{i,k}$.

Firstly, the scheduler initializes two vectors: $\mathbf{U'_C} = \mathbf{U_C}$ and $\mathbf{U_B} = \emptyset$.

The first exception to be observed is when $|\mathbf{G'}| < |\mathbf{U_C}|$, which means that no RBGs will be allocated to $(|\mathbf{U_C}| - |\mathbf{G'}|)$ DRBs. In this case, if the scheduler operates as described in the previous example, then it

could allocate sub-bands to DRBs with a high deviation value, but with bad channel conditions. For this reason, we introduce a new criterion to choose the proper subset of $\mathbf{U_C}$ containing $|\mathbf{G'}|$ DRBs to be served with one RBG, before calculating the deviation values. It is a pseudo-optimal solution, with the aim of maintaining a low computational load. Rather than eliminating from the allocation procedure those DRBs with the worst average CQI value (in the same way as QGRBA), we consider that the BestCQI_HD strategy aims at allocating for each DRB the RBG corresponding to the maximum or the second maximum CQI value. Keeping this in mind, for each DRB the scheduler evaluates the average value between its maximum and its second maximum CQI. Then, it selects those $(|\mathbf{U_C}| - |\mathbf{G'}|)$ DRBs with the lower average value previously calculated and deletes them from $\mathbf{U'_C}$ and the related rows from matrix $\tilde{\mathbf{C}}$, as reported from Line 2 to Line 11. Next, in Line 13, the controller calculates a matrix $\mathbf{D}$ with the same size of $\tilde{\mathbf{C}}$, which represents the location of the maximum CQI values of each DRB. Binary elements $d_{i,k}$ of $\mathbf{D}$, for each row i^*, are calculated as follows:

$$d_{i^*,k} = \begin{cases} 1, & \text{if } k \in \underset{k \in \{1,\dots,|\mathbf{G'}|\}}{\arg\max} \{\tilde{c}_{i,k}\}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.17}$$

The second exception to be observed is when one UE experiences its best channel quality in more than one RBG, termed Best RBGs, as reported in Line 15. i.e., $\exists$ at least one $i \in \mathbf{U'_C}$ so that $\sum_{k=1}^{|\mathbf{G'}|} d_{i,k} > 1$. In this case, the scheduler has the possibility to assign to the related DRB any RBG out of its Best RBGs, while maintaining the same performance for it in terms of throughput. However, the choice of one RBG rather than another one belonging to the same set of Best RBGs could affect the performance of other DRBs. For this reason, rather than making a random choice, we introduce a further criterion to choose, for each DRB, the proper RBG among the set of Best RBGs to be marked as its first maximum value (i.e., the best sub-band candidate to be assigned to it). Our objective for this criterion is to minimize the number of DRBs competing for the same RBG, and its implementation is equivalent to minimizing the amount of 1s per column in matrix $\mathbf{D}$. The details are reported in Pseudo-code 2, from Line 19 to 31. As output, we obtain a new matrix $\mathbf{D}$ in which, for each row $i \in \mathbf{U'_C}$, it follows $\sum_{k=1}^{|\mathbf{G'}|} d_{i,k} = 1$.

Once the above operations have been carried out, the scheduler checks if $\forall k^* \in \{1, \ldots, |\mathbf{G'}|\}$, it follows $\sum_{i=1}^{N_{DRB}} d_{i,k^*} \leq 1$, as reported in Line 32. If so, the controller can immediately allocate to each DRB i the RBG k so that $d_{i,k} = 1$, because no DRB competes for the same sub-band. Otherwise, more DRBs are competing for the same RBG(s). In this case, the scheduler applies the strategy described below for the entire matrix $\tilde{\mathbf{C}}$, even in the situation in which only two DRBs compete for a single RBG. We have made this choice to fully exploit the complete vision of matrix $\tilde{\mathbf{C}}$, by taking into account the impact of each choice on the next ones.

At this point, the controller calculates a Deviation Matrix $\mathbf{X} = \mathbf{Q}\mathbf{I}_{\mathbf{U'_C}} - \tilde{\mathbf{C}}$, where $\mathbf{I}_{\mathbf{U'_C}}$ is a row vector of all ones with length $\left|\mathbf{U'_C}\right|$ and $\mathbf{Q} = [q_1, \ldots, q_{|\mathbf{U'_C}|}]^T$ is a column vector whose element q_{i^*} is:

$$q_{i^*} = \max_{k \in \{1,\ldots,|\mathbf{G'}|\}} \left\{\tilde{c}_{i^*,k}\right\}. \tag{2.18}$$

Then, the deviation value between the maximum and the second maximum of each DRB are inserted into a vector $\mathbf{E} = [e_1, \ldots, e_{|\mathbf{U'_C}|}]^T$. Each element e_{i^*} is calculated as follows:

$$e_{i^*} = \min_{\substack{k \in \{1,\ldots,|\mathbf{G'}|\} \\ k \neq m_{i^*}}} \left\{x_{i^*,k}\right\}, \tag{2.19}$$

where $x_{i,k}$ is an element of matrix $\mathbf{X}$ and:

$$m_{i^*} = \arg\max_{k \in \{1,\ldots,|\mathbf{G'}|\}} \left\{d_{i^*,k}\right\}. \tag{2.20}$$

Finally, the controller selects the DRB corresponding to row i' of matrix $\tilde{\mathbf{C}}$, where:

$$i' = \arg\max_{i \in \{1,\ldots,|\mathbf{U'_C}|\}} \left\{e_i\right\} \tag{2.21}$$

In case of parity, a random choice of i' is made. RBG $\mathbf{G'}[m_{i'}]$ is assigned to DRB $\mathbf{U'_C}[i']$, with a CQI value equal to $q_{i'}$ (i.e., $a_{\mathbf{U'_C}[i'],\mathbf{G'}[m_{i'}]} = 1$). Because one DRB has been served with one RGB, row i' is deleted from Matrix $\tilde{\mathbf{C}}$ and from vector $\mathbf{U'_C}$. Column $m_{i'}$ is deleted from Matrix $\tilde{\mathbf{C}}$ and from vector $\mathbf{G'}$.

The overall procedure is iteratively performed until Matrix $\tilde{\mathbf{C}}$ is empty (i.e., all DRBs are served or all available RBGs are allocated).

Pseudo-code 2 Channel aware scheduling strategies

Definitions:

- $\mathbf{U_C}$: vector of DRBs to be scheduled;
- **G'**: vector of all available RBGs;
- **C**: CQI matrix whose elements are $c_{i,k}$;
- $\tilde{\mathbf{C}}$: sub-matrix of **C**;
- **A**: Allocation Matrix whose elements are $a_{i,k}$.

Iteration:

$\mathbf{U'_C} = \mathbf{U_C}$; $\mathbf{U_B} = \emptyset$;
if $|\mathbf{G'}| < |\mathbf{U'_C}|$ **then**
  $\mathbf{U_A} = \mathbf{U'_C}$
  for each DRB $i \in \mathbf{U'_C}$ **do**
    Extract from matrix $\tilde{\mathbf{C}}$ the elements of row i and insert them into a vector $\tilde{\mathbf{c}}_i$;
    Sort $\tilde{\mathbf{c}}_i$ in non-increasing order and calculate $h_i = \frac{\tilde{\mathbf{c}}_i[1]+\tilde{\mathbf{c}}_i[2]}{2}$;
  end for
  Sort $\mathbf{U_A}$ in such a way that, $\forall$ index j of $\mathbf{U_A}$, it follows: $h_{\mathbf{U_A}[j-1]} \leq h_{\mathbf{U_A}[j]} \leq h_{\mathbf{U_A}[j+1]}$;
  Remove the last $|\mathbf{G}|$ elements from vector $\mathbf{U_A}$;
  Delete from vectors $\mathbf{U'_C}$ the elements whose indices are in $\mathbf{U_A}$ and from matrix $\tilde{\mathbf{C}}$ the rows in $\mathbf{U_A}$;
end if
while $\tilde{\mathbf{C}} \neq \emptyset$ **do**
  Calculated Matrix **D** by using (2.17);
  for each DRB $i \in \mathbf{U'_C}$ **do**
    if $\sum_{k \in \{1,\ldots,|\mathbf{G'}|\}} d_{i,k} > 1$ **then**
      Insert i into vector $\mathbf{U_B}$
    end if
  end for
  if $\mathbf{U_B} \neq \emptyset$ **then**
    repeat
      for each DRB $i \in \mathbf{U_B}$ **do**
        if $\exists$ at least one k^* so that $d_{i,k^*} = 1 \wedge \sum_{i \in \{1,\ldots,|\mathbf{U'_C}|\}} d_{i,k^*} = 1$ **then**
          Get k^*. In case of parity, a random choice of k^* is made. Set $d_{i,k} = 0, \forall k \neq k^*$. Remove DRB i from $\mathbf{U_B}$
        end if
      end for
    until new changes in **D** are made or $\mathbf{U_B}$ is empty
    while $\mathbf{U_B} \neq \emptyset$ **do**
      Get $k^* = \underset{k \in \{1,\ldots,|\mathbf{G}|\}}{\arg\min} \sum_{i \in \{1,\ldots,|\mathbf{U'_C}|\}} d_{i,k}$. In case of parity, a random choice of k^* is made.
      A random value of $i^* \in \mathbf{U_B}$ so that $d_{i^*,k^*} = 1$ is chosen. Set $d_{i^*,k} = 0, \forall k \neq k^*$ and remove DRB i^* from $\mathbf{U_B}$.
    end while
  end if
  if $\forall k^* \in \{1,\ldots,|\mathbf{G'}|\}$, it follows $\sum_{i=1}^{N_{DRB}} d_{i,k^*} \leq 1$ **then**
    $\forall i \in \{1,\ldots,|\mathbf{U'_C}|\}$, calculate m_i in (2.20), and set $a_{\mathbf{U'_C}[i],\mathbf{G}[m_i]} = 1$, $\tilde{\mathbf{C}} = \emptyset$.
  else
    Calculate $\begin{cases} \text{vector } \mathbf{E} \text{ by using (2.19) and } i' \text{ in (2.21)} & \text{if BestCQI\_HD is selected} \\ \text{vector } \mathbf{E} \text{ by using (2.22) and } i' \text{ in (2.23)} & \text{if BestCQI\_LS is selected} \end{cases}$
    Set $a_{\mathbf{U'_C}[i'],\mathbf{G'}[m_{i'}]} = 1$.
    Remove row i' from Matrix $\tilde{\mathbf{C}}$ and from vector $\mathbf{U'_C}$.
    Remove Column $m_{i'}$ from Matrix $\tilde{\mathbf{C}}$ and from vector **G'**.
  end if
end while

2.5.3 BestCQI Lowest Second

In this subsection, we describe the BestCQI_LS scheduling strategy which can be adopted in Lines 4, 15 and 19 of pseudo-code 1.

For greater clarity, the explanation follows pseudo-code 2 as a reference. The exceptions to be observed are the same of BestCQI_HD. For this reason, we adopt the same criteria from Line 2 to 31. As regards the remaining operations to be carried out, since the approach of BestCQI_LS takes into account the CQI values of the second maximum, instead of the deviation values, pseudo-code for BestCQI_LS is exactly the same of BestCQI_HD, with the exception of vector **E** and i' in Line 35. Each element of **E** is calculated as follows:

$$e_{i^*} = \max_{\substack{k \in \{1,\dots,|\mathbf{G'}|\} \\ k \neq m_{i^*}}} \{\tilde{c}_{i^*,k}\} \tag{2.22}$$

The value i' is calculate as:

$$i' = \arg\min_{i \in \{1,\dots,|\mathbf{U_C}|\}} \{e_i\} \tag{2.23}$$

In case of parity, a random choice of i' is made.

2.5.4 Enhanced Joint Scheduling Scheme for a non-ideal CAC

As mentioned at the beginning of this Section, our control scheme can work well under the assumption of an ideal Connection Admission Control (CAC) which guarantees that the GBR value of each admitted GBR DRB can be satisfied by efficiently allocating RBGs. However, in a more realistic scenario with a non-ideal CAC, the number of admitted GBR DRBs could be larger than the maximum amount of those which can be fulfilled, even in the case of adopting an ideal scheduler. In this scenario, the proposed scheduling scheme could be not efficient. In fact, on the one hand, serving cyclically all GBR DRBs with one RBG at a time until their requirement is satisfied leads to an improved fairness among DRBs. On the other hand, in the case of a too large number of GBR DRBs, this strategy can lead to distributing the available RBGs among all GBR DRBs, without satisfying anyone. To solve this issue, we need to adapt our scheme to the case of a realistic CAC. To better understand

Table 2.3: *Output of Scheduling Strategies*

	Allocation Matrix	CQI values	satisfied GBR DRBs
QGRBA	$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$	$\mathbf{c}'_1 = \emptyset$ $\mathbf{c}'_2 = [5, 9, 5]$ $\mathbf{c}'_3 = \emptyset$ $\mathbf{c}'_4 = [11, 9]$	$\{2, 4\}$
BestCQI_HD (First TTI)	$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$	$\mathbf{c}'_1 = [13]$ $\mathbf{c}'_2 = [9]$ $\mathbf{c}'_3 = [14]$ $\mathbf{c}'_4 = [8, 6]$	$\{1, 3\}$
eJS scheme BestCQI_HD (Next TTI)	$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$	$\mathbf{c}'_1 = [13]$ $\mathbf{c}'_2 = \emptyset$ $\mathbf{c}'_3 = [14]$ $\mathbf{c}'_4 = [8, 5, 6]$	$\{1, 3, 4\}$

our solution, we provide an example which describes our proposal in comparison to the one adopted by the QGRBA algorithm, which supports GBR DRBs.

Let us consider the following Matrix $\tilde{\mathbf{C}}$, containing the CQI values of 4 admitted GBR DRBs (with GBR of 2Mbps) to be scheduled by proper allocating the 5 available RBGs.

$$\tilde{\mathbf{C}} = \begin{bmatrix} 13 & 12 & 2 & 4 & 4 \\ 9 & 8 & 5 & 9 & 5 \\ 3 & 14 & 4 & 4 & 5 \\ 11 & 9 & 8 & 5 & 6 \end{bmatrix}. \tag{2.24}$$

As regards the QGRBA, first the scheduler calculates the DRBs' priority on basis of the average CQI values and the required transmission rates, as reported in Section 3.1. In this example, all GBR DRBs have the same GBR value, therefore the priority depends only on their average CQI value. The priority order is $\{4, 2, 1, 3\}$. Then, RBGs are allocated to DRB 4, until the GBR requirement is met. Then, next DRB, in order of priority, is served. In this example, we obtain Allocation Matrix and $\mathbf{c}'_i$ vectors reported in Table 2.3. As shown, the scheduler serves only two out of four DRBs, by fulfilling their requirements.

As regards our proposed scheme, we analyze the solution of example (2.24) obtained by adopting the BestCQI_HD strategy. The scheduler serves all DRBs, but only two DRBs satisfy their GBR requirements. In fact, considering the CQI values in $\mathbf{c}'_2$ and $\mathbf{c}'_4$ and the related throughput values in Fig. 2.3, we observe that DRBs 2 and 4 achieve 1.224 Mbps and 1.616 Mbps, respectively. So, three RBGs have been wasted, because have been allocated to two GBR DRBs which do not meet their minimum data rate requirements.

In order to overcome this issue, our eJS scheme works as follows. At the end of the scheduling procedure in a TTI, the scheduler verified if one or more GBR DRBs are unsatisfied. If so, for the next TTI, one GBR DRB must be deleted from the set of admitted GBR DRB, aiming to successfully serve all the remaining GBR DRBs. Among all unsatisfied GBR DRBs, the scheduler selects as GBR DRB to be discarded the one with the worst sum of CQI values related to the sub-bands allocated to it. Let us note that our strategy is different from the one adopted by QGRBA scheme, where the scheduler determines which DRB should not be served on the basis of the CQI value averaged over the whole band. In example (2.24), at the end of the first TTI, the DRB 2 is discarded and the scheduler allocates radio resources only to DRB 1, 3, and 4. As reported in Table 2.3, the eJS scheme satisfies three out of four GBR DRBs, outperforming the QGRBA. In general, the procedure of deleting one unsatisfied GBR DRB per TTI is executed iteratively until the requirement of each accommodated GBR DRB is fulfilled.

2.6 Performance Evaluation

In this section, we evaluate the performance of the proposed scheduling scheme by simulations in MATLAB environment. Results are averaged over 50 independent simulations.

2.6.1 Simulation Assumptions

The simulation scenario is composed of a downlink transmission system of a single gNB. We assume CQI feedback values of all N_{UE} UEs are known at the gNB for all available RBGs for each TTI. The CQI values in Matrix $\mathbf{C}$ are uniformly distributed uniformly from CQI_{min} to CQI_{max}.

Table 2.4: *System Parameters*

Parameter	Symbol	Value
Number of RBGs	N_{RBG}	25
Number of PRBs per RBG	r	4
Sub-carrier spacing	Δf	$15kHz$
Time slot length	t_s	$0.5ms$
Proportional fair weight	β	0.001
Number of symbols per time slot	N_{symbol}	7
TTI length	TTI	$1ms$
Number of UEs	N_{UE}	$2, 5, 10, 15, 20, 25$
Number of DRBs per UE		1
Guaranteed data rate	R	$0.768, 2, 4Mbps$
Simulation length	t_{SIM}	$1000TTI$
CQI values	CQI_{min}	2
	CQI_{max}	14

As reported in Section 4.3, the gNB needs to define a proper mapping between CQI values received by each UE and the related MCS index. To establish what MCS index should be used in the PDSCH, we adopt the CQI-MCS index mapping Table A.4-14 in [2], which is shown in Table 2.1. On basis of the MCS Index, the modulation order and code rate are derived by means of Table 5.1.3.1-1 in [6], which supports QPSK, 16-QAM or 64-QAM. Then, the Transport Block Size (TBS_i) related to the ith DRB, that is the number of information bits transmitted in 1 TTI, is estimated on basis of the number of allocated PRBs and the MCS adopted for each RBG, as reported in [6]. The resulting throughput experienced in one RBG is shown in Fig. 2.3.

We consider that each UE is associated with a single DRB ($N_{UE} = N_{DRB}$), which can be either a GBR or non-GBR DRB. As regards the traffic model, for each DRB we assume a greedy source that always has data to transmit and generates them at the maximum rate possible [37]. The set of parameters used in simulations is provided in Table 5.1. However, the control scheme works in general, even if a different numerology and/or a larger available bandwidth are adopted.

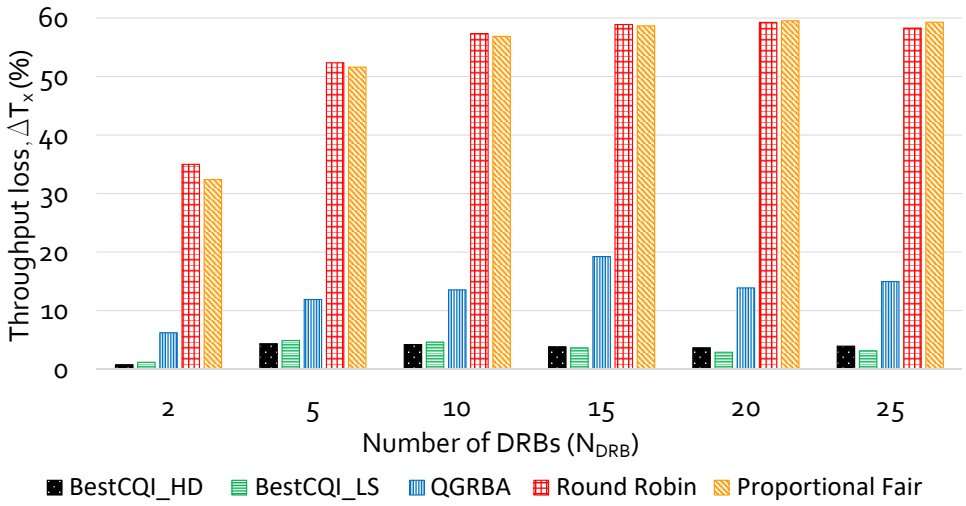


Figure 2.4: *Cell Throughput Loss compared to BestCQI, under different amount of DRBs.* $R_i = 2Mbps$ *for each DRB* i.

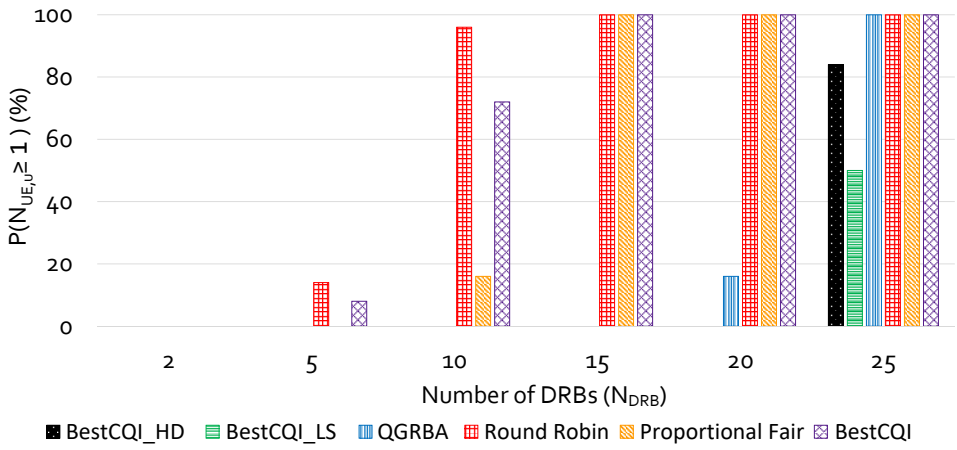


Figure 2.5: *Probability that at least a GBR service does not reach the minimum guaranteed data rate.* $R_i = 2Mbps$ *for each DRB* i.

2.6.2 Performance Metrics

- Cell Throughput

$$T_x = \sum_{j=1}^{N_{DRB}} T_i. \tag{2.25}$$

It is the cell throughput achieved by means of scheduling algorithm x, with $x = \{$BestCQI_HD, BestCQI_LS, QGRBA, RR, PF, BestCQI$\}$ and T_i the throughput of DRB i.

resume Cell Throughput Loss

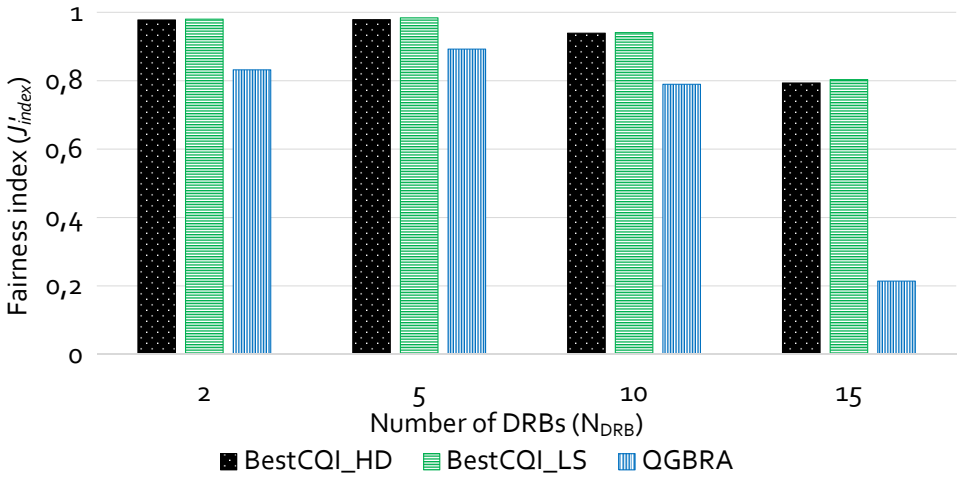


Figure 2.6: *Fairness index under different number of DRBs.* $R_i = 2Mbps$ *for each DRB* i.

$$\Delta T_x(\%) = \frac{T_{BestCQI} - T_x}{T_{BestCQI}} 100, \tag{2.26}$$

where $T_{BestCQI}$ is the maximum cell throughput achievable in the simulated channel conditions by means of BestCQI. ΔT_x is the percentage throughput loss compared to $T_{BestCQI}$. The smaller the value of ΔT_x, the better the performance of x.

resume Number of unsatisfied GBR DRBs and number of satisfied GBR DRBs.

The number of unsatisfied GBR DRBs $(N_{GBR,u})$ is the amount of GBR DRBs that do not meet their stringent GBR requirement, that is:

$$N_{GBR,u} = \sum_{j=1}^{N_{GBR}} h, \text{ where } h = \begin{cases} 1, & \text{if } T_j < R_j \\ 0, & \text{otherwise.} \end{cases} \tag{2.27}$$

Obviously, the number of satisfied GBR DRBs $(N_{GBR,s})$ is the complementary number for N_{GBR}.

resume Probability of unsatisfied GBR DRBs

Let P $(N_{GBR,u} \geq 1)$ be the probability that at least one GBR DRB does not meet its GBR requirement.

resume Fairness Index

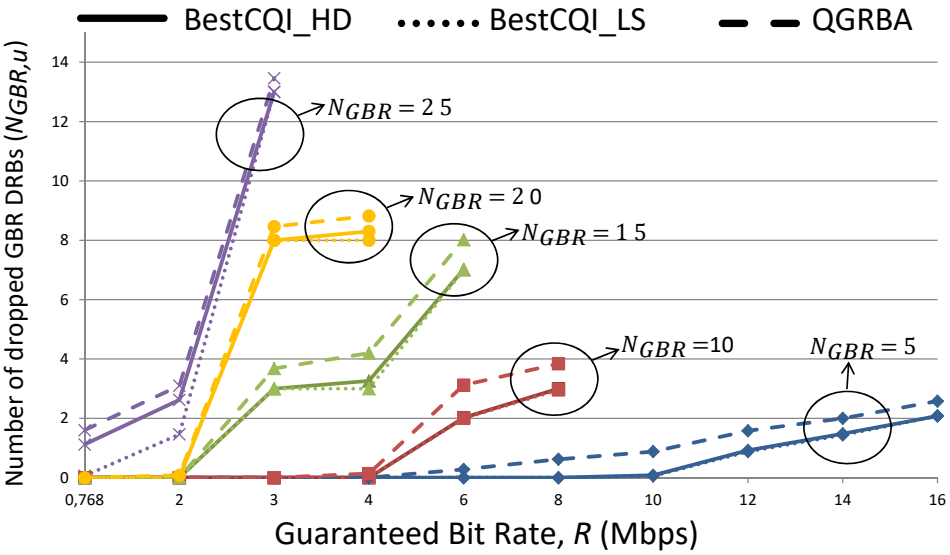


Figure 2.7: *Number of dropped GBR DRBs under different minimum required data rate (R) and number of DRBs (N_{DRB}). $R_i = R$ for each DRB i.*

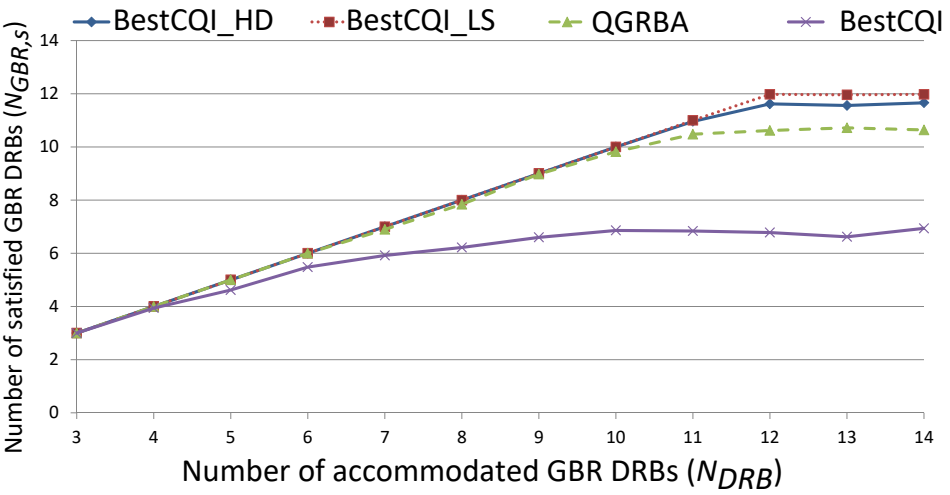


Figure 2.8: *Number of satisfied GBR services under different number of DRBs. $R_i = 4Mbps$ for each DRB i.*

As fairness index, we consider J_{index} in (3.25), where we assume $x_i = T_i - R_i$, $\forall i \in \{1, ..., N_{DRB}\}$. Because J_{index} varies from $\frac{1}{N_{DRB}}$ to 1, in order to correctly compare J_{index} under different number of users, we derive from (3.25) a modified fairness parameter J'_{index} which varies from 0 to 1 regardless of the number of DRBs, as follows:

$$J'_{index} = \left(J_{index} - \frac{1}{N_{DRB}} \right) \frac{N_{DRB}}{N_{DRB} - 1} \tag{2.28}$$

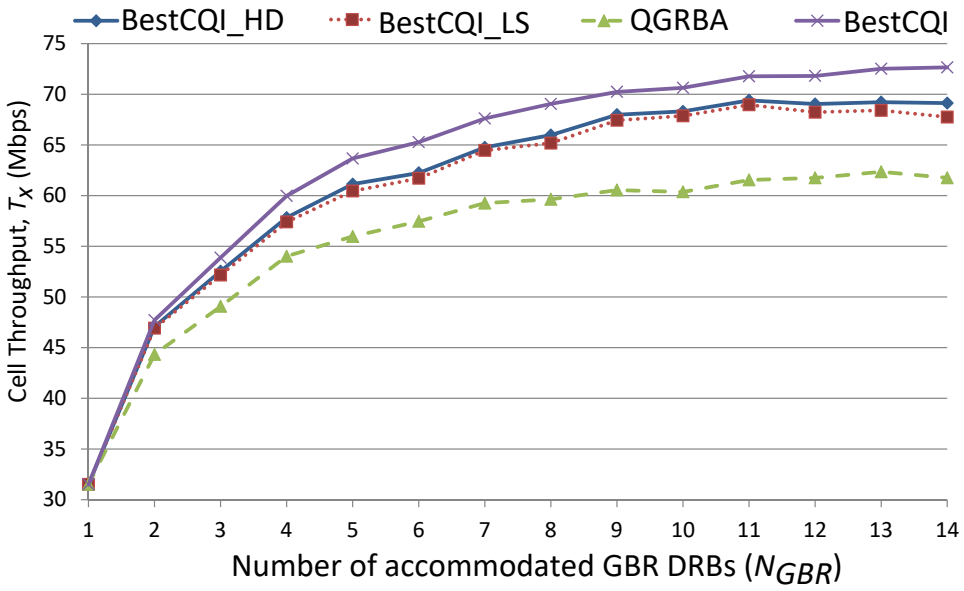


Figure 2.9: *Cell Throughput under different number of DRBs.* $R_i = 4Mbps$ *for each DRB* i.

2.6.3 Performance Analysis

We compare the eJS scheme with RR, BestCQI, PF, and QGRBA. In order to achieve a fair evaluation, we adapt these reference algorithms to our System Model, by using the high-layer configured sub-band CSI reporting method and the 5G RA Type 0, as described in Section 3.1. In the following, we analyze two Case Studies. In the first one, we evaluate the proper functioning of the eJS scheme when any DRB is a GBR DRB and a non-ideal CAC is adopted. In the second case, we focus on the performance in presence of both GBR and non-GBR services in order to assess the effectiveness of our approach based on a joint control.

Case Study 1: only GBR DRBs

In this Case Study, we assume $N_{UE} = N_{DRB} = N_{GBR}$, $R_i = R \quad \forall i \in \mathbf{U_{GBR}}$, and all GBR DRBs have been accommodated by the CAC.

In Figs. 2.4, 2.5, and 2.6, we assume $R = 2Mbps$. Fig. 2.4 shows $\Delta T_x(\%)$ of all the analyzed scheduling algorithms vs N_{DRB}. Both the proposed strategies show the same optimal performance, in fact the throughput loss compared to the maximum value achievable by means of BestCQI remains below 5%. QGRBA achieves in any case worse performance compared to our strategies, with throughput loss which fluctuates be-

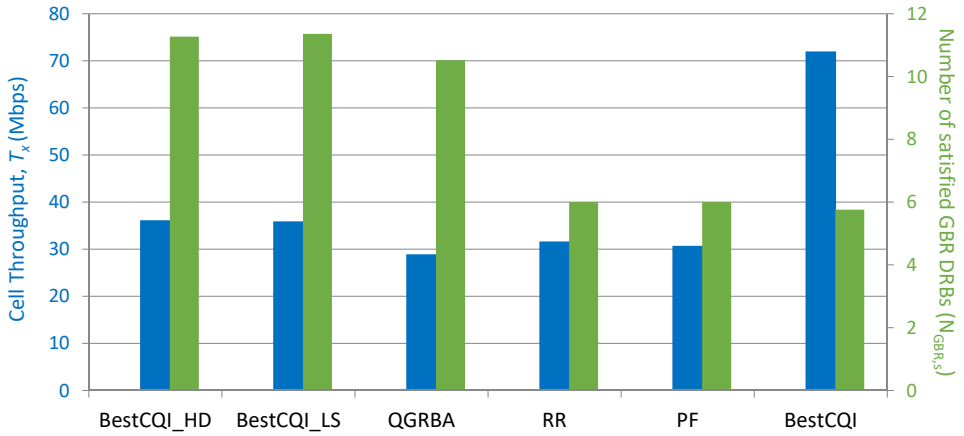


Figure 2.10: *Cell Throughput and number of satisfied GBR DRBs. 12 GBR DRBs, where $R_i = 2Mbps$ for each DRB i. CQI $\in [8, 14], \forall i \in \{1, \ldots, 6\}$, and CQI $\in [2, 7], \forall i \in \{7, \ldots, 12\}$.*

tween 5% and 20%. Moreover, the throughput loss of RR and PF is significantly larger and, already in the presence of a moderate load, remains approximately constant (60%) regardless of the amount of UEs.

Fig. 2.5 shows the probability that at least a GBR service does not reach the minimum guaranteed data rate, under different amount of DRBs. RR, BestCQI and PF exhibit the worst performances, as expected, because these algorithms do not belong to the class of QoS aware schedulers, unlike the other ones. As regards QGRBA, in the case of 20 UEs, the probability of not satisfying all UEs is equal to 18%, while both BestCQI_HD and BestCQI_LS strategies still show a probability of 0%. Finally, we analyze the overload case, that is, when on the basis of the channel conditions, the amount of the available radio resources may not be enough to satisfy 25 GBR DRBs. In this case, for both our scheduling strategies, the probability of not serving at least one GBR service inevitably increases. More specifically, BestCQI_LS reports the lowest probability. This result is expected, since this strategy allocates radio resources by aiming to do not severely penalize the throughput achieved by the single DRB, thus increasing the probability of achieving its minimum required data rate.

Fig. 2.6 depicts the Fairness Index under different number of DRBs. For a fair comparison with the QGRBA scheduler, we consider only $N_{DRB} = \{2, \ldots, 15\}$ for which all GBR services are satisfied. Both the proposed strategies show significantly better results than QGRBA,

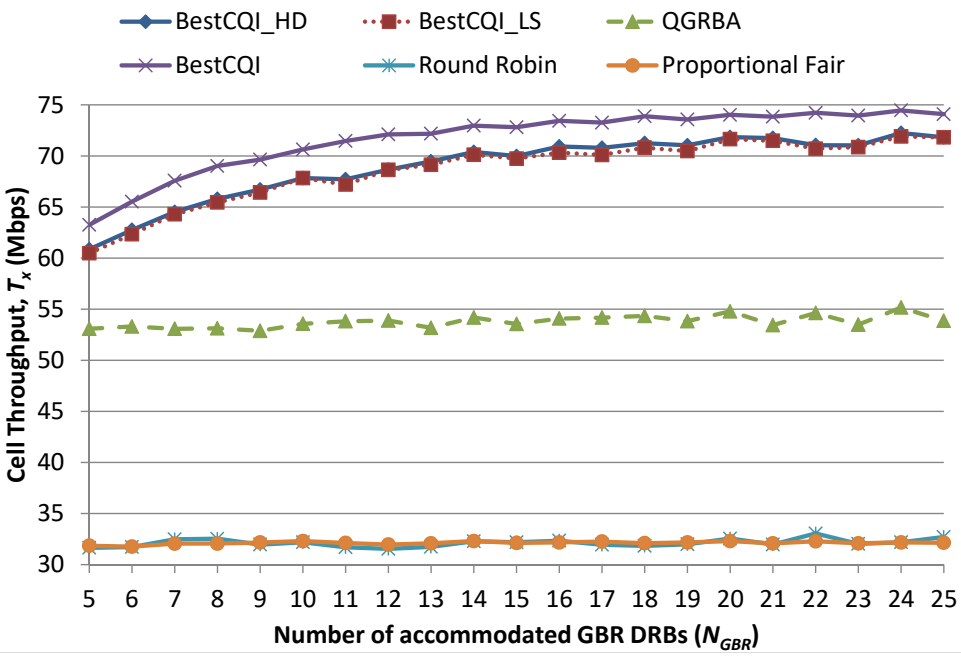


Figure 2.11: *Cell Throughput under different number of DRBs. 2 GBR DRBs with $R = 2Mbps$, 2 GBR DRBs with $R = 4Mbps$ and $(N_{DRB} - 4)$ non-GBR DRBs.*

especially in the case of 15 users. In addition, we note that, due to the adoption of 5G Resource Allocation Type 0, which is based on groups of RBs (RBGs), the level of granularity is low. Therefore, the ideal value of Fairness Index, that is 1, may be unreachable, especially when the number of DRBs is large.

Finally, results in Figs. 2.4 and 2.6 prove that both proposed strategies exhibit the same optimal trade-off between the goal of maximizing the system throughput and the one of achieving the fairness in throughput among DRBs.

Now, we assess the eJS scheme behavior when the adopted CAC accommodates GBR DRBs until the theoretical limit of the cell capacity is reached, i.e., $N_{GBR} \cdot R \simeq 100$Mbps. In Fig. 2.7 we show the number of dropped GBR DRBs when $R_i = R \quad \forall i \in \mathbf{U_{GBR}}$, under different N_{GBR} and R values. In the case of very light traffic load, the proposed strategies and QGRBA scheme show good performances. However, when the traffic load increases, the eJS scheme outperforms the QGRBA. More specifically, eJS scheme serves, on average, a number of GBR DRBs equal to the maximum possible value. We show it through the following two examples. First, when $N_{GBR} = 15$ and $R = 6Mbps$, considering

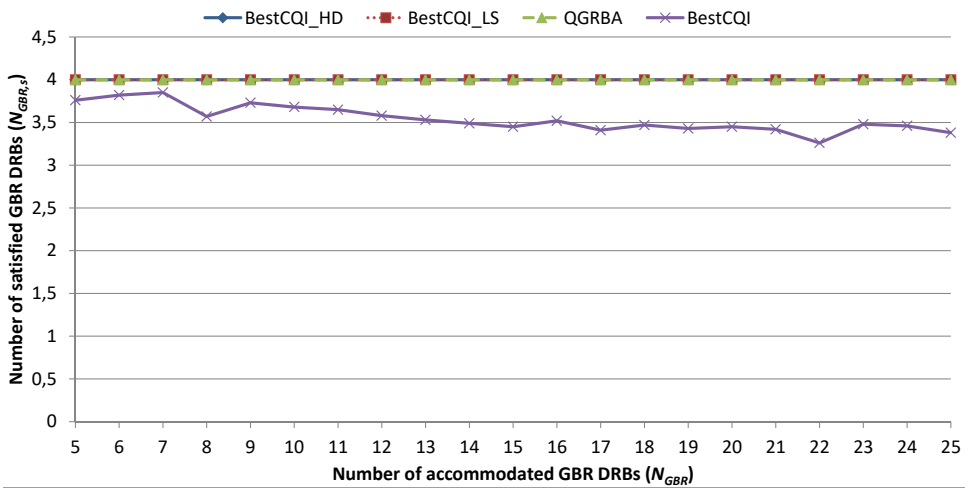


Figure 2.12: *Number of satisfied GBR DRBs under different number of DRBs. 2 GBR DRBs with $R = 2Mbps$, 2 GBR DRBs with $R = 4Mbps$ and ($N_{DRB} - 4$) non-GBR DRBs.*

throughput values in Fig. 2.3, each GBR DRB typically requires at least 3 RBGs, so the maximum amount of GBR DRBs which can be satisfied is 8. In this case, our scheduling strategies drop exactly 7 out of 15 GBR DRBs, while QGRBA drops one more. Second, when $N_{GBR} = 5$ and $R = 10Mbps$, each GBR DRB requires at least 4 RBGs, then 6 GBR DRBs can be satisfied at most. In this case, our scheduling strategies serve all GBR DRBs, while QGRBA serves only 4 out of 5 GBR DRBs.

We also evaluated, for any considered R value, whether the above eJS scheme behavior, close to the ideal one, occurred at the expense of the cell throughput, and here we report the results for $R = 4Mbps$. More specifically, we show in Figs. 2.8 and 2.9 the number of satisfied services ($N_{GBR,s}$) and the cell throughput (T_x), when the number of accommodated GBR DRBs ranges from 1 to 14, respectively. Both our scheduling strategies significantly outperform the QGRBA, because not only do they guarantee, on average, one GBR DRB more, but they also exhibit a cell throughput which remains about 10% higher (see Fig. 2.9). Moreover, the proposed strategies achieve a cell throughput slightly lower than that obtained by the BestCQI, which is the theoretical upper-bound, but with the benefit of almost doubling the number of satisfied GBR DRBs. Figs. 2.8 and 2.9 also allow us to assess the performance differences between the two proposed strategies. As the number of GBR DRBs increases, T_x is slightly higher by adopting the BestCQI_HD com-

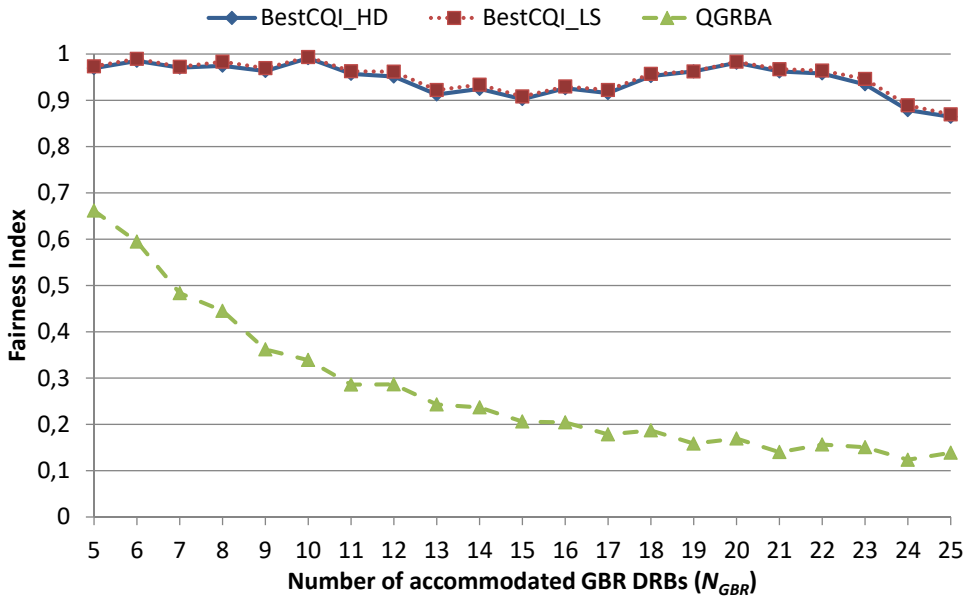


Figure 2.13: *Fairness Index under different number of DRBs. 2 GBR DRBs with* $R = 2Mbps$*, 2 GBR DRBs with* $R = 4Mbps$ *and* $(N_{DRB} - 4)$ *non-GBR DRBs.*

pared to the BestCQI_LS, while $N_{GBR,s}$ is slightly lower, as expected.

Finally, we evaluate the performance of BestCQI Highest Deviation and BestCQI Lowest Second in the realistic case in which some users experience better channel quality (e.g., because they are near the base station), while other ones experience worse channel quality (e.g., because they are at the cell edge). Fig. 2.10 depicts the Cell Throughput and the number of satisfied GBR DRBs obtained by each analyzed scheduling scheme, when $N_{DRB} = 12$, $R_i = 2Mbps \quad \forall i \in \mathbf{U_{GBR}}$, $CQI_{i,k} \in [8, 14] \quad \forall i \in \{1, \ldots, 6\} \wedge \forall k \in \{1, \ldots, N_{RBG}\}$, and $CQI_{i,k} \in [2, 7] \quad \forall i \in \{7, \ldots, 12\} \wedge \forall k \in \{1, \ldots, N_{RBG}\}$. As expected, only the GBR DBRs with better channel conditions are satisfied by adopting RR, PF and BestCQI, because these scheduling schemes are not designed to meet GBR service requirements. In particular, BestCQI achieves the maximum throughput by assigning RGBs to only the DRBs with better CQI values. On the other hand, both the proposed eJS scheme and QGRBA satisfy a larger amount of GBR DRBs, even those that experiment a bad channel quality. For this reason, despite the latter strategies serving more GBR DRBs, the average throughput per DRB is lower than the one achieved by adopting the BestCQI. More specifically, BestCQI_HD and BestCQI_LS satisfy, on average, an additional

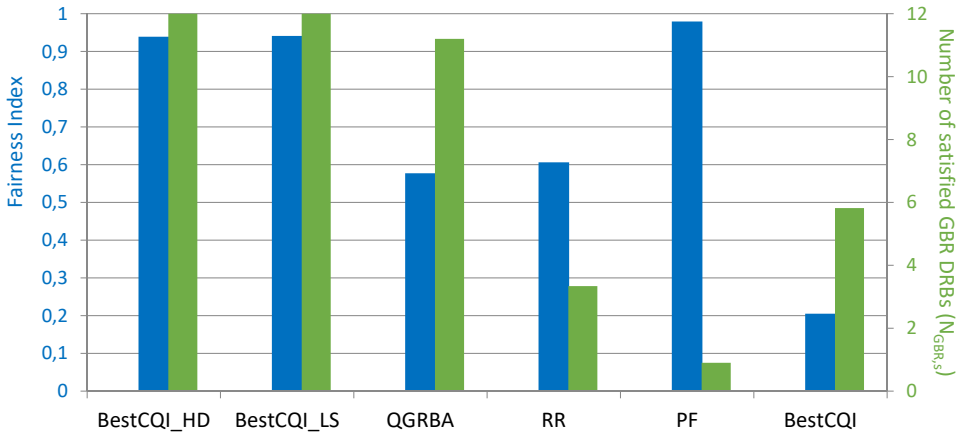


Figure 2.14: *Fairness Index and number of satisfied GBR DRBs. 6 UEs, each one associated with 2 GBR DRBs with $R = 2Mbps$ and 2 non-GBR DRBs.*

GBR DRB compared to the QGRBA scheme, while achieving a system throughput better than QGRBA, RR and PF. These results demonstrate the adaptability of our eJS scheme to different channel conditions.

Case Study 2: GBR and non-GBR DRBs

In this subsection, we assess the quality of our eJS scheme in managing heterogeneous traffic for eMBB scenarios. At this aim, we consider 2 GBR DRBs requiring a minimum guaranteed bit rate equal to $R = 2Mbps$, 2 GBR DRBs requiring $R = 4Mbps$ and the remaining $(N_{DRB} - 4)$ DRBs are non-GBR DRBs. We analyze the performance when N_{DRB} ranges from 5 to 25. Fig. 2.11 depicts the cell throughput achieved by means of each analyzed scheduling scheme under different number of DRBs. We note that RR, and PF achieve very low cell throughput values, while QGRBA exhibits only 54Mbps. On the other hand, BestCQI_HD and BestCQI_LS show performance near to the BestCQI. Next, we evaluate in Fig. 2.12 the number of satisfied GBR DRBs for the proposed scheduling schemes, QGRBA and BestCQI. We note that, even in the presence of only 4 GBR DRBs, the BestCQI scheme fails to satisfy all GBR DRBs, while the QoS aware schedulers satisfy any of them, as expected. For this reason, in Fig. 2.13 we analyze the Fairness Index of the QoS aware schedulers only. It is evident that, as the percentage of non-GBR DRBs increases, QGRBA performance deteriorates significantly. On the other hand, both BestCQI_HD

and BestCQI_LS exhibit a Fairness index that fluctuates between 0.85 and 0.98 due to the low granularity of the radio resources to be assigned. These results show clearly that QGRBA is not designed to schedule radio resources when there are also best effort services. On the other hand, in Figs. 2.11 and 2.13, our eJS scheme exhibits optimal performance both in terms of cell throughput and fairness, demonstrating the effectiveness of the joint control of non-GBR DRBs and GBR DRBs with different minimum required data rate values.

Finally, we analyze the quality of the proposed scheme in a realistic eMBB scenarios, in which more than one GBR or non-GBR DRB is assigned to a UE. At this aim, we consider a scenario composed of 6 UEs, each one associated with 2 GBR DRBs requiring a minimum guaranteed bit rate equal to $R = 2Mbps$, and 2 non-GBR DRBs. Fig. 2.14 shows the number of satisfied GBR DRBs and the Fairness Index for the proposed scheduling schemes and all the reference works. As expected, the results conform to the previously analyzed cases in which each UE is associated with one GBR or non-GBR DRB.

2.7 Conclusion

In this chapter, we addressed the radio resource scheduling problem encountered by a gNB which allocates PRBs in downlink to users requiring GBR and/or non-GBR services, following 5G NR guidelines in a sub-6GHz band. By adopting the standardized radio Resource Allocation Type 0 and the higher layer configured sub-band CSI reporting method, we proposed a new channel and QoS aware scheduling scheme, which jointly supports non-GBR and GBR DRBs in an eMBB scenario, guaranteeing for the latter ones the minimum data rate required. Also, the proposed scheme was extended to work well in realistic scenarios, where non-ideal CACs were adopted. At the aim of being able to choose a different trade-off between the two objectives of maximizing the system throughput and reaching the fairness in throughput among DRBs, in our scheduling scheme, we proposed two different strategies. The first one, called BestCQI_HD, was proposed to improve the system throughput at the expense of a slight loss in terms of fairness. The second strategy, called BestCQI_LS, aimed at improving the fairness among DRBs at the

expense of a slight loss in terms of system throughput. We conducted a simulation study under different traffic types and channel conditions. The results show that the proposed scheduling scheme always outperforms the other benchmark algorithms, by exhibiting a higher degree of fairness and larger amount of satisfied GBR DRBs, at the cost of a slight throughput loss with respect to the maximum achievable throughput.

CHAPTER 3

Radio Resource Management in D2D-enabled mmWave Mobile Broadband (MMB)

In this chapter, we focus on D2D communications, mmWave radio transmissions and adaptive beamforming techniques. Enabling D2D communications can increase significantly the overall system capacity and data rates, while reducing eNodeB data traffic, end-to-end latency and power consumption [9]. The main problem is the management of interferences, not only with traditional cellular communications, but also among D2D communications themselves. Since traditional cellular communications typically take place in licensed band, the first problem can be overcome by adopting Outband D2D communications in appropriate unlicensed spectrum bands. At this regard, we observe that the 2.4 GHz and 5.8 GHz unlicensed bands, currently used for massive Wi-Fi and Bluetooth services, are characterized by an uncontrollable interference. Moreover, the 5.8 GHz spectrum can be also adopted for Licensed Assisted Access (LAA) and the extended version (eLAA) introduced by 3GPP in Release 13 and 14, respectively, which are Carrier Aggregations between the licensed LTE carrier and the unlicensed LTE carriers. In this context, we

take into account the second key technology: transmission in millimeter-Wave band (i.e., the spectrum between 6 GHz and 300 GHz) [38]. The 60 GHz unlicensed band is largely uncongested compared to the 2.4 GHz and 5.8 GHz bands, and has not yet been investigated by 3GPP Release 16 to be supported by 5G New Radio-Unlicensed (NR-U), whose current target is to extend the applicability of 5G NR to sub 7 GHz unlicensed spectrum bands. So, we choose the 60 GHz band because of the ability to utilize a huge unlicensed bandwidth, so to provide multiple Gbps transmission rates [39]. However, because of the high frequency, 60 GHz transmissions are characterized by strong propagation loss. Although it is a drawback for long-range communications, on the other hand very short-range communications benefit from a reduced interference power of farther interfering transmitters. Finally, in order to further reduce interference from close communication links, we can use high-gain adaptive antenna array supporting directional transmission with narrow beam, exploiting adaptive beamforming techniques. Due to the smaller size of antennas at 60 GHz, in the future it will be feasible to enclose ultra-dense array structures in smartphones [40], so we can use adaptive beamforming techniques in both User Equipments (UEs) and Base Stations.

For the above reasons, the symbiosis of D2D communications, 60 GHz band and directional transmissions is regarded as one of the most promising solution to increase significantly the overall system capacity, by enhancing spatial reuse, and to reduce the end-to-end latency and the power consumption. However, despite these benefits, in environments with high density of UEs, most likely one or more D2D communications will be inside the beamwidth range of another direct communication. Therefore, it is necessary to develop proper radio access control schemes and efficient scheduling algorithms to manage the interference among directional D2D communication, in order to achieve high performance in terms of transmission efficiency, throughput, and end-to-end delay. As reported in [41], an analytical solution will take long computation time that is unacceptable for mmWave cells where the duration of one time slot is only a few microseconds. For this reason, a heuristic approach which does not take long computation time is needed.

3.1 Related Work and Motivation

The issues described above have already received attention, and several access control schemes for D2D communications in 60 GHz band are available in literature. 60 GHz standards have been defined for indoor Wireless Personal Area Networks (WPANs), such as IEEE 802.15.3c [42], and for Wireless Local Area Networks (WLANs), such as IEEE 802.11ad [43]. An accurate description and comparative analysis of both protocols can be found in [44]. Since these standards adopt a slotted Time-Division Multiple Access (TDMA), a lot of research works propose a centralized control scheme based on slotted TDMA radio access with concurrent transmission support [10–12, 41, 45–50].

In [10, 45] the frame structure is based on the IEEE 802.15.3c, and each frame has a fixed length equal to 1000 time slots. During a time interval called Contention Access Period, each UE sends the transmission request to the PicoNet Controller (PNC) using the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) technology. After that, during a frame time, the controller runs a scheduling algorithm, and then delivers the relative scheduling information (i.e., time slots allocated) to UEs. In [45] concurrent transmission are allowed on the basis of an Exclusive Region (ER), centered on each receiver device, as function of the estimated interfering transmission power, without considering the effective signal-to-interference-plus-noise ratio (SINR). The ER is also adopted by Rehman *et al.* [10]. In order to improve the overall system throughput, they propose a graph-based scheduling algorithm, which favors short-range flows at the expense of a strong unfairness. The main weakness of these approaches is that each UE, once the request has been sent, has to wait for the next frame to start transmissions.

In [46] the authors adopt a frame structure based on several scheduling periods, each one composed on a beacon period (for network synchronization), a control message exchanging period (in which each transmitter sends its requests and the PNC decides the allocation method for the transmitters), and a data transfer interval. At the aim of maximizing the network throughput per time slot, they present a vertex coloring based resource allocation algorithm. Like [10, 45], they propose to enable concurrent transmission considering the ER and the main lobe interference.

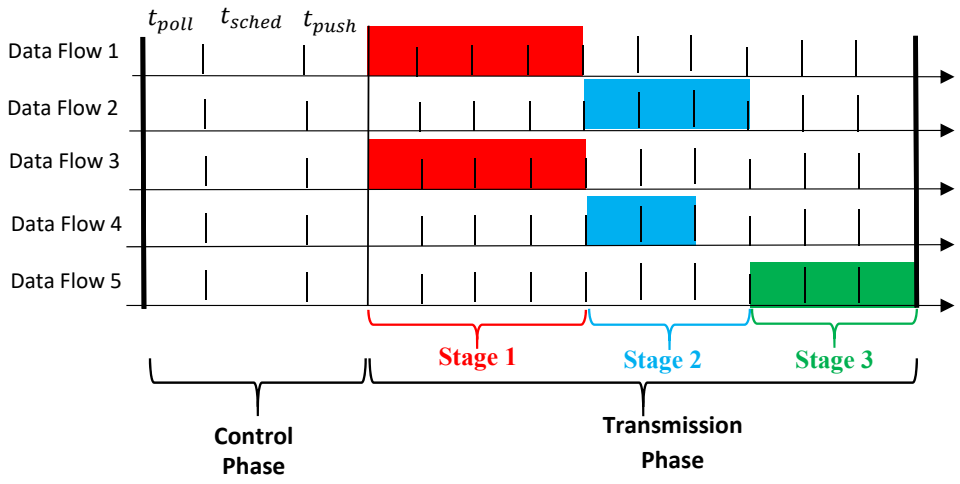


Figure 3.1: *Time-line illustration of frame structure and an example of concurrent data flow transmissions*

In addition, they take into account the interference from the sidelobes, by defining a new power decision threshold in order to reduce this sidelobe interference.

Unlike the approaches analyzed above, the authors of [11, 12, 41, 47, 48] adopt a variable-length frame structure, which depends on the user demands. The frame structure is shown in Fig. 3.1. The central control unit receives transmission requests by means of a polling operation, takes scheduling decisions during the scheduling time with the aim of minimizing the length of transmission phase, and sends transmission rules to UEs during the pushing time of the same frame. During the transmission phase, slots are organized into one or more stages of variable length, in which non-severe interfering data flows are concurrent transmitted. Despite the frame overhead being increased, in this way UE requests and transmissions occur in the same frame, thus reducing the wait before transmitting. Son *et al.* [11] propose an iterative graph-based scheduling algorithm, in which the mmWave network is considered as "pseudo-wired". For this reason, the concurrent transmission condition takes only into account that a device can transmit or receive to\from only one device at the same time. Starting from the algorithm in [11], Niu *et al.* [12] assume that the transmission rates of each wireless link is well-known and fixed at three discrete values (2, 4, and 6 Gbps) on the basis of the distance between UEs. Their scheduling algorithm allows concurrent transmissions only if the SINR of each link is larger than the minimum

SINR to support its transmission rate.

Finally, we focus on some works adopting multi-hop mmWave D2D Communications among UEs. In [47], the authors propose a scheduling algorithm for popular content downloading in a mmWave small cell. They consider a scenario in which a single Access Point (AP) has to deliver the same packages to all UEs belonging to its coverage area. The AP, rather than delivering packets to each UE, sends packets only to nearby UEs; then, they will forward packets to the other UEs by means of multi-hop D2D communications. In the same context, Giatsoglou *et al.* [49] propose to increase popular content exchanges between the paired UEs by using a new policy for device caching. Wei *et al.* [50] investigate the relay transmission through other UEs if an entire light-of-sight (LOS) path is available, when the direct mmWave links are subject to random Bernoulli blockages.

Let us note that all above works consider a scenario limited to the coverage area of a single PNC or an AP, whose coverage radius is typically equal to $15m$. Instead, in [41] Niu *et al.* propose a centralized control scheme for a wider indoor scenario with a high density of UEs. In order to extend the advantages of 60 GHz D2D communications to the whole environment, they assume that several Access Points (APs) are installed inside the coverage area of a Macro eNodeB and interconnected among themselves through a high speed wireless connection at 60 GHz, as shown in Fig. 3.2. This solution implies that the access and the backhaul network use the same resources, therefore resource allocation management is more difficult than a single PNC/AP scenario because it should take into account also the interference between the access and the backhaul network. At the aim of maximizing the system efficiency by having the ability to allocate all available resources without restriction, they propose a central controller which considers both the access and the backhaul network jointly. The authors consider the Central Controller (CC) located inside one AP, so in this way, the proposed architecture acts as a stand-alone Millimeter-Wave Mobile Broadband (MMB) system [51], that is, a 5G system operating exclusively on mmWave bands. The controller also acts as a gateway, because it is connected to the Internet via a direct and high speed wired connection. The remaining APs in the scenario communicate with the gateway to send

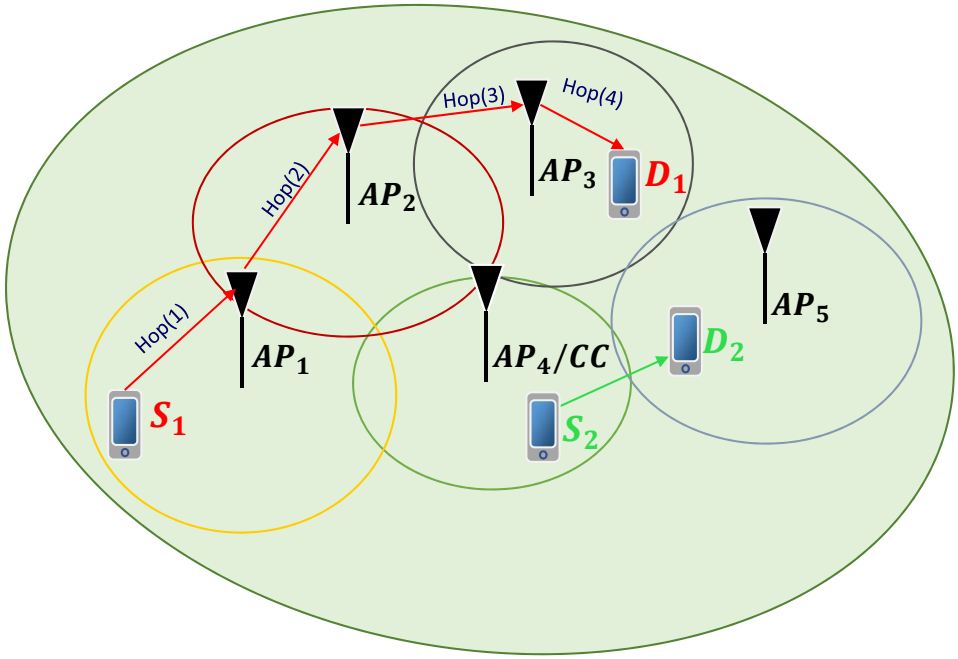


Figure 3.2: *An MMB system for a wide indoor scenario with a high density of UEs. The communication between each pair of UEs inside the whole scenario can take place without traversing the core network via a multi-hop path (e.g., communication between Source S_1 and Destination D_1) or directly (e.g., communication between S_2 and D_2). The Central Controller (CC) is inside AP_4.*

(receive) data to (from) Internet. The communication between each pair of UEs inside this scenario can take place without traversing the core network via a multi-hop path (through one or more APs) or directly (traditional D2D communication), as shown in Fig. 3.2. On the basis of their previous work [12], in [41] Niu *et al.* present an improved access scheme termed D2DMAC, which consists of a path selection criterion to determine if a communication takes place through a multi-hop path or directly, and a concurrent transmission scheduling algorithm, which supports multi-hop communications. The basic strategy adopted in their scheduling algorithm is to deliver all packets in a frame time, whether delivery occurs directly or via a multi-hop path. Gao *et al.* in [48] propose an improved D2DMAC scheme, termed directional D2D medium access control (D3MAC). Compared to the D2DMAC scheme, in [48] the authors introduce a contention graph in order to depict the contention relationship between flows. It is constructed in the way that two vertices are connected with an edge only if the maximum interference between

them is greater than or equal to a fixed threshold value. On the basis of this graph, the heuristic transmission scheduling algorithm of D3MAC, which is the same of [41], reduces the number of iterations.

Although D3MAC scheme has addressed the multi-hop D2D transmission problem in a wide scenario, it shows two significant weaknesses. First, in the presence of one or more multi-hop paths, the packet delivery in a single frame may increase the frame length, so much that the next polling time could occur after several milliseconds. Consequently, this causes delays for new transmission requests, that could be too long for delay-sensitive services. Second, the D3MAC scheme is not very efficient because it does not fully exploit the opportunities of concurrent transmission as we will show later on.

3.2 Contributions

In this chapter, we investigate the access control problem in an indoor D2D-enabled stand-alone MMB system with a high density of UEs for the purpose of interference mitigation. We aim to enhance the transmission efficiency, by exploiting concurrent transmissions, to maximize the throughput, to minimize the end-to-end delay, and to improve the fairness among users, while taking into account to keep the computational load as low as possible. To this end, we propose a new centralized access control scheme composed of a data flow management strategy and a multi-criteria scheduling algorithm based on a greedy graph vertex-coloring technique. Like D3MAC, our access control scheme is based on the variable-length frame structure of Fig. 3.1 and jointly manages D2D communications and transmissions in the access and the backhaul network. In addition to the D3MAC, our scheduling algorithm takes also into account the source rate and the service delay requirements.

Furthermore, on basis of the proposed Centralized Access Control scheme, we focus on cell planning for a MMB system. This is not an easy procedure, because planning a cellular system depends on a large number of aspects, such as signal propagation, expected traffic load and types of service to be guaranteed, transmitter and receiver antenna parameters (power, gains, directivity), interference control, transmission scheduling, and so on. Since a mmWave network has very different characteristics

compared to previous generation networks, we need to introduce an analysis specific for the peculiarities of this network technology. Our aim is to provide some guidelines on how to approach cell planning problem in a MMB scenario.

The main contributions can be summarized as follows:

- As regards multi-hop transmissions, unlike D3MAC, our resource management strategy is based on performing only one hop per frame, with the aim of maximizing the number of possible concurrent transmissions, thus improving the transmission efficiency and the delay-sensitive service performance.

- For each transmission request, the Central Controller chooses the proper path type (i.e., direct or multi-hop path) through a new path selection criterion, which takes into account also the source rate and the performance differences between the direct path and the multi-hop path, by means of new distance-based metrics.

- We define a new multi-criteria scheduling algorithm, in order to reach the fairness, enhance transmission efficiency, maximize the system throughput, and reduce the end-to-end delay. These criteria are grouped in three phases. First, analysis of the interference relations among data transmissions, and creation of an interference graph. Second, assignment of different transmission stages to intolerant interfering sub-flows, by using a new graph multi-coloring method. Third, a new criterion for minimizing the number of time slots to be allocated to each stage.

- To further enhance transmission efficiency, the controller splits, wherever necessary, a data flow transmission in more transmission stages inside the same frame without increasing the number of stages; this minimizes the frame length for the same amount of data traffic transmitted.

- By simulation, we benchmark our approach against the D3MAC scheme [48]. The comparative analysis shows that our control scheme outperforms the reference scheme in terms of throughput, end-to-end delay and fairness in any simulated traffic condition.

- We provide a radio planning procedure organized in two phases: coverage planning and capacity planning. As for the first one, we derive coverage planning constraints, in terms of maximum supported distance between transmitter and receiver (UE and/or AP) in non-line-of-sight (NLOS) and line-of-sight (LOS) condition, in order to provide full coverage of the service area and to ensure the functioning of the radio access scheme. As for capacity planning, because system capacity depends heavily on the network topology, we first derive a relation between the network parameters to be configured and the maximum number of users to be managed by the system, in a generic network topology fulfilling coverage constraints. Then, we analyze a Case Study at the aim of configuring appropriate network parameters to achieve good system performance.

3.3 System Overview

We consider a wide indoor D2D-enabled MMB system, operating at 60 GHz, inside the coverage area of a M-eNB with a high density of UEs. The MMB system consists of N_{AP} APs and N_U UEs operating on the same spectrum. We assume that UEs are equipped with multi-standard technology transceivers (e.g., mmWave and LTE Advanced). So, a UE, while is connected at a low frequency with the macro eNodeB for traditional services, can communicate at 60 GHz with a device (UE or AP) inside the same MMB system through a direct communication or a multi-hop path across APs without traversing the Core Network[1].

On the basis of the described scenario and the adopted architectural solution, in addition to the interference between D2D communications, the interference between mmWave access and backhaul network arises. So, in order to efficiently manage these interferences, as in [48], we choose a centralized control approach, considering both radio access and backhaul networks jointly. More specifically, we assume a local Central Controller (CC), inside an AP, which is the one which reduces the average one-way latency between each pair of APs. Other assumptions are listed below:

[1]The proposed control scheme could also work with traditional communications to the Internet, by considering an AP gateway as the destination device.

1. APs are in LOS between them and their position is fixed.

2. The Central Controller knows the static topology of APs and the up-to-date location of UEs. The latter information can be obtained by a maximum-likelihood classifier based on beam-space channel matrix [52].

3. Each UE and AP is equipped with ideal directional adaptive antennas [45] with the same beam angle: the antenna gain is constant within the beam angle θ ($G = \frac{2\pi}{\theta}$) and zero outside; the transmitter antenna is able to automatically adapt its radiation pattern, by putting the main lobe in the direction of the receiver antenna.

4. Each UE can be served by only one Access Point and is associated with the nearest AP.

3.3.1 Basics of the Radio access scheme

As in [11, 12, 48], we adopt a slotted TDMA with concurrent transmission support. A fundamental constraint of TDMA approach is the high synchronization accuracy needed in the system. The clock of all APs is synchronized by the controller; then, each AP synchronizes the clock of UEs associated with it. However, to ensure time synchronization taking into account propagation delays, a radio network planning is needed. We provide it at the end of this chapter.

In the considered scenario, where the beamforming technology is a key point, the management of UE request packets is not based on the CSMA/CA standard, but on the polling technique, because the conventional carrier sensing medium contention schemes do not work well with directional antennas, as outlined in [53].

Time is divided into time slots of equal length and organized in non-overlapping frames with variable length [11]. As shown in Fig. 3.1, a frame consists of two phases: control phase and transmission phase. The first phase contains three time intervals: polling time, scheduling time and pushing time.

During the polling time (t_{poll}), each AP sequentially polls all UEs associated with it to check whether any UE has data to transmit. Each UE must respond after a fixed interval (Short Inter Frame Space, SIFS)

with a request packet or with a 'keep alive' response message if it does not have any data to transmit. Once this procedure is completed, each AP will report the information received from UEs to the Central Controller through the wireless backhaul network.

During the scheduling time (t_{sched}), the Central Controller applies an efficient scheduling algorithm based on received requests and location information, to meet the traffic demands.

Next, during the pushing time (t_{push}), the CC distributes scheduling rules to all APs through the wireless backhaul network. Afterwards, each AP forwards the scheduling information to UEs associated with it. Upon receipt of the information from its AP, each UE sends an acknowledgment message. The time required for the pushing operation is the same as the polling.

Finally, during the transmission phase, slots are organized into one or more stages of variable length. During a stage, non-severe interfering flows are concurrently transmitted. Let us note that each wireless device can not transmit and receive at the same time, as described in Section 3.5.

We set slot time $t_{SLOT} = 5\mu s$ and its payload time $t_{PAY} = 4\mu s$, as in [48]. In addition, in [19] we derived the maximum distance for mmWave communication in non-line-of-sight (NLOS) conditions, that is, $d_{\text{max,NLOS}} = 16.97m$. This choice guarantees all overhead parameters and a proper guard interval, to prevent the use of compensation techniques for time slot synchronization (e.g., timing advance).

3.3.2 60 GHz channel model

The main characteristics of 60 GHz communications are high frequency and large bandwidth, so high path loss and transmission rate values. We adopt the path loss model for frequency spectrum above 6 GHz in Indoor-office scenarios provided by the 3GPP standard in [54]. The path loss in dB at distance d can be expressed as follows:

$$PL_{[dB]}(d) = PL_{[dB]}(d_0) + 10\alpha \log_{10}\left(\frac{d}{d_0}\right) + X_{\sigma_{[dB]}}, \tag{3.1}$$

where $PL_{[dB]}(d_0) = -10\log_{10}\left[\left(\frac{\lambda}{4\pi d_0}\right)^2 G_T G_R\right]$, λ is the wavelength, G_T and G_R are transmitter and receiver antenna gain, α is the path loss exponent, d_0 is the reference distance, d is the distance between transmitter and receiver, $X_{\sigma_{[dB]}}$ represents the large-scale fading, modeled as a log-normal distribution which in the dB-domain corresponds to a zero-mean Gaussian distribution with standard deviation $\sigma_{[dB]}$. The values of α and $\sigma_{[dB]}$ depend on the visibility condition, i.e., Line-Of-Sight (LOS) or Non-Line-Of-Sight (NLOS), and are reported in Table 3.1. The LOS state is considered when no buildings or walls block the direct path between TX and RX. The impact of objects, such as chairs, desks, office furniture, and so on, is modeled using the shadowing term $X_{\sigma_{[dB]}}$ [55].

In order to estimate the maximum achievable transmission rate of data flow i from transmitter TX_i to receiver RX_i, we adopt Shannon's theoretical formula[2], as in [10]:

$$R_i = \eta W \log_2\left(1 + \text{SINR}_i\right), \tag{3.2}$$

where η represents the transceiver efficiency ($\eta \in [0, 1]$), W is the system bandwidth, and SINR_i is the signal to interference plus noise ratio, calculated as:

$$\text{SINR}_i = \frac{P_{T_i}/PL(d(TX_i, RX_i))}{N_0 W + \sum_{j \neq i} P_{T_j}/PL(d(TX_j, RX_i))}, \tag{3.3}$$

where P_{T_i} is the signal power of TX_i, PL is the path loss in (3.1) converted from the dB-domain to the linear domain, $d(TX_i, RX_i)$ is the distance between TX_i and RX_i, and N_0 is the power spectral density of the white Gaussian noise.

Let us note that $\sum_{j \neq i} P_{T_j}/PL(d(TX_j, RX_i))$ is the power from all interfering data flows.

3.4 Our Access Control Strategy

For the mmWave network architecture and the basic radio access scheme introduced above, we propose a new resource management strategy aiming to fulfill the following objectives:

[2] A future implementation of a proper modulation and code scheme will set accurate transmission rate values. However, the framework for our analysis would remain identical.

- enhance transmission efficiency, taking fully advantage of concurrent transmissions;
- maximize throughput;
- minimize end-to-end delay;
- keep computational complexity as low as possible.

These targets are often in opposition to each other. On the one hand, we try to offer best efficiency, enabling as many as possible concurrent transmissions. On the other hand, we try to maximize transmission rates, by minimizing the interference (i.e., by reducing the amount of concurrent transmissions). In addition, we need to establish the best order of transmission to minimize the end-to-end delay.

In this chapter, several aspects are considered and several criteria are developed in order to propose a new Access Control scheme which aims to best accomplish each target. In the next subsection, before defining the Central Controller operations, the approach taken for the management of multi-hop flows is described and the motivations are explained.

3.4.1 Multi-hop transmission management

In Fig. 3.3 we show a data session between source UE, $S(n)$, and destination UE, $D(n)$, termed "active communication(n)"; it starts at time t_0 and ends at time t_1. Also, we denote with $t_{start}(i)$ the beginning time of frame i, and with "data flow(n, i)", $df(n, i)$, the amount of traffic data generated by source $S(n)$ during frame $(i - 1)$, i.e., from $t_{start}(i - 1)$ to $t_{start}(i)$. During the polling time of the ith frame, $S(n)$ will send to the related AP a request packet, containing, inter alia, $D(n)$ address and the amount of traffic data to be transmitted, $df(n, i)$. On the basis of the decisions made by the controller, $df(n, i)$ can be transmitted directly (direct path) or via Access Points (multi-hop path). In the latter case, $df(n, i)$ needs to be transmitted sequentially, hop-by-hop, as depicted in Fig. 3.4. We define "sub-flow(n, i, j)", $sf(n, i, j)$, as the transmission of $df(n, i)$ related to hop (j). The transmission of a direct data flow is a single sub-flow($n, i, 1$).

As regards multi-hop transmissions, two different strategies can be considered:

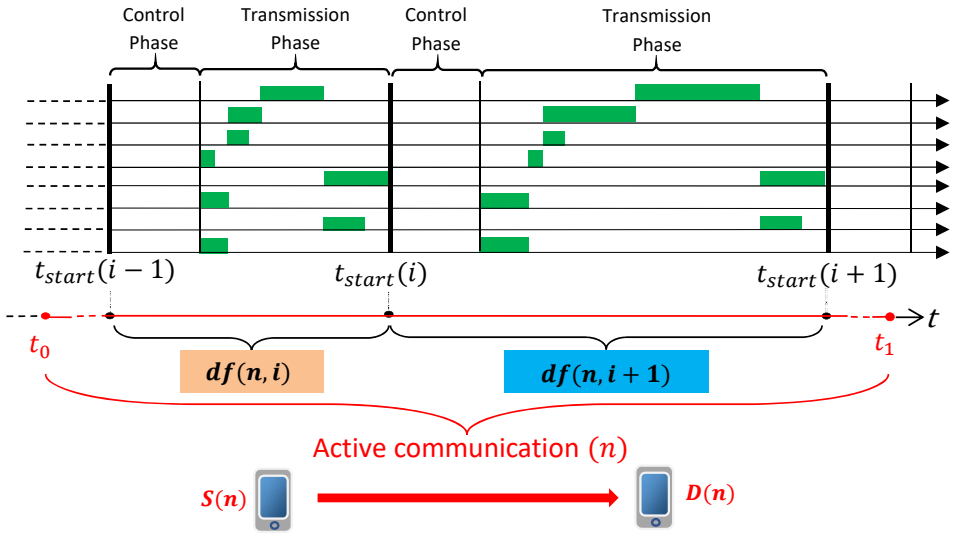


Figure 3.3: *Graphic representation of active communication(n) and the related data flows.*

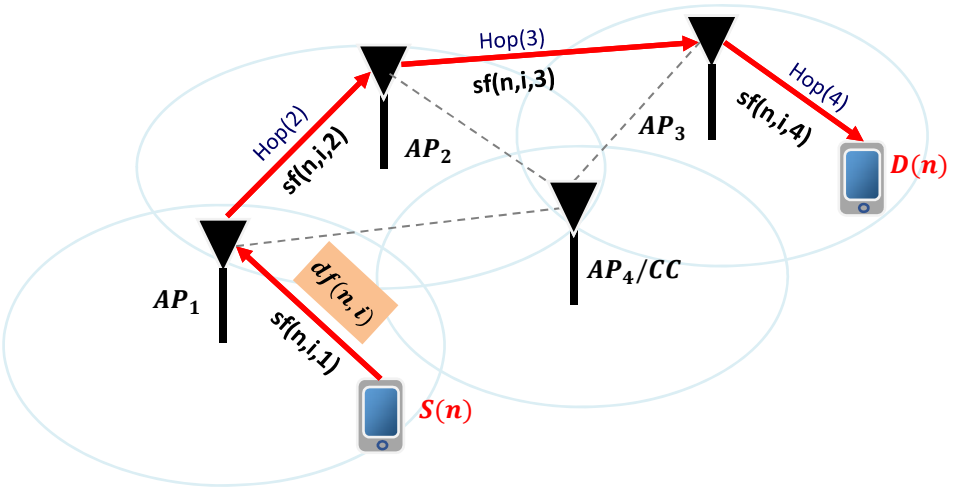
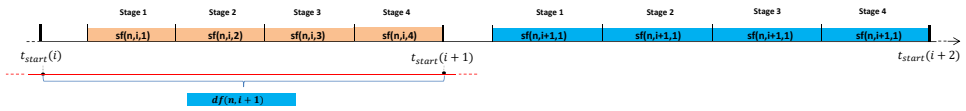


Figure 3.4: *Graphic representation of data flow(n,i) and the related sub-flows.*

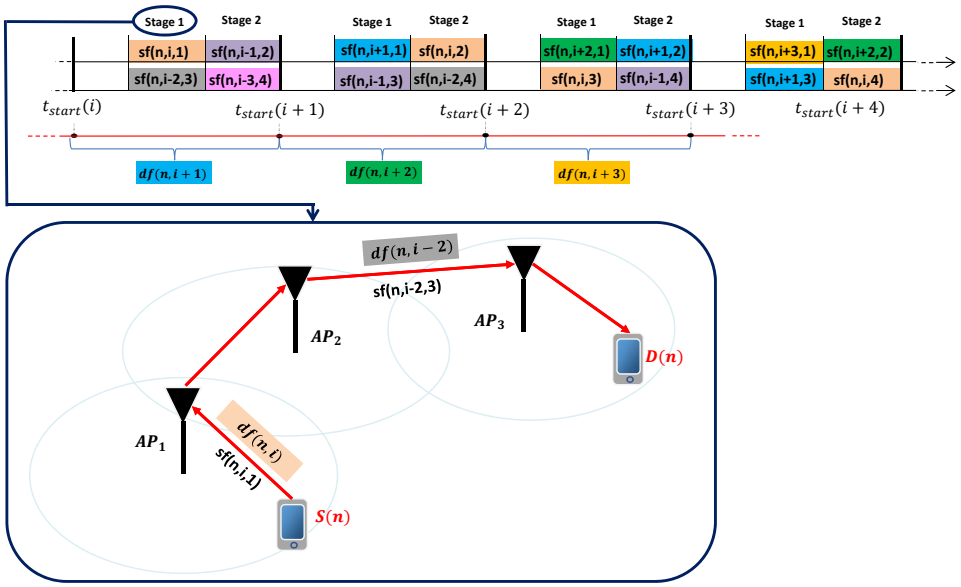
1. complete data flow transmission from source to destination UE, during a single frame;

2. perform only one hop per frame.

D3MAC scheme adopts strategy 1, i.e., each AP stores the received $df(n,i)$ and forwards it in the next stage of the same frame (see Fig. 3.5a). Instead, with strategy 2, each AP stores the received $df(n,i)$ and forwards it in the next frame. The choice of strategy is relevant because significantly affects the system performance, as we show below.



(a) *Strategy 1*



(b) *Strategy 2*

Figure 3.5: *Active communication consisting of multi 4-hop data flows. Transmission in steady state with two strategies.*

Firstly, we consider a simple example of an active communication consisting of a single data flow, which is delivered through 4 hops (sub-flows). With strategy 1, the total overhead time is related only to the control phase of a single frame. Instead, with the second strategy, because only one sub-flow can be delivered during a frame, the total overhead time is related to four control phases. Therefore, if we restrict our analysis to just one multi-hop data flow, the best strategy seems to be the first one.

However, it is very likely that during the transmission phase, the source UE of the considered active communication continues to generate packets. Because sub-flows belonging to the same data flow are sequential transmissions, as shown in Fig. 3.5a, the frame length can be long. The longer the frame length, the greater the amount of traffic generated by source UE, i.e., the data flow size. The process of lengthening continues until the length of the frame reaches the steady state with a long duration. Conversely, with strategy 2, because in each frame all

sub-flows belong to different data flows of the same active communication, some of these sub-flows can be scheduled concurrently, as shown in Fig. 3.5b. The frame length will be shorter compared to strategy 1 and, consequently, the traffic amount generated during a frame is reduced. As a result, an active communication is expected to be split into more data flows with a small number of packets that can exploit concurrent transmissions.

Consequently, with strategy 2, the time necessary to transmit the same amount of data is less than with the first one (i.e., the efficiency increases and the packet delay is reduced). In addition, we note that the control phase frequency is greater, so other new source UEs can benefit from a reduced wait for the next polling time (i.e., a reduced accumulated initial delay and an improved fairness among UEs). For these reasons, unlike D3MAC [48], in our control access scheme we adopt strategy 2.

3.4.2 Central Controller Operations

A high-level block diagram of the access control strategy is shown in Fig. 3.6.

During the polling time, each AP obtains request packets from its UEs with new data flows to be transmitted, characterized by: destination UE; amount of data in terms of number of packets ($n_{packets}$), assuming a fixed packet size of $N_{\text{bytes,packet}}$ bytes; source rate (R_{source}), e.g., 2 or 4 Gbps; end-to-end delay threshold (d_{thr}), if any. Then, each AP will report to the AP/CC the requests for transmission of new data flows. In addition, each AP, which has received and stored a data flow in the previous frame, will send to the Central Controller a transmission request to forward it to the next hop. As regards active communications with delay-sensitive services, we assume that each AP discards all related packets which have accumulated a delay larger than the delay threshold, before sending its requests to the Central Controller.

Therefore, at the end of each polling time, the Central Controller knows all UEs and APs requests. For all new data flows, the controller selects the path type: it establishes whether data transmission occurs by a direct path (direct sub-flow) or a multi-hop path (more sub-flows). The multi-hop path is determined by a static routing. If the controller chooses the multi-hop path, then it selects only the first sub-flow for the current

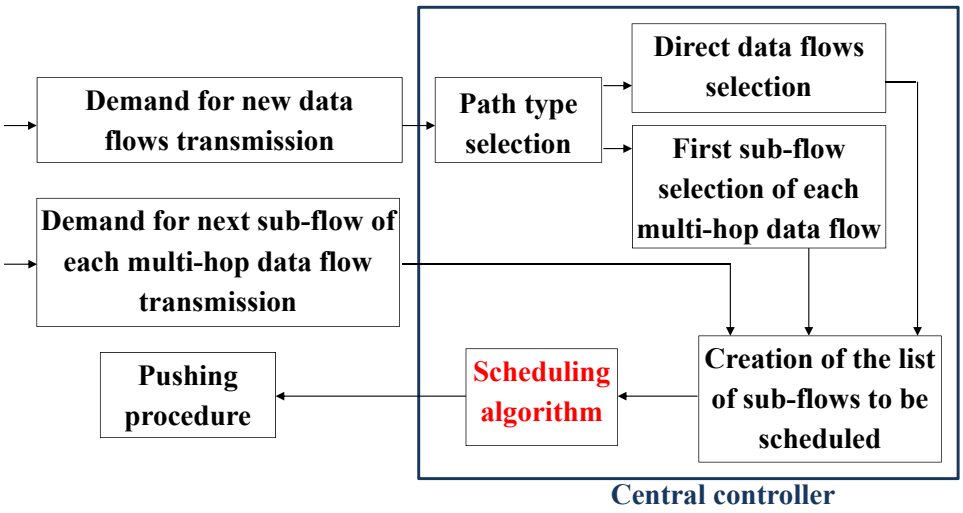


Figure 3.6: *High-level block diagram of the access control strategy.*

scheduling procedure. The path selection criterion is described in the next subsection.

Finally, the list of sub-flows to be scheduled is created. It is the input of the scheduling algorithm described in Section 3.6. This list contains all direct sub-flows (single hop), the first sub-flow related to each new multi-hop data flow, and the next sub-flow related to each multi-hop data flow stored in APs.

3.4.3 Path type selection

The objective of the path selection criterion is to choose the proper path type (i.e., direct or multi-hop path) for new data flows. Here, we propose an improved criterion in comparison with [18]. In addition to the constraint on the maximum communication distance in NLOS conditions ($d_{\text{max,NLOS}}$), which ensures the time slot synchronization, the new criterion takes also into account R_{source} and the performance differences between the direct path and the multi-hop path in terms of achieved transmission rate. In order to meet the requirement of low computational complexity for this path type selection algorithm, we introduce new parameters based only on the distance metric.

For each new data flow(n, i), by using (3.2), we derive the minimum SINR value supporting the source rate $R_{\text{source},n}$, as follows:

$$\text{SINR}_{\text{min},n} = 2^{\frac{R_{\text{source},n}}{\eta W}} - 1. \tag{3.4}$$

So, we define $d_{\text{max,source},n}$ as the maximum distance between $S(n)$ and $D(n)$ supporting $R_{\text{source},n}$ in absence of interference. By using (3.3) in which $\sum_{j \neq i} P_{T_j}/PL(d(TX_j, RX_i)) = 0$ and $\text{SINR}_i = \text{SINR}_{\text{min},n}$, we obtain $d_{\text{max,source},n}$ as follows:

$$d_{\text{max,source},n} = d_0 \left(\frac{P_{T_n}/PL(d_0)}{\text{SINR}_{\text{min},n} N_0 W} \right)^{\frac{1}{\alpha}}, \tag{3.5}$$

where P_{T_n} is the transmission power of $S(n)$.

Also, we aim to compare the expected performance of the direct path against the one of the multi-hop path. We observe that, due to the direct visibility among APs, the performance of a multi-hop data flow is main limited by the first and the last hop, which are generally in NLOS. Between these two, it is likely that the one at greater distance has less favorable characteristics. For this reason, we define a multi-hop distance value as:

$$d_{\text{mh},n,i} = \max \left[d\left(S(n), AP_{S(n)}\right), d\left(AP_{D(n)}, D(n)\right) \right], \tag{3.6}$$

where $AP_{S(n)}$ and $AP_{D(n)}$ represent the Access Points to which $S(n)$ and $D(n)$ are attached, respectively. Finally, in order to consider the performance of the whole multi-hop path, we introduce a weight equal to the number of hops ($N_{\text{hop},n,i}$).

On the basis of $d_{\text{max,NLOS}}$ and the new parameters (3.5) and (3.6), we define a new path type selection algorithm, shown in pseudo-code 3. Therein, the distance of the direct path, $d\left(S(n), D(n)\right)$, is compared with $d_{\text{max,NLOS}}$ to ensure the time slot synchronization, and with (3.5) and (3.6) to guarantee $R_{\text{source},n}$, as far as possible.

3.5 Transmission Scheduling Problem Formulation

In this Section, we present the problem formulation of the optimal scheduling decisions. The scheduling function has a key role in the Central Controller. As depicted in Fig. 3.6, a list of sub-flows to be scheduled in the current frame is the input of the scheduling algorithm. For each sub-flow i related to active communication(n), the Central Controller knows the related hop (i.e., transmitting device TX_i and receiving device RX_i), the amount of packets to be transmitted ($n_{\text{packets},i}$), and the source rate $R_{\text{source},n}$.

Pseudo-code 3 Path Type Selection

Definitions:

- $\mathbf{F_N}$: set of new requests for transmission (from UEs)
- $d_{\text{max,NLOS}}$: maximum NLOS distance
- $d\left(S(n), D(n)\right)$: distance from $S(n)$ to $D(n)$ of active communication(n)
- $d_{\text{max,source},n}$: maximum distance supporting the source rate
- $d_{\text{mh},n,i}$: multi-hop distance
- $N_{\text{hop},n,i}$: number of hops of multi-hop path

Iteration:

for each element $\in \mathbf{F_N}$ **do**
  if $d\left(S(n), D(n)\right) > d_{\text{max,NLOS}}$ **then**
    choose multi-hop path {time slot synchronization is not guaranteed with the direct path}
  else
    if $d\left(S(n), D(n)\right) > d_{\text{max,source},n}$ **then**
      if $d\left(S(n), D(n)\right) < d_{\text{mh},n,i}$ **then**
        choose direct path {source rate is not guaranteed with the direct path, but the multi-hop path has worse performance}
      else
        choose multi-hop path
      end if
    else
      if $d\left(S(n), D(n)\right) > N_{\text{hop},n,i} d_{\text{mh},n,i}$ **then**
        choose multi-hop path {despite source rate being guaranteed with the direct path, the weighted multi-hop path still performs better}
      else
        choose direct path
      end if
    end if
  end if
end for

We consider N_{SF} sub-flows to be scheduled in the transmission phase of a single frame. As reported in Section 3.3.1, the transmission phase is divided into N_{stage} non-overlapping stages, and the kth stage lasts N_{ts_k} time slots. In each stage, multiple sub-flows could be scheduled to transmit concurrently. Given the traffic demand, in order to maximize the transmission efficiency, the optimal scheduling decisions should accommodate the traffic demand ($n_{\text{packets},i}$) of each sub-flow i with the minimum number of time slots. This objective can be formulated as follows:

$$\min \sum_{k=1}^{N_{\text{stage}}} N_{\text{ts}_k}. \tag{3.7}$$

Let us underline that N_{stage} and $N_{\text{ts},k}$ are both unknowns of the problem. $N_{\text{stage}} \in \{1, \ldots, N_{SF}\}$, i.e., $N_{\text{stage}} = 1$ when all sub-flows are transmitted concurrently, and $N_{\text{stage}} = N_{SF}$ when no sub-flow is transmitted concurrently. $N_{\text{ts}_k} \in \{1, \ldots, \max_{\forall i} [n_{\text{packets},i}]\}$, because in one time slot each sub-flow can transmit one or more packets, based on its transmission rate.

Now, we analyze the system constraints. Let s_i^k be a Boolean variable which indicates whether sub-flow i is scheduled in stage k. If it is, $s_i^k = 1$; otherwise $s_i^k = 0$.

Because transmitting and receiving antennas of a wireless device (UE or AP) operate at the same carrier frequency, each wireless device can only transmit or receive at the same time from at most one device. This constraint can be formulated as follows.

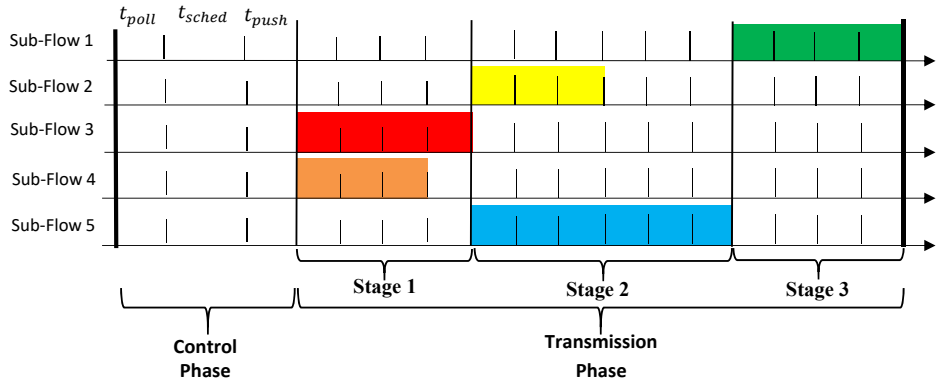
$$s_i^k + s_j^k \leq 1, \forall k, \forall i \neq j \text{ if } RX_i = TX_j; \tag{3.8}$$

$$s_i^k + s_j^k \leq 1, \forall k, \forall i \neq j \text{ if } RX_i = RX_j; \tag{3.9}$$

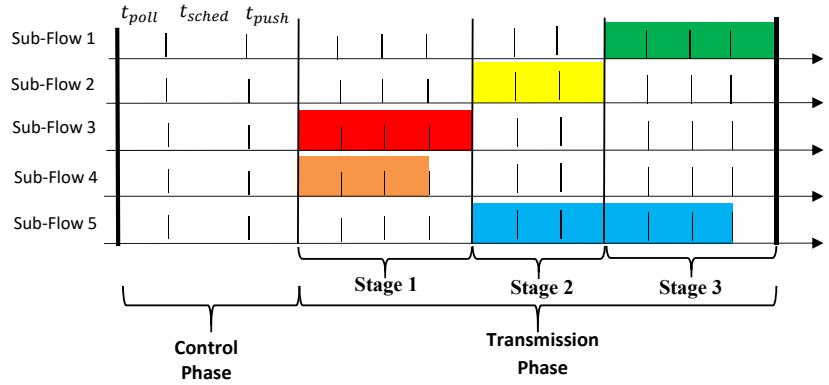
$$s_i^k + s_j^k \leq 1, \forall k, \forall i \neq j \text{ if } TX_i = TX_j. \tag{3.10}$$

To exploit concurrent transmissions, the signal to interference plus noise ratio experienced by each flow i in stage k (SINR_i^k) should be at least equal to $\text{SINR}_{\text{min},i}$, which is the minimum SINR value supporting the source rate $R_{\text{source},n}$ of the related active communication(n). This constraint is formulated as follows:

$$\text{SINR}_i^k \geq \text{SINR}_{\text{min},i}, \quad \forall i, k \text{ so that } s_i^k = 1, \tag{3.11}$$



(a) *Only one transmission stage per sub-flow*



(b) *More transmission stages per sub-flow*

Figure 3.7: *Benefit to transmit sub-flows into more stages.*

where

$$\text{SINR}_i^k = \frac{s_i^k P_{T_i} / PL(d(TX_i, RX_i))}{N_0 W + \sum_{j \neq i} s_j^k b_i^j P_{T_j} / PL(d(TX_j, RX_i))}, \tag{3.12}$$

in which $b_i^j = 1$ if RX_i is inside the beamwidth range of TX_j; otherwise, $b_i^j = 0$.

The traffic demand ($n_{\text{packets},i}$) of each sub-flow i must be fully accommodated in the current transmission phase. It means that:

$$\sum_{k=1}^{N_{\text{stage}}} \left\lfloor \frac{s_i^k N_{\text{ts}_k} t_{PAY} R_i^k}{8 N_{\text{bytes,packet}}} \right\rfloor \geq n_{\text{packets},i}, \forall i, \tag{3.13}$$

where:

$$R_i^k = \eta W \log_2 \left(1 + \text{SINR}_i^k\right), \forall i, k. \tag{3.14}$$

As regards constraint (3.13), we need to make the following considerations. Let us note that in [48], the authors assume that each sub-flow

is activated just once in the transmission phase, that is,

$$\sum_{k=1}^{N_{\text{stage}}} s_i^k = 1, \forall i. \tag{3.15}$$

By adopting this assumption, as depicted in Fig. 3.7a, each stage k ends only when each sub-flow i, so that $s_i^k = 1$, has transmitted all of its packets, e.g., in the figure N_{ts_2} ends when sub-flow 5 is completed. This means that, to satisfy constraint (3.13), for each stage k the N_{ts_k} value is established by the sub-flow requesting more time slots in it (i.e., the sub-flow with a large amount of data and/or a low transmission rate).

Unlike [48], in this chapter, to better achieve objective (3.7), we propose to split, wherever necessary and possible, a sub-flow into more transmission stages. For instance, if sub-flows 1 and 5 can be transmitted concurrently in stage 3 (i.e., if $\text{SINR}_1^3 \geq \text{SINR}_{\text{min},1}$ and $\text{SINR}_5^3 \geq \text{SINR}_{\text{min},5}$), then N_{ts_2} can be properly reduced by enhancing concurrent transmissions, as shown in Fig. 3.7b. This assumption can be formulated as follows.

$$\sum_{k=1}^{N_{\text{stage}}} s_i^k \leq N_{\text{stage}}, \forall i. \tag{3.16}$$

It is clear that assumption (3.16) favors a more efficient transmission than assumption (3.15), at the expense of a greater computational complexity. In fact, considering assumption (3.15), constraint (3.13) corresponds to N_{SF} inequalities each one in a single unknown, while considering assumption (3.16), constraint (3.13) can be written as a system of

N_{SF} inequalities each one in N_{stage} unknowns:

$$\begin{cases} \left\lfloor \frac{s_1^1 N_{\text{ts}_1} t_{PAY} R_1^1 + \cdots + s_1^{N_{\text{stage}}} N_{\text{ts}_{N_{\text{stage}}}} t_{PAY} R_1^{N_{\text{stage}}}}{8 N_{\text{bytes,packet}}} \right\rfloor \geq n_{\text{packets},1} \\ \left\lfloor \frac{s_2^1 N_{\text{ts}_1} t_{PAY} R_2^1 + \cdots + s_2^{N_{\text{stage}}} N_{\text{ts}_{N_{\text{stage}}}} t_{PAY} R_2^{N_{\text{stage}}}}{8 N_{\text{bytes,packet}}} \right\rfloor \geq n_{\text{packets},2} \\ \cdots \\ \left\lfloor \frac{s_{N_{SF}}^1 N_{\text{ts}_1} t_{PAY} R_{N_{SF}}^1 + \cdots + s_{N_{SF}}^{N_{\text{stage}}} N_{\text{ts}_{N_{\text{stage}}}} t_{PAY} R_{N_{SF}}^{N_{\text{stage}}}}{8 N_{\text{bytes,packet}}} \right\rfloor \\ \qquad \geq n_{\text{packets},N_{SF}}. \end{cases} \tag{3.17}$$

In addition, this system in general has not a unique solution, because $N_{\text{stage}} \leq N_{SF}$. Therefore, it is evident that the degrees of freedom of our problem formulation are greatly increased compared to [48] as well as the computational complexity of the problem.

In summary, the objective of the optimal scheduling is (3.7) s.t. constraints (3.8)-(3.11), and (3.13). It is easy to observe that the formulated problem is a Mixed Integer Nonlinear Program (MINLP), which is generally NP-hard. Using the branch-and-bound algorithm, it will take significantly long computation time [11], and it is unacceptable for practical mmWave cells where the duration of one time slot is only a few microseconds. For this reason, in the next Section we propose a multi-criteria heuristic scheduling algorithm to obtain near-optimal solutions which will not take significantly long computation time.

3.6 Our proposed Scheduling Algorithm

In this section, we propose a greedy algorithm that breaks up the global scheduling problem into a series of non-overlapping sub-problems, to make the locally optimal choice at each phase, with the intent of finding a global optimum. Our multi-criteria scheduling algorithm consists of three phases, as shown in Fig. 3.8, and in Lines 11-19 of pseudo-code 4, which summaries the overall access control procedure. For each phase, the details are described in the following subsection.

Pseudo-code 4 Overall access control procedure

Definitions:

- $\boldsymbol{F_N}$: set of new requests for transmission (from UEs)
- $\boldsymbol{F_O}$: set of requests for next sub-flow of multi-hop data flow transmission (from APs)

Initialization:

1: $\boldsymbol{F_D} = \emptyset$, $\boldsymbol{F_{MH}} = \emptyset$, $\boldsymbol{L} = \emptyset$
2: **for** each data flow $\in \boldsymbol{F_N}$ **do**
3: select the path type by following pseudo-code 3.
4: **if** direct path has been selected **then**
5: Insert it in vector $\boldsymbol{F_D}$.
6: **else**
7: Insert it in vector $\boldsymbol{F_{MH}}$.
8: **end if**
9: **end for**
10: Insert each direct data flow $\in \boldsymbol{F_D}$, the first sub-flow of each multi-hop data flow $\in \boldsymbol{F_{MH}}$, and the sub-flows $\in \boldsymbol{F_O}$ in List $\boldsymbol{L}$.
11: **for** each subflow $\in \boldsymbol{L}$ **do**
12: Create lists $\boldsymbol{L_{1,i}}$, $\boldsymbol{L_{2,i}}$, and $\boldsymbol{L_{3,i}}$ by following the criteria of Step 1.1.
13: **end for**
14: Create an undirected interference graph, G(V,E), by following Step 1.2.
15: Color the graph G(V,E) by following pseudo-code 5.
16: Apply the multi-color criterion to graph G(V,E) by following Step 2.2.
17: Establish Color Transmission Order by following Step 2.3.
18: Calculate the initial stage lengths by following pseudo-code 6.
19: Extend the stage lengths, if necessary, by following pseudo-code 7.
20: The scheduling rules are pushed to all APs.

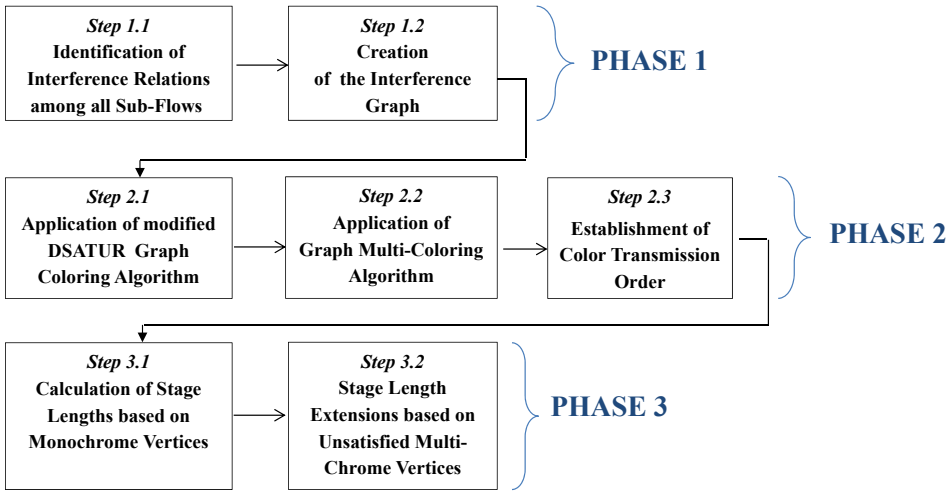


Figure 3.8: *Multi-criteria scheduling algorithm phases.*

3.6.1 Phase 1

The objective of this phase is to identify the critical interference relations among all sub-flows in input.

Step 1.1. The first step is to get for each sub-flow two final lists: the

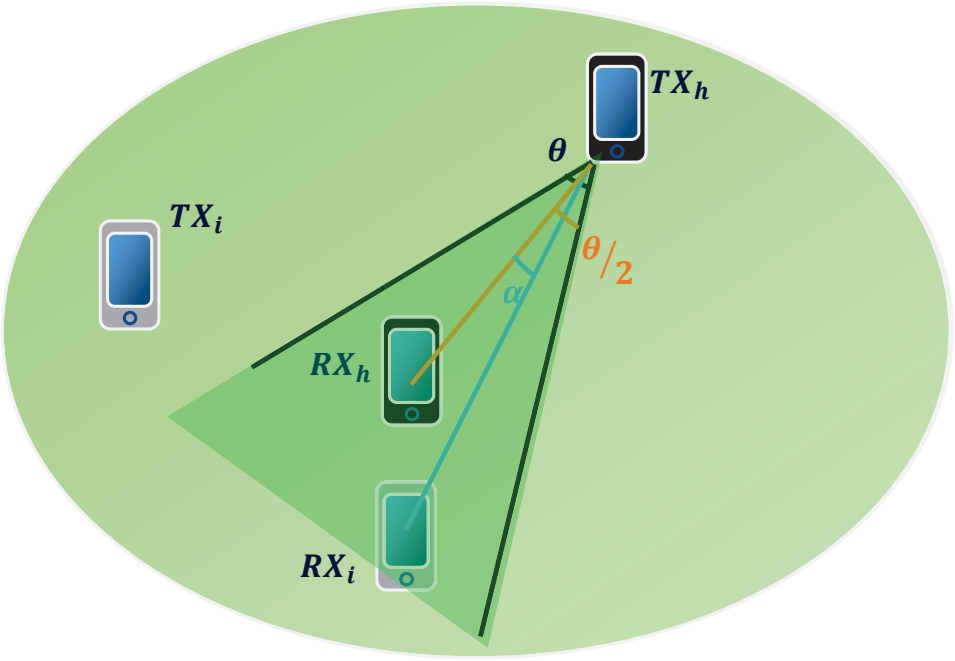


Figure 3.9: *Second interference criterion.*

list of intolerable interfering sub-flows and the list of tolerable ones. At this aim for the ith sub-flow, firstly the Central Controller derives $\boldsymbol{L_{1,i}}$, $\boldsymbol{L_{2,i}}$ and $\boldsymbol{L_{3,i}}$ as the lists of interfering sub-flows, by using three criteria described in the following.

Keeping in mind constraint (3.8)-(3.10),if receiver RX_i or transmitter TX_i of sub-flow i matches with RX_j or TX_j of sub-flow j, then sub-flows i and j must be scheduled in different stages. So, sub-flow j is inserted into list $\boldsymbol{L_{1,i}}$ and i into list $\boldsymbol{L_{1,j}}$.

The second criterion takes into account the antenna directivity, as in [10]. According to Fig. 3.9, the CC evaluates an angle α, by using the cosine law, as a function of distance metric. If $\alpha \leq \frac{\theta}{2}$, then RX_i is inside the beamwidth range of TX_h (i.e., h is an interfering sub-flow for i), consequently sub-flow h is inserted into list $\boldsymbol{L_{2,i}}$, otherwise not.

Unlike list $\boldsymbol{L_{1,i}}$, list $\boldsymbol{L_{2,i}}$ may contain sub-flows that are **non-severe** interfering with sub-flow i. The second criterion, in fact, takes into account only the antenna directivity, but does not quantify the received interference power. Because the high path loss at 60 GHz greatly reduces the interference level, it could be possible to transmit more interfering sub-flows concurrently, thus increasing the transmission efficiency.

For the above reason, we define a third criterion to establish what in-

terfering sub-flows can be classified as tolerable ones for sub-flow i (i.e., their interference is so non-severe that they can transmit concurrently). The CC allows concurrent transmissions only if $R_{\text{source},i}$ is guaranteed, as far as possible. At this aim, for sub-flow i, the SINR_i at receiver RX_i is estimated by using (3.3) where $j \in \boldsymbol{L_{2,i}}$. Next, the obtained SINR_i value is compared with $\text{SINR}_{\text{min},n}$, that is the value which guarantees $R_{\text{source},n}$, already defined in the previous section. If $\text{SINR}_i < \text{SINR}_{\text{min},n}$, the j-th sub-flow with the highest interference power is removed from list $\boldsymbol{L_{2,i}}$ and inserted into list $\boldsymbol{L_{3,i}}$. Then, SINR_i is recalculated. This process continues until $\text{SINR}_i \geq \text{SINR}_{\text{min},n}$.

Finally, for the ith sub-flow, we define $\boldsymbol{L_i} = \boldsymbol{L_{1,i}} \cup \boldsymbol{L_{3,i}}$ as the list of intolerable interfering sub-flows and $\boldsymbol{L_{2,i}}$ as the final list of tolerable interfering sub-flows.

Step 1.2. The second step is to create an undirected interference graph, G(V,E), where each sub-flow is a vertex and the edges represent the intolerable interference relations among all sub-flows. More specifically, if a sub-flow j is present in list $\boldsymbol{L_i}$, then an edge interconnects vertex i and vertex j.

3.6.2 Phase 2

The objective of the second phase is to assign different transmission stages to intolerant interfering sub-flows and to sort the transmission stages.

Step 2.1. Firstly, it is important to minimize the number of stages (N_{stage}) exploiting concurrent transmissions. To solve this problem, we adopt a vertex coloring approach. As a matter of the fact, graph coloring has considerable application to a large variety of complex problems involving optimization. In particular conflict resolution can often be accomplished by means of graph coloring, as shown in [56]. Our target to minimize N_{stage} corresponds to color vertices of the interference graph G(V,E) by using the minimum number of colors such that no two adjacent vertices share the same color. Therefore, different colors (stages) must be assigned to vertices (sub-flows) interconnected between them in the graph. Despite the graph coloring problem being known to be NP-complete, in literature much attention has been focused on the development of heuristic algorithms which will usually produce a good,

though not necessarily optimal, coloring for any graph in a reasonable amount of time. Among several algorithms available in literature, we consider the DSATUR graph-coloring method [57] because of its computational efficiency and low complexity. Starting with a single color, the original DSATUR algorithm assigns a new color to the selected vertex only if there are no available colors among those already assigned. It runs iteratively until all vertices are colored.

The main drawback is that the DSATUR algorithm considers all vertices having the same transmission priority. Instead, in our system, a vertex representing the first or an intermediate hop of a multi-hop data flow could be transmitted with low priority, because packets will not be delivered to the destination UE in the current frame, but will only be stored in an AP. In order to minimize the end-to-end delay, we take into account the priority of sub-flows, by proposing a modified version of the DSATUR algorithm, as described in pseudo-code 5. Our modification favors to assign the same color to high priority sub-flows, as far as possible, and this will positively affect the reduction of the end-to-end delay, as we will see in step 2.3.

At the output of the modified DSATUR algorithm we have a color pool $\boldsymbol{C}$ and one color assigned to each vertex. The pool size $|\boldsymbol{C}|$ represents the number of stages.

Step 2.2. As described in Section 3.5, in order to further improve the transmission efficiency, our innovative idea is to split transmission of sub-flows into more stages, as formulated in assumption (3.16). Therefore, the next step is to color vertices with more than one color, while guaranteeing the restrictions imposed by the interference graph. Let us note that this criterion must not increase the number of stages (i.e., the color pool size), already determined in the previous step.

Fig. 3.10 shows an example of the benefit to apply more than one color to a vertex (i.e., to split data transmission of a sub-flow in more stages). We consider to schedule five sub-flows, whose intolerant interference relations are shown in the interference graph. We suppose that the sub-flow corresponding to vertex 5 needs a large number of time slots to transmit its data. As shown in the interference graph of Fig. 3.10a, yellow-colored vertex 5 could also be colored with green, because its neighbors (intolerable interfering sub-flows) use only the red color, con-

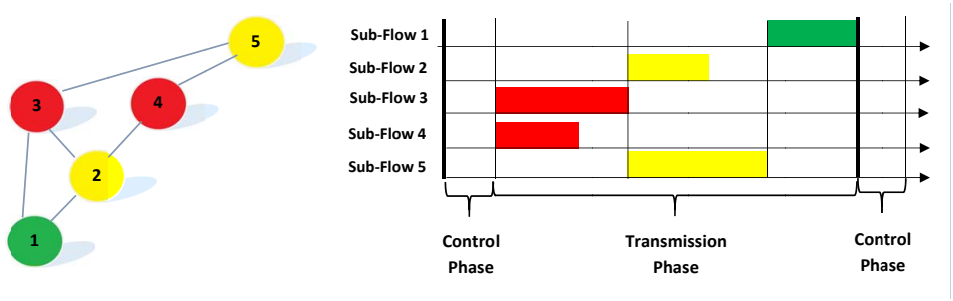
Pseudo-code 5 *Step 2.1.* Color graph criterion (Modified DSATUR Algorithm)

Definitions:

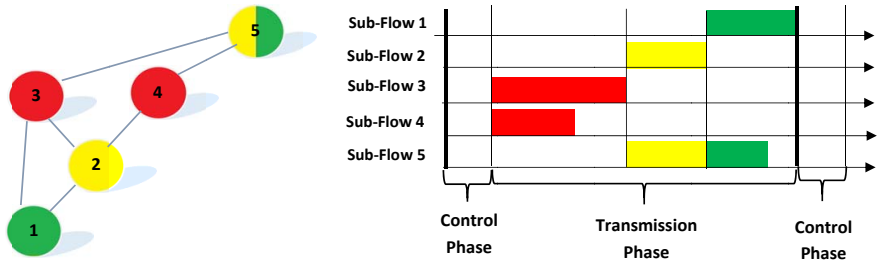
- C: Color pool, containing initially only one color ($|C| = 1$)
- $\theta_{f,i}$: saturation degree of vertex i (total number of different colors to which the vertex is connected)

Iteration:

while all vertices are colored **do**
  sort all uncolored vertices by decreasing order of $\theta_{f,i}$
  select the vertex (vertices) having the maximum $\theta_{f,i}$ value
  if there are more vertices having the same $\theta_{f,i}$ value **then**
    select among them the vertex (vertices) having the maximum number of uncolored neighbors
    if there are more vertices with the above characteristics **then**
      select among them (if present) a vertex representing a high-priority sub-flow (i.e., the last hop of a multi-hop data flow or a direct sub-flow); otherwise a random choice is made
    end if
  end if
  if there are color(s) in C different from those assigned to the neighbors **then**
    color the vertex with one of the available colors
  else
    color the vertex with a new color and increase the size of C
  end if
end while



(a) *Only one color per sub-flow*



(b) *More colors per sub-flow*

Figure 3.10: *Benefit to apply more colors to a vertex.*

sequently its data transmission can be split, as in Fig. 3.10b, reducing the current frame length.

In this example, we also note that vertex 4 could be two-colored (red

and green), instead of vertex 5. But vertex 4 completes its transmission by using less time slots than the only-red-colored vertex 3, so the frame length would remain the same of Fig. 3.10a. This means that we need to select, among all vertices that are candidates for multi-coloring, the ones corresponding to the sub-flows that require more time slots to transmit their data. With this aim in mind, we want to associate an easy-to-calculate weight to each vertex, thus to create a weighted colored graph. Because the amount of time slots depends on the number of packets to be transmitted ($n_{packets,i}$) by sub-flow i, the distance between transmitter and receiver, and also on the visibility conditions (α), for the ith vertex, the weight proposed is:

$$w_i = \frac{1}{|\boldsymbol{C_i}|} \left(\frac{dist\,(TX_i, RX_i)}{d_0} \right)^{\alpha} n_{packets,i}, \tag{3.18}$$

where $\boldsymbol{C_i}$ is the set of colors assigned to vertex i (at the beginning, $|\boldsymbol{C_i}| = 1$). In this way, a high-weighted vertex represents a sub-flow requiring a large number of time slots.

In conclusion, in Step 2.2, we define a new criterion aiming to multi-color one or more vertices, always guaranteeing that adjacent vertices cannot share the same color. This criterion works as follows. First, the controller creates a weighted colored graph and a list of vertices ($\boldsymbol{V}$) in descending order by weights. Next, starting from the first element of $\boldsymbol{V}$, if in $\boldsymbol{C}$ there are colors available for it (i.e., one or more colors in $\boldsymbol{C}$ are not assigned to any neighbor and to the vertex itself), then one additional color is assigned to the vertex. After that, its weight value and list $\boldsymbol{V}$ are updated so that its priority to obtain another color is reduced. Instead, if there are no available colors for it, the vertex is deleted from the list. This process continues until $\boldsymbol{V}$ is empty.

Step 2.3. Now, we have a list of stages (colors) and a many-to-many correspondence between colors and vertices. The last step in the second phase is to establish the transmission order of stages (colors), which influences the end-to-end delay. At this regard, we remember that in our modified DSATUR algorithm, we have given high priority to direct or last-hop sub-flows. For this reason, we establish to transmit the colors (stages) containing more high priority sub-flows first, the other ones then. In the event that there is an equality, the CC makes a random choice.

As output of the entire second phase, we have $|\boldsymbol{C}|$ colors in transmission order ($k = 1, 2, ..., |\boldsymbol{C}|$) and, for each color k, a list of vertices using this color ($\boldsymbol{L_{col_k}}$).

3.6.3 Phase 3

The third and last algorithm phase focuses on calculating the number of time slots to be allocated to each color. It is important to make the most efficient allocation with the minimum number of time slots, while fulfilling all sub-flows. The issue is that the amount of required resources depends on transmission rate, which in turn depends on current interference.

Step 3.1. In order to establish in the first instance the duration of the first stage (color $k = 1$) in terms of time slots, among all vertices in list $\boldsymbol{L_{col_1}}$ the CC selects only the monochrome ones, because they cannot split the related data in more stages. For monochrome vertex i, the Central Controller calculates the estimated SINR$_i$, evaluated by using (3.3), where $j \in \boldsymbol{L_{2,i}} \cap \boldsymbol{L_{col_1}}$, i.e., by taking into account only the sub-flows transmitting during stage 1, including the multi-color ones. Then, the transmission rate R_i can be estimated by using (3.2). Finally, the number of time slots necessary to transmit the data related to sub-flow i is:

$$n_{ts_{i,1}} = \left\lceil \frac{n_{packets,i} 8 N_{\text{bytes,packet}}}{t_{PAY} R_i} \right\rceil . \tag{3.19}$$

After calculating (3.19) for each monochrome vertex in $\boldsymbol{L_{col_1}}$, the length of stage 1 (n_{ts_1}) is set equal to $\max_i \left\{ n_{ts_{i,1}} \right\}$.

Before moving on the second color, the CC needs to analyze the multi-colored vertices $\in \boldsymbol{L_{col_1}}$, if any. They ensure more concurrent transmissions, as previously shown, but on the other hand they complicate the estimation of interference. For example, if sub-flow h corresponding to a two-colored vertex $\in \boldsymbol{L_{col_1}} \cap \boldsymbol{L_{col_j}}$ could meet its demand by using only color 1, then the CC deletes it from $\boldsymbol{L_{col_j}}$ because it will cause no interference to sub-flows transmitting during the j-th stage. Instead, if sub-flow h needs to use more than one color, then the CC evaluate how many sub-flow packets can be transmitted during stage (color) 1 (i.e., in n_{ts_1} time slots):

$$n_{packets_{h,col_1}} = \left\lfloor \frac{n_{ts_1} R_h t_{PAY}}{8 N_{\text{bytes,packet}}} \right\rfloor . \tag{3.20}$$

Therefore, the Central Controller estimates the amount of remaining packets to be transmitted by sub-flow h during stage (color) j.

Once all vertices in $\boldsymbol{L_{col_1}}$ are analyzed, the CC can establish the duration of the second color and so on, in a similar manner to the first one. The main difference consists in the multi-colored vertices management. If the considered color represents the last available color of a multi-colored vertex, then the CC needs to evaluate whether all its remaining packets can be transmitted during this last stage. If it does not succeed in meeting its demands, then the CC includes this vertex (sub-flow) in a list ($\boldsymbol{L_{NotSatisfied}}$) containing all sub-flows needing to extend the length of one or more colors (stages) to transmit all the related data. The details of all operations described above are shown in pseudo-code 6.

Step 3.2. After the length of each stage has been established in the first instance by the monochrome vertices, the CC needs to analyze the elements in $\boldsymbol{L_{NotSatisfied}}$.

If the list contains a single sub-flow i, the CC selects the color whose length needs to be extended, as following:

- if it is a high-priority sub-flow, then its first color is selected;
- if it is a low-priority sub-flow, then its last color is selected.

Then, the length of the selected color is set out as the number of time slots necessary to fulfill data transmission of sub-flow i.

Finally, we describe the strategy adopted if $\boldsymbol{L_{NotSatisfied}}$ contains more than one sub-flow. We consider the graph in Fig. 3.11. We suppose that the transmission order derived in Step 2.3 is {RED, YELLOW, GREEN}, $\boldsymbol{L_{NotSatisfied}} = \{5, 6\}$, and both vertices correspond to direct sub-flows (high-priority). If the CC applies the above strategy for them, then we obtain that sub-flow 5 requires to extend color YELLOW, and sub-flow 6 color RED.

We note that this strategy is not efficient, because the best solution is to extend the length of color GREEN they have in common. In a more complex situation, there may be more colors in common with all

Pseudo-code 6 *Step 3.1.* Initial estimate of stage lengths

Definitions:

- $\boldsymbol{C}$: Color Pool, in transmission order
- $\boldsymbol{L_{col_k}}$: list of vertices using color (stage) k
- n_{ts_k}: number of time slots to be allocated to color (stage) k
- $\boldsymbol{C_i}$: set of assigned color(s) to vertex (sub-flow) i, in transmission order
- $n_{packets,i}$: number of packets tof vertex (sub-flow) i
- $n_{ts_{i,k}}$: number of time slots required by vertex (sub-flow) i in color (stage) k

Iteration:

for each color $k \in C$ **do**
  for each vertex $i | i \in \boldsymbol{L_{col_k}} \wedge |\boldsymbol{C_i}| = 1$ **do**
    calculate $n_{ts_{i,k}}$
  end for
  set $n_{ts_k} = \max_i \left\{ n_{ts_{i,k}} \right\}$
  for each vertex $i | i \in \boldsymbol{L_{col_k}} \wedge |\boldsymbol{C_i}| > 1$ **do**
    if color k is the first element of $\boldsymbol{C_i}$ **then**
      calculate $n_{ts_{i,k}}$
      if $n_{ts_{i,k}} < n_{ts_k}$ **then**
        delete vertex i from $\boldsymbol{L_{col_m}}, \forall m \in \boldsymbol{C_i}, m > k$
      else
        evaluate how many packets i can transmits in n_{ts_k}
      end if
    else if k is the last element of $\boldsymbol{C_i}$ **then**
      estimate the number of time slots necessary to fulfill data transmission of i in color k: $\widehat{n}_{ts_{i,k}}$
      if $\widehat{n}_{ts_{i,k}} > n_{ts_k}$ **then**
        insert i in list $\boldsymbol{L_{NotSatisfied}}$
      end if
    else
      calculate $\widehat{n}_{ts_{i,k}}$
      if $\widehat{n}_{ts_{i,k}} < n_{ts_k}$ **then**
        delete vertex i from $\boldsymbol{L_{col_m}}, \forall m \in \boldsymbol{C_i}, m > k$
      else
        evaluate how many packets i can transmits in n_{ts_k}
      end if
    end if
  end for
end for

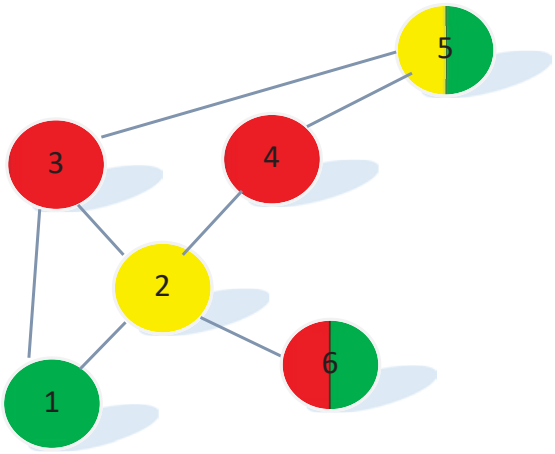


Figure 3.11: *Example of more colors in* $L_{NotSatisfied}$.

$$M_C = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$M'_C = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{L_{ColorExt}} = \{7,4\}$$

Figure 3.12: *Color Matrix Procedure.*

sub-flows in $\boldsymbol{L_{NotSatisfied}}$. In order to extend the minimum number of colors ensuring transmission efficiency, we propose the following greedy criterion. It is a sub-optimal solution to keep the computational load low.

We introduce a color matrix $\boldsymbol{M_C}$ of size $|\boldsymbol{L_{NotSatisfied}}| \times |\boldsymbol{C}|$, where each element $a_{i,j} = 1$ if the ith sub-flow ($i \in \boldsymbol{L_{NotSatisfied}}$) uses color j ($j \in \boldsymbol{C}$), $a_{i,j} = 0$ otherwise. Fig. 3.12 shows an example of this procedure. The controller selects the column with the maximum number of 1s (random choice between columns 5 and 7). In our example, column 7 is selected. Next, the controller deletes all rows containing 1 in the selected column to obtain a row-reduced matrix, and inserts color

7 in list $\boldsymbol{L_{ColorExt}}$. This process continues until Color Matrix is empty. As output of the criterion, we obtain a list of color(s) to be extended ($\boldsymbol{L_{ColorExt}}$).

Then, we perform the following steps:

1. for each color j in $\boldsymbol{L_{ColorExt}}$, we select each sub-flow $i \in \boldsymbol{L_{NotSatisfied}}$ using **only** color j, and estimate the amount of time slots necessary to fulfill the relative data transmission ($\widehat{n}_{ts_{i,j}}$). Then, we remove them from $\boldsymbol{L_{NotSatisfied}}$, and set out the new length of color j as the maximum value of $\widehat{n}_{ts_{i,j}}$.

2. We verify whether sub-flow $i \in \boldsymbol{L_{NotSatisfied}}$ (if any) can now complete its transmission with the new color length values. If so, then we remove it from $\boldsymbol{L_{NotSatisfied}}$. After that, if $\boldsymbol{L_{NotSatisfied}}$ is not empty, $\boldsymbol{M_C}$ is recalculated and the process starts again, otherwise it ends.

The details on extending the stage lengths of sub-flows in $\boldsymbol{L_{NotSatisfied}}$ are shown in the pseudo-code 7.

3.7 Performance Evaluation

In this section, we evaluate the performances of the proposed access control scheme in MATLAB environment.

3.7.1 Simulation assumptions

In our simulation scenario, as in [48] we consider a flat indoor area of $50m \times 50m$, covered with 9 APs uniformly distributed to form a regular grid. The CC is located inside the AP in the center of the scenario. We consider N_U UEs uniformly distributed within the whole area. We assume that transmissions among APs occur in LOS conditions, whereas any other ones in NLOS.

The set of parameters used in simulations is provided in Table 3.1. Each result is averaged over 100 independent simulations. To take into account the frame overhead, we set each control phase length (t_{cont}) equal to 10 time slots [19].

Pseudo-code 7 *Step 3.2.* Stage lengths to be extended

Definitions:

- $\boldsymbol{L_{NotSatisfied}}$: list of vertices (sub-flows) not satisfying their demands
- n_{ts_k}: number of time slots allocated to color (stage) k
- $\boldsymbol{C_i}$: set of color(s) assigned to vertex (sub-flow) i, in transmission order
- $n_{packets,i}$: number of packets of vertex (sub-flow) i

Iteration:

while $\boldsymbol{L_{NotSatisfied}}$ is not empty **do**
  if $|\boldsymbol{L_{NotSatisfied}}| == 1$ **then**
    if $i \in \boldsymbol{L_{NotSatisfied}}$ represents a high-priority sub-flow **then**
      select the first color $m \in \boldsymbol{C_i}$
    else
      select the last color $m \in \boldsymbol{C_i}$
    end if
    estimate $\widehat{n}_{ts_{i,m}}$ (i.e., the number of time slots necessary to fulfill data transmission of sub-flow i in color m) and set $n_{ts_m} = \widehat{n}_{ts_{i,m}}$
  else
    create the Color Matrix $\boldsymbol{M_C}$
    repeat
      select the column j with the maximum number of 1s;
      delete all rows containing 1 in column j and insert j in list $\boldsymbol{L_{ColorExt}}$
    until $\boldsymbol{M_C}$ is empty
    for each color $j \in \boldsymbol{L_{ColorExt}}$ **do**
      for each sub-flow $l \in \boldsymbol{L_{NotSatisfied}}$ such that $j \in \boldsymbol{C_l} \wedge |\boldsymbol{C_l} \cap \boldsymbol{L_{ColorExt}}| = 1$ **do**
        remove l from $\boldsymbol{L_{NotSatisfied}}$ and estimate $\widehat{n}_{ts_{l,j}}$
      end for
      set $n_{ts_j} = \max_l \left\{\widehat{n}_{ts_{l,j}}\right\}$
    end for
    for each sub-flow $h \in \boldsymbol{L_{NotSatisfied}}$ (i.e., $|\boldsymbol{C_h} \cap \boldsymbol{L_{ColorExt}}| > 1$), if any **do**
      if h can now complete its transmission with the new color length values **then**
        remove h from $\boldsymbol{L_{NotSatisfied}}$
      end if
    end for
  end if
end while

Table 3.1: *System Parameters*

Parameter	Symbol	Value
System bandwidth [10]	W	$1600MHz$
Carrier frequency	f_c	$60GHz$
Beam angle	θ	$45°$
UE and AP TX Power	P_T	$10mW$
Background Noise	N_0	$-114dBm/MHz$
PL exponent [54]	α	1.73 (LOS) 3.19 (NLOS)
Shadowing Standard deviation [54]	σ	$3dB$ (LOS) $8.29dB$ (NLOS)
Maximum NLOS distance	$d_{\text{max,NLOS}}$	$16.97m$
Control phase length	t_{con}	10 time slots
TX and RX Gain	G_R, G_T	8
Reference distance	d_0	$1m$
Slot time [48]	t_{SLOT}	$5\mu s$
Payload time [48]	t_{PAY}	$4\mu s$
Physical overhead [53]	t_{PHY}	$250ns$
SIFS interval [53]	t_{SIFS}	$100ns$
Probability	P	0.99
Packet size [48]	$N_{\text{bytes,packet}}$	1000 bytes
Transceiver efficiency	η	1
Number of UEs	N_U	$20, 25, \ldots, 50$
Number of active commun.	N_C	$2, 5, 10, 15$
Delay threshold [48]	d_{thr}	$50ms$
Source bit rate	R_{source}	2 Gbps, 4 Gbps
Simulation time [48]	t_{SIM}	$500ms$

3.7.2 Traffic Model

We assume that a source UE generates an active communication for a destination UE at a random time (t_0) uniformly distributed in the continuous range $[0, t_{SIM}]$. An active communication consists of a variable number of packets, generated at a constant peak rate (R_{source}). We introduce 6 traffic classes and assume that the number of generated packets for an active communication of class h is uniformly distributed in a discrete range $\Delta_h = [5000(h-1), 5000h]$, with $h = \{1, 2, \ldots, 6\}$.

For each simulation, we set the number of active communications (N_C). Source and destination UEs are randomly chosen on N_U UEs in such a way that:

- a UE may be only source or destination UE;
- a source UE is, at most, the source of a single active communication;
- a destination UE may be the destination of more active communications.

3.7.3 Performance metrics

We compare our solution with the D3MAC scheme [48]. Unlike our control scheme, Gao *et al.* consider that the transmission rate of each wireless link is fixed and known (equal to 2, 4 or 6 Gbps, according to the distance between devices). In order to fairly assess the two control schemes, we adapt D3MAC scheme to our more accurate approach in which the transmission rates are variable according to the interference level. We analyze the system performance in several traffic conditions and any active communication requiring a stringent delay threshold, $d_{thr} = 50ms$, as in [48]. We introduce several metrics aiming to analyze not only performances of the overall system (packet level performances) but also of each active communication (session level performances).

Let P be the set of all generated packets, and P_d the set of packets delivered within the delay threshold, i.e., $P_d = \{p | p \in P \wedge d_{e2e,p} < d_{thr}\}$, where $d_{e2e,p}$ is the end-to-end delay of packet p. We measure:

$$\text{Successfully delivered packets}(\%) = \frac{|P_d|}{|P|}100, \tag{3.21}$$

$$\text{Average packets delay} = \frac{1}{|P_d|} \sum_{\forall p \in P_d} d_{e2e,p}. \tag{3.22}$$

In order to compare the behavior of two control schemes in the same traffic conditions, we define "Throughput Gain" of the ith active communication (ΔTh_i) as the difference between the throughput achieved by our control scheme (Th_i) and the one achieved by D3MAC (Th_i^{D3MAC}), normalized to the source rate ($R_{\text{source},i}$).

$$\Delta Th_i = \frac{Th_i - Th_i^{D3MAC}}{R_{\text{source},i}}, \quad \Delta Th_i \in [-1, 1]. \tag{3.23}$$

where Th_i is calculated as the ratio between the number of successfully delivered packets and the time interval from the first packet generation time to the last packet delivery time.

We also estimate the Average Throughput Gain, as:

$$\Delta Th = \frac{1}{N_C} \sum_{i=1}^{N_C} \Delta Th_i. \tag{3.24}$$

Finally, we estimate the Fairness among active communications, which is an important performance criterion in all resource allocation schemes. In literature the Jain's index [33] is a parameter which quantitatively measures the degree of fairness offered by a system allocating resources to N_C contending requests. The Jain's Index is defined as follows:

$$J_{index} = \frac{\left(\sum_{i=1}^{N_C} x_i\right)^2}{N_C \sum_{i=1}^{N_C} x_i^2}, \tag{3.25}$$

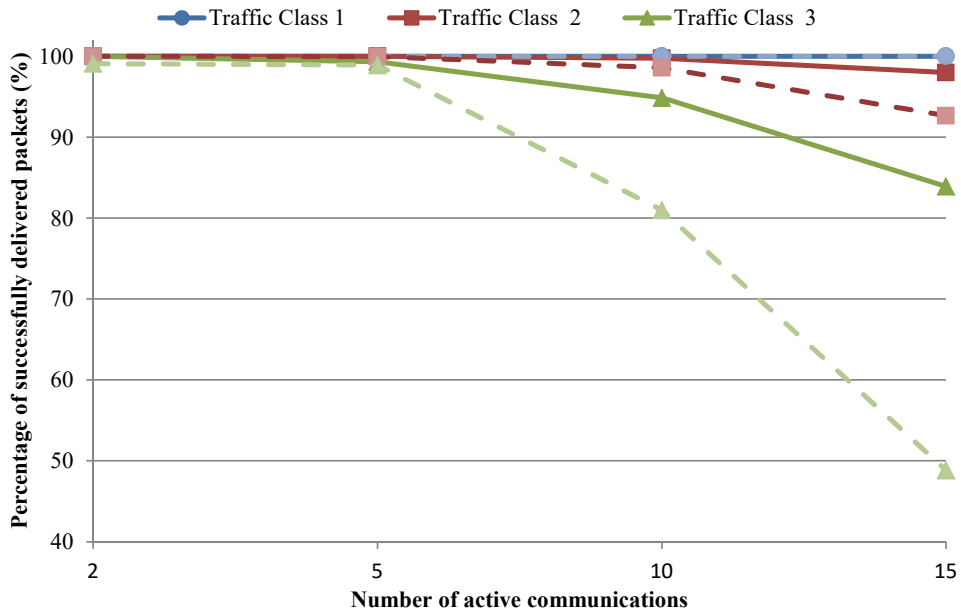
where x_i represents the resources allocated to ith contender. The results range from $1/N_C$ (i.e., all resources allocated to a single active connection) to 1 (i.e., each request receives the same amount of resources).

We can assume $x_i = |P_{d,i}|$, that is, the number of packets related to active communication i delivered within the delay threshold. However, J_{index} is a good fairness parameter only if all requests contend the same resources. Instead, in our system there may be some sub-flows that can use the entire resources without contenting with the other ones (because they do not have interfering sub-flows), while other sub-flows need to share some radio resources.

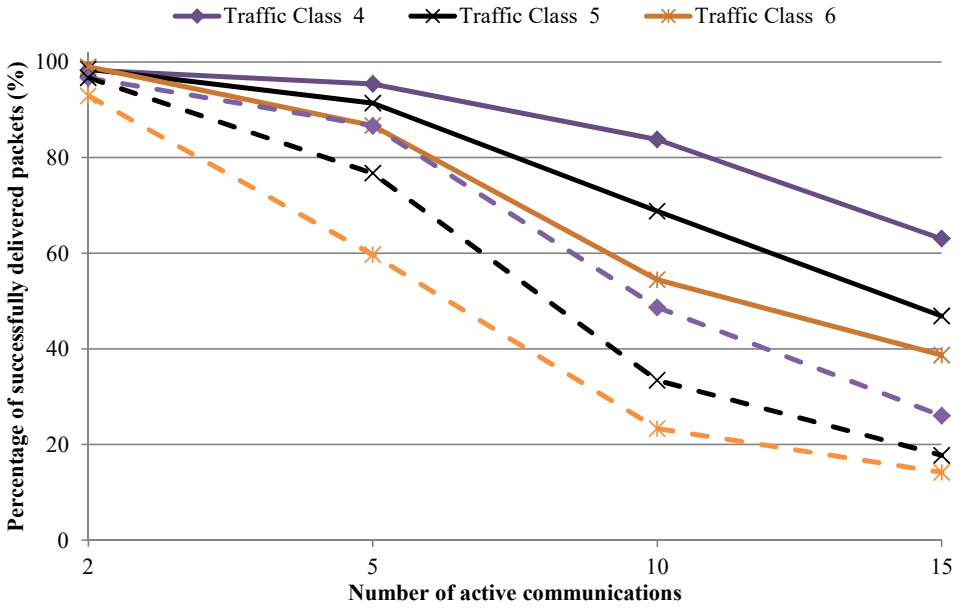
This means that in our scenario the J_{index} values are not representative in absolute terms. For this reason, we assume as fairness index "ΔJain's Index", the difference between the fairness achieved by our proposal (J_{index}) and the one achieved by D3MAC scheme (J_{index}^{D3MAC}):

$$\Delta J_{index} = J_{index} - J_{index}^{D3MAC}, \tag{3.26}$$

where $\Delta J_{index} \in \left[\frac{1}{N_C} - 1, 1 - \frac{1}{N_C}\right]$

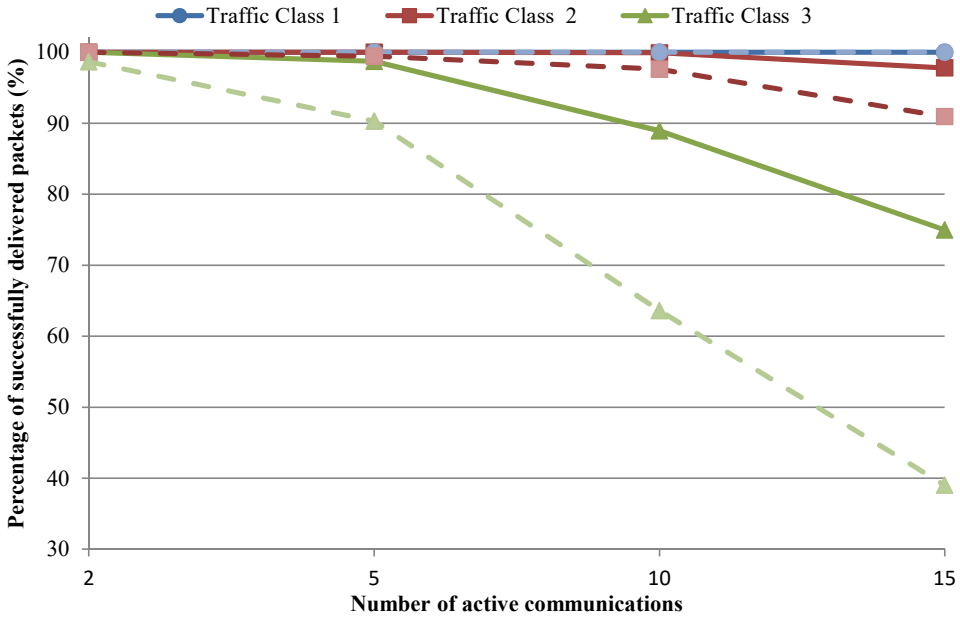


(a) *Traffic class 1, 2, and 3*

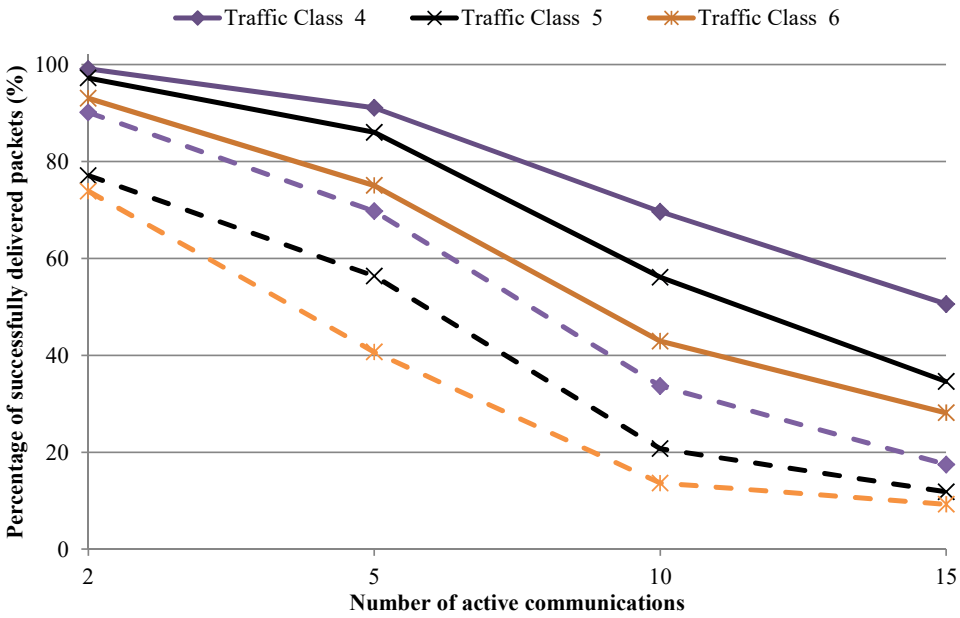


(b) *Traffic class 4, 5, and 6*

Figure 3.13: *Percentage of successfully delivered packets, under different traffic classes and number of active communications. Source rate of 2 Gbps for each active communication. Solid line represents our control scheme performance, dashed line D3MAC performance.*



(a) *Traffic class 1, 2, and 3*



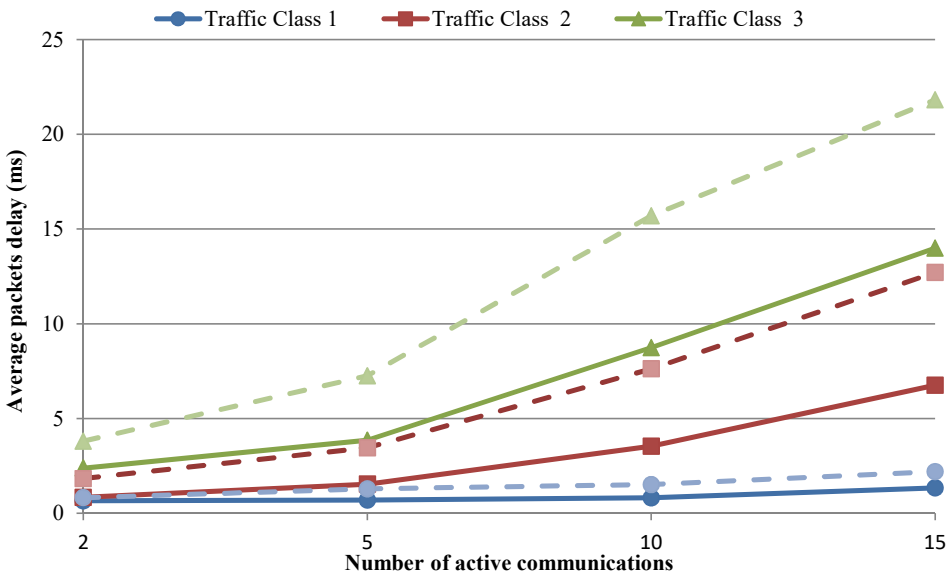
(b) *Traffic class 4, 5, and 6*

Figure 3.14: *Percentage of successfully delivered packets, under different traffic classes and number of active communications. Source rate of 4 Gbps for each active communication. Solid line represents our control scheme performance, dashed line D3MAC performance.*

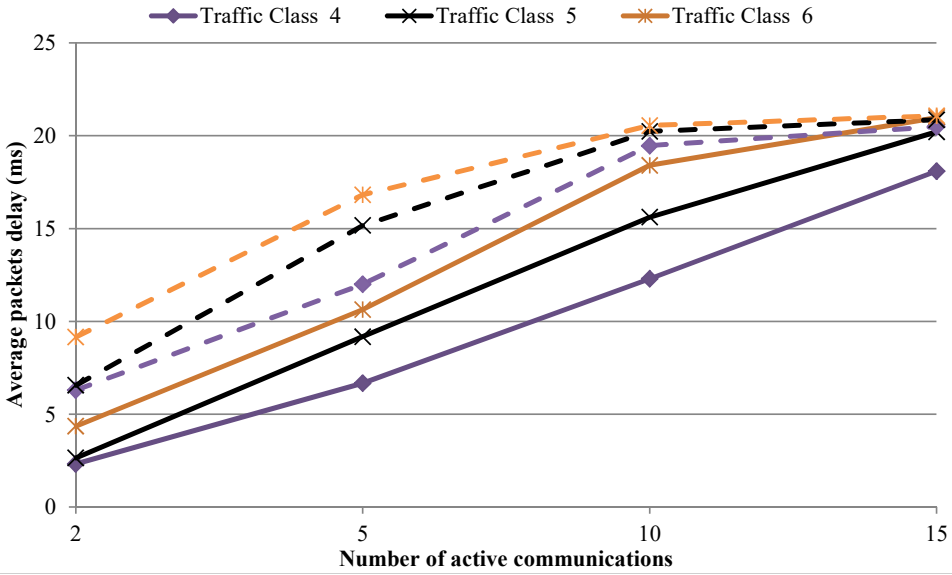
3.7.4 Performance Analysis

Packet Level Performance

Under several traffic conditions ($N_U = 30$, $N_C = \{2, 5, 10, 15\}$ and Traffic class $h = \{1, 2, ..., 6\}$), the percentage of successfully delivered



(a) *Traffic class 1, 2, and 3*

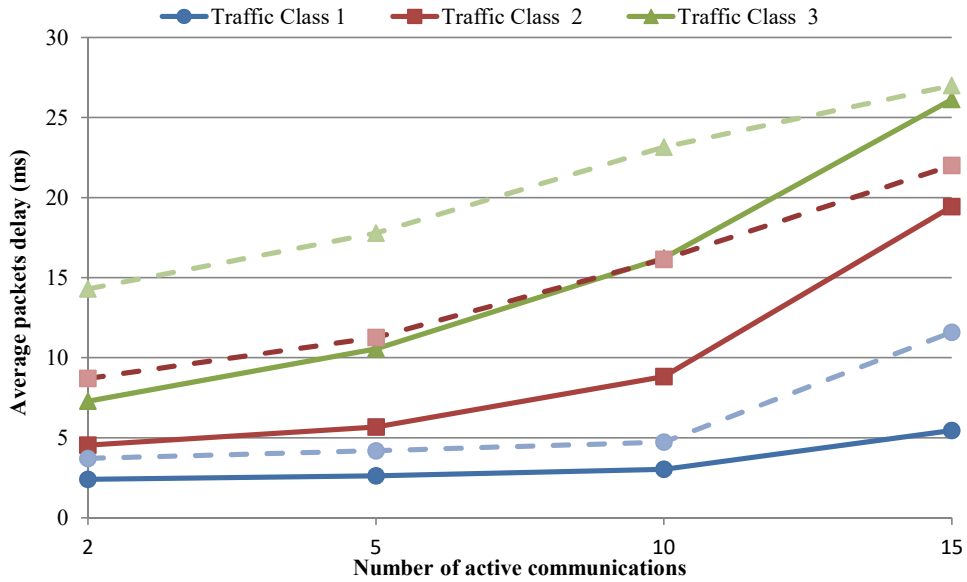


(b) *Traffic class 4, 5, and 6*

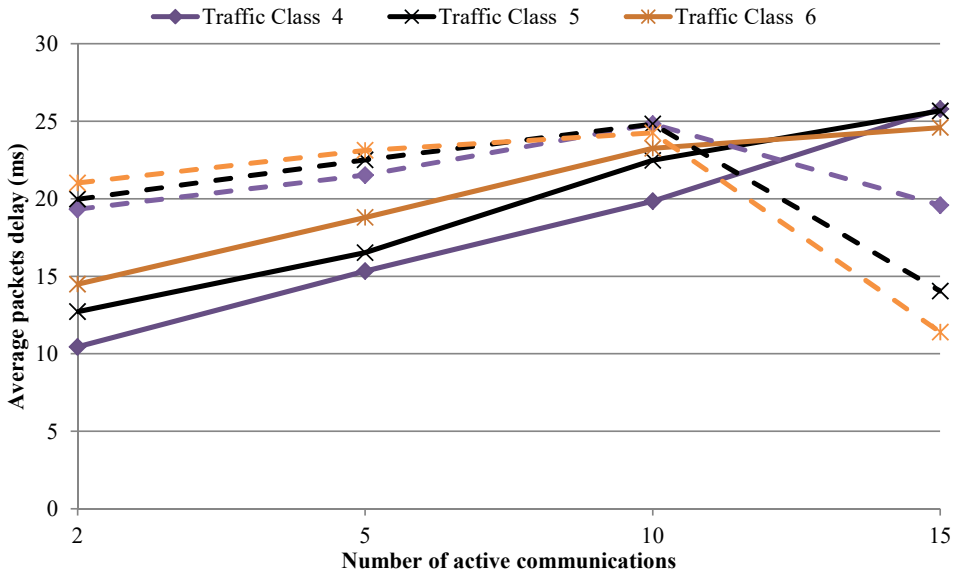
Figure 3.15: *Average packets delay, under different traffic classes and number of active communications. Source rate of 2 Gbps for each active communication. Solid line represents our control scheme performance, dashed line D3MAC performance.*

packets by adopting our control scheme and D3MAC scheme are shown in Figs. 3.13 and 3.14, where source rate of any active communication is equal to 2 Gbps and 4 Gbps, respectively.

In the case of very light traffic load (traffic class 1), both our algorithm and the D3MAC seem to show the same excellent performances. However, when the traffic load increases our solution outperforms the



(a) *Traffic class 1, 2, and 3*



(b) *Traffic class 4, 5, and 6*

Figure 3.16: *Average packets delay, under different traffic classes and number of active communications. Source rate of 4 Gbps for all active communications. Solid line represents our control scheme performance, dashed line D3MAC performance.*

D3MAC scheme, the more the traffic load the more the difference in performance. Of course, in the case of very heavy traffic load, both algorithms show worse performance because the network is close to saturation, but our proposal is still better.

We also report the average delay measures in Figs. 3.15 and 3.16, when source rate of any active communication is equal to 2 Gbps and 4

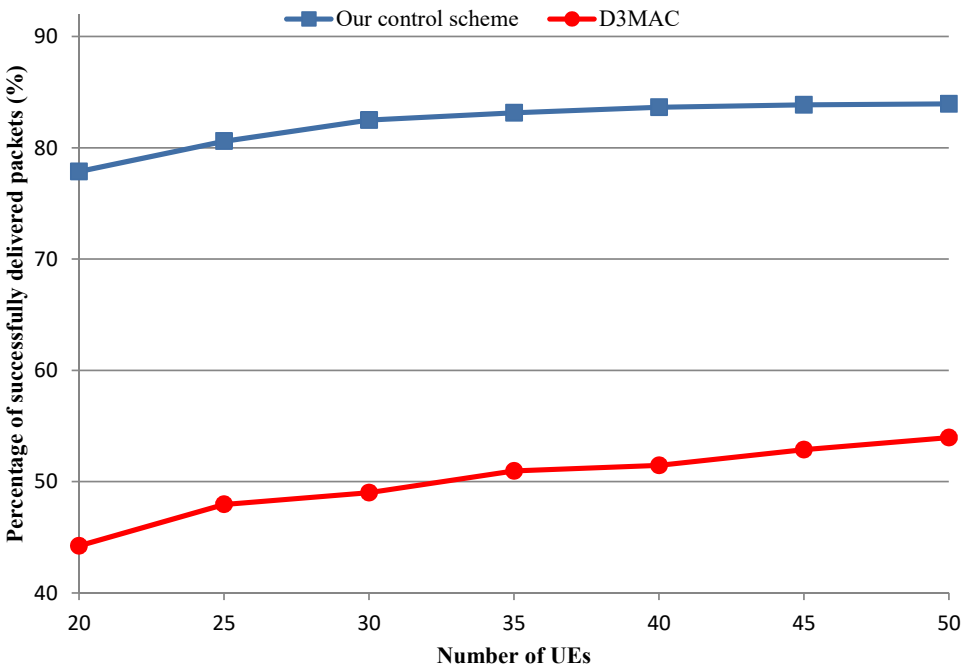
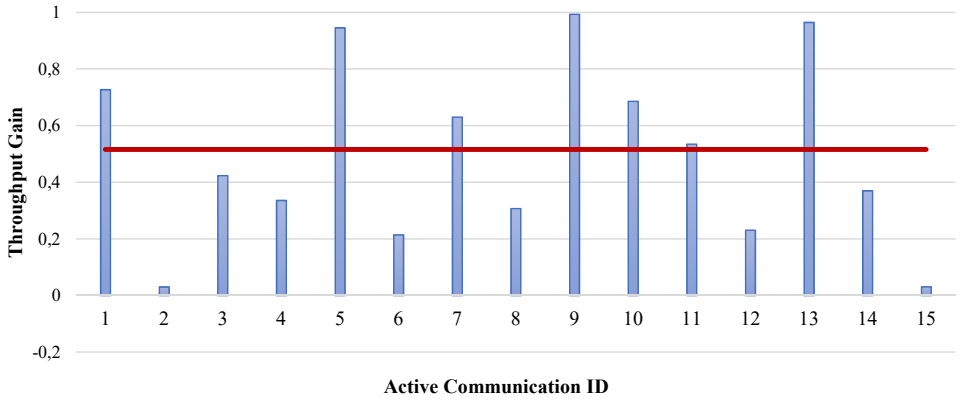


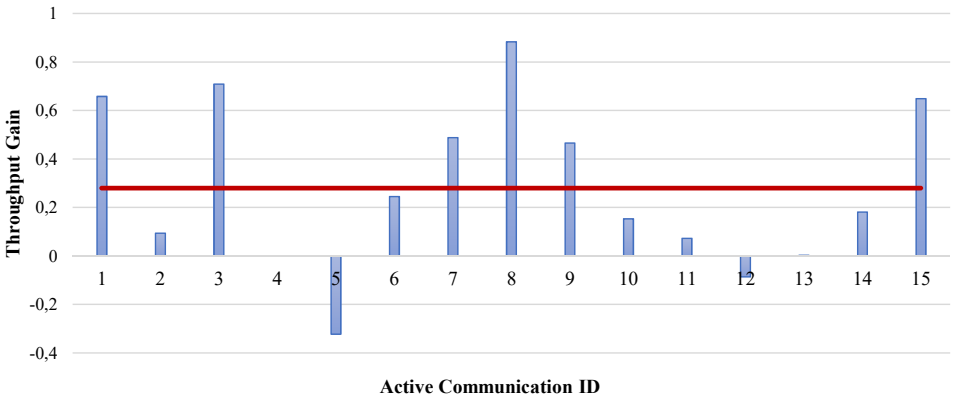
Figure 3.17: *Percentage of successfully delivered packets, under different number of UEs. 10 active communications, $R_{source} = 2$ Gbps and Traffic class 3 for any active communication.*

Gbps, respectively. We note that already under very light load conditions (traffic class 1) our control scheme is better than D3MAC. The delay difference increases as the traffic load increases. In the case of heavy traffic conditions ($N_C = 15$, $R_{\text{source}} = 2$ Gbps and traffic class 6) this difference comes down. Let us note that the measured statistics take into account only packets delivered within the threshold, so in heavy traffic conditions our control scheme achieves similar performance in average delay compared to D3MAC but with a very larger number of delivered packets (about 150% more). In the case of very heavy traffic load (e.g., $N_C = 15$, $R_{\text{source}} = 4$ Gbps and traffic class 4 to 6) the delay performance of D3MAC seems to be better, but only because the network is close to saturation and only a small amount of packets (20%) has been successfully delivered.

Now, as in [48], we assess the control schemes under several number of UEs, uniformly distributed in the whole area. This simulation is characterized by an invariable total offered load: 10 active communications, $R_{\text{source}} = 2$ Gbps and traffic class 3. Fig. 3.17 shows that the percentage of successfully delivered packets increases with N_U for both algorithms.



(a) *Simulation test 1*



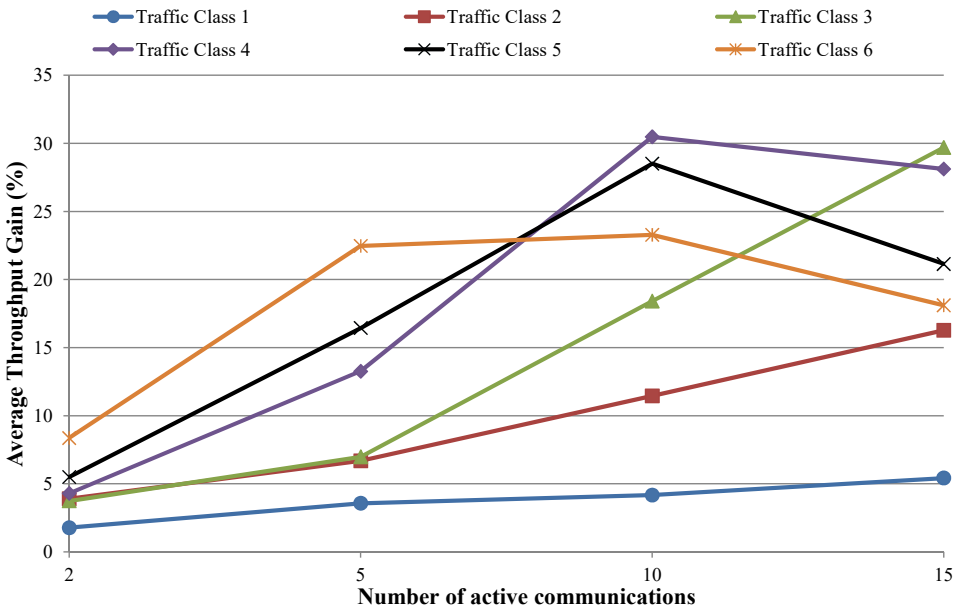
(b) *Simulation test 2*

Figure 3.18: *Throughput Gain for each active communication and average value (horizontal red line). 15 active communications, $R_{source} = 2$ Gbps and Traffic class 3 for any active communication.*

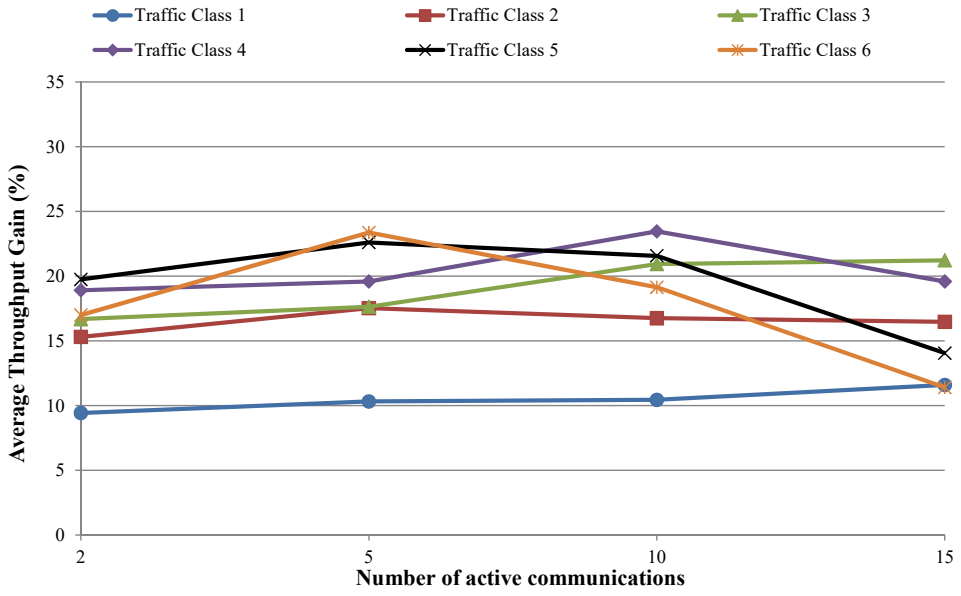
The explanation for this is that, for equal N_C, the smaller the number of UEs, the higher the probability that the same UE will be the destination of two or more active communications, thus increasing the number of intolerable interfering sub-flows, so the percentage of successfully delivered packets within the threshold is reduced. This phenomenon is all the more evident when the number of UEs is less than or closer to three times the number of active connections. Similar results are obtained by considering different total offered traffic loads.

Session Level Performance

In order to better understand the behavior of our control scheme compared to the one of D3MAC, in Fig. 3.18, we show the throughput gain of each active communication in two single representative simulation



(a) *Source rate of 2 Gbps for any active communication*

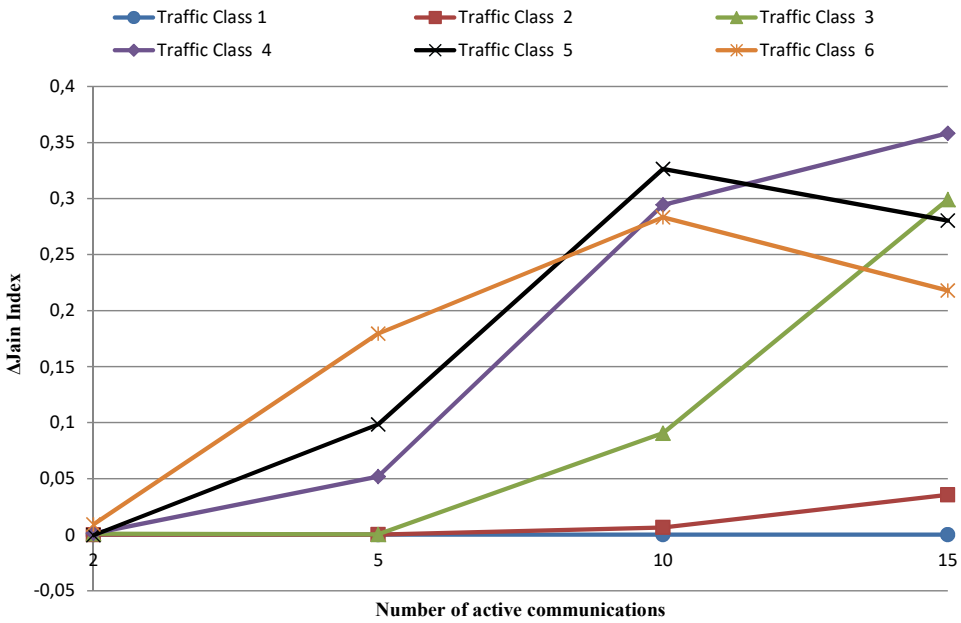


(b) *Source rate of 4 Gbps for any active communication*

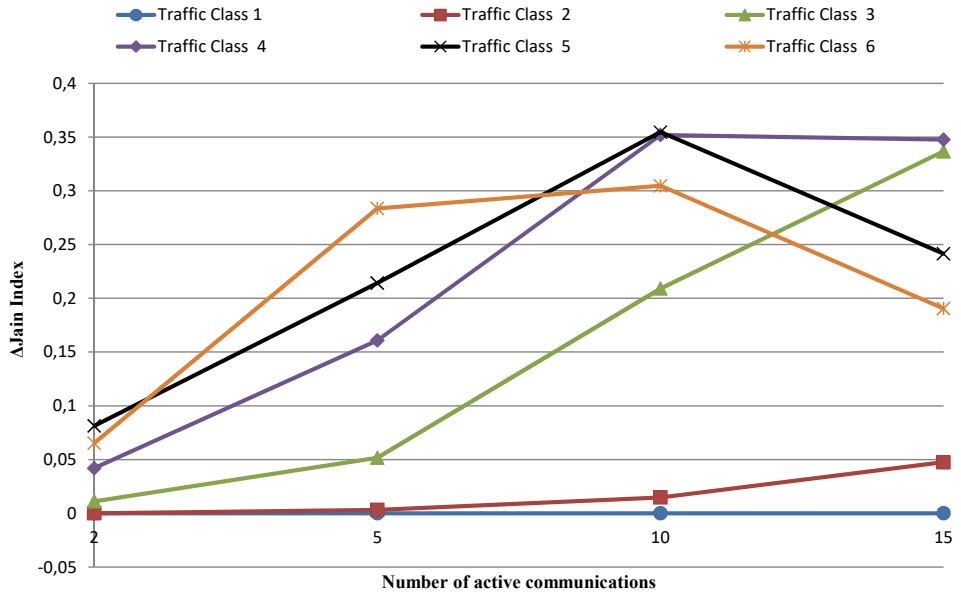
Figure 3.19: *Average Throughput Gain under different Traffic Classes and number of active communications.*

tests, with fixed traffic conditions, consisting on $N_U = 30$, $N_C = 15$, $R_{\text{source}} = 2$ Gbps, and traffic class 3.

In the first simulation test, our control scheme achieves an Average Throughput Gain value (represented as the horizontal red line) of about 50 percentage points compared to D3MAC scheme, and all active com-



(a) *Source rate of 2 Gbps for all active communications*



(b) *Source rate of 4 Gbps for all active communications*

Figure 3.20: ΔJ_{index} *under different traffic loads and number of active communications.*

munications are characterized by $\Delta Th_i \geq 0$. More specifically, three active communications $(i = \{5, 9, 13\})$ show $\Delta Th_i \approx 1$. This means that with our control scheme these three active communications achieve almost the maximum throughput, while with D3MAC their throughput is approximately zero (i.e., their transmission is inhibited and a very small

number of packets is delivered on time).

In the second simulation test, we achieve an Average Throughput improvement of 27.9% compared to D3MAC. Unlike the first one, the improvement does not occur on each single active connection, so we experience $\Delta Th_i < 0$ for two active communications ($i = \{5, 12\}$). However, let us note that in the worst case $\Delta Th_5 \approx -0.32$, it means that our control strategy has not inhibited any communication, while allowing three active connections ($i = \{3, 8, 15\}$) to achieve significantly better performance than the D3MAC scheme and improving the Average Throughput Gain.

These two simulation tests show also that the system throughput depends strongly on the randomness of configurations (i.e., the distribution of source and destination UEs). However, our control schemes always outperforms the D3MAC, as shown in Fig. 3.19, where the Average Throughput Gain, ΔTh $(\%)$, in average over 100 independent simulations, is always positive, whatever the traffic class and the number of active communications.

The analysis related to Fig. 3.18 suggests that our scheme may have higher degree of fairness than the benchmark scheme. At this regard, in Fig. 3.20, we report ΔJ_{index} under the traffic conditions previously considered. We note that under very light load conditions (traffic class 1), regardless of the number of active communications, $\Delta J_{index} = 0$, that is, J_{index} is about 1 for both control scheme. As the traffic class and the number of active communications increase, the probability of interfering sub-flows increases. So, the probability of many sub-flows competing for the same resources increases. In these conditions, ΔJ_{index} is always greater than zero, so our control scheme results more fair than D3MAC and the advantage is getting higher as the traffic load rises, up to a medium-heavy load conditions. Finally, in heavy traffic conditions, this difference comes down because the network is close to saturation.

3.8 Radio Network Planning

In this section, we propose a Radio Network Planning for the proposed D2D-enabled MMB system working in 60 GHz band. Since mmWave networks have very different characteristics compared to previous gen-

eration networks, we provide an analysis specific for the peculiarities of mmWave technology and the adopted Centralized Access Control scheme.

The planning procedure is organized in two steps: coverage planning and capacity planning.

3.8.1 Coverage Planning

In this section, we derive the coverage planning constraints based on our centralized mmWave Access Control, in terms of maximum distance between transmitter and receiver (UE and/or AP) and minimum transmission rate value for the related wireless link. These constraints are derived under the assumptions of not introducing any compensation technique for time slot synchronization (e.g., timing advance) and of transmitting at least one packet of 1000 bytes in a time slot.

The time to send a packet is [53]:

$$t_{packet} = t_{PHY} + t_{HEAD} + \delta_p + t_{p,PAY} \tag{3.27}$$

where t_{PHY} is the physical overhead time required for physical layer control protocols, the beamformer training, and signal acquisition/synchronization; $t_{p,PAY}$ is the payload time, which depends on the transmission rate R, $t_{HEAD} = 56\frac{8}{R}$ is the overhead to transmit the header fields of MAC, IP and UDP Protocol Data Units, $\delta_p = \frac{d}{c}$ is the propagation delay, which depends on the distance d and the speed of light c. Taking into account the above assumption, from (3.27), it follows:

$$t_{PHY} + 56\frac{8}{R} + \frac{d}{c} + 1000\frac{8}{R} \leq t_{SLOT} \tag{3.28}$$

where the transmission rate R in (3.2) depends on the SINR, that is a function of the received power ($P_R = P_T k(d)$) and the interference level ($\sum_j I_j$). The latter parameter can be estimated only after the scheduling rules have been set, whereas the received power depends on the transmitter power and the path loss, therefore on the distance between wireless devices, their visibility condition, and antenna gains. Let us note that the value of the first member of (3.28) increases with decreasing transmission rate (R) and increasing distance (d), while the second member is fixed.

Now, we derive the minimum value of R and the maximum value of d that satisfy the inequality (3.28), in the ideal assumptions of absence of interfering signals and no guard time (T_G). The main issue is that the path loss is a non-deterministic variable. In fact, the large-scale fading $X_{\sigma_{[dB]}}$ (i.e., the shadowing phenomenon) produces variations in the estimated path loss value, therefore R in (3.2) is a function of (d, σ). For this reason, we want to derive the guaranteed minimum values of transmission rate (R_g) with probability P ($P \in [0, 1]$), under different values of d and visibility conditions, in absence of interference, that is:

$$Pr\left(R(d, \sigma) \geq R_g\right) = P \tag{3.29}$$

Since $R(d, \sigma)$ in (3.2) is a monotonically decreasing function of path loss, Eq. (3.29) corresponds to:

$$Pr\left(PL_{[dB]}\left(d, \sigma_{[dB]}\right) \leq PL_{max,[dB]}\right) = P \tag{3.30}$$

Since shadowing $X_{\sigma_{[dB]}}$ has a normal distribution in the dB-domain, on the basis of (3.1), the path loss at distance d is a random variable characterized by the following probability density function (PDF):

$$f_{PL_{[dB]}(d)}(x) = \frac{1}{\sqrt{2\pi}\sigma_{[dB]}} e^{-\frac{\left(x - \overline{PL}_{[dB]}(d)\right)^2}{2\sigma^2_{[dB]}}} \tag{3.31}$$

where $\overline{PL}_{[dB]}(d)$ is the average value of the path loss at distance d (i.e., the path loss in (3.1) where $X_{\sigma_{[dB]}} = 0$), and $\sigma_{[dB]}$ is the standard deviation. By using (3.31), the condition (3.30) can be expressed as:

$$\int_{PL_{max,[dB]}}^{+\infty} f_{PL_{[dB]}(d)}(x)\, dx = 1 - P \tag{3.32}$$

From which, it follows:

$$1 - P = \frac{1}{2} erfc\left(\frac{PL_{max,[dB]} - \overline{PL}_{[dB]}(d)}{2}\right) \tag{3.33}$$

Finally, we obtain:

$$PL_{max,[dB]} = \overline{PL}_{[dB]}(d) + \sqrt{2}\sigma_{[dB]} erfc^{-1}[2\,(1 - P)] \tag{3.34}$$

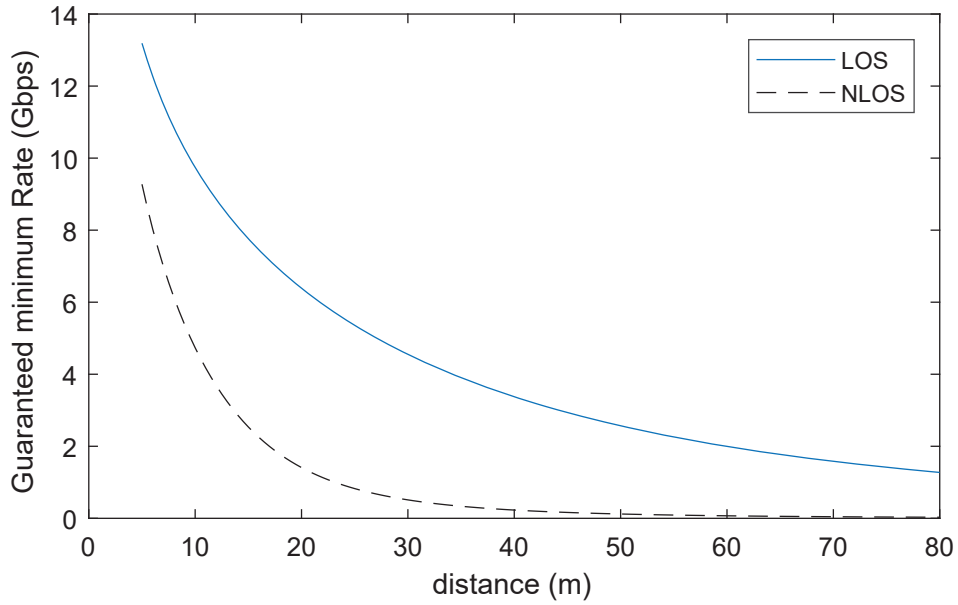


Figure 3.21: *Guaranteed minimum transmission values (with a fixed probability $P = 0.99$) under different distance values and visibility conditions, in absence of interference.*

This result means that, in order to guarantee R_g at distance d with probability P, we need to add to $\overline{PL}_{[dB]}(d)$ (i.e., the deterministic value of $PL_{[dB]}(d)$) a margin value M_{PL} depending on P and $\sigma_{[dB]}$:

$$M_{PL} = \sqrt{2}\sigma_{[dB]} erfc^{-1}[2\,(1-P)] \tag{3.35}$$

Under the assumption of guaranteeing the minimum transmission rate with a probability of 99%, by using the system parameters in Table 3.1, we get two margin values: one in LOS condition ($M_{PL,LOS} = 2.04718dB$) and one in NLOS condition ($M_{PL,NLOS} = 3.60584dB$). So, by using (3.34), in absence of interfering signals, we get R_g values, under different values of d and visibility conditions, as shown in Fig. 3.21

Taking into account this relationship between R_g and d, we can get the pair of values (R, d) that meets condition (3.28) when the first member is equal to $t_{SLOT} = 5\mu s$ (i.e., the limit values). More specifically, in NLOS condition we obtain a minimum transmission rate equal to $1.8086Gbps$ and a maximum distance of $23.49m$, whereas in LOS condition a minimum transmission rate of $1.8827Gbps$ and a maximum distance of $77.6m$. These pairs of values guarantee the slot synchronization in the ideal conditions of $\sum_j I_j = 0$ and $t_G = 0$.

Now, taking into account that the scheduling algorithm rules are derived with the aim of making the interference negligible (by enabling concurrent transmissions only if the interference level is tolerated), we can add a small margin to the coverage pre-planning constraints derived above in ideal conditions. More specifically, we set the minimum transmission rate equal to $R_{min} = 2Gbps$, in any visibility condition. This means that the payload time ($t_{p,PAY}$) of a time slot is equal to $4\mu s$, and the overhead time ($t_{OVER} = t_{PHY} + t_{HEAD} + \delta_p$) depends only on the distance d. Then, taking into account the relationship between R_g and d in Fig. 3.21, we can get the relative distance values in NLOS and in LOS condition. More specifically, in NLOS condition we obtain a maximum distance of $d_{max,NLOS} = 16.97m$, that is the more stringent condition, whereas in LOS condition a maximum distance $d_{max,LOS} = 59.93m$. Therefore, the overhead time in the worst case ($d = 59.93m$) is $t_{OVER,w} = 674ns$, and in a time slot we have a guard time equal to:

$$t_G = t_{SLOT} - t_{p,PAY} - t_{OVER,w} = 326ns \tag{3.36}$$

which is useful as implementation margin.

In conclusion, since in general UEs and APs could be in NLOS each other, in order to ensure the proper functioning of the system, APs must be placed in such a way as to guarantee a coverage radius of no more than $16.97m$. On the other hand, since APs should be strategically in LOS each other, the constraint of distance between two neighboring APs is less stringent, more specifically they may be distant no more than $59.93m$. These choices guarantee all overhead parameters with a proper guard interval t_G, without the introduction of any compensation techniques for time slot synchronization (e.g., timing advance).

3.8.2 Capacity Planning and Parameter Configuration

For the considered access control scheme, it is not possible to define a maximum capacity in terms of global throughput, because it depends heavily on the single analyzed configuration. More specifically, it may significantly vary not only on the basis of the network topology, the amount of UEs, and the number of active communications, but also with the variation of destination UEs and mostly with their location. In fact,

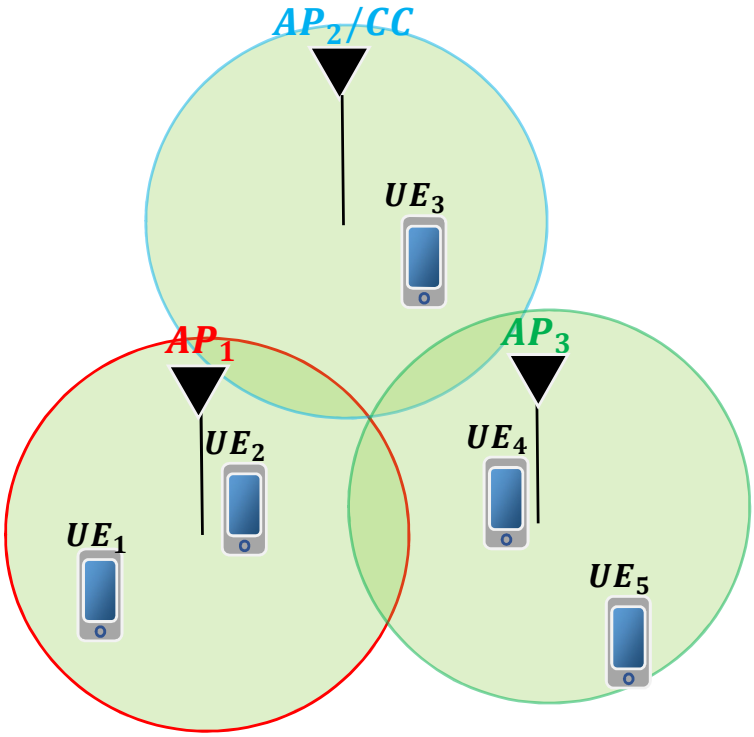


Figure 3.22: *Polling Scenario.*

concurrent transmission cannot occur if a UE is the same destination of different active communications or if the receiver UE is within the range of the main-lobe beamwidth of interfering transmitter devices.

For this reason, we only estimate the system capacity in terms of the maximum number of users that can be managed by the system ($N_{UE,max}$), in a fixed network topology that respects the coverage constraints. $N_{UE,max}$ is strongly related to the length of the control phase. In fact, let us note that the adopted variable-length frame structure is efficient as long as the control phase has a "reasonable" duration to guarantee delay-sensitive services and to improve fairness, in other words, its length needs attention to avoid an uncontrolled increase.

The objective of this section is to determine the appropriate length of the control phase to achieve good performance in ultra-dense scenarios. As described in Section 3.3.1, the control phase is composed of three time intervals: polling time (t_{poll}), scheduling time (t_{sched}), and pushing time (t_{push}). The time required to carry out the scheduling task depends on the number of requests to be scheduled and the controller computing power, whereas polling and pushing time are variable according to

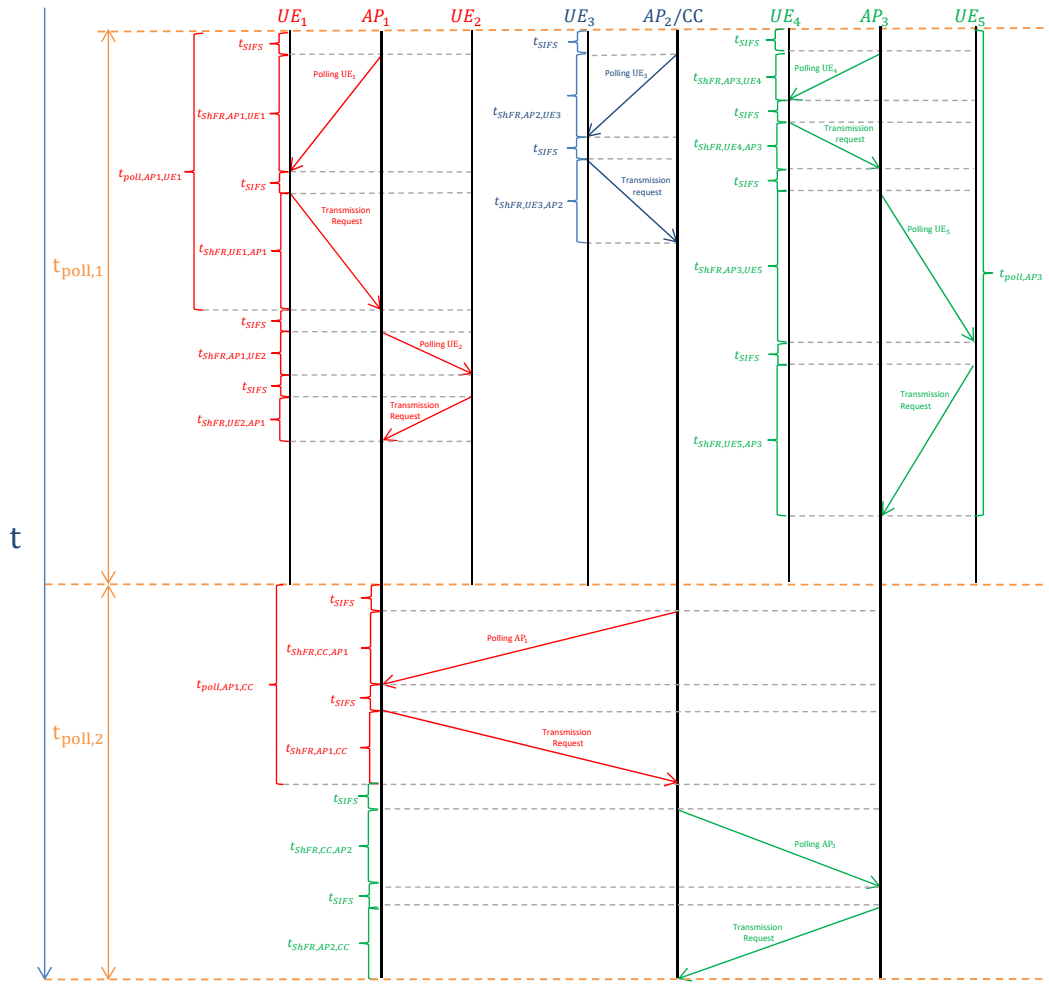


Figure 3.23: *Polling procedure.*

users' density, their distance from the nearest AP, and the area to be covered (i.e., the number of APs needing to communicate with the Central Controller). As for t_{push}, since the scheduling rules are sent only to source and destination UEs of the scheduled active communications, their amount is less than or at least equal to the total number of polled UEs inside the area, therefore:

$$t_{push} \leq t_{poll} \tag{3.37}$$

For these reasons, among the control phase time intervals, we focus on the polling time because it has the most stringent role in determining $N_{UE,max}$.

N_{UE} vs Polling Time

Referring to the simplified scenario in Fig. 3.22, we present an example to better illustrate the polling procedure (Fig. 3.23), in which there are 3 APs, 5 UEs, and the Central Controller is located inside the AP2.

The polling operation consists of two sub-steps: $t_{poll,1}$ and $t_{poll,2}$. During the first one, each AP sequentially polls all UEs inside its coverage area. We consider that the i-th AP sends a short frame message of length 14 byte [53] to the j-th UE. After a Short Interframe Space (SIFS), the j-th UE sends its transmission request to the AP through a short frame message. The time required to perform the described operations is:

$$t_{poll,AP_i,UE_j} = 2t_{ShFR} + 2t_{SIFS} \tag{3.38}$$

where t_{SIFS} is the SIFS interval, and t_{ShFR} is the time necessary to send a short frame message:

$$t_{ShFR} = t_{PHY} + 14\frac{8}{R} + \delta_p \tag{3.39}$$

Consequently, the time it takes the i-th AP to poll any attached UE is:

$$t_{poll,AP_i} = \sum_j t_{poll,AP_i,UE_j}, \forall UE_j \in AP_i \tag{3.40}$$

The above time depends on the number of UEs attached to AP_i and the relative distance, therefore it depends strongly on the distribution of UEs inside the scenario. Because this operation is executed at the same time by all APs, the overall time required to achieve this sub-step depends on the AP requesting the longest time. So, in order to poll all UEs inside the analyzed scenario, $t_{poll,1}$ must comply with the following condition:

$$t_{poll,1} \geq \max_i(t_{poll,AP_i}), \forall i = 1, 2, ..., N_{AP} \tag{3.41}$$

where N_{AP} is the number of APs in the area. In the example in Fig. 3.23, $t_{poll,1}$ must be greater or equal to t_{poll,AP_3}.

The second sub-step, instead, is related to the information exchange between all APs and the Central Controller. The time depends on the number of APs and the distances between APs and the Central Controller:

$$t_{poll,2} = \sum_{j=1}^{N_{AP}-1} t_{poll,AP_j,CC} \tag{3.42}$$

where $t_{poll,AP_j,CC}$ is the time required for the polling operation from the Central Controller to the j-th AP:

$$t_{poll,AP_j,CC} = 2t_{ShFR} + 2t_{SIFS} \tag{3.43}$$

Unlike the first sub-step, this time is fixed once all APs are positioned. In summary, the overall polling operation lasts:

$$t_{poll} = t_{poll,1} + t_{poll,2} \tag{3.44}$$

Fig. 3.23 shows clearly that $t_{poll,2}$ is fixed, whereas $t_{poll,1}$ depends strongly on the amount of UEs and their distances from the serving APs, therefore on the coverage area of the APs. This analysis allows us to derive a relationship between N_{UE} and t_{poll}. We define $t_{poll,AP,UEw}$ as the time it takes an AP to poll one UE in the worst condition, that is, UE is at the maximum distance from the AP and the interference level is severe. So, in the worst case, i.e., at least one AP has all the related UEs in the worst condition, by using (3.41), we obtain:

$$t_{poll,1,w} = N_{UEw} t_{poll,AP,UEw} \tag{3.45}$$

Finally, we derive the maximum number of UEs that an AP can poll, as follows:

$$N_{UEw,max} = \left[\frac{N_{slot,poll} t_{SLOT} - t_{poll,2}}{t_{poll,AP,UEw}} \right] \tag{3.46}$$

where $N_{slot,poll}$ is the length of t_{poll} in terms of number of time slots.

3.8.3 A Case Study

Starting from the above results, with the following analysis, we want to set up the control phase parameters in an ultra-dense scenario respecting the coverage constraints derived in Section 3.8.1. We aim to evaluate if the proposed access control can have a concrete implementation, in other words, if the wait to transmit is reasonable.

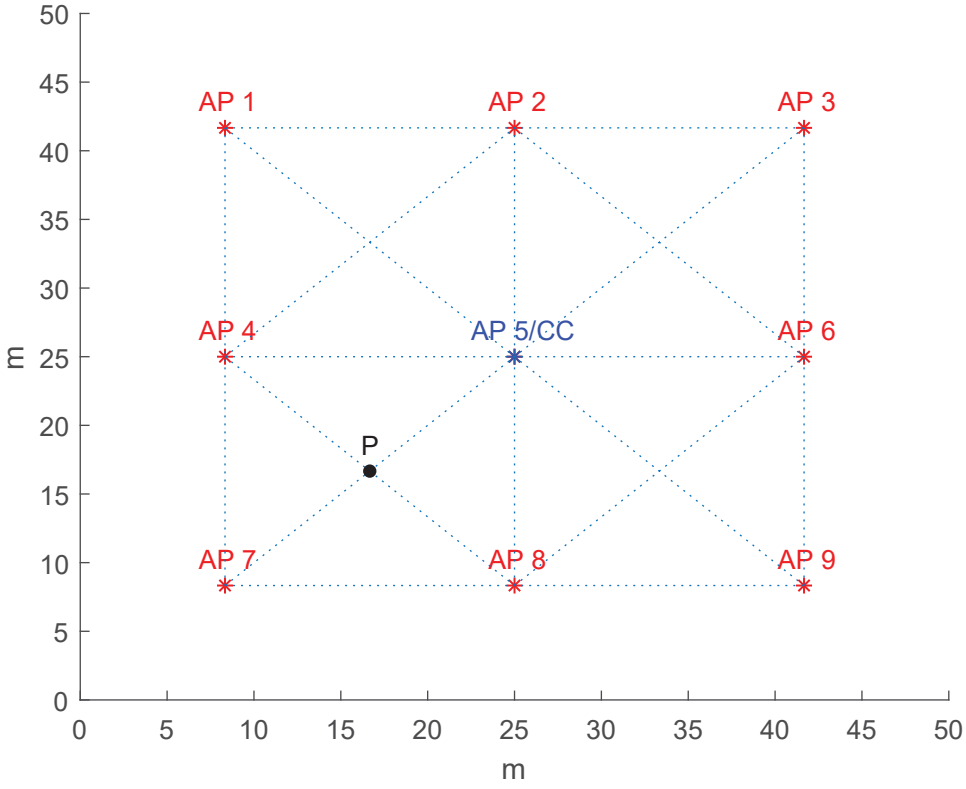


Figure 3.24: *Analyzed Scenario.*

As in [18, 48], we consider to cover a flat indoor area of $50m$ x $50m$ and we assume that the area is covered with 9 Access Points distributed to form a regular grid, as shown in Fig. 3.24. We assume that direct communication between UEs and communication between UE and AP occur in NLOS conditions. In this network topology, the maximum distance between UE and AP is $11.8m$, therefore it respects the constraint of $16.97m$ in NLOS condition. The Central Controller is located inside the Access Point at the center of the scenario (AP5), so that four APs are at $23.57m$ from the controller and the other four ones are at $17.5m$. We assume that APs are strategically placed in LOS condition between each other, therefore the constraint of $59.93m$ in LOS is widely respected.

In order to evaluate $N_{UEw,max}$ in this Case Study, by using (3.46), we need to evaluate $t_{poll,AP,UEw}$ and $t_{poll,2}$. We begin our analysis with $t_{poll,AP,UEw}$. In Fig. 3.2 the worst condition is represented by point P, in the event that the main antenna beam of all APs is directed towards point P. A UE placed on point P is about $11.8m$ distant from its serving AP (e.g., AP7), and suffers interference from AP4, AP5, and AP6 (in-

terference from the other APs can be neglected, thanks to the huge path loss). Then, the transmission rate of the wireless link between AP7 and UEw is evaluated by using (3.2) and (3.34) in NLOS condition. We obtain $R_{AP7,UEw} = 620.55Mbps$ and $t_{ShFR} = 0.4698\mu s$. Consequently, the time required for the polling operation from AP7 to one UEw is $t_{poll,AP,UEw} = 1.139\mu s$. Let us note that $R_{AP,UEw}$ is less than the R_{min} value, derived in Section 3.8.1, but this constraint must be respected only during the transmission phase, since polling packets (short frame messages) are not of length 1000 byte. The fulfillment of the constraint on the minimum rate during the transmission phase is guaranteed by the scheduling procedure.

As for $t_{poll,2}$, in the considered scenario, there are 8 APs having to communicate with the central one (AP5). The transmission rates are estimated like above: we obtain $5.6363Gbps$ for APs at $23.57m$ (AP1, AP3, AP7, AP9), and $7.0198Gbps$ for the other ones at $17.5m$. Consequently, by using (3.42), the time necessary for the second sub-step is $t_{poll,2} = 6.98\mu s$. Finally, by using (3.46), we derive the relationship between $N_{slot,poll}$ and $N_{UEw,max}$ in the analyzed scenario. Fig. 3.25 shows the amount of time slots necessary for the polling time under different number of UEs in the worst condition.

For example, if we consider a polling time equal to 3 time slots (i.e., $N_{slot,poll} = 3$), then this parameter configuration allows the system to manage at most 7 UEs per AP, each one in the worst condition. So, it is possible to poll 63 UEs in the whole area, which corresponds to a UE density of $25200UEs/km^2$. This value almost doubles the maximum UE density considered in [48], that is $16000UEs/km^2$, in a scenario similar to the one analyzed here. For this reason, in [18] we chose a polling time of 3 time slots, which we think of as high enough to ensure the proper system functioning.

To complete the analysis of this Case Study, with the aim of assessing performance of our access control scheme by simulations, in [18] we fixed the pushing time equal to 3 time slots, in order to fulfill also the condition (3.37) in the worst case. As for the scheduling time, because it depends strongly on the computing power of the control unit, we assumed a computational unit that typically performs the scheduling operations in at least $20\mu s$ (i.e., 4 time slots).

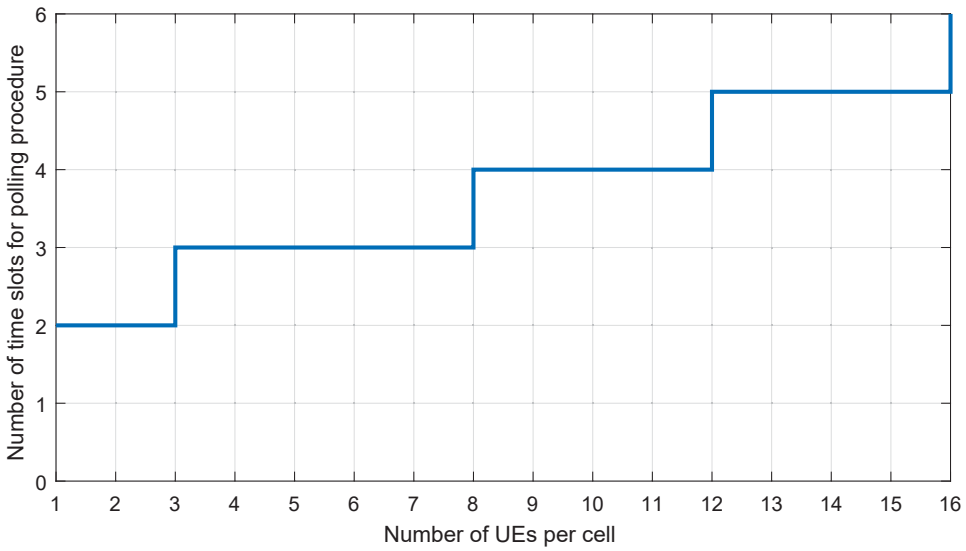


Figure 3.25: *Number of time slots required for polling procedure under different amount of UEs in the worst condition.*

In conclusion, we estimated the control phase equal to 10 time slots (i.e., $50\mu s$) on average; this is a reasonable value that allows the system to manage all UEs in the area with the expected UE density, and has provided optimal performance in terms of throughput and end-to-end delay, under different traffic loads, as shown in [18].

3.9 Conclusion

In this chapter, we analyzed the symbiosis of mmWave transmission, D2D communications, beamforming technique, the adoption of the 60 GHz unlicensed band. We proposed a new TDMA-based centralized access control for 60 GHz D2D communications in a high UE density indoor scenario, covered by some Access Points interconnected through mmWave links. In this scenario, in addition to the interference between D2D communications themselves, the interference between the mmWave access network and the mmWave backhaul network arises. With the aim of managing the above interferences, enhancing transmission efficiency and achieving high performance in terms of throughput and end-to-end delay, a new access strategy and a scheduling algorithm were defined. In order to meet the above targets, multiple aspects affect the best decision to be made. For this reason, our scheduling algorithm is composed of various criteria grouped in three phases. In the first phase, three criteria

are introduced to determine the interference relations among sub-flows. In the second phase, concurrent transmission rules are determined by using two criteria based on multi graph-coloring techniques. Finally, a last criterion is defined to establish the minimum number of time slots to be allocated to each stage of the transmission phase. We benchmarked our centralized control scheme against the D3MAC scheme [48], one of the most complete works available in literature. Extensive simulations, under various traffic classes and number of active communications, demonstrate that our access scheme has notably improved the concurrent transmission efficiency and outperforms the considered reference scheme in terms of throughput, end-to-end packet delay, and fairness.

Furthermore, we proposed a Radio Network Planning for the proposed architecture organized in two steps. In the first one, under the assumptions of not introducing any compensation technique for time slot synchronization, we derived the coverage planning constraints, in terms of maximum distance between transmitter and receiver in any visibility condition. In the second phase, the capacity planning was derived, in terms of the maximum number of users that can be managed by the system with good performance.

CHAPTER 4

Dynamic Uplink Radio Resource Dimensioning and NOMA-based PUSCH resource allocation for mMTC scenarios

4.1 Introduction

In this chapter, we address the radio resource allocation problem for mMTC scenarios. As reported in the Introduction chapter, this scenario is characterized by a large number of low-complexity and energy-constrained MTC devices sending periodically very short packets with relaxed delay requirement, without or with very little human interventions. Examples include smart grids, environment monitoring, industry automation, home automation, and so on. In specific, mMTC usage scenario requires more than 1 million connections within 1 square kilometer [58, 59], while traditional 4G mobile networks support up to several thousand connections, which often limits their use in mobile phones, computers and similar smart devices [60].

In the mMTC scenario, because of the relaxed delay requirements, a contention-based Random Access (RA) procedure can be adopted. However, a huge number of MTC devices might simultaneously initiate their procedure to get access to the network causing a severe congestion prob-

lem in the system if conventional LTE RA procedure is adopted [61,62]. Currently, in LTE / LTE-Advanced (LTE-A) uplink communication, the MTC device firstly performs a contention-based RA procedure by transmitting a preamble in the Physical Random Access Channel (PRACH). Then, if the RA attempt has been successful and there are available radio resources into the Physical Uplink Shared Channel (PUSCH), the MTC device transmits its data in it. Otherwise, the MTC device re-attempts the access procedure in successive RA cycles for a maximum number of times.

When thousand or more devices simultaneously request to send data, massive RA attempts increase the preamble collision probability, thus decreasing the number of successful accesses. This problem could be alleviated by increasing the resources allocated to PRACH. However, due to limited uplink resources, the higher the amount of resources available in the PRACH, the lower the ones available in the PUSCH. Therefore, many MTC devices which have successfully completed their RA attempt, could not find enough transmission resources in the PUSCH. One way to increase the transmission efficiency in the PUSCH resources is the adoption of a NOMA technique. In fact, by adopting a NOMA technique, more than one MTC device can share the same time-frequency resource by mainly exploiting the power-domain (e.g., PD-NOMA), or the code-domain (e.g., Sparse Code Multiple Access, SCMA), or both the power and the code domains (e.g., Power Domain Sparse Code Multiple Access, PSMA [63]). In the traditional PD-NOMA, two or more devices with large channel gain difference (e.g., large path-loss difference) are normally paired in the same radio resource. At the receiver side, the successive interference cancellation (SIC) algorithm is utilized to detect the desired signals. The basic idea of SIC is that different signals are successively decoded. After one user's signal is decoded, it is subtracted from the combined signal before the next user's signal is decoded. The main drawback is that, in order for the algorithm to work properly at the receiver side, a proper uplink power allocation control based on the knowledge of the channel conditions needs to be applied [64]. As regards the SCMA, two or more MTC devices are superposed to the same radio resource by adopting sparse multidimensional codewords. At the receiver side, the Message Passing Algorithm (MPA) is applied to detect

the different transmitted signals in an iterative manner based on Maximum Likelihood (ML) algorithms. The main drawback is high computational burden of the MPA receiver, especially when the number of multiplexed communications is large. However, in the Uplink case considered here, the heavy computation load of the MPA is exclusively on the base station, while the encoding procedures at the mobile stations are not computationally intensive [65]. As regards the PSMA receiver, although it permits to increase the spectral efficiency about 50% compared with the previous ones, it inherits not only the drawbacks from both the SCMA and PD-NOMA receivers but also it increases the system complexity of about one order of magnitude with respect to the SCMA one. In this chapter, we consider a scenario with low-complexity MTC devices that do not forward the measurement reports to the gNB. So, PD-NOMA and PSMA techniques cannot be implemented at no cost, in terms of signaling overhead, complexity and energy consumption. For these reasons, SCMA is one of the most promising techniques to support the data transmission from a large amount of low-complexity MTC devices, so we adopt it in the PUSCH. In this chapter, we analyzed also the feasibility of our SCMA-based PUSCH resource allocation framework by considering both the complexity and the signaling overhead points of view.

However, adopting the SCMA technique in the PUSCH is not enough to achieve good performance in an mMTC scenario. Another point is to define an innovative grant-based RA procedure tailored for mMTC scenario. In this context, the academic community is studying and proposing new optimized RA procedures at the aim of avoiding the RRC connection setup overhead, e.g., the 2-phase connectionless RA procedure described in [66, 67]. By adopting this procedure, immediately after the reception of the RAR message (Phase 2), the MTC device transmits its data packet. However, the main disadvantage is that the device sends the data packet regardless of successful or unsuccessful access attempt, wasting energy in the latter case. In parallel, other works [14, 68–70] followed a different approach, modifying the first phase of the traditional 4-phase RA procedure to carry additional information to the gNB. In [68] the authors suggest a new access scheme to simultaneously transmit both preambles and a small-sized message in the PRACH as a message em-

bedded preamble sequence by using distinct root numbers. However, we found the innovative approach proposed in [69, 70] very interesting. These works are based on the usage of tagged preamble sequences, i.e., each device sends the sum of a randomly selected preamble sequence with root r and a randomly selected tag sequence with root $k \neq r$. In [69] the authors adopt the tagged preamble sequences to capture multiple Timing Advance (TA) values for a single detected preamble and propose a PUSCH resource allocation method based on the multiple TA values captured. In [70], and also in [14], these tag sequences allow the gNB to detect the preamble collision during the first phase. In this way, the device reduces the number of signaling transmissions per failed access attempt to one. Specifically, in [70] the authors still adopt the conventional 4-phase RA procedure, while in this chapter the strategy of transmitting tagged preambles is also used to apply a 2-step RRC connectionless RA procedure, achieving a further reduction of the energy consumption.

Despite the numerous benefits, the advanced SCMA-based PUSCH resource allocation and the 2-step RRC connectionless RA procedure is not enough to support a massive MTC scenario.

In literature, several works (e.g., [71–73]) take into account only the limit of PRACH resources and define new RA procedures to decrease the preamble collision probability. These works are based on Access Class Barring (ACB) approaches [13], i.e., congestion control schemes designed for limiting the number of simultaneous access attempts, thus reducing the number of RA failed attempts. References [71,72] are based on the conventional ACB scheme, where the Base Station (BS) periodically broadcasts an ACB factor $p \in [0, 1]$ to the MTC devices. Every MTC device having data to be transmitted draws a uniform random number $q \in [0, 1]$. If $q \leq p$, then the MTC device becomes eligible to make the contention-based RA procedure. Otherwise, it is barred for a time period. The main challenge of these ACB schemes is to adapt the ACB factor p according to the traffic load in the PRACH. In fact, p should be a small value in the case of bursty and heavy-loaded scenario in order to relieve congestion, while it should be a large value in the case of light traffic condition in order to efficiently use the uplink resources and do not delay inappropriately the access attempts. In [71] the authors pro-

pose a dynamic adaptation of the ACB factor p based on the estimation of the MTC devices that are in backoff, i.e., those which will re-attempt their RA procedure. In [72] the authors design a Q-learning algorithm to dynamically tune the ACB factor p such that it can rapidly react to the traffic changes using local information available at the BS. In [73], a different fully distributed ACB scheme based on the non-cooperative game theory is proposed. Rather than being based on the p value sent in broadcast by the BS, in that access scheme, each MTC device calculates its own activation probability based on the acknowledgments sent by the BS. Despite these works aiming to maximize the number of successes in the PRACH, they do not consider the opportunity to dynamically dimensioning the uplink resources of the PRACH and the impact of the limited resources of the PUSCH. In [74] the authors, taking into account that the uplink resources are limited, propose a new control scheme which, before a RA cycle begins, allocates radio resources between PRACH and PUSCH, and broadcasts this configuration to all MTC devices. The main weakness of [74] is that the number of MTC devices attempting to access is considered well-known at the BS.

In this chapter, starting from [74], we address on the issue to find a good trade-off between the amount of radio access resources allocated to the PRACH and the ones needed for data transmission in the PUSCH. Specifically, we show that an optimal PRACH and PUSCH resource allocation should be based on current traffic. For this reason, we propose a dynamic load-aware control scheme based on the access attempt number, termed as Dynamic Uplink Resource Dimensioning (DURD) By simulations in MATLAB environment, the performance of our dynamic control has been compared with static dimensioning. The results show that our control achieves the best performance in terms of succeeded communications, while guaranteeing lower energy consumption.

Moreover, in light of the performance achieved, we applied to the dynamic uplink resource dimensioning the innovative idea to exploit the unused PUSCH resources, if any, to serve also some MTC devices that have failed their access attempt. This system is termed as Enhanced Dynamic Uplink Resource Dimensioning (EDURD). More specifically, for a given RA cycle j, instead of limiting the number of attempting MTC devices by means of an ACB scheme, our idea was to serve a larger num-

ber of attempting MTC devices in j. This approach mitigates the congestion since it reduces the amount of re-attempting devices in the successive RA cycles. In particular, for a subset of collided preambles, the scheduler allocates a pool of resources to each of them in a contention-based mode. This resource allocation in the PUSCH does not replace the conventional RA procedure in the PRACH, but it is an additional operation for increasing the number of succeeded communications, if possible. The main drawback of this approach is that contention-based transmission in the PUSCH may collide causing vain additional energy consumption. To attenuate this issue, we determined the optimal size of the subset of collided preambles that maximizes the number of successes in the PUSCH, thus reducing the number of collisions in it. At this aim, we present an analytic problem formulation and, accordingly, we proposed an iterative algorithm, called PUSCH Resource Reallocation Algorithm (PRRA), to implement our strategy. The proposed PRRA was applied assuming both a static and dynamic uplink PRACH/PUSCH dimensioning. The main drawback of these proposed load-aware dynamic uplink resource dimensioning schemes is that they require as input the number of attempting MTC device at each RA cycle that is an information not available at the gNB. To make the solution viable, we introduce a predictive formula for estimating the expected traffic based only on information available at the gNB : the number of succeeded preambles, the number of collided preambles, and the number of available preambles.

4.2 Background

4.2.1 LTE Uplink

Fig. 4.1 shows a typical RA cycle of 5 ms[1], where Uplink radio resources are divided into a PRACH subset and a PUSCH subset.

The PRACH is 1.08 MHz wide in frequency and has a time duration which depends on the RA Preamble Format (e.g., by using the common Preamble Format 0, the PRACH is 1 ms in time). The PRACH access resources consist of 64 orthogonal preamble sequences which are

[1]In LTE systems different RA cycle lengths has been standardized based on the PRACH Configuration Index [75]. In particular, the typical PRACH Configuration Index ranges from 6 to 8, corresponding to RA cycle of 5ms [76, 77].

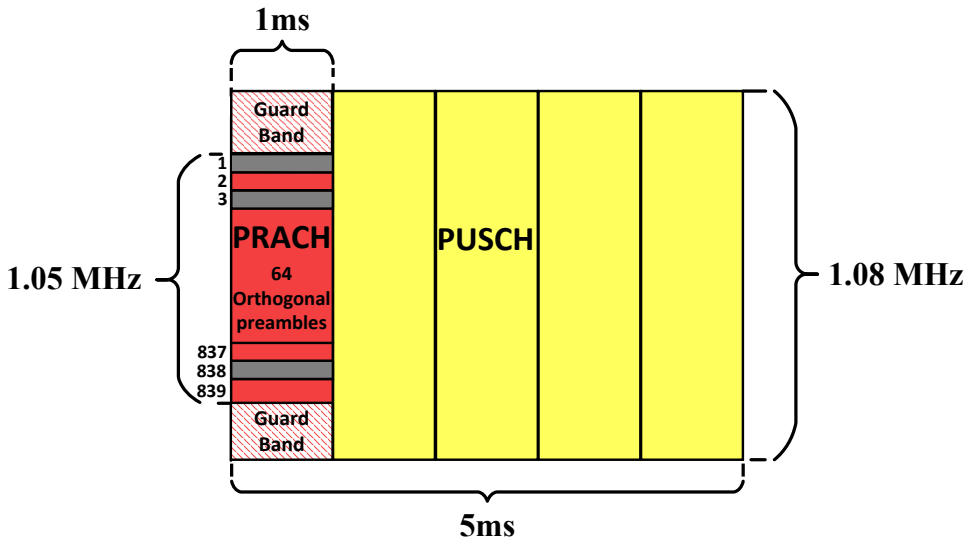


Figure 4.1: *A typical RA cycle of 5ms.*

mapped to 839 subcarriers of 1.25 kHz, and are generated from a reference Zadoff-Chu (ZC) sequence. These sequences are analyzed in detail in the next chapter. The definition is reported in (5.1). The PUSCH consists of 72 sub-carriers of 15 kHz, is used to transmit user data and occupies the remaining available radio resources.

In the following, we describe the conventional contention-based RA procedure in the LTE/LTE-A system and show it in Fig. 4.2, in case the communication is successful on the second attempt. Before initiating their RA procedures, the devices receive periodically, from the Evolved Node B (eNode B), the System Information Block (SIB) which contains, inter alia, broadcast information related to PRACH structure, the Backoff Window (B_W) size for random backoff procedure, and the maximum number of access attempts (M_A). The legacy RA procedure involves the following four-step message handshake between each device and the eNodeB.

Step 1. Preamble transmission. Each device randomly selects a preamble sequence out of the ones available for the contention-based procedure with equal probability and transmits it on the PRACH. Obviously, there is a non-zero preamble collision probability, since the same preamble can be selected by more than one device. Once the preamble transmission has been completed, the device increases its counter N_A by one.

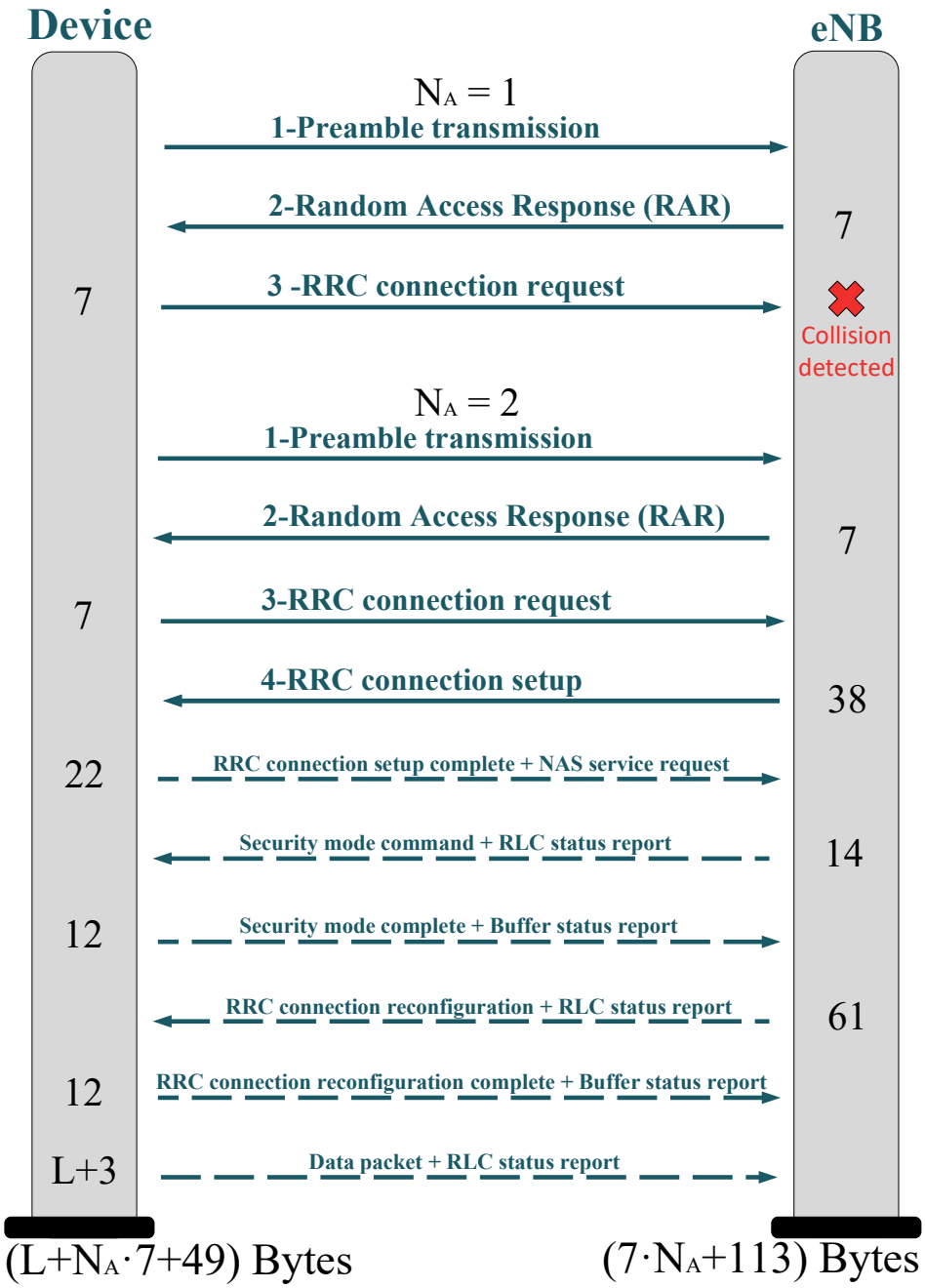


Figure 4.2: *Conventional 4-step RA procedure, when the communication is successful on the second attempt, and signaling messages for data packet transmission in LTE/LTE-A and eMTC technologies.*

Step 2. RAR message. After detecting the received preambles, the eNodeB can only detect whether a specific preamble has been transmitted or not, but it cannot recognize how many devices have transmitted it, i.e., if a collision has occurred. For each detected preamble, the eN-

odeB transmits the Random Access Response (RAR), which contains the Timing advance command, the temporary C-RNTI and the UL Grant. In particular, the latter field is of 27 bits, 18 of which are reserved for time and frequency resource allocation of the message transmission in Step 3 [78].

Step 3. RRC connection request transmission. After the RAR reception, each MTC device transmits on the scheduled radio channel the Radio Resource Control (RRC) connection request to set up the RRC connection with the eNodeB. However, if more than one UE has selected the same preamble in Step 1, they will receive the same RAR and send their scheduled messages on the same radio channel, which makes the eNodeB hard to decode the received message correctly. In this case, the eNodeB recognizes a preamble collision.

Step 4. RRC connection setup reception. After correctly decoding the RRC connection requests, if there are available resources in the PUSCH for the device transmission, the eNodeB transmits the RRC connection setup message to the corresponding device.

We note that if the device does not receive either the RAR message or the RRC connection setup within the related predetermined time window, termed W_{RAR} and W_{CR}, respectively, then it reattempts the RA procedure inside the backoff window (B_W) only if $N_A \leq M_A$.

If the 4-step RA procedure was successful, further signaling messages have to be exchanged between the UE and the eNB in order for the UE to initiate the transmission of the data packet. In detail, the device sends the RRC connection setup complete message and initializes the Non-Access Stratum (NAS) procedures to the Mobility Management Entity (MME), including security. Then, the UE will receive the RRC connection reconfiguration message, which is used to establish the data radio bearer. Finally, it transmits the data packet.

4.2.2 eMTC and NB-IoT

Since the LTE is highly inefficient for supporting the MTC traffic characterized by sporadic infrequent transmission of small packets, the 3GPP has already introduced, in Release 13, a suite of two complementary technologies adapted for this type of traffic, denoted as enhanced Machine Type Communication (eMTC) and NarrowBand IoT (NB-IoT) [15].

Both technologies are optimized for granting lower complexity, and providing longer battery life, while seamlessly coexisting with other LTE services. As regards some technical specifications for the eMTC, it adopts 1.4 MHz receiver bandwidth, a reduced maximum transmission power (20 dBm), and extended discontinuous reception (eDRX) modes. However, the RA procedure and the data transmission are completely inherited from the conventional LTE/LTE-A shown in Fig. 4.2.

As regards the NB-IoT technology, it further pursues the goals of providing a cost-effective solution. In fact, some relevant NB-IoT features are: 200 kHz receiver bandwidth, half-duplex operation, new narrowband (NB) physical channels for downlink and uplink, and improved power saving modes. In addition, the 4-step RA procedure has been inherited from the LTE/LTE-A, while the data packet transmission has been deeply modified to support small-sized data in a more efficient way. In particular, a Cellular IoT (CIoT) EPS optimization in the Control Plane (CP) has been designed to allow a piggyback data packet into the NAS service request, which is sent together with the RRC connection setup complete message. Thus, the NB-IoT CP CIoT EPS optimization allows the UE to transmit the data packet immediately after the 4-step RA procedure [79].

We emphasize that both the aforementioned RA procedures and data transmissions are connection-oriented. Therefore, although NB-IoT proposes a re-engineering of the conventional RRC procedures, it suffers from the RRC connection setup overhead.

4.3 System Model

As reported in the 5G implementation guidelines provided by GSMA in [80], 5G networks can be deployed in different deployment options, where StandAlone (SA) options consist of only one generation of radio access technology (i.e., LTE or NR), and Non-StandAlone (NSA) options consist of NR radio cells combined with LTE radio cells using dual connectivity to provide radio access. In this chapter, we consider a transmission scenario based on SA Option 2, where the radio access network consists of only one Next Generation NodeB (gNB) connected to the 5G Core and N_{MTC} MTC devices. Each MTC device performs a contention-

based access procedure to request resources for data transmission. We do not adopt a NSA option because the dual connectivity technology is not suitable for supporting low-complexity and energy-constrained MTC devices. Moreover, among the SA options, we choose option 2 because it is the only one that adopts the NR radio access, and therefore it is able to fully support the mMTC scenario.

As regards the radio interface, we adopt the smallest 5G NR numerology (i.e., subcarrier spacing of 15 kHz, which corresponds to $T_s = 1$ ms), because small subcarrier spacing results in longer symbol duration and lower overhead. Therefore, delay-tolerant services, such as MTC services, can benefit from small subcarrier spacing to reduce bandwidth consumption. In addition, we assume that each RA cycle occupies the same amount of uplink resources of LTE, i.e., 72 subcarriers of 15 kHz and a fixed time length, $T_{ra} = 5$ time slots (see Fig. 4.1). Also, we adopt Preamble Format 0 (i.e., a preamble lasts 1 T_s). In the frequency domain, PRACH and PUSCH occupy the entire considered bandwidth, while in the time domain PRACH lasts $T_{pr} \in \{1, 2, \ldots, T_{ra} - 1\}$ time-slot units and PUSCH T_{pu} slots. In summary, for each RA cycle we have:

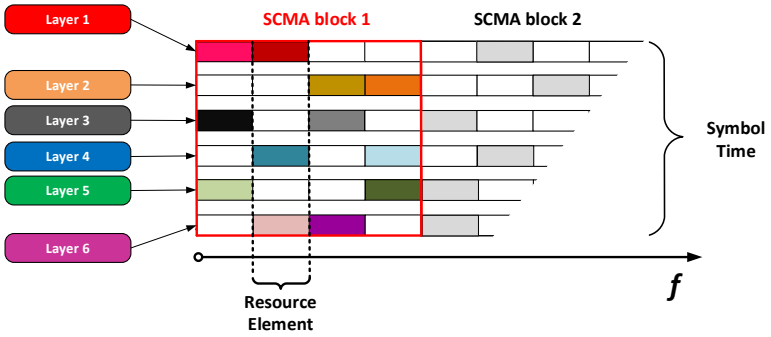
$$T_{ra} = T_{pr} + T_{pu}. \tag{4.1}$$

We consider only L_0 out of 64 preambles available for the RA contention-based procedure in 1 T_s. So, the total number of preamble sequences (L) in a RA cycle is $L = L_0 T_{pr}$.

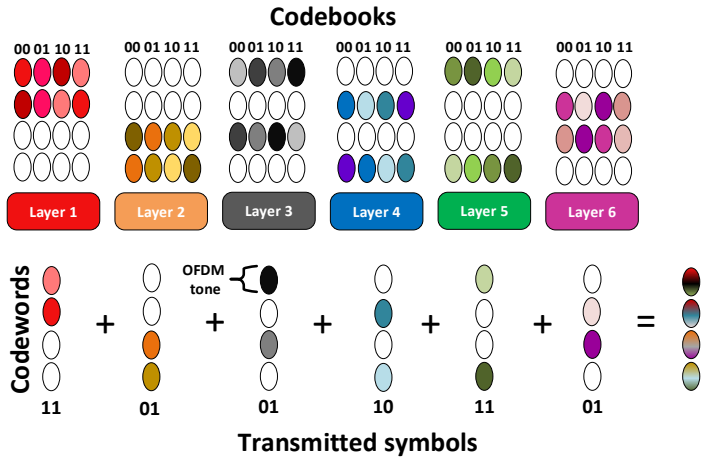
4.3.1 SCMA Overview

To further increase the transmission efficiency, we adopt the SCMA technique for PUSCH resources. The SCMA technique, defined in [81], can be regarded as a code division Non-Orthogonal Multiple Access (NOMA) scheme, i.e., it allows multiple transmissions on the same Resource Element (RE), as shown in Fig. 4.3a. Multiple SCMA layers are defined, and one or more layer can be assigned to a device or data stream. In Fig. 4.3a, on the first RE, the first part of each transmitted symbol belonging to Layers 1, 3, and 5 are superposed.

The superposition of different levels in a single RE is allowed by the use of different codebooks, as shown in Fig. 4.3b, which are built



(a) *Multiple Access in one SCMA block.*



(b) *Bit-to-codeword mapping.*

Figure 4.3: *Example of multiple access and bit-to-codeword mapping of an SCMA encoder with* $Q = 4$, $S = 2$, $K_{max} = 3$, $L_{SB} = 6$, $I = 4$.

on multidimensional constellations, instead of the conventional linear spreading, typical of the traditional Code Division Multiple Access (CDMA). In addition, SCMA uses sparse spreading to reduce the number of symbol collisions. For instance, in Fig. 4.3a the number of superposed layers in a RE is 3 instead of 6 (traditional case with non-sparse spreading). The superposition pattern per RE is typically statically configured and indicated by the factor graph $\mathbf{F}$, which is a binary matrix of size $Q \times L_{SB}$, whose element $f_{i,j} = 1$ if and only if Layer j transmits inside RE i. For instance, the factor graph $\mathbf{F}$ in Fig. 4.3a is:

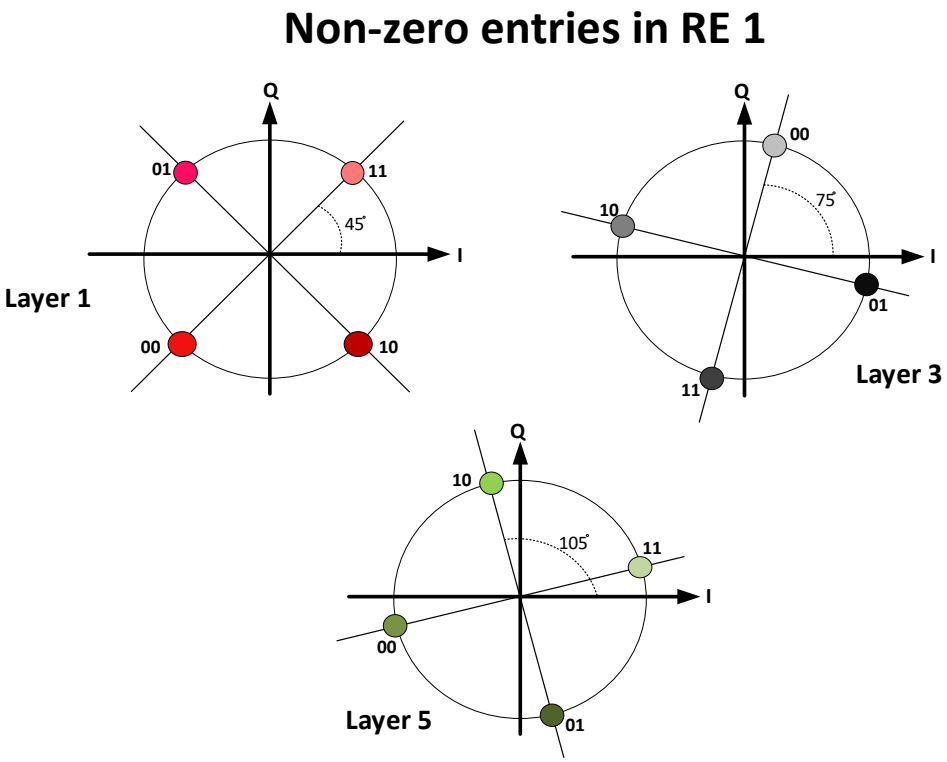


Figure 4.4: *SCMA Codebooks. Example of the first dimension of the codebook in RE1.*

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}. \tag{4.2}$$

According to SCMA encoder, one $SCMA_{block}$ is the minimum resource quantity which can be shared by different layers. Inside a single $SCMA_{block}$, a layer can be associated to the transmission of one symbol. One $SCMA_{block}$ occupies Q subcarriers in one symbol time, so Q is the number of REs in one $SCMA_{block}$. Also, we denote with S the number of REs which a layer occupies respect to Q, and K_{max} the maximum number of overlapped layers with different codebooks in one RE. So, the number of different layers per $SCMA_{block}$ is:

$$L_{SB} = \binom{Q}{S}, \tag{4.3}$$

and

$$K_{max} = \binom{Q-1}{S-1} = \frac{L_{SB}S}{Q}. \tag{4.4}$$

The gNB receives, over each Resource Element, the sum of the symbols sent by K_{max} MTC devices, each one by using the codebook assigned to it. To detect the signals transmitted by the MTC devices, the iterative Message Passing Algorithm (MPA) is performed [63]. The complexity order of this method is given by $O(I^{K_{max}})$ per RE for each iteration, where I is the number of constellation points [82].

Let us analyze an example of an SCMA encoder in Fig. 4.3a, where $Q = 4$, $S = 2$, $L_{SB} = 6$, and $K_{max} = 3$. In order for the receiver to can decode the superposed data, in transmission at least K_{max} different constellations must be adopted. Each constellation should be S-dimensional and each dimension should be composed of I points. Several constellations are available in literature [83, 84] which typically are designed on basis on a multi-dimensional mother constellation with a good Euclidean distance profile. The mother constellation is then rotated to achieve a reasonable product distance. A simple example is shown in Fig. 4.4, where we show a constellation adopted for the first non-zero entry (i.e., the first dimension of the constellation) of Layers 1, 3, and 5, which are superposed in the same RE 1. The three constellations are generated from the mother constellation QPSK by applying the phase rotations $\left\{0, \frac{\pi}{6}, \frac{\pi}{3}\right\}$, as reported in [81]. Then, all symbols transmitted in the same RE do not match each other. Also, we note that, in this example, the same information bits are retransmitted in a second RE ($S = 2$) to reduce the bit error probability at the receiver side.

Since the PUSCH is allocated over 72 subcarriers and each time slot contains 14 OFDM symbols per subcarrier, the number of $SCMA_{block}$ per time slot is:

$$N_{SB} = \left\lfloor \frac{72}{Q} \cdot 14 \right\rfloor. \tag{4.5}$$

Finally, the number of layers in one RA cycle, L_{RAC} is:

$$L_{RAC} = N_{SB} L_{SB} T_{pu}. \tag{4.6}$$

4.3.2 The proposed basic two-step RA procedure

In parallel to both the eMTC and NB-IoT technologies analyzed in Section 4.2.2, that adopt a connection-oriented approach, connectionless transmission protocols for machine-type traffic have been proposed in literature at the aim of avoiding the RRC connection setup overhead. These protocols allow the MTC device to transmit small packets without the establishment of radio bearers [66,67,85]. So, the signaling message exchange between the network and the MTC device that is used to set up the RRC connection and to establish device and security contexts is deleted, implying a device battery power saving.

Among the available proposals, we consider the 2-step connectionless packet transmission procedure described in [67]. By adopting this procedure, the MTC device transmits, immediately after the reception of the RAR message (Step 2), its data packet together with an UL context containing all necessary information related to the device identity, PDN-ID, and security. Thus, once the gNB receives the data packet, it forwards the packet to a connectionless access gateway, which inspects the context header, verifies integrity, performs decryption, and, based on stored state information, forwards the packet to the expected network entity. The main drawback is that the collision detection occurs after the packet has been transmitted, i.e., regardless of whether the access attempt has been successful or not, the device sends the data packet. Taking into account this procedure, we propose a new improved connectionless 2-step RA procedure and data transmission shown in Fig. 4.5. Like in [67], the MTC device transmits, after Step 2, the data packet piggybacked with the UL context. Conversely, at the aim of overcoming the issue of sending the data packed regardless of whether the access attempt has been successful or not, we adopt the early Preamble Collision Detection (e-PACD) technique, proposed in [70], where the gNB can detect in Step 1 whether a preamble has been affected by collision or not. In detail, each device randomly selects one preamble among those available for contention-based procedure and transmits a tagged preamble, consisting of both the selected preamble and a tag sequence. By exploiting the received tagged preambles, the gNB can detect, for each received preamble, whether a collision has been occurred by extrapolating the tags associated to it and verifying if more than one tag has been transmitted. We

note that the feasibility of the tagged preamble detection is not a trivial problem. For this reason, the next chapter is entirely dedicated to this purpose. In the following, we consider an ideal gNB receiver.

Moreover, because for the machine type communications a very small amount of data is expected [86], we assume for each transmission request the same upper bound value, sent in broadcast, denoted as θ_{max} bits. So, in Step 2, the gNB assigns to each successful access attempt, the PUSCH resources needed to satisfy the maximum data transmission (θ_{max}), if available.

Consequently, the MTC device which has received the message response from the gNB, within the W_{RAR} window, transmits its data in the PUSCH of the next RA cycle. We note that the data can be transmitted in the next RA cycle since we set W_{RAR} =5ms. In fact, thanks to the two-step procedure, this time is sufficient to guarantee the transmission of the tagged preambles (including cyclic prefix and propagation delay), the processing delay at the gNB side, the transmission of the RAR message, and 1ms of margin time. Conversely, if the message has not been received from the gNB, within the W_{RAR} window, then it will reattempt only if $N_A \leq M_A$.

The proposed 2-step RA procedure results not only in very few messages being exchanged but also in a very low signaling overhead, as depicted in Fig. 4.6 and described in the following. In order to reduce the signaling overhead, we assume that SCMA blocks are assigned to MTC device in multiples of SCMA Block Groups (SBGs), i.e., one SBG is the smallest unit of resources that can be allocated to an MTC device. An SBG corresponds to one SCMA block assigned for the time of 1 TTI. Moreover, we assume that Matrix $\mathbf{F}$, a set of L_{SB} codebooks and the mapping between codebook and the assigned layer are static and known by both sides of the communication link (e.g., sent in broadcast in the SIB).

As regards the number of bits associated to the SCMA allocation information for the MTC device, it results that, when $S = 2$, the 18 bits reserved for time and frequency allocation in the conventional RAR (as reported in Section 4.2.1) are enough to deliver the information, regardless of the value of Q. Therefore, no additional bits need to be sent by the gNB in DL. The proof is reported in Section 4.8.1. Furthermore,

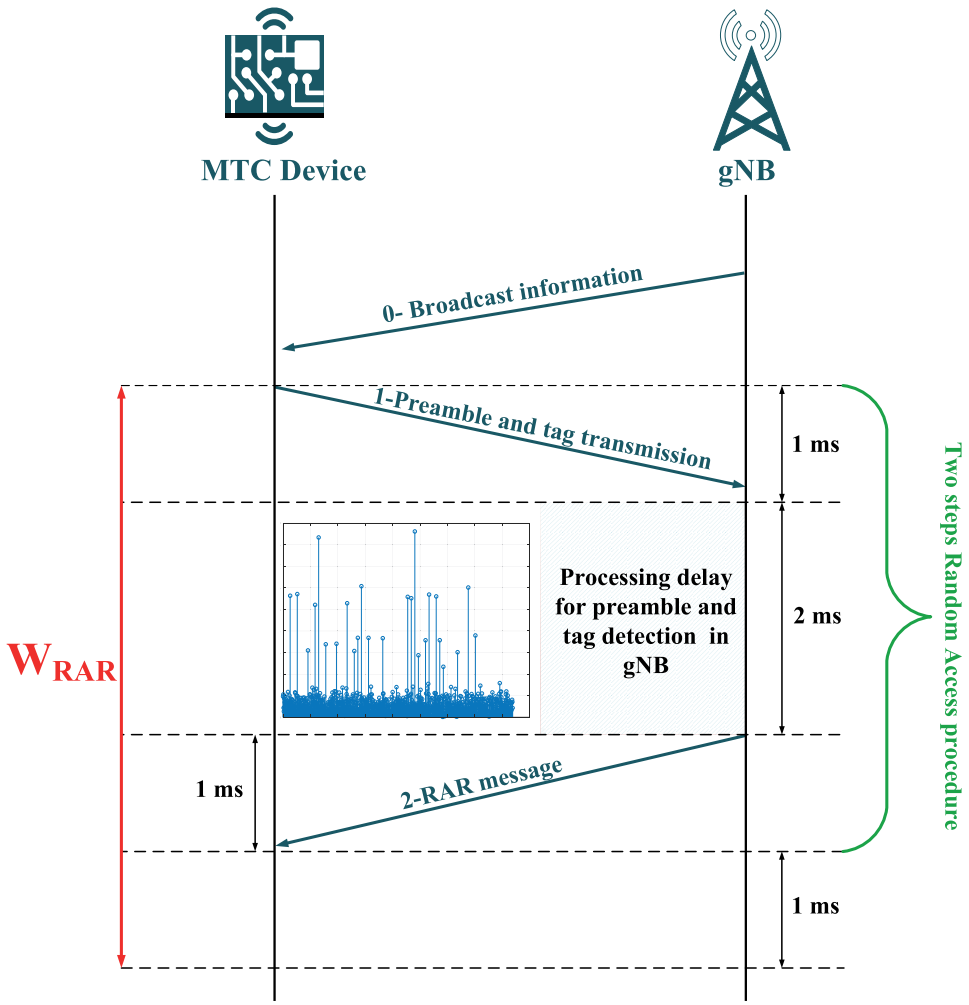


Figure 4.5: *The proposed basic two-step RA procedure.*

as in [66], we assume that the UL context header sent together with the data packet is of 9 bytes. Globally, the total amount of bytes necessary to transmit L bytes of data is equal to $L + 9$ bytes in UL and 7 bytes in DL, regardless the number of access attempts, i.e., N_A.

In order to underline the advantages of our two-step RA procedure, we compare the signaling overhead with the traditional four-step RA procedure reported in Section 4.2.1, and the one adopted by NB-IoT. By adopting the LTE procedure depicted in Fig. 4.2, the amount of bytes needed to transmit L bytes of data is equal to $(L + N_A \cdot 7 + 49)$ in UL and $(N_A \cdot 7 + 113)$ in DL. We note that this legacy RA procedure is the same adopted by the enhanced MTC (eMTC) technology. Conversely, the NB-IoT technology adopts a slightly simplified 4-step procedure consisting of $(N_A \cdot 8 + L)$ bytes in UL and $(N_A \cdot 7 + 8)$ in DL [66]. In conclusion,

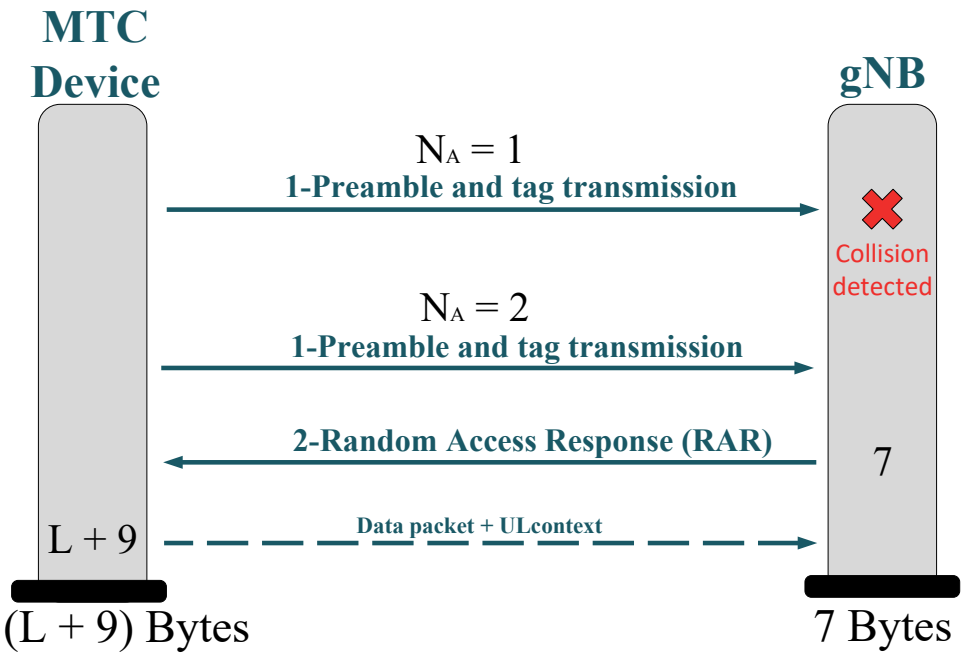


Figure 4.6: *Two-step RA procedure and data packet transmission.*

our proposal not only shows a lower amount of signaling with respect to both eMTC and NB-IoT (if $N_A > 1$). In addition, by adopting both the eMTC and the NB-IoT RA procedure, the MTC device for each RA attempt needs to perform 2 transmissions , independently whether the procedure has been succeeded or not. Instead, by adopting our proposal, the number of transmission per RA attempt is reduced to 1. So, the connectionless concept helps to reduce energy consumption in the MTC device, mainly because in this case there is a lower number of transmissions performed by the radio module in comparison with the eMTC and NB-IoT [87].

4.3.3 Traffic Model

The Beta traffic model is related to a scenario where a large number of MTC devices access the network in a highly synchronized manner [88]. Therefore, we assume that each MTC device generates data at time $t \in [0, T_{arrival}]$ following the Beta distribution [89] and that each activation time is independent of each other. The Beta distribution is

characterized by the following Probability Density Function (PDF):

$$f(t) = \frac{t^{\alpha-1}\left(T_{arrival}-t\right)^{\beta-1}}{T_{arrival}^{\alpha+\beta-1} Beta(\alpha,\beta)}, t \in [0, T_{arrival}], \tag{4.7}$$

where

$$Beta(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt. \tag{4.8}$$

Given N_{MTC} devices, we derive a discrete random process which counts the number of data arrivals per RA cycle inside the time period $[0, T_{arrival}]$. Let $\{N(\omega, n)\}$ denote this discrete random process, where $\omega \in \Omega$ is an experiment and $n = 1, 2, \ldots, \left\lfloor \frac{T_{arrival}}{T_{ra}} \right\rfloor$ is a point in time representing the nth RA cycle. Then, the function $N(\overline{\omega}, n)$ is a realization of the random process related to the outcome $\overline{\omega}$. For each n value, $N(\overline{\omega}, n)$ is the sum of the arrivals in the interval time $[(n-1)T_{ra}, nT_{ra}[$.

We can define the temporal auto-correlation function $r_{NN}(\overline{\omega}, l)$ at the time-lag l between two points in time.

$$r_{NN}(\overline{\omega}, l) = \sum_{n=1}^{\left\lfloor \frac{T_{arrival}}{T_{ra}} \right\rfloor - l} N(\overline{\omega}, n)N(\overline{\omega}, n+l), \tag{4.9}$$

for $l = -\left\lfloor \frac{T_{arrival}}{T_{ra}} \right\rfloor + 1, \ldots, \left\lfloor \frac{T_{arrival}}{T_{ra}} \right\rfloor - 1$. By simulation, we estimated that $r_{NN}(\overline{\omega}, l) = r_{NN}(l)$, and the mean value $\mu_n(\overline{\omega}) = \mu_n$, for a very large number of experiments, thus we assume that the process is ergodic in the wide sense.

The auto-covariance can be also expressed as a function of the time-lag l as follows:

$$c_{NN}(l) = r_{NN}(l) - \mu_n^2, \tag{4.10}$$

The Pearson Auto-Correlation Coefficient (PAC) is defined as

$$\rho(l) = \frac{c_{NN}(l)}{c_{NN}(0)}. \tag{4.11}$$

As it is known, the range of the PAC is $-1 \leq \rho(l) \leq 1$. In particular, when $|\rho(l)| \geq 0.8$ the sequence is highly correlated [90]. Fig. 4.7 shows $\rho(l)$ when Simulation Parameters in Table 4.1 are adopted. As shown, $|\rho(l)| \geq 0.8$ when $|l| \leq 296$, i.e., the number of arrivals is highly correlated in 296 consecutive RA cycles.

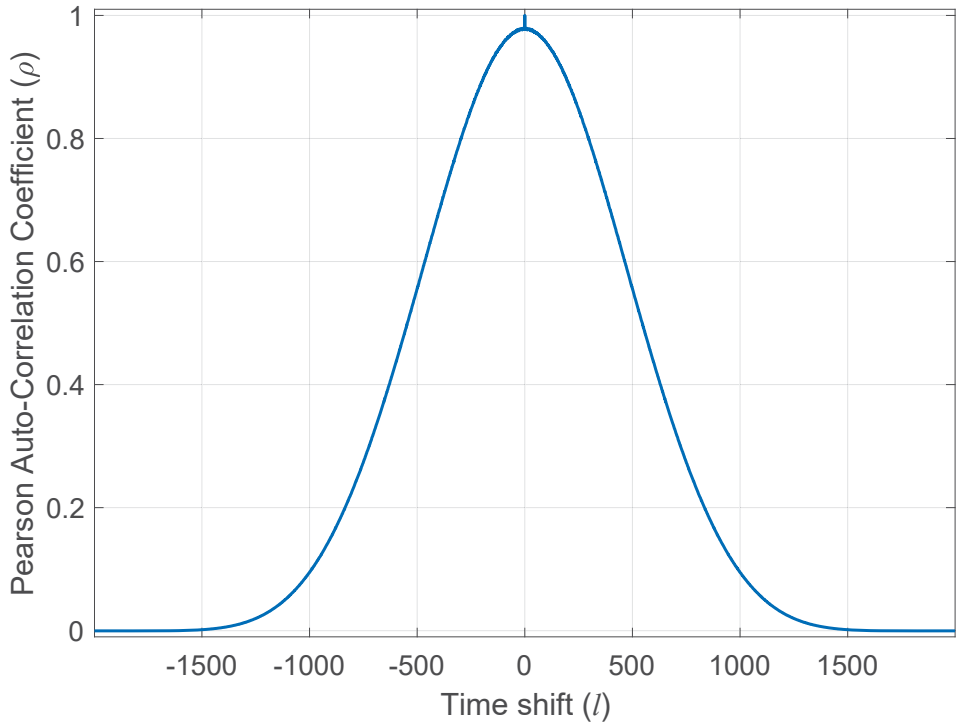


Figure 4.7: *Pearson Auto-Correlation Coefficient (PAC) of the random process* $\{N(\omega, n)\}$.

4.4 Uplink resource dimensioning

In this section, we formulate the optimal uplink resource dimensioning problem between the PRACH and the PUSCH in a single RA cycle known the number of MTC devices (M) which are performing the contention-based RA procedure, i.e., to derive the optimal T_{pr}^{*} value. For deriving the optimal T_{pr}^{*}, we need to estimate the number of succeeded preambles (P_S) and the number of available resources in the PUSCH based on SCMA technique (DT_{max}) as function on T_{pr}. As regards P_S, we note that it is a random variable whose mean value ($\bar{P}_S$) is derived on basis of T_{pr} and M, as follows:

$$\bar{P}_S = M \left(1 - \frac{1}{L_0 T_{pr}}\right)^{M-1}. \tag{4.12}$$

The proof is reported in Section 4.8.2. Clearly, given M, the higher T_{pr}, the higher $\bar{P}_S$.

As regards DT_{max}, we derive the maximum number of available transmissions in the PUSCH, each one consisting of θ_{max} bits, as follows:

$$DT_{max} = \left\lfloor \frac{L_{RAC}}{\left\lceil \frac{\theta_{max}}{r \log_2(I)} \right\rceil} \right\rfloor = \left\lfloor \frac{N_{SB} L_{SB}}{\left\lceil \frac{\theta_{max}}{r \log_2(I)} \right\rceil} \right\rfloor (T_{ra} - T_{pr}), \tag{4.13}$$

where $\log_2(I)$ is the number of information bits sent for each symbol of the constellation and r is the code rate. It is evident that, the higher T_{pr}, the lower N_{ST}.

The above results confirm the following major problems.

1. The PRACH resources are not sufficient when thousand or more devices simultaneously request to send data. In fact, as reported in (4.12), for very high M values, the smaller T_{pr}, the faster $\bar{P}_S$ approaches zero. This problem could be alleviated by increasing the access resources, i.e., T_{pr}.

2. The number of resources available for the PUSCH decreases with increased T_{pr}. So, also if we have a high value of successful accesses, by adopting a high T_{pr} value, a part of these ones could not be satisfied due to lack of resources in the PUSCH.

Considering the previous issues there is the need to find a good trade-off between the amount of access resources, contained in T_{pr}, and the ones used for data transmission, contained in T_{pu}.

A communication is succeeded if both the following conditions occur. First, the MTC device transmits with success its preamble sequence in the PRACH, i.e., it is the only MTC device that has transmitted the selected preamble sequence. Second, there are resources available in the PUSCH for its data transmission.

We report an example in Fig. 4.8. This figure shows, in Steps 1 and 2, the connectionless RA procedure, while in Step 3 the related data transmission, when either a conventional or an enhanced PUSCH resource allocation is applied. In the following we focus on the conventional resource allocation. In Step 1, $M = 9$ MTC devices transmit one randomly chosen preamble in the PRACH. Let $\mathbf{P_T} = [p_1, \ldots, p_M]$ denote the vector containing the preambles transmitted by the MTC devices. Specifically, in the example $\mathbf{P_T} = [2, 10, 10, 1, 3, 3, 1, 10, 4]$, i.e., preambles 2 and 4 have been transmitted by only one MTC device, whereas preambles 1,

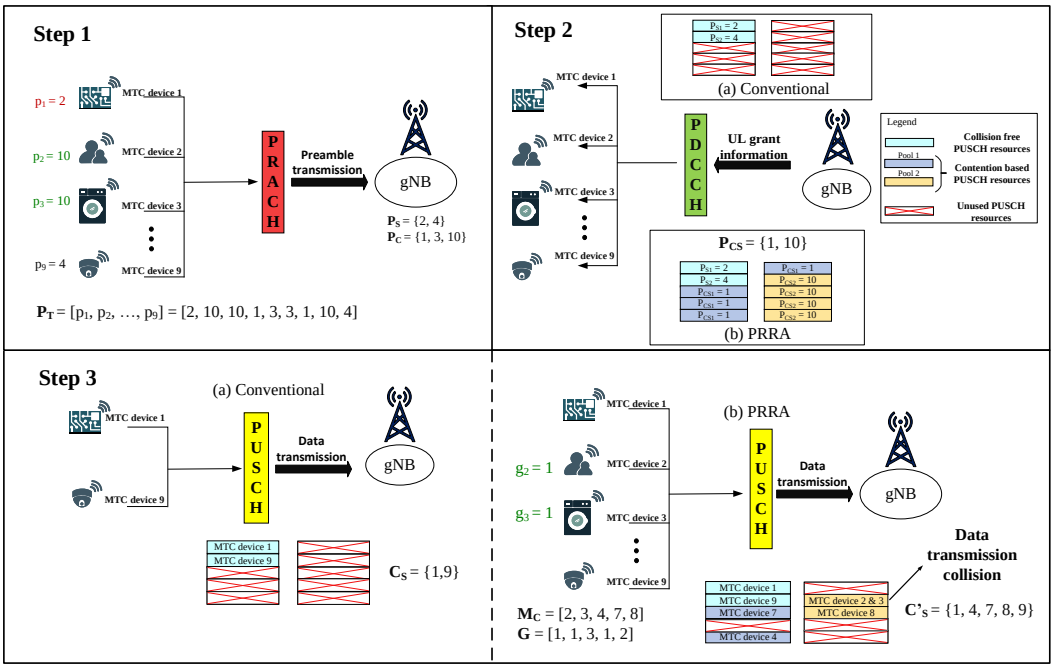


Figure 4.8: *Conventional PUSCH resource allocation vs PRRA-based PUSCH allocation.*

3, and 10 by two or more MTC devices. After the reception of all transmitted preambles, the gNB detects the preambles 2 and 4 as succeeded, whereas preambles 1, 3, and 10 as collided. We denote $\mathbf{P_S}$ as the set of succeeded preambles with cardinality P_S, and $\mathbf{P_C}$ as the set of collided preambles with cardinality P_C. In the example, we have $\mathbf{P_S} = \{2, 4\}$ and $\mathbf{P_C} = \{1, 3, 10\}$. It follows Step 2a. In this Step, the gNB reserves one data transmission out of $DT_{max} = 10$ for each preamble in $\mathbf{P_S}$, and then the gNB sends in the Physical Downlink Control Channel (PDCCH) the RAR message containing, *inter alia*, the UL grant information. We note that if $P_S > DT_{max}$, the number of succeeded communication is limited by the DT_{max} resources. Next, in Step 3a, all the MTC devices, which have transmitted a preamble, decode the RAR message received from Step 2 and only the ones which have transmitted a preamble in $\mathbf{P_S}$ (i.e., MTC devices 1 and 9) will transmit the data in the dedicated data transmission.

For the sake of simplicity, in the problem formulation we assume that the MTC devices, which have successfully passed their access attempt in the PRACH, will transmit their data in the PUSCH of the same RA cycle (if available), instead of the next RA cycle, as reported in Section 4.3.

In light of the example below reported and this assumption, the amount of successful communications (C_S) inside an RA cycle can be derived as:given a RA cycle the number of succeeded communications is:

$$C_S = \min(P_S, DT_{max}). \tag{4.14}$$

Since P_S is a random variable, we approximate (4.14) as:

$$\tilde{C}_S = \min(\bar{P}_S, DT_{max}). \tag{4.15}$$

We note that DT_{max} is a deterministic variable, given θ_{max}, T_{pr}, Q, K_{max}, I, and T_{ra}, that are all information known at the gNB. Conversely, the P_S value is available at the gNB once all the preamble sequences have been received and correctly detected.

Obviously, the average number of succeeded communications, is

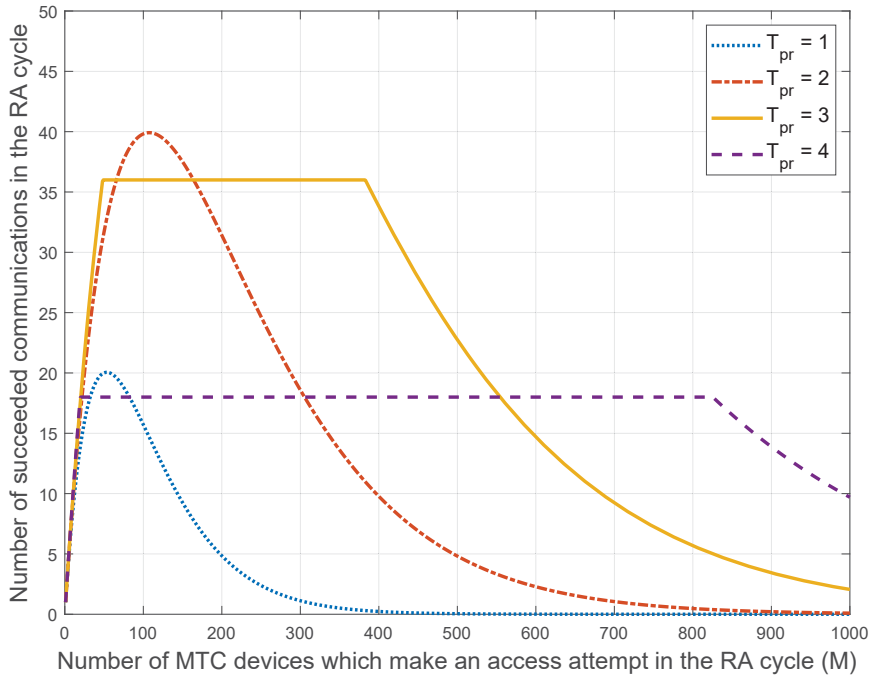
$$\bar{C}_S = \min(\bar{P}_S, DT_{max}). \tag{4.16}$$

In Fig. 4.9a we plot $\bar{C}_S$ versus M for any T_{pr} value. We observe that the straight lines (e.g., M ranging from 50 to 380 for T_{pr} = 3) are caused by the saturated PUSCH resources. As shown, there is not a single optimal T_{pr} value, since it depends strongly on the offered load, i.e., on the value of M.

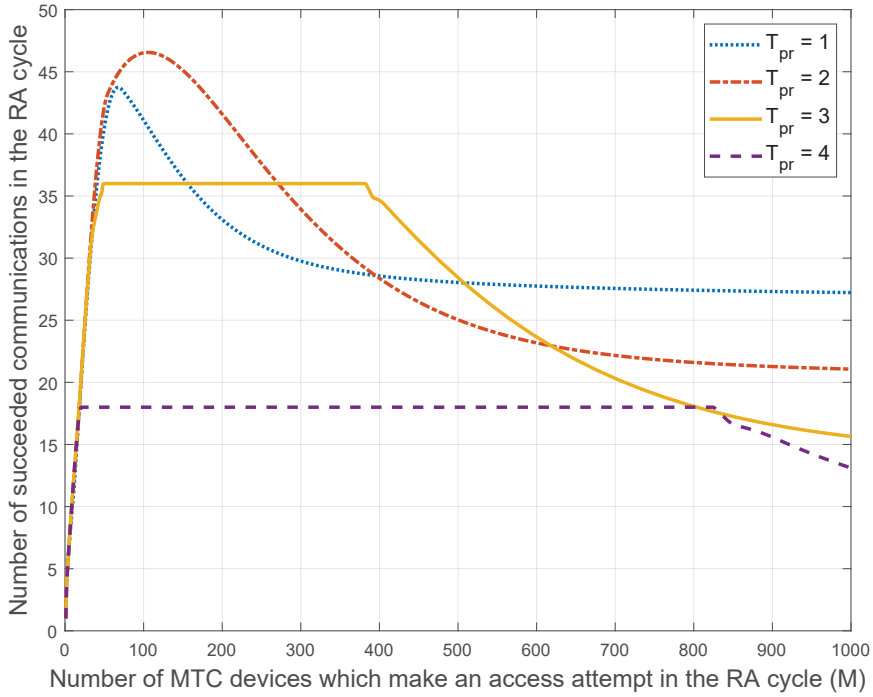
For this reason, at the aim of maximizing (4.16), we proposd the DURD scheme, that works as follows. Given RA cycle j, the gNB knows for each selected preamble, thanks to the 2-step RA procedure adopted, whether a collision has been occurred or it has successfully transmitted immediately after the reception of the preamble. In other words, during RA cycle j, the following information are available at the gNB:

- $L^j = L_0 T_{pr}^j$, i.e., the number available preambles in the PRACH of RA cycle;
- P_S^j, i.e., the number of successful access attempts;
- P_C^j, i.e., the number of collided preambles.

Despite the gNB knowing the amount of collided preambles, it does not know how many access attempts occurred in the collided preambles. However, in this section we assume M^j perfectly known on the basis



(a) $\bar{C}_S$



(b) $\bar{C}'_S$ *by using (4.25)*

Figure 4.9: *Number of successful transmissions in an RA cycle vs the number of MTC devices (M) which carry out the RA procedure for different T_{pr} values.*

of the information above reported. This information will be used for correctly dimensioning the T_{pr} value of the next RA cycle. The starting

point is to exploit that the arrival process of total access attempts in two consecutive RA cycles is strongly correlated, since the arrival process of the access re-attempts follows the uniform distribution in a backoff window and the arrival process of new access attempts follows the Beta distribution. The proof is reported in [23]. Thus, we can assume that

$$M^{j+1} \cong M^{j}. \tag{4.17}$$

So, the T_{pr} dimensioning for the next RA cycle $j+1$ is calculated as follows:

$$T_{pr}^{j+1} = \underset{T_{pr}\in\{1,\ldots,T_{ra}-1\}}{\arg\max} \left\{\bar{C}_S(M^j)\right\}. \tag{4.18}$$

We assume that the MTC device which has received the RAR message from the gNB in the RA cycle $j-1$, within the W_{RAR} window, transmits its data in the PUSCH of the RA cycle j. We note that the data can be transmitted in the next RA cycle since we set $W_{RAR} = 5$ ms. In fact, thanks to the two-step procedure, this time is sufficient to guarantee the transmission of the tagged preambles (including cyclic prefix and propagation delay), the processing delay at the gNB side, the transmission of the RAR message, and 5 ms of margin time.

So, given RA cycle j, knowing M^{j-1}, Q, K_{max}, S, θ_{max}, T_{pr}^{j-1}, P_S^{j-1}, and P_C^{j-1}, the scheduler calculates DT_{max} by using (4.13). If $P_S^{j-1} \leq DT_{max}$, the scheduler assigns in the PUSCH P_S data transmissions in a collision free mode. Otherwise, if $P_S^{j-1} > DT_{max}$, the scheduler assigns DT_{max} data transmissions in a collision free mode to a randomly chosen subset of succeeded preambles with cardinality DT_{max}.

To evaluate the goodness of the proposed scheme, we made a long term analysis, and we report a single simulation test consisting of a fixed distribution of the new access attempts in the space of 10 s, i.e., 2000 RA cycles, when $N_{MTC} = 100000$. The other simulation parameters are reported in Table 4.1.

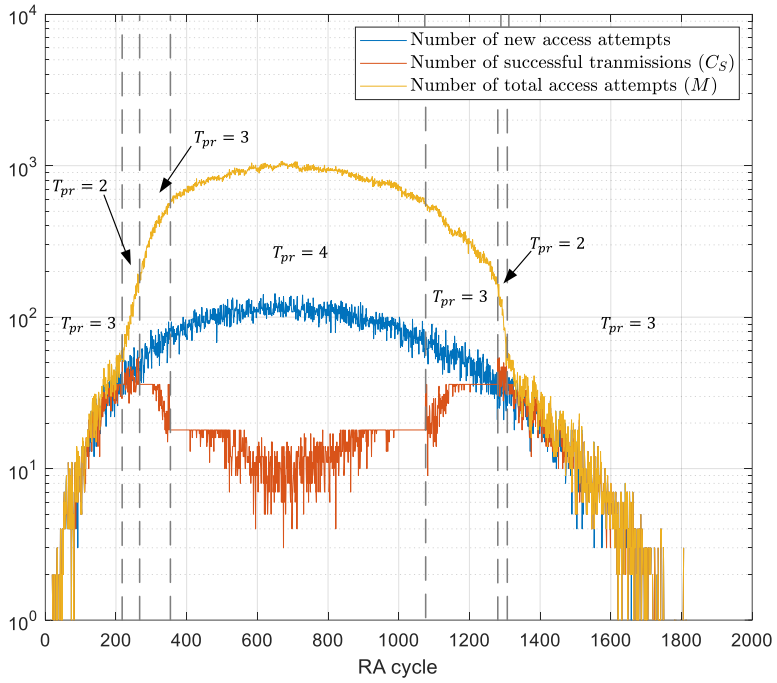
Fig. 4.10a depicts the temporal distributions of the new access attempts, the total access attempts (M) including the reattempts, and the successful communications C_S per RA cycle, with the optimal T_{pr} value calculated by using (4.18). We notice seven zones, each one showing a different T_{pr} value. As shown in Fig.4.9a, the optimal dimensioning is $T_{pr} = 3$ from a light load ($M = 1$) to medium-high load ($M = 554$),

except for the interval $M \in [66, \ldots, 163]$, where the optimal dimensioning is $T_{pr} = 2$. Conversely, in the presence of high load ($M > 555$), the optimal T_{pr} dimensioning becomes $T_{pr} = 4$. In light of this analysis, in Fig. 4.10a, as the offered load increases, the optimal dimensioning varies from $T_{pr} = 3$ (light load) to $T_{pr} = 4$ (high load), according to the given M value. Once the peak has been reached, the traffic load decreases and, consequently, the optimal dimensioning changes from $T_{pr} = 4$ (high load) to $T_{pr} = 3$ (light load).

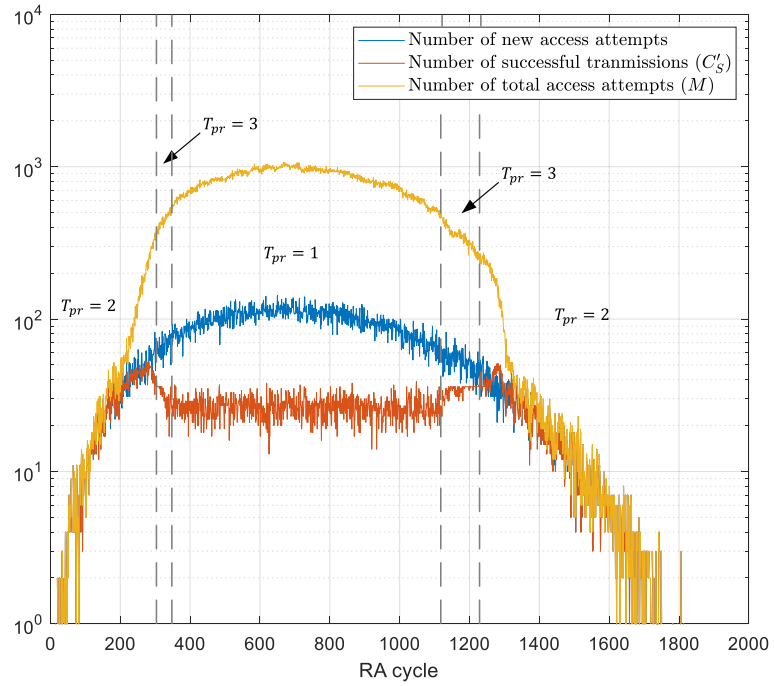
However, we note that, this represents an ideal dynamic uplink resource dimensioning since the M value is considered well-known by the gNB for each RA cycle.

4.5 Enhanced Dynamic Uplink Resource Dimensioning

In light of the results obtained by the DURD scheme, we propose a new strategy to raise the number of succeeded communications, when $P_C > 0$ and $P_S < DT_{max}$, i.e., P_S data transmissions are scheduled in the PUSCH, while $DT_U = DT_{max} - P_S$ data transmissions remain unused in the PUSCH. For example, in Fig. 4.8, in Step 3a, by adopting the conventional resource allocation, only two communications are succeeded, and $DT_U = 8$ data transmissions are wasted. To explain the enhanced strategy, we consider Step 2b and 3b. In Step 2b, the gNB assigns $P_S = 2$ data transmissions to the preambles in $\mathbf{P_S}$ in a collision free manner, as in the conventional resource allocation. In addition, the gNB allocates also the $DT_U = 8$ data transmissions to a subset of collided preambles, denoted as $\mathbf{P_{C_S}} \subseteq \mathbf{P_C}$, with cardinality P_{C_S}. Since there is a 1 to many relationship between each collided preamble and the related MTC devices, the gNB cannot allocate the DT_U data transmissions in a collision free mode. For this reason, for each collided preamble in $\mathbf{P_{C_S}}$, the scheduler needs to allocate a pool of R_θ resources consisting of DT_{P_C} data transmissions, with $DT_{P_C} > 1$. In the example, $\mathbf{P_{C_S}} = \{1, 10\}$, and $DT_{P_C} = 4$. Then, as in the conventional procedure, the gNB sends in the PDCCH the RAR message. Next in Step 3b, all the MTC devices which have transmitted a preamble decode the message received from Step 2. The MTC devices which have transmitted a preamble in $\mathbf{P_S}$ (i.e., MTC devices 1 and 9) will transmit the data in the dedicated data trans-



(a) *Dynamic Uplink Resource Dimensioning*



(b) *Enhanced Dynamic Uplink Resource Dimensioning*

Figure 4.10: *Temporal distributions of the new access attempts, the total access attempts (M) including the reattempts, and the successful communications (C'_S) per RA cycle for different resource dimensioning*

mission, as in the conventional procedure. At the same time, each MTC device that has transmitted in the PRACH a preamble belonging to $\mathbf{P_{C_S}}$, will draw independently a uniform random number $g \in \{1, \ldots, DT_{P_C}\}$, and will send its data in the gth data transmission of the resource pool. Clearly, if the data transmission g of the same pool of resources has been selected by only one MTC device, then the communication is succeeded. Otherwise, all the MTC devices that have selected the same data transmission will experience a collision. In the example, we denote with $\mathbf{M_C} = [2, 3, 4, 7, 8]$ the vector of MTC devices that draw a value $g \in \{1, \ldots, 4\}$, and with $\mathbf{G} = [1, 1, 3, 1, 2]$ the vector containing the extracted g values. The collided preamble 1 has been transmitted by the MTC devices 4 and 7. On the basis of the related g value extracted, they will send their data in the data transmission 3 and 1 of the pool of resources reserved for the collided preamble 1, respectively. As regards the collided preamble 10, both the MTC devices 2 and 3 will send their data in the data transmission 1 of the related pool of resources, while MTC device 8 will send its data in the data transmission 2. So, the MTC devices 2 and 3 experience a collision, while 4, 7, and 8 are successful additional data transmissions compared to the conventional resource allocation.

Obviously, the number of successful additional data transmission (A_S) is a random variable that depends on the number of successful additional data transmission per collided preamble (A_{S,P_C}) and on $P_{C_S} \leq P_C$. Since A_{S,P_C} and P_{C_S} are independent random variables, the average number of total successful additional data transmissions is

$$\bar{A}_S = \bar{A}_{S,P_C} \bar{P}_{C_S}. \tag{4.19}$$

As regards $\bar{A}_{S,P_C}$, it depends on DT_{P_C}, and on the number of MTC devices that have transmitted the same collided preamble. We denoted it as K_C and it is termed as "collision coefficient". The relationship between $\bar{A}_{S,P_C}$, DT_{P_C}, and K_C is:

$$\bar{A}_{S,P_C} = \begin{cases} K_C \left(1 - 1/DT_{P_C}\right)^{K_C - 1} & \text{if } DT_{P_C} > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.20}$$

To calculate $\bar{A}_{S,P_C}$, we need to derive both K_C and DT_{P_C}. However, also K_C and DT_{P_C} are random variables, depending on M and L. The

expected value of K_C can be calculated as

$$\bar{K}_C = \frac{M\left(\frac{1}{L}\right)\left[1-\left(1-\frac{1}{L}\right)^{M-1}\right]}{1-\left(1-\frac{1}{L}\right)^{M}-M\left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)^{M-1}}. \tag{4.21}$$

The proof is reported in Section 4.8.3. As regards DT_{P_C}, its expected value is calculated as

$$\overline{DT}_{P_C} = \begin{cases} \frac{DT_{max}-\bar{P}_S}{\bar{P}_{C_S}} & \text{if } \bar{P}_S < DT_{max}-1 \\ 0 & \text{otherwise,} \end{cases} \tag{4.22}$$

where $\bar{P}_{C_S}$ ranges in $\left[1, \min\left(DT_{max}-\bar{P}_S, \bar{P}_C\right)\right]$, and $\bar{P}_C$ is the average cardinality of $\mathbf{P_C}$. This last parameter is equal to

$$\bar{P}_C = L\left[1-\left(1-\frac{1}{L}\right)^{M}-M\left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)^{M-1}\right], \tag{4.23}$$

and the proof is reported in Section 4.8.2.

So, we derived $\bar{K}_C$ and a range of values assumed by $\overline{DT}_{P_C}$. For each pair of values $\bar{K}_C$ and $\overline{DT}_{P_C}$, we can calculate $\bar{A}_{S,P_C}$ by using (4.20), where we set $K_C = \bar{K}_C$, and $DT_{P_C} = \overline{DT}_{P_C}$. This treatment is valid when using the drift approximation [91]. Then, by using (4.19) we can calculate $\bar{A}_S$. Finally, we redefined the average number of succeeded communications ($\bar{C}'_S$) as:

$$\bar{C}'_S = \begin{cases} \bar{P}_S + \bar{A}_S & \text{if } \bar{P}_S < DT_{max}-1 \\ \min(\bar{P}_S, DT_{max}) & \text{otherwise.} \end{cases} \tag{4.24}$$

We note that, given a fixed T_{pr} value, M, and θ_{max}, it follows that the values of $\bar{P}_S$ and DT_{max} are determined by (4.12) and (4.13), respectively, while $\bar{A}_S$ assumes a range of values as $\bar{P}_{C_S}$ varies. In other words, in (4.24) the only term that can be maximized is $\bar{A}_S$. So, the enhanced strategy computes the optimal $\bar{P}_{C_S}$ value, denoted as $\bar{P}^*_{C_S}$, as follows

$$\bar{P}^*_{C_S} = \underset{\bar{P}_{C_S}\in\left[1,\min\left(\overline{DT}_U,\bar{P}_C\right)\right]}{\arg\max} \left\{\bar{A}_S\right\}. \tag{4.25}$$

The related $\overline{DT}_{P_C}$ and $\bar{A}_S$ values are denoted as $\overline{DT}^*_{P_C}$ and $\bar{A}^*_S$, respectively. By simulation, given $M \in \{1, \ldots, 1000\}$ and $T_{pr} \in \{1, \ldots, 4\}$, we calculated the maximum value of $\bar{C}'_S$ as

$$\bar{C}'_{S,max} = \begin{cases} \bar{P}_S + \bar{A}^*_S & \text{if } \bar{P}_S < DT_{max} - 1 \\ \min(\bar{P}_S, DT_{max}) & \text{otherwise.} \end{cases} \tag{4.26}$$

In Fig. 4.9b, we plot the variation of $\bar{C}'_{S,max}$ versus M for any T_{pr} value. Compared to Fig. 4.9a, the $T_{pr} = 1$ and $T_{pr} = 2$ dimensionings achieve a remarkable improvement for any value of M. Specifically, the peak of the improvement of $T_{pr} = 2$ is about 20 for $M \geq 480$, while the peak of $T_{pr} = 1$ is about 28 for $M \geq 180$. As regards $T_{pr} = 3$, a light improvement occurs when $M \geq 500$ and the peak is about 13 for $M \geq 900$. $T_{pr} = 4$ has no significant room for improvement. The lower the T_{pr} value, the larger $\overline{DT}_U$, the higher the improvement.

In light of the advantages depicted in Fig. 4.9b we propose a new Enhanced Dynamic Uplink Resource Dimensioning (EDURD) that works as follows. Given RA cycle j, the gNB knows L^j, P^j_S, and P^j_C, but it does not know M^j. However, in this section we assume M^j perfectly known. This information is used for correctly dimensioning the T_{pr} value of the next RA cycle $j + 1$ as follows

$$T^{j+1}_{pr} = \underset{T_{pr} \in \{1, \ldots, T_{ra}-1\}}{\arg\max} \left\{ \bar{C}'_{S,max}(M^j) \right\}. \tag{4.27}$$

In addition, the knowledge of the traffic load is also used to apply the enhanced resource allocation in the PUSCH. So, knowing M^{j-1}, Q, K_{max}, S, θ_{max}, T^{j-1}_{pr}, P^{j-1}_S, and P^{j-1}_C, the scheduler calculates DT_{max} by using (4.13). If $P^{j-1}_S < DT_{max} - 1$, the scheduler applies the PUSCH Resource Reallocation Algorithm (PRRA) for the enhanced resource allocation in the PUSCH. The algorithm calculates $DT^j_U = DT_{max} - P^{j-1}_S$, and K^{j-1}_C as follows

$$K^{j-1}_C = \frac{M^{j-1} - P^{j-1}_S}{P^{j-1}_C}. \tag{4.28}$$

Then, after some iterations, the algorithm provides as output the $P^*_{C_S}$ and the $DT^*_{P_C}$ values that maximize $\bar{A}_S$ inside RA cycle j. The algorithm is described in pseudo-code 8, and its time complexity is $O(n)$, where $n = \min\{DT_U, P_C\}$.

Pseudo-code 8 PUSCH Resources Reallocation Algorithm

Inputs: M, Q, K_{max}, S, θ_{max}, T_{pr}, T_{ra}, P_S, and P_C
Iteration:
1: calculate DT_{max} by using (4.13), $K_C = {}^{(M-P_S)}/_{P_C}$, and $DT_U = DT_{max} - P_S$;
2: create an auxiliary vector $\mathbf{A} = [1, \ldots, \min\{DT_U, P_C\}]$;
3: calculate $\mathbf{B} = \lfloor DT_U \mathbf{A}^{\circ(-1)} \rfloor$, where $(\cdot)^\circ$ is the Hadamard power operator;
4: calculate $\mathbf{C} = K_C \left(\mathbf{J}_{1,|\mathbf{B}|} - \mathbf{B}^{\circ(-1)} \right)^{\circ(K_C-1)}$ where $\mathbf{J}_{1,|B|}$ is the all-ones matrix of size $1 \times |\mathbf{B}|$;
5: calculate $\mathbf{D} = \mathbf{A} \circ \mathbf{C}$, where $\circ$ is the Hadamard product;
6: calculate $\bar{A}_S = \max\{\mathbf{D}\}$ and $P^*_{C_S} = \arg\max\{\mathbf{D}\}$;
7: calculate $DT^*_{P_C} = \mathbf{B}[P^*_{C_S}]$.

Given $P^*_{C_S}$, the subset $\mathbf{P^*_{C_S}}$ is randomly extracted among all the possible subsets of $\mathbf{P_C}$, having a cardinality equal to $P^*_{C_S}$. Finally, the gNB assigns in the PUSCH P_S data transmissions in a collision free mode, and it reserves $DT^*_{P_C}$ data transmissions to each collided preamble in $\mathbf{P^*_{C_S}}$. Otherwise, if $P_S^{j-1} \geq DT_{max} - 1$, the scheduler assigns DT_{max} data transmissions in a collision free mode to a randomly chosen subset of succeeded preambles with cardinality DT_{max}. We underline that expression (4.28) is different from (4.21), since it is calculated *a posteriori* given P_S^{j-1}, and P_C^{j-1}, while the first one is the average value of K_C calculated *a priori* on the basis of M and L.

Also for the EDURD system, we evaluate the goodness of the proposed scheme by making a long term analysis, with the same simulation parameters adopted for the DURD, and we depict in Fig. 4.10b the temporal distributions of the new access attempts, M, and C'_S per RA cycle. As shown in Fig. 4.9b, the optimal dimensioning results $T^*_{pr} = 2$ for $M \leq 271$, $T^*_{pr} = 3$ for $272 \leq M \leq 507$, and $T^*_{pr} = 1$ for $M \geq 508$. So, in Fig. 4.10b, when the offered load increases, the initial optimal dimensioning is $T_{pr} = 2$, and it remains the same up to a medium load condition. After that, the optimal dimensioning becomes $T_{pr} = 2$, for a few RA cycles and then changes in $T_{pr} = 1$, being in a high load condition. The dual behavior is valid for the unloading part of the simulation.

As expected, comparing Fig. 4.10a with Fig. 4.10b, the main advantage is obtained in the high load condition, corresponding to the middle zone of the curves. In fact, in this zone, the optimal dimensioning of the DURD is $T_{pr} = 4$, and consequently it achieves a maximum number of succeeded communication equal to 18 for each RA cycle far from the

peak, and even lower in the proximity of the peak. Conversely, as regards the EDURD, the optimal dimensioning is $T_{pr} = 1$, and the average number of succeeded communications is equal to about 27, in the entire zone. The high load of M corresponds to $T_{pr} = 1$ because a large portion of the preambles in the PRACH are collided and, consequently, the succeeded transmissions are related to the re-allocation in the PUSCH, that works better with $T_{pr} = 1$. We underline that the advantages of adopting EDURD instead of DURD are achieved also for $M \in \{60, \ldots, 255\}$, where the EDURD adopts a $T_{pr} = 2$ dimensioning which leads to a higher number of succeeded communications, even if less evident.

Finally, we underline that also in this case we need to estimate M^j in each RA cycle j. But, in the EDURD system, the knowledge of M is needed not only for the dynamic uplink dimensioning but also to properly apply the PRRA. The analysis concerning the estimation of M is reported in the next Section.

4.6 Predictive estimation of access attempt number

For each jth RA cycle, at the aim of optimizing the resource allocation between PRACH and PUSCH, the gNodeB should know the total amount of active MTC devices (M^j) attempting to access, considering both the reattempting ones which failed in the previous RA cycles and the new arrived ones in the jth RA cycle. Given an RA cycle j, the gNB knows T^j_{pr}, P^j_S, and P^j_C. From an empirical study, we set as estimated value of M^j, denoted as $\tilde{M}^j$, the following equation:

$$\tilde{M}^j = P^j_S + 2 \cdot 1.97^{\frac{P^j_C}{L_0 T^j_{pr}}} P^j_C, \tag{4.29}$$

Equation (4.29) resulted sufficiently accurate for the DURD purpose, as demonstrated by the results shown in the next Section. However, it resulted not accurate for the EDURD system. This is related to the limits of the estimate of M^j as a function of P^j_S, P^j_C, and L^j, regardless of the estimation methodology.

In Fig. 4.11 we plot $\bar{P}_C$, and $\bar{P}_S$ vs M, when $L = 54$, and $L = 216$. As shown, for high load scenario and $L = 54$, it follows that $\bar{P}_C = L$, and $\bar{P}_S = 0$. So, the value of M cannot be determined as function of the

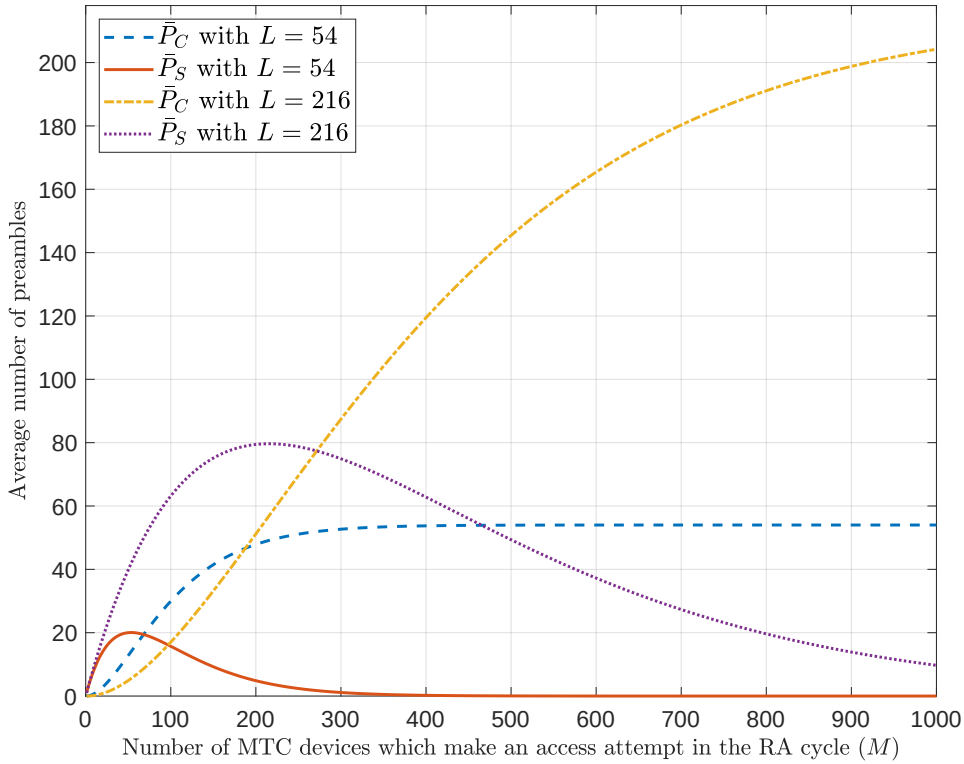


Figure 4.11: *collision graph*

above parameters, since it is not unique. Conversely, for $L = 216$, given a pair of values $\bar{P}_C$, and $\bar{P}_S$ we can obtain a single value of M.

As regards the DURD system, we underline that the dimensionings reported in Fig. 4.9a show that when M^j is large, it follows that the optimal dimensioning $T_{pr}^{j+1} = 4$ is adopted, i.e., $L^{j+1} = 216$, and consequently M^{j+1} could be estimate as a function of P_S^{j+1}, and P_C^{j+1}.

Conversely, as regards the EDURD, in the presence of high traffic load, the optimal dimensioning $T_{pr}^{j+1} = 1$ is adopted, i.e., $L^{j+1} = 54$ (See Fig. 4.9b). This means that the described estimation methods cannot be adopted. Now, we are working on adopting an advanced DNN-based traffic load etimation method to fix this issue.

4.7 Performance Evaluation

In this section, we consider the following uplink dimensioning systems, where the PUSCH resources are assigned on basis of the SCMA technique. We denote them as:

- S_i, with $i \in \{1, \ldots, 4\}$, as the system with static uplink dimension-

Table 4.1: *Simulation Parameters*

Parameter	Symbol	Value
Preambles reserved for contention-based procedure	L_0	54
Number of subcarriers	N_{SC}	72
Subcarrier bandwidth	b	15 kHz
Number of total MTC devices	N_{MTC}	40000, 50000, 75000, 100000
Number of data transmission requests per MTC device		1
Maximum number of RA attempts	M_A	10
Data transmission size	θ_{max}	160 bits
Arrival distribution	Beta	α = 3, β = 4 $T_{Arrival}$ = $10s$
RA cycle duration	T_{ra}	5 time slots
Simulation Length	T_{sim}	$10s$
SCMA parameters	Q K_{max} S	4 3 2
Backoff Window	B_W	20ms
Constellation Points	I	4

ing $T_{pr} = i$;

- $DURD$ [21] as the system with optimal dynamic uplink dimensioning based on Fig. 4.9a;
- $eDURD$ [21] as the $DURD$ system based on the predictive estimation of M by using (4.29).
- Ex, with $x = \{S_i, DURD, eDURD\}$, as the enhanced system x with the contention-based additional transmissions in the PUSCH, by using the PRRA.

We compare the performance of above systems by simulation in MATLAB environment. The simulation parameters are provided in Table 4.1, and results are averaged over 50 independent simulations. We introduce two metrics.
1) *System Throughput Gain*. It is the percentage throughput gain compared to S_1, i.e., $\Delta Th_x = \left[{}^{(Th_x - Th_{S_1})}/_{Th_{S_1}}\right] \cdot 100$, where Th_x is the system

throughput in terms of number of successful communications achieved by means of system x.
2) *Energy Consumption Indicator*. It is the average number of times in which an MTC device enters in the RX/TX active state [22], denoted as $\bar{E}_I$. For each MTC device, E_I is calculated as $E_I = E_{RA} + E_T$, where $E_{RA} \in \{1, \ldots, M_A\}$ is the number of access attempts, and E_T counts the total number of contention-based and contention-free transmissions in the PUSCH. In particular, $E_T \in \{0, 1\}$ for the non-Enhanced systems, while $E_T \in \{0, \ldots, M_A\}$ for the other ones.

Figs. 4.12 and 4.13 show $\Delta Th_x(\%)$ and the Energy Consumption Indicator $\bar{E}_I$, respectively. Let us start considering a moderate load, i.e., $N_{MTC} = 40000$. Among all the static dimensioning systems, only ES_1 and ES_2 show a throughput gain with respect to the related standard systems. Instead, as expected, ES_3 does not show improvement with respect to S_3 because, in this light traffic conditions, $M \leq 500$ for each RA cycle. By comparing ES_2 with DURD, we underline that the PRRA improves the performance of a static system also compared to an optimized dynamic solution. As regards the EDURD, the results are the same of the ones obtained by ES_2, being the optimal dimensioning $T^*_{pr} = 2$ for any RA cycle. Finally, EeDURD obtains the same results as EDURD, confirming that the predictive estimation of M is efficient in the considered working zone. As regards a medium-high load ($N_{MTC} = 50000$), compared to the previous case, also ES_3 shows a gain with respect to S_3. As regards the dynamic systems, EDURD shows the similar performance of the ES_3, because the optimal dimensioning is equal to $T^*_{pr} = 3$ for almost the totality of the RA cycles. In addition, it achieves the best performance in terms of energy consumption. Also in this case, the predictive estimation of M is effective. In the high load scenarios ($N_{MTC} \in \{75000, 100000\}$), among the static systems the best performances in throughput are obtained with ES_1. The EDURD system has a good gain with respect to any static system. In particular, the highest gain is achieved for $N_{MTC} = 75000$ because the optimal dimensioning ranges from $T^*_{pr} = 3$ to $T^*_{pr} = 1$, while for $N_{MTC} = 100000$ the gain is lower, because the optimal dimensioning is $T^*_{pr} = 1$ for many RA cycles. The energy consumption is slightly higher with respect to the other dynamic systems, but in line with the best static energy consumption. As

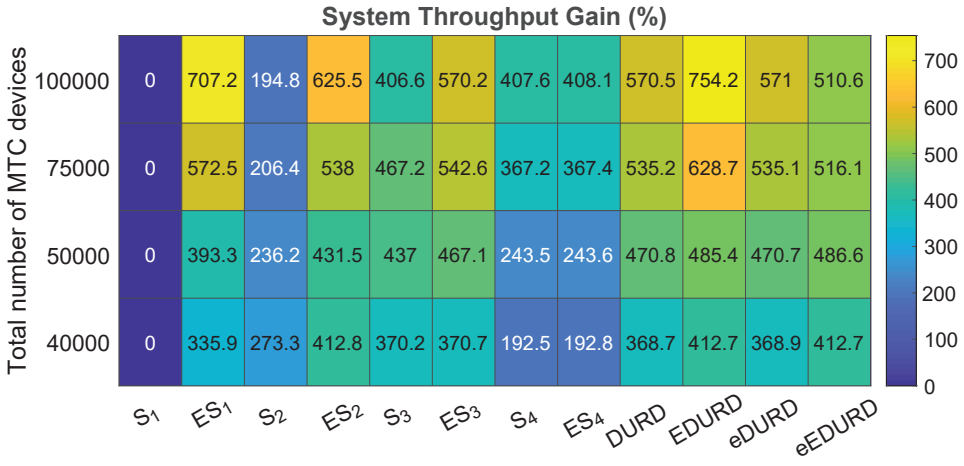


Figure 4.12: *Throughput Gain under different systems and M*

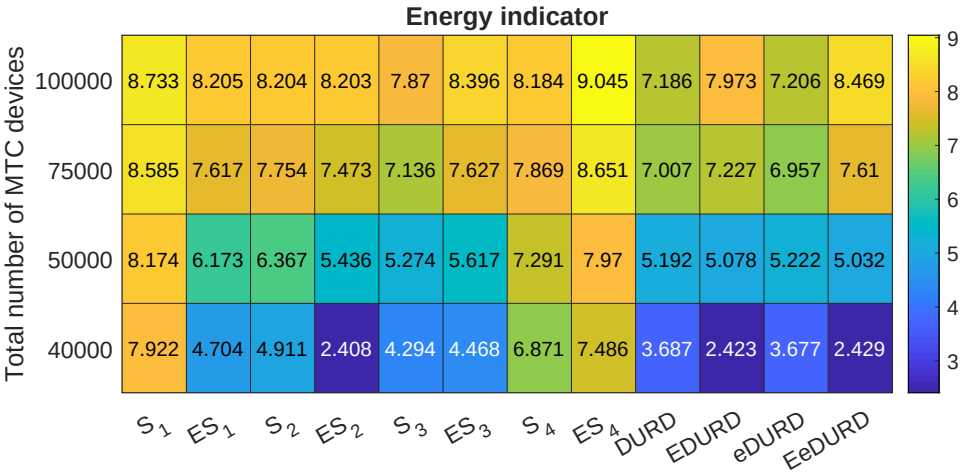


Figure 4.13: *Energy Consumption Indicator under different systems and M*

regards the estimated versions under very high load, for EeDURD the predictive estimation does not work so well, because the optimal dimensioning is $T^*_{pr} = 1$ and only $L = 54$ preambles are available. The L value is too much low with respect to the number of attempting MTC devices and the estimated value of K_C, as a function of P_C and L, is more approximated. In stead, the predictive estimation in the eDURD system shows good performance because the optimal dimensioning is $T^*_{pr} = 4$, see Fig. 4.9a.

4.8 Appendix

4.8.1 Signaling Overhead for the SCMA allocation

In this section, we derive the number of bits needed by the gNB for delivering to the MTC device the SCMA allocation information. We remind that in an SCMA system, the physical channel corresponds to one or more transmission layers to be used inside the allocated SCMA blocks. In this paper, we assume that the radio resources are allocated in multiples of SBGs, each one consisting of Q subcarriers for the time of 1 TTI. So, considering that the amount of subcarriers allocated to the PUSCH is 72 and that the maximum T_{pu} value is 4 time slot units, the maximum number of available SBGs in one RA cycle is $\frac{72}{Q} \cdot \frac{4T_s}{TTI}$. We also assume that, for each MTC device, the gNB can assign a range of consecutive SBGs and one layer to be used for each SCMA blocks allocated to it. Under this assumption, the total amount of bits required for delivering the SCMA allocation information is equal to:

$$bit_{req} = 2 \cdot \left\lceil \log_2 \left(\frac{72}{Q} \cdot \frac{4T_s}{TTI} \right) \right\rceil + \lceil \log_2 L_{SB} \rceil . \tag{4.30}$$

By considering (4.3) and that $T_s = TTI = 1$ ms, it follows:

$$bit_{req} = 2 \cdot \left\lceil \log_2 \left(\frac{72}{Q} \cdot 4 \right) \right\rceil + \left\lceil \log_2 \binom{Q}{S} \right\rceil . \tag{4.31}$$

Since $2 \cdot \lceil x \rceil \leq \lceil 2 \cdot x \rceil + 1$, it follows:

$$bit_{req} \leq \left\lceil 2 \cdot \log_2 \left(\frac{72}{Q} \cdot 4 \right) \right\rceil + \left\lceil \log_2 \binom{Q}{S} \right\rceil + 1. \tag{4.32}$$

By applying the property $\lceil x \rceil + \lceil y \rceil \leq \lceil x + y \rceil + 1$, it results:

$$bit_{req} \leq \left\lceil 2 \cdot \log_2 \left(\frac{72}{Q} \cdot 4 \right) + \log_2 \binom{Q}{S} \right\rceil + 2. \tag{4.33}$$

After some algebraic manipulations, (4.33) can be written as:

$$\begin{aligned} bit_{req} &= \left\lceil \log_2\left(\frac{72}{Q}\cdot 4\right)^2 + \log_2\left(\frac{Q!}{S!\,(Q-S)!}\right)\right\rceil + 2 \\ &= \left\lceil \log_2\left(\frac{72}{Q}\cdot 4\right)^2 + \log_2\left(\frac{\prod_{i=0}^{S-1}(Q-i)}{S!}\right)\right\rceil + 2 \\ &= \left\lceil \log_2\left(\frac{(72\cdot 4)^2}{S!}\cdot\frac{\prod_{i=1}^{S-1}(Q-i)}{Q}\right)\right\rceil + 2. \end{aligned} \tag{4.34}$$

Finally, by applying the property $\lceil x+y\rceil \leq \lceil x\rceil + \lceil y\rceil$, the number of required bits as function of S and Q is equal to:

$$bit_{req} \leq \left\lceil \log_2\left(\frac{(72\cdot 4)^2}{S!}\right)\right\rceil + \left\lceil \log_2\left(\frac{\prod_{i=1}^{S-1}(Q-i)}{Q}\right)\right\rceil + 2. \tag{4.35}$$

When $S=2$, the number of required bits is given by:

$$bit_{req} \leq \left\lceil \log_2\left(1-\frac{1}{Q}\right)\right\rceil + 18. \tag{4.36}$$

It is obvious that, for each value of $Q>2$, $bit_{req}\leq 18$. This means that the 18 bits reserved for time and frequency allocation in the conventional RAR, are enough to deliver the SCMA allocation information, regardless of the value of Q.

4.8.2 Calculation of $\bar{P}_S$ and $\bar{P}_C$

We define the random variables X_i, with $i\in\{1,\ldots,L\}$, as:

$$X_i = \begin{cases} 1 \text{ if the } i\text{th preamble has been selected only by} \\ \quad \text{one MTC device out of the M} \\ 0 \text{ otherwise} \end{cases} \tag{4.37}$$

Therefore:

$$P_S = \sum_{i=1}^{L} X_i. \tag{4.38}$$

The mean value of P_S, $\bar{P}_S$, is calculated as follows:

$$\bar{P}_S = E\{P_S\} = E\left\{\sum_{i=1}^{L} X_i\right\} = \sum_{i=1}^{L} E\{X_i\}. \tag{4.39}$$

Since X_i are L random variables identically distributed, the mean value $E\{X_i\}$, for each i, is equal to:

$$E\{X_i\} = \Pr(X_i = 1) = M\left(\frac{1}{L}\right)\left(1 - \frac{1}{L}\right)^{M-1}. \tag{4.40}$$

The last step has been calculated by applying the total probability theorem for M independent extractions. So:

$$\bar{P}_S = M\left(1 - \frac{1}{L}\right)^{M-1} = M\left(1 - \frac{1}{L_0 T_{pr}}\right)^{M-1}. \tag{4.41}$$

Now, in order to calculate $\bar{P}_C$, let us start by calculating the total number of preambles out of L transmitted by M MTC devices. It is denoted as P_T. At this aim, we define Y_i, with $i \in \{1, \ldots, L\}$, as Boolean random variables equal to 1 if and only if the ih preamble has been selected by at least one MTC device. Therefore:

$$P_T = \sum_{i=1}^{L} Y_i. \tag{4.42}$$

The mean value of P_T is calculated as follows:

$$\bar{P}_T = E\left\{\sum_{i=1}^{L} Y_i\right\} = \sum_{i=1}^{L} E\{Y_i\}. \tag{4.43}$$

Since Y_i are L random variables identically distributed, the mean value of P_T can be calculated as $\bar{P}_T = LE\{Y_i\}$, where $E\{Y_i\} = \Pr(Y_i = 1) = 1 - (1 - 1/L)^M$, for each i. Then, the average number of collided preambles can be easily calculated as

$$\bar{P}_C = \bar{P}_T - \bar{P}_S = L\left[1 - (1 - 1/L)^M\right] - M(1 - 1/L)^{M-1}. \tag{4.44}$$

4.8.3 Calculation of $\bar{K}_C$

Let $|\mathcal{M}_p|$ be the cardinality of preamble index p, that is, the number of devices that have chosen preamble index p. Firstly, we derive the probability that the preamble p has been selected by exactly h devices, given that it has been selected at least by i MTC devices, with $i = \{1, \ldots, M\}$ and $h \geq i$. Let us define A as the event $|\mathcal{M}_p| = h$ and B as the event $|\mathcal{M}_p| \geq i$. The probability that $|\mathcal{M}_p| = h$, given that $|\mathcal{M}_p| \geq i$, results:

$$Pr\{A \mid B\} = \frac{Pr\{A \cap B\}}{Pr\{B\}}. \tag{4.45}$$

Since $h \geq i$, it follows that the event A implies the event B, so $Pr\{A \cap B\} = Pr\{A\}$. So:

$$Pr\{A \mid B\} = \frac{\binom{M}{h}\left(\frac{1}{L}\right)^h\left(1-\frac{1}{L}\right)^{M-h}}{1-\sum_{j=0}^{i-1}\binom{M}{j}\left(\frac{1}{L}\right)^j\left(1-\frac{1}{L}\right)^{M-j}}. \tag{4.46}$$

Then, we determine the average value of the collision coefficient, $\bar{K}_C$, as:

$$\bar{K}_C = \mathbb{E}\{K_C\} = \sum_{h=2}^{M} hPr\left\{K_C = h\right\}, \tag{4.47}$$

where $Pr\left\{K_C = h\right\}$ is the probability that the collision coefficient K_C assumes the value h. This probability corresponds to $Pr\{A \mid B\}$, where A is the event $|\mathcal{M}_p| = h$ and B is the event $|\mathcal{M}_p| \geq 2$, i.e., the preamble is collided. Therefore:

$$Pr\{K_C = h\} = \frac{\binom{M}{h}\left(\frac{1}{L}\right)^h\left(1-\frac{1}{L}\right)^{M-h}}{1-\sum_{j=0}^{1}\binom{M}{j}\left(\frac{1}{L}\right)^j\left(1-\frac{1}{L}\right)^{M-j}}. \tag{4.48}$$

Let us put $D = 1-\sum_{j=0}^{1}\binom{M}{j}\left(\frac{1}{L}\right)^j\left(1-\frac{1}{L}\right)^{M-j}$, $p = \frac{1}{L}$, and $q = 1-p$. It follows:

$$Pr\{K_C = h\} = \frac{1}{D}\binom{M}{h}p^h q^{M-h}. \tag{4.49}$$

By substituting (4.49) in (4.47), it follows:

$$\bar{K}_C = \frac{1}{D}\sum_{h=2}^{M} h\binom{M}{h}p^h q^{M-h} = \frac{1}{D}\left[\sum_{h=0}^{M} h\binom{M}{h}p^h q^{M-h} - \sum_{h=0}^{1} h\binom{M}{h}p^h q^{M-h}\right] = \frac{1}{D}\left[\sum_{h=0}^{M} h\binom{M}{h}p^h q^{M-h} - Mpq^{M-1}\right]. \tag{4.50}$$

The first addend inside the squared brackets is the average value of a binomial distribution with parameters M and p, so, it is equal to Mp. Then,

$$\bar{K}_C = \frac{Mp}{D}[1-q^{M-1}] = \frac{M\left(\frac{1}{L}\right)\left[1-\left(1-\frac{1}{L}\right)^{M-1}\right]}{1-\left(1-\frac{1}{L}\right)^{M} - M\left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)^{M-1}}. \tag{4.51}$$

$$Pr\{K_C = h\} = \binom{M-2}{h-2}\left(\frac{1}{L}\right)^{h-2}\left(1-\frac{1}{L}\right)^{M-h}, \tag{4.52}$$

with $h = \{2, \ldots, M\}$.

Let us put $j = h-2$, $n = M-2$, $p = \frac{1}{L}$, and $q = 1-p$. Then:

$$\bar{K}_C = \sum_{j=0}^{n}(j+2)\binom{n}{j}p^j q^{n-j} = \sum_{j=0}^{n} j\binom{n}{j}p^j q^{n-j} + 2\sum_{j=0}^{n}\binom{n}{j}p^j q^{n-j}. \tag{4.53}$$

By exploiting the binomial formula, it follows:

$$\bar{K}_C = \sum_{j=0}^{n} j\binom{n}{j}p^j q^{n-j} + 2(p+q)^n. \tag{4.54}$$

Since the first addend of the second member is the average value of a binomial distribution with parameters n and p, and the second added is

equal to 2, it follows:

$$\bar{K}_C = np + 2 = \frac{M-2}{L} + 2. \tag{4.55}$$

4.9 Conclusion

In this chapter, we analyzed the radio resource allocation problem in an mMTC scenario in order to manage a massive number of MTC devices requiring the transmission of small-sized data. We proposed a dynamical uplink radio resource allocation, termed DURD, which optimizes the resource allocation between the PRACH and the PUSCH according to the traffic load condition in every RA cycle. In order to reduce the energy consumption of the MTC devices, we proposed an advanced two-step RA procedure, which permits to the gNB to detect the preamble collisions. This two-step RA procedure reduces the energy consumption per RA attempt in comparison to the legacy 4-step one. As regards the PUSCH, at the aim of properly multiplexing a large number of small-sized data, we adopted the SCMA technique to properly allocated radio resources. Simulation results showed that the proposed dynamic dimensioning control allows to achieve significant improvements in terms of both system throughput and MTC device energy consumption compared to a traditional static dimensioning.

Moreover, in light of this promising performance, we proposed an enhanced version of the DURD, denoted as EDURD, to exploit also the unused PUSCH resources, if any, in a contention-based mode. We presented an analytic problem formulation for determining the optimal number of collided preambles to which allocate the unused PUSCH resources in contention-based mode, and then we proposed an iterative algorithm, termed PRRA, to implement our strategy. Simulation results showed that EDURD exhibits better performance, in terms of throughput gain, with respect to the classic ones and the DURD at the cost of negligible increment of the energy consumption in very high load condition.

Finally, in order to make our solution viable, we propose a predictive estimate of expected traffic based only on information available at the gNodeB. This estimate resulted sufficiently accurate for the DURD, but not for the EDURD.

CHAPTER 5

Modeling and Analysis of Tagged Preamble Transmissions for mMTC scenarios

In the previous chapter, we adopt a 2-step RRC connectionless RA procedure based on the transmission of tagged preambles. This evolved grant-based RA procedure can allow to reduce the number of signaling transmissions per access attempt and reduce to 2 the number of steps in the RA procedure. However, the expected improved performance for these advanced RA procedures is achieved only if the gNB receiver detector works accurately in the first step, that is, if all and only the preamble-tag pairs transmitted by the devices are detected. For this reason, in this chapter we analyze the reception of both the preambles and tags.

In this regard, we observe that unlike the conventional preamble transmission in which all the transmitted sequences have the same root r, here the gNB receives the sum of several sequences composed of different roots. This strongly affects the correct detection of both preambles and tags, because preambles and tags generated from different root sequences are not orthogonal. The higher the number of attempting devices, the higher the average number of selected preambles, the higher

the number of different ZC roots. Consequently, in a mMTC scenario the interference can be so high that the performance of the gNB receiver is compromised. The tagged preamble sequence detection is hard to model analytically, so its performance evaluation is often reported by means of simulation results [14, 69, 70]. In [14] we assumed an ideal preamble-tag detector, whereas in [69, 70] the authors formulate the problem of preamble-tag detection by introducing the concept of detection thresholds for both preambles and tags. However, they just derive an analytical expression for the preamble and tag detection probabilities, without providing a solution either in closed form or numerically, and based on *a posteriori* information. In fact, the detection performance is only provided by means of average values over a large number of simulations. As reported in [92], simulation results are typically highly time-consuming in a massive scenario and the obtained results are not easily reproducible. For this reason, an analytical model that provides a closed-form solution is valuable. To our best knowledge, in literature no one has yet dealt with evaluating it. The contributions of our chapter can be summarized as follows.

1. As the main contribution, we present a rigorous methodology for modeling and analyzing the signal processed by the gNB receiver at the first step of the RA procedure, when tagged preamble sequences are transmitted in mMTC scenarios. We derive an *exact* expression for the detection probability distributions of both the preamble and tag at the receiver, in presence of Noise and Interference due to other preambles and tags.

2. A strength of our modeling, in addition to its demonstrated accuracy, is that the expressions obtained are in closed form, and we use them to analytically derive the threshold values to detect preambles and tags with a given probability.

3. The proposed models are powerful and computationally simple tools that can be used for determining issues and limits of tagged preamble detection strategies in order to investigate and suggest new ones. For example, we limit the working zones in which a simple detection strategy can work well; we also evaluate the impact of failed

tagged preamble detection rates on the signaling overhead and energy consumption for the MTC devices.

5.1 Background

5.1.1 Conventional Preamble

In LTE, the RA procedure is done by means of ZC sequences [75]. They are defined as:

$$z_r[n] = \exp\left[-j\frac{\pi r n(n+1)}{N_{ZC}}\right], \tag{5.1}$$

for $n = 0, ..., (N_{ZC} - 1)$, where N_{ZC} denotes the ZC sequence length, and r is the root index. This latter is an integer value less than N_{ZC} which is chosen to be relatively prime respect to N_{ZC}. Let us note that if N_{ZC} is an odd prime value, then $r \in \{1, \ldots, (N_{ZC} - 1)\}$. For this reason, typically N_{ZC} is chosen to be an odd prime value, as we assume in the following.

A preamble sequence is generated cyclically shifting the reference ZC sequence (5.1) by pN_{CS}. It is denoted as:

$$z_r^p[n] = z_r[(n + pN_{CS})_{N_{ZC}}], \tag{5.2}$$

where p is the preamble index randomly selected in the set $\{1, ..., N_{P_T}\}$, $N_{P_T} = \lfloor N_{ZC}/N_{CS} \rfloor$, N_{CS} is the preamble cyclic shift, and $(\cdot)_{N_{ZC}}$ denotes the modulo-N_{ZC} operation. We note that N_{CS} and r are broadcast by the eNodeB as part of the system information. Among all N_{P_T} available preambles, $N_P \leq N_{P_T}$ are reserved for the contention-based RA procedure, while the remaining ones are reserved for the collision-free Access procedure. The ZC sequences have the following properties:

1. Ideal cyclic auto-correlation. The correlation function, denoted as $C_{z_r^p, z_r}[\tau]$, between a sequence $z_r^p[n]$ and the reference sequence (5.1) with the same root r and cyclic shift equal to 0, is non-zero only in $\tau = pN_{CS}$, in which it assumes a positive real value.

$$\begin{aligned} C_{z_r^p, z_r}[\tau] =& \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} z_r^p[n] z_r^*[n+\tau] = \\ =& \sqrt{N_{ZC}} \quad \delta[\tau - pN_{CS}], \end{aligned} \tag{5.3}$$

where $(\cdot)^*$ denotes the complex conjugation and $\delta[\tau]$ is the discrete-time unit impulse function. We exploit this property to derive how much a received sequence $z_r^p[n]$ is shifted (i.e., which preamble index p has been selected), by calculating the real part of the correlation function $C_{z_r^p,z_r}[\tau]$, denoted as $\Re\left\{C_{z_r^p,z_r}[\tau]\right\}$.

2. Cyclic cross-correlation. The absolute value of the cyclic cross-correlation between two ZC sequences with different root numbers, denoted as $C_{z_r^p,z_k}$, is equal to 1 for each τ.

$$|C_{z_r^p,z_k}[\tau]| = \frac{1}{\sqrt{N_{ZC}}}\left|\sum_{n=0}^{N_{ZC}-1} z_r^p z_k^*[n+\tau]\right| = 1, \tag{5.4}$$

where $k \in \{1,\ldots,(N_{ZC}-1)\}$, $k \neq r$. So, let us underline that $\Re\left\{C_{z_r^p,z_r}[\tau]\right\}$, for each τ, assumes a generally different value belonging to the interval $[-1,1]$.

5.1.2 Tagged preambles

The tagged preamble sequence consists of both a preamble and a tag ZC sequence, which are transmitted together [69, 70]. In particular, the tagged preamble sequence is defined as:

$$x_{r,k_p}^{p,t}[n] = P_T(z_r^p[n] + z_{k_p}^t[n]), \tag{5.5}$$

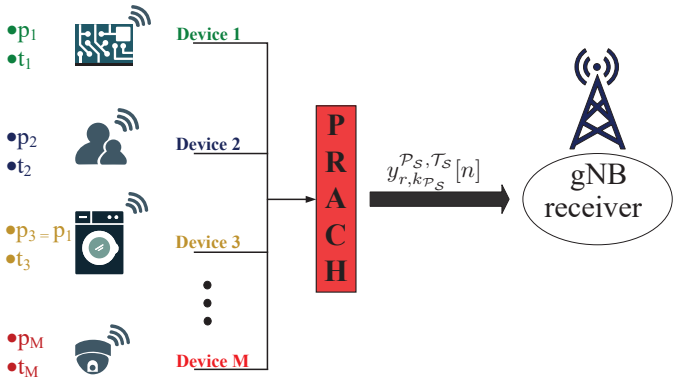
where P_T denotes the transmit power, and $z_{k_p}^t[n]$ the tag sequence. It is expressed as:

$$z_{k_p}^t = z_{k_p}[(n+tN_{TS})_{N_{ZC}}], \tag{5.6}$$

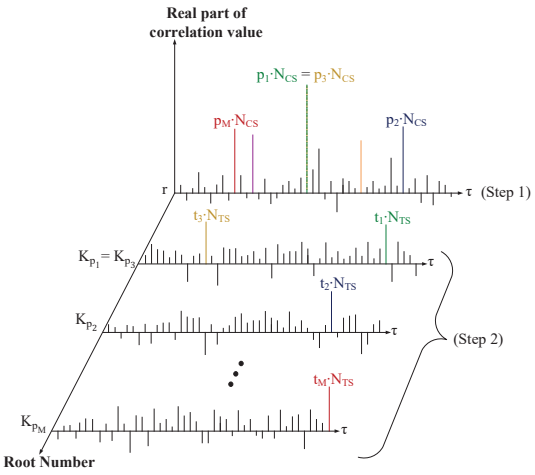
where k_p indicates the tag root number related to the pth preamble index, N_{TS} the tag cyclic shift value, and t denotes the tag randomly selected in $\mathcal{T} = \{1,...,(N_T)\}$, in which $N_T = \lfloor N_{ZC}/N_{TS} \rfloor$ is the overall number of tags. The root k_p is determined by a fixed mapping function of the selected preamble index p, i.e., each preamble is mapped into a specific tag root number $k_p \neq r$.

5.2 System Model and Issues

We are considering the contention-based RA procedure in a PRACH in a given instant time, as shown in Fig. 5.1a. Let $\mathcal{M} = \{1,\ldots,M\}$ denote



(a) *First step of the Random Access procedure.*



(b) *Representation of the real part of the correlations calculated at the receiver.*

Figure 5.1: *An example of the actions done for detecting preambles and tags.*

the set of attempting devices[1] using the same preamble root number r on the same PRACH, and $\mathcal{P} = \{1, \ldots, N_P\}$ denotes the set of total available preambles for the contention-based procedure.

Each device i, with $i \in \mathcal{M}$, selects a preamble $p_i \in \mathcal{P}$ and a tag $t_i \in \mathcal{T}$. Also, let us define:

- **A** as the preamble association matrix of size $M \times N_P$, where each element $a_{i,j}$ is a boolean equal to 1 if and only if device i has selected preamble j,

- **B** as the tag association matrix of size $M \times N_T$, where each element

[1] $\mathcal{M}$ includes the devices that perform the first access attempt and the re-attempting ones that collided in the past RA cycles.

$b_{i,j}$ is a boolean value equal to 1 if and only if device i has selected tag j,

- $\mathcal{P}_\mathcal{S} \subseteq \mathcal{P}$ as the subset of preambles that have been selected by the M devices to perform the RA procedure, i.e., $\mathcal{P}_\mathcal{S} = \{j \mid j \in \mathcal{P} \text{ and } \sum_{i=1}^{M} a_{i,j} > 0\}$,
- $\mathcal{T}_\mathcal{S} \subseteq \mathcal{T}$ as the subset of tags that have been selected by the M devices, i.e., $\mathcal{T}_\mathcal{S} = \{j \mid j \in \mathcal{T} \text{ and } \sum_{i=1}^{M} b_{i,j} > 0\}$.

For each device i, $\sum_{j=1}^{N_P} a_{i,j} = 1$, and $\sum_{j=1}^{N_T} b_{i,j} = 1$. In addition, for each preamble $p \in \mathcal{P}$,

- $\mathcal{M}_p \subseteq \mathcal{M}$ denotes the subset of attempting devices that have chosen the same preamble index p, i.e., $\mathcal{M}_p = \{i \mid i \in \mathcal{M} \text{ and } a_{i,p} = 1\}$.
- $|\mathcal{M}_p|$ as the cardinality of preamble index p, that is, the number of devices that have chosen preamble index p. Let us underline that $\mathcal{M}_p$ can be empty, so $|\mathcal{M}_p|$ can be equal to 0.

For each pair (p, t), with $p \in \mathcal{P}$ and $t \in \mathcal{T}$, $\mathcal{M}_{(p,t)} \subseteq \mathcal{M}$ denotes the subset of attempting devices that have chosen the same pair (p, t), i.e., $\mathcal{M}_{(p,t)} = \{i \mid i \in \mathcal{M}, a_{i,p} = 1 \text{ and } b_{i,t} = 1\}$.

We define $\mathcal{P}_h \subseteq \mathcal{P}$, with $h = 0, 1, \ldots, M$, as the subset of preambles that have been chosen by an amount of devices equal to h, i.e., $\mathcal{P}_h = \{p \mid p \in \mathcal{P} \text{ and } |\mathcal{M}_p| = h\}$. Also in this case, $\mathcal{P}_h$ can be empty. It follows that

$$\mathcal{P} = \bigcup_{h=0,1,\ldots,M} \mathcal{P}_h. \tag{5.7}$$

Finally, for each selected preamble $p_s \in \mathcal{P}_\mathcal{S}$, we define $\mathcal{M}_{p_s} \subseteq \mathcal{M}$ as the subset of attempting devices that have chosen the preamble p_s, and $\mathcal{T}_{p_s} \subseteq \mathcal{T}$ as the subset of selected tags that are associated to the preamble p_s, i.e., $\mathcal{T}_{p_s} = \{j \mid j \in \mathcal{T}_\mathcal{S} \text{ and } b_{i,j} = 1, \text{ with } i \in \mathcal{M}_{p_s}\}$.

5.2.1 Issues on the preamble and tag detection procedure

In order to identify the issues for a correct preamble and tag detection, we provide here an overview of how the gNB receiver works. As shown

in Fig. 5.1a, the attempting device i choses in a random way the pair (p_i, t_i) and transmits in the PRACH the following tagged preamble:

$$x_{r,k_{p_i}}^{p_i,t_i}[n] = P_{T_i}(z_r^{p_i}[n] + z_{k_{p_i}}^{t_i}[n]). \tag{5.8}$$

We denote the total received sequence at the gNB with $y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n]$, since it is a function of the set of selected preambles $\mathcal{P}_\mathcal{S}$ and tags $\mathcal{T}_\mathcal{S}$.

$$y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n] = \sum_{i=1}^{M}\sum_{g=1}^{G_i} h_{i,g} x_{r,k_{p_i}}^{p_i,t_i}[(n + d_{i,g})_{N_{ZC}}] + N[n], \tag{5.9}$$

where G_i denotes the number of multi-paths for each device i, $h_{i,g}$ denotes the channel gain of the gth path related to device i, and $d_{i,g}$ the propagation delay of the gth path related to device i. $N[n] = \Re\{N[n]\} + j\Im\{N[n]\}$ represents the circular symmetry complex Gaussian noise, where $\Re\{N[n]\} \sim \mathcal{N}(0, \sigma_N^2)$ and $\Im\{N[n]\} \sim \mathcal{N}(0, \sigma_N^2)$.

For the sake of simplicity, in the following we assume $P_{T_i} = 1$, $G_i = 1$, $h_{i,g} = 1$, and $d_{i,g} = 0, \forall i \in \mathcal{M}$. The impact of the multi-path propagation, the non-ideal channel conditions, and the propagation delay on the preamble and tag detection are analyzed in the last Section.

We denote as SNR, the following Signal-to-Noise Ratio measured at the gNB receiver:

$$SNR = \frac{\sum_{n=0}^{N_{ZC}-1} \left| \sum_{i=1}^{M} x_{r,k_{p_i}}^{p_i,t_i}[n] \right|^2}{2\sigma_N^2}. \tag{5.10}$$

Once the sequence (5.9) has been received by the gNB, it should detect the transmitted M preamble-tag pairs by performing some actions that can be divided into 2 steps.

Step 1. Since each device $i \in \mathcal{M}$ uses the same root number r to transmit its preamble sequence, the gNB computes the correlation value $C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau]$ between the total received sequence $y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n]$ and the reference sequence (5.1), in order to detect all the transmitted preambles $p_s \in \mathcal{P}_\mathcal{S}$. In particular, by exploiting the property (i), we consider the real part of the correlation (5.3), denoted as $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau]\right\}$.

At the aim of clarifying this step, we show in Fig. 5.2 an example of the correlation values assumed when $M = 15$. Fig. 5.2a shows

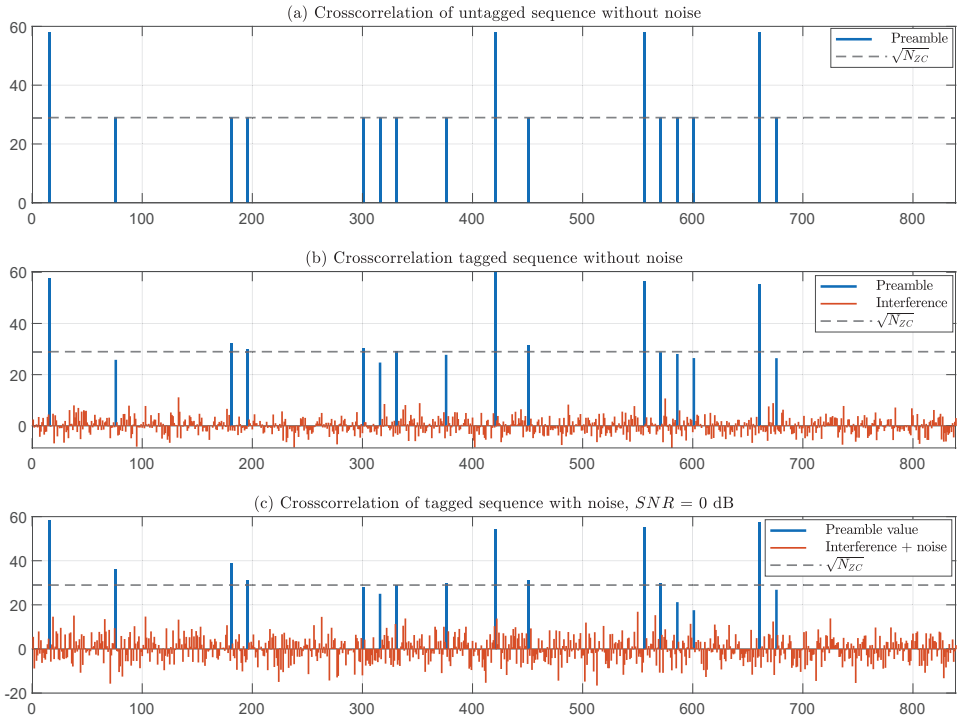


Figure 5.2: *Example of the real part of the correlation* $C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_r}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]$ *when* $M = 15$ *and* $N_{ZC} = 839$.

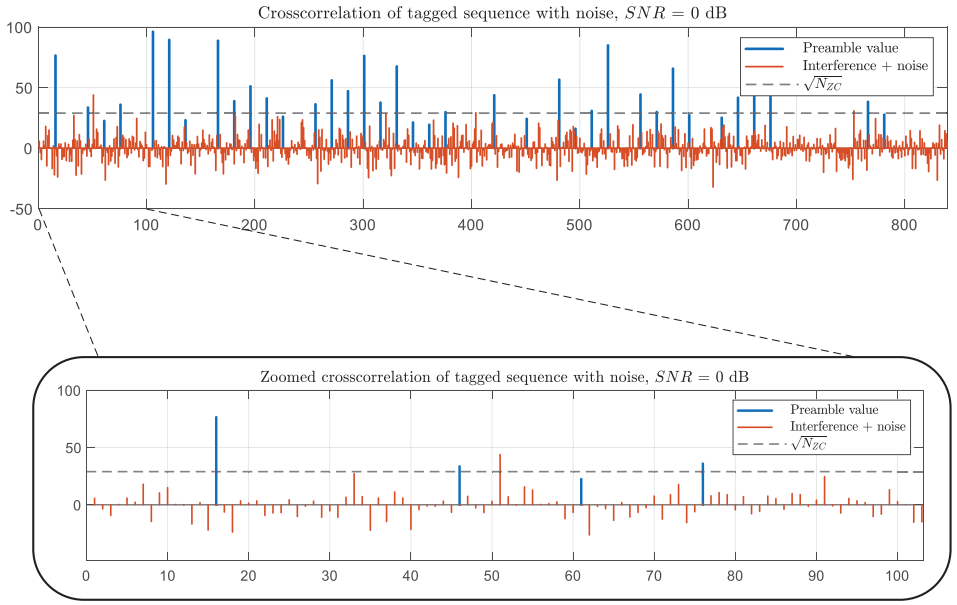


Figure 5.3: *Example of the real part of the correlation* $C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_r}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]$ *when* $M = 50$ *devices attempt the access with tagged preambles and in the presence of noise.*

$\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_r}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]\right\}$ in the event that the sequences $x_{r,k_{p_i}}^{p_i,t_i}[n]$ were transmit-

ted without tags, i.e., $z^{t_i}_{k_{p_i}}[n] = 0$, $\forall i \in \mathcal{M}$, and in the absence of noise, i.e., $N[n] = 0$. As expected, in each point $\tau = p_i N_{CS}$ the correlation value is exactly equal to $|\mathcal{M}_{p_i}|\sqrt{N_{ZC}}$, while 0 in all the other points. In Fig.5.2b each device transmits a tagged sequence, i.e., $z^{t_i}_{k_{p_i}}[n] \neq 0$, $\forall i \in \mathcal{M}$, in the absence of noise. Because of the property (ii), the sum of M cross-correlations produces non-zero values at each point n of the sequence with values that range, in general, in $[-M, M]$. We underline that these values are the interference that affects the ability to correctly detect the preambles. Consequently, not only we obtain non-zero values at the points where no preamble has been selected, but also at $\tau = p_i N_{CS}$ the cross-correlation value can be higher or lower than $|\mathcal{M}_{p_i}|\sqrt{N_{ZC}}$. Finally, in Fig.5.2c we consider the case in which also $SNR = 0$ dB. In this case, the interference plus noise causes the correlation $\Re\left\{C_{y^{\mathcal{P}_S,\mathcal{T}_S}_{r,k_{\mathcal{P}_S}},z_r}[\tau]\right\}$ to assume values that could be very different from the ideal ones shown at the top of the figure.

The latter effect is even more evident as the attempting device number increases, as shown in Fig. 5.3, when $M = 50$. In particular, the zoom at the bottom shows that the correlation value in some points without any selected preamble could exceed the value $\sqrt{N_{ZC}}$ due to the interference, while the correlation value could be lower than $\sqrt{N_{ZC}}$ in some points related to a selected preamble. It can be easily understood that this effect affects the number of preambles correctly detected, both in terms of undetected transmitted preambles (i.e., false negative preambles) and in terms of detected non-transmitted preambles (i.e., false positive preambles). This phenomenon is the focus of next Section in order to obtain adequate probabilistic tools for the correct detection of the transmitted preambles.

Once the preamble detection procedure has been completed, the gNB obtains a new set of detected preambles denoted as $\mathcal{P}_\mathcal{D} \subseteq \mathcal{P}$, and in the ideal condition $\mathcal{P}_\mathcal{D} = \mathcal{P}_\mathcal{S}$.

Step 2. For each preamble $p_s \in \mathcal{P}_\mathcal{D}$ that has been detected by the gNB, a new correlation between the received sequence $y^{\mathcal{P}_S,\mathcal{T}_S}_{r,k_{\mathcal{P}_S}}[n]$ and the reference ZC sequence with root k_{p_s}, as function of p_s, should be calculated to detect all tags in $\mathcal{T}_{p_s}$. The aim is to check whether the preamble p_s has been selected by more than one device, i.e., to detect immediately

whether it has collided. Also in this case, the noise and the interference due to the cross-correlations cause that the correlation values oscillate respect to the ideal values. In Section 5.4, we study in detail this phenomenon and propose an analytical model.

5.2.2 An example of the detection procedure

In Fig. 5.1a we illustrated an example of the first step of the RA procedure. Each MTC device selects in a random way one preamble-tag pair (p_i, t_i). We note that the MTC device 1 selects the same preamble of MTC device 3, but a different tag. In Fig. 5.1b we show step 1 and 2 of the detection procedure in successful conditions, i.e., all preambles and all tags sent are correctly detected. During step 1, the gNB calculates the real part of the correlation between the received sequence $y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n]$ and the reference sequence $z_r[n]$. In the point $p_1 N_{CS} = p_3 N_{CS}$ the correlation value is higher than the one assumed by the other preambles that have been selected once. After the gNB has detected all transmitted preambles, step 2 begins. It calculates, for each $p_s \in \mathcal{P}_\mathcal{D}$, the real part of the correlation between the received sequence and the sequence $z_{k_{p_s}}[n]$. We note that, since $p_1 = p_3$, two different tags are found in the correlation with the sequence $z_{k_{p_1}} = z_{k_{p_3}}$, then the collision is detected. Instead, in the correlations with $z_{k_{p_2}}[n]$ and $z_{k_{p_M}}[n]$ only one tag is found, i.e., preambles p_2 and p_M were not collided.

5.3 Modeling for the Preamble Detection

In this section, we derive the analytical model useful for the detection of the preambles transmitted in the presence of noise and interference due to the tagged sequences.

To detect whether a preamble $p \in \mathcal{P}$ has been selected by at least one device (i.e., if $p \in \mathcal{P}_\mathcal{S}$), the gNB calculates the correlation value between the received sequence (5.9) and the reference sequence (5.1). The correlation value is:

$$C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau] = \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n] z_r^*[n+\tau] =$$
$$= \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} \left(\sum_{i=1}^{M} x_{r,k_{p_i}}^{p_i,t_i}[n] + N[n] \right) z_r^*[n+\tau] = \tag{5.11}$$
$$= \frac{1}{\sqrt{N_{ZC}}} \sum_{m=1}^{M} \sum_{n=0}^{N_{ZC}-1} x_{r,k_{p_i}}^{p_i,t_i}[n] z_r^*[n+\tau] + N[\tau].$$

Let us put

$$C_{z_r^{p_i},z_r}[\tau] = \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} z_r^{p_i}[n] z_r^*[n+\tau], \tag{5.12}$$

and

$$C_{z_{k_{p_i}}^{t_i},z_r}[\tau] = \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} z_{k_{p_i}}^{t_i}[n] z_r^*[n+\tau]. \tag{5.13}$$

Then:

$$C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau] = \sum_{i=1}^{M} (C_{z_r^{p_i},z_r}[\tau] + C_{z_{k_{p_i}}^{t_i},z_r}[\tau]) + N[\tau]. \tag{5.14}$$

Let us analyze the real part of (5.14):

$$\Re\left\{ C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau] \right\} = \sum_{i=1}^{M} \Re\left\{ C_{z_r^{p_i},z_r}[\tau] \right\} + \\ + \sum_{i=1}^{M} \Re\left\{ C_{z_{k_{p_i}}^{t_i},z_r}[\tau] \right\} + \Re\{N[\tau]\}. \tag{5.15}$$

By applying property (5.3), it follows

$$\Re\left\{ C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau] \right\} = \sum_{p_s \in \mathcal{P}_\mathcal{S}} |\mathcal{M}_{p_s}| \sqrt{N_{ZC}} \quad \delta[\tau - p_s N_{CS}] + \\ + \sum_{i=1}^{M} \Re\left\{ C_{z_{k_{p_i}}^{t_i},z_r}[\tau] \right\} + \Re\{N[n]\}. \tag{5.16}$$

The term $\Re\left\{C_{z_{k_{p_i}}^{t_i},z_r}[\tau]\right\}$ can be written as:

$$\Re\left\{C_{z_{k_{p_i}}^{t_i},z_r}[\tau]\right\} = \frac{1}{\sqrt{N_{ZC}}}\sum_{n=0}^{N_{ZC}-1}\cos\left(\theta_{k_{p_i},r}^{t_i}[n,\tau]\right), \tag{5.17}$$

where $\theta_{k_{p_i},r}^{t_i}[n,\tau] = \frac{\pi}{N_{ZC}}\{k_{p_i}(n+t_iN_{TS})_{N_{ZC}}((n+t_iN_{TS})_{N_{ZC}}+1)) - r(n+\tau)_{N_{ZC}}((n+\tau)_{N_{ZC}}+1))\}$. At a general time instant τ, the cross-correlation $\Re\left\{C_{z_{k_{p_i}}^{t_i},z_r}[\tau]\right\}$ becomes a random variable, denoted in the following as H_i, for $i = 1,\ldots,M$ with expected value $\mu_{H_i} = 0$, and variance $\sigma_{H_i}^2 = \frac{1}{2}$.

Let us note that each random variable H_i depends on the pair (p_i,t_i) that has been selected by the device i. Since the probability that at least two devices select the same pair (p,t) can be neglected, the sequence of random variables $\{H_1,\ldots,H_M\}$ is a sequence of independent and identically distributed (i.i.d.) variables drawn from the same distribution with finite variance. Then, by exploiting the central limit theorem, the distribution of $H = \sum_{i=1}^{M} H_i$ approaches a Gaussian distribution with $\mu_H = M\mu_{H_i} = 0$ and $\sigma_H^2 = M\sigma_{H_i}^2 = \frac{M}{2}$, as M gets larger. Let us note that the above analysis would not work properly when the number of devices M is low; however, in this case, there would be no significant interference and collision issues. Instead, the target of this chapter is to support an mMTC scenario, so the M value is enough large that the hypothesis of convergence of the central limit theorem is valid.

Assessing the relationship (5.16) as a whole, at each time $\tau' \neq p_sN_{CS}$, $\forall p_s \in \mathcal{P}_\mathcal{S}$, we note that the first addend of the second member of (5.16) is equal to 0, the second addend (which corresponds to H) follows the distribution $\mathcal{N}\left(0,\frac{M}{2}\right)$ and the third addend follows the distribution $\mathcal{N}\left(0,\sigma_N^2\right)$. Thus, the Cumulative Distribution Function (CDF) of the random variable $\Phi = \Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau']\right\}$, for $\tau' \neq p_sN_{CS}, \forall p_s \in \mathcal{P}_\mathcal{S}$, is equal to:

$$F_\Phi(\phi) = Pr\{\Phi \leq \phi\} = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\phi}{\sqrt{M+2\sigma_N^2}}\right)\right], \tag{5.18}$$

where $\mathrm{erf}(\cdot)$ is the Gauss error function.

Now, let us analyze (5.16), when $\tau'' = p_s N_{CS}, \forall p_s \in \mathcal{P}_\mathcal{S}$. The first addend of the second member of (5.16) is equal to $|\mathcal{M}_{p_s}|\sqrt{N_{ZC}}$, while the second and the third one follow the distribution $\mathcal{N}\left(0, \frac{M}{2} + \sigma_N^2\right)$. Thus, the CDF of the random variable $\Phi_P = \Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}, z_r}[\tau'']\right\}$, for $\tau'' = p_s N_{CS}, \forall p_s \in \mathcal{P}_\mathcal{S}$, is equal to:

$$F_{\Phi_p}(\phi) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\phi - |\mathcal{M}_{p_s}|\sqrt{N_{ZC}}}{\sqrt{M + 2\sigma_N^2}}\right)\right]. \tag{5.19}$$

5.4 Modeling for the Tag Detection

In this section, we derive the analytical model for the detection of the tags transmitted in the presence of noise and interference due to the other preambles and related tags. For each preamble $p_s \in \mathcal{P}_\mathcal{D}$ that has been correctly detected by the gNB, a new correlation between the received sequence $y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n]$ and the reference ZC sequence with root k_{p_s}, as function of p_s, is calculated to detect tags in $\mathcal{T}_{p_s}$. The aim is to check whether the preamble p_s has been selected by more than one device, i.e., it has collided. The cross-correlation is:

$$C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}, z_{k_{p_s}}}[\tau] = \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[n] z_{k_{ps}}^*[n+\tau]. \tag{5.20}$$

Let us put

$$C_{z_{k_{p_i}}^{t_i}, z_{k_{p_s}}}[\tau] = \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} z_{k_{p_i}}^{t_i}[n] z_{k_{ps}}^*[n+\tau], \tag{5.21}$$

and

$$C_{z_r^{p_i}, z_{k_{p_s}}}[\tau] = \frac{1}{\sqrt{N_{ZC}}} \sum_{n=0}^{N_{ZC}-1} z_r^{p_i}[n] z_{k_{ps}}^*[n+\tau]. \tag{5.22}$$

Then

$$C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}, z_{k_{p_s}}}[\tau] = \sum_{i=1}^{M} C_{z_{k_{p_i}}^{t_i}, z_{k_{p_s}}}[\tau] + \sum_{i=1}^{M} C_{z_r^{p_i}, z_{k_{p_s}}}[\tau] + N[\tau]. \tag{5.23}$$

Now, let us analyze the real part of (5.23), which can be rewritten as:

$$\begin{aligned}\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_{k_{p_s}}}[\tau]\right\} = \sum_{t_i\in\mathcal{T}_{ps}}\sqrt{N_{ZC}}\quad \delta[\tau - t_i N_{TS}] + \\ + \sum_{i\in\mathcal{M}-\mathcal{M}_{p_s}}\Re\left\{C_{z_{k_{p_i}}^{t_i},z_{k_{p_s}}}[\tau]\right\} + \\ + \sum_{i=1}^{M}\Re\left\{C_{z_r^{p_i},z_{k_{p_s}}}[\tau]\right\} + \Re\{N[\tau]\}.\end{aligned} \tag{5.24}$$

Let us underline that in the first addend of the second member we have neglected the possibility that more than one device has selected the same preamble-tag pair.

The term $\Re\left\{C_{z_{k_{p_i}}^{t_i},z_{k_{p_s}}}[\tau]\right\}$ can be written as:

$$\Re\left\{C_{z_{k_{p_i}}^{t_i},z_{k_{p_s}}}[\tau]\right\} = \frac{1}{\sqrt{N_{ZC}}}\sum_{n=0}^{N_{ZC}-1}\cos\left(\theta_{k_{p_i},k_{p_s}}^{t_i}[n,\tau]\right), \tag{5.25}$$

where $\theta_{k_{p_i},k_{p_s}}^{t_i}[n,\tau] = \frac{\pi}{N_{ZC}}\{k_{p_i}(n+t_iN_{TS})_{N_{ZC}}((n+t_iN_{TS})_{N_{ZC}}+1)) - k_{p_s}(n+\tau)_{N_{ZC}}((n+\tau)_{N_{ZC}}+1))\}$. At a general time instant τ, the cross-correlation $\Re\left\{C_{z_{k_{p_i}}^{t_i},z_{k_{p_s}}}[\tau]\right\}$ becomes a random variable denoted as I_i with expected value $\mu_{I_i} = 0$, and variance $\sigma_{I_i}^2 = \frac{1}{2}$. By exploiting the central limit theorem, and considering that no device has selected the same pair (p,t) of another device, the distribution $I = \sum_{i\in\mathcal{M}-|\mathcal{M}_{p_s}|} I_i$ approaches a Gaussian distribution with $\mu_I = 0$ and $\sigma_I^2 = \frac{1}{2}\left(M - |\mathcal{M}_{p_s}|\right)$, as M gets larger. The main problem of this result is that $|\mathcal{M}_{p_s}|$ is not known a priori. To overcome this issue, we replaced the variance σ_I^2 with its expected value $\mathbb{E}\{\sigma_I^2\} = \frac{M}{2} - \frac{\mathbb{E}\{|\mathcal{M}_{ps}|\}}{2}$, where $\mathbb{E}\{|\mathcal{M}_{p_s}|\}$ becomes a known value given the number of devices M and the number of available preambles N_P, as proved in Section 5.6.1. Under the above approximation, it follows that

$$I \sim \mathcal{N}\left(0, \frac{M}{2} - \frac{M-1}{2N_P} - \frac{1}{2}\right). \tag{5.26}$$

As regards the term $\Re\left\{C_{z_r^{p_i},z_{k_{p_s}}}[\tau]\right\}$ related to the third addend of

(5.24), it can be written as:

$$\Re\left\{C_{z_r^{p_i},z_{k_{ps}}}[\tau]\right\} = \frac{1}{\sqrt{N_{ZC}}}\sum_{n=0}^{N_{ZC}-1}\cos\left(\theta_{r,k_{ps}}^{p_i}[n,\tau]\right), \tag{5.27}$$

where $\theta_{r,k_{ps}}^{p_i}[n,\tau'] = \frac{\pi}{N_{ZC}}\{r(n+p_iN_{CS})_{N_{ZC}}((n+p_iN_{CS})_{N_{ZC}}+1)) - k_{ps}(n+\tau)_{N_{ZC}}((n+\tau)_{N_{ZC}}+1))\}$. At a general time instant τ, the cross-correlation $\Re\left\{C_{z_r^{p_i},k_{ps}}[\tau]\right\}$ becomes a random variable denoted as L_{p_i}, with expected value $\mu_{L_{p_i}} = 0$, and variance $\sigma^2_{L_{p_i}} = \frac{1}{2}$.

However, for the random variable $L = \sum_{i\in\mathcal{M}} L_{p_i}$ the hypotheses of the central limit theorem are no longer valid. In fact, in this case the value assumed by L_{p_i} depends only on the preamble p_i selected, and the probability that at least two devices have chosen the same preamble p_i cannot be neglected. Consequently, we can no longer consider all the variables L_{p_i}, with $i = 1,\ldots,M$, as independent variables. To overcome this problem, we rewrite the third addend of the second member of (5.24) as follows:

$$\begin{aligned} L &= \sum_{i=1}^{M}\Re\left\{C_{z_r^{p_i},z_{k_{ps}}}[\tau]\right\} = \sum_{p\in\mathcal{P}}|\mathcal{M}_p|\Re\left\{C_{z_r^{p},z_{k_{ps}}}[\tau]\right\} = \\ &= \frac{1}{\sqrt{N_{ZC}}}\sum_{p\in\mathcal{P}}\sum_{n=0}^{N_{ZC}-1}|\mathcal{M}_p|\cos\left(\theta_{r,k_{ps}}^{p}[n,\tau]\right) = \\ &= \sum_{p\in\mathcal{P}}|\mathcal{M}_p|L_p. \end{aligned} \tag{5.28}$$

where we imposed $L_p = \Re\left\{C_{z_r^{p},z_{k_{ps}}}[\tau]\right\}$, with $p\in\mathcal{P}$. By applying (5.7) in (5.28), it follows:

$$L = \sum_{h=0}^{M}\sum_{p\in\mathcal{P}_h}hL_p = 0 + \sum_{p\in\mathcal{P}_1}L_p + \sum_{p\in\mathcal{P}_2}2L_p + \cdots\sum_{p\in\mathcal{P}_M}ML_p. \tag{5.29}$$

Each random variable hL_p is characterized by the expected value $\mu_{hL_p} = 0$, and variance $\sigma^2_{hL_p} = \frac{h^2}{2}$. So, if each set $\mathcal{P}_h$ is either empty or with great cardinality, then the central limit theorem can be applied for each addend of (5.29), i.e., $\sum_{p\in\mathcal{P}_h}hL_p$ approaches a Gaussian distribution with mean

value 0 and variance equal to $|\mathcal{P}_h|\frac{h^2}{2}$. The random variable L can be written as:

$$L = \sum_{h=1}^{M} \mathcal{N}\left(0, |\mathcal{P}_h|\frac{h^2}{2}\right) = \mathcal{N}\left(0, \frac{1}{2}\sum_{h=1}^{M} h^2|\mathcal{P}_h|\right). \tag{5.30}$$

Otherwise, if at least one set $\mathcal{P}_h$ has a low cardinality, then the central limit theorem cannot be applied for the related addend. In particular, we examine the worst case, that is, when the cardinality of each set is $\mathcal{P}_h = 1$, with $h = 1, \ldots, N < M$, and we obtain the sum of N independent random variables each one not equally distributed with the other ones. However, in Section 5.6.2 we verify that the sequence of independent random variables $\{L_p, 2L_p, \ldots, NL_p\}$ satisfies the Lindeberg's condition [93]. Accordingly, the central limit theorem can be applied, and the random variable $L = \sum_{h=1}^{N} hL_p$ follows $\mathcal{N}\left(0, \frac{1}{2}\sum_{h=1}^{L} h^2\right)$, if N is large enough.

Globally, the random variable $L \sim \mathcal{N}\left(0, \frac{1}{2}\sum_{h=1}^{M} |\mathcal{P}_h|h^2\right)$, whatever the size of the sets $\mathcal{P}_h$. However, $|\mathcal{P}_h|$ is not known a priori. To overcome this issue, we replaced the variance σ_L^2 with its expected value $\mathbb{E}\{\sigma_L^2\}$, that is a fixed and known value given the number of devices M and the number of available preambles N_P, as proved in Section 5.6.3. So, by applying (5.57), the random variable L can be approximated as follows:

$$L \sim \mathcal{N}\left(0, \frac{1}{2}\left[\frac{M(M-1)}{N_P} + M\right]\right). \tag{5.31}$$

Assessing the relationship (5.24) as a whole, at each time $\tau' \neq t_i N_{TS}$, $\forall t_i \in \mathcal{T}_{p_s}$, we obtain a random variable $\Omega = \Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}, z_{k_{p_s}}}[\tau']\right\}$, which assumes the values related to the interference and noise in the tag detection procedure. We note that in $\tau = \tau'$ the first addend of the second member is equal to 0, while the remaining addends are three independent Gaussian variables: I, L, and $\mathcal{N}(0, \sigma_N^2)$. Therefore, it follows that the random variable Ω is:

$$\Omega \sim \mathcal{N}\left(0, \frac{1}{2}\left[\frac{(M-1)^2}{N_P} + 2M - 1\right] + \sigma_N^2\right). \tag{5.32}$$

The CDF of Ω is equal to:

$$F_{\Omega}(\omega) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\omega}{\sqrt{\frac{(M-1)^2}{N_P} + 2M - 1 + 2\sigma_N^2}}\right)\right]. \quad (5.33)$$

Now, let us analyze (5.24), when $\tau'' = t_i N_{TS}$, $\forall t_i \in \mathcal{T}_{p_s}$. We obtain a random variable, denoted as Ω_T, which assumes the values related to the transmitted tags. It follows $\Omega_T = \Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_{k_{ps}}}[\tau'']\right\} = \sqrt{N_{ZC}} + \Omega$. So, the CDF of Ω_T is equal to:

$$F_{\Omega_T}(\omega) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\omega - \sqrt{N_{ZC}}}{\sqrt{\frac{(M-1)^2}{N_P} + 2M - 1 + 2\sigma_N^2}}\right)\right]. \quad (5.34)$$

5.5 Performance Analysis

In this Section, we show the accuracy of proposed analytical approach by comparing our results with those obtained by simulation in MATLAB Environment. The parameters adopted are enlisted in Table 5.1. Simulations were run N_S times until all the results averaged up to the N_Sth simulation differ from those averaged up to the $(N_S - 1)$th simulation by less than 0.01%. Furthermore, the proposed model allows us to estimate the application limits (i.e., the working zone) of an efficient tag-preamble detector, in terms of maximum number of attempting devices M_{max} and SNR requirement at the gNB receiver.

5.5.1 Preamble Detection Analysis and Thresholds

Let us remember that the random variable Φ_p represents the correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau'']\right\}$ in a point $\tau'' = p_s N_{CS}$, $\forall p_s \in \mathcal{P}_\mathcal{S}$ (i.e., a point in which the preamble has been transmitted by at least one device), while $F_\Phi(\phi)$ represents the correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau']\right\}$ in a point $\tau' \neq p_s N_{CS}$, $\forall p_s \in \mathcal{P}_\mathcal{S}$ (i.e., a point in which no preamble has been transmitted).

Table 5.1: *SIMULATION PARAMETERS*

Parameter	Symbol	Value
Zandoff-Chu sequence length	N_{ZC}	839
Preamble Cyclic Shift	N_{CS}	13
Number of preambles available for the contention-based procedure	N_P	54
Tag Cyclic Shift	N_{TS}	13
Preamble root index	r	1
Tag root index	k_p	$N_{CS} \cdot p$
Total number of attempting MTC devices in one RA cycle	M	$1:100$[a]
Signal-Noise Ratio at the gNB receiver	SNR	$0:20$ dB
Detection probability	ϵ	$[0,1]$

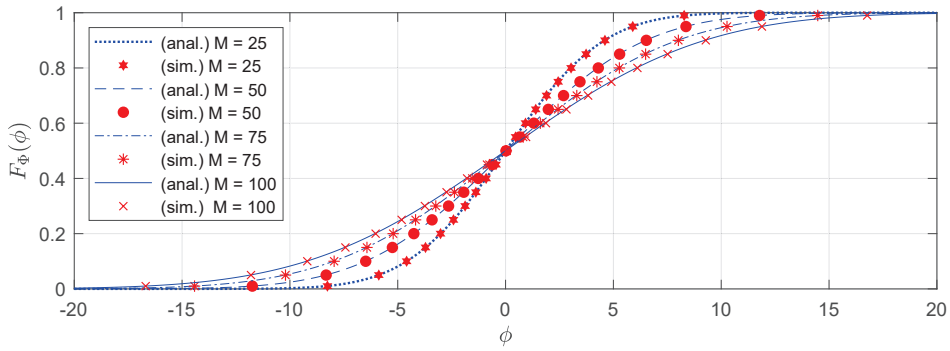
[a] These values are consistent with the massive MTC scenario adopted in [14].

Fig. 5.4a shows the comparison between the analytical CDF $F_\Phi(\phi)$, derived in (5.18), with the numerical results obtained by simulations, for $M = 25, 50, 75, 100$ and $SNR = 20$ dB; while Fig. 5.4b shows an analogous comparison for the CDF of Φ_P, derived in (5.19), when $|\mathcal{M}_{p_s}| = 1$, i.e., the worst case for the successful preamble detection. In both figures, the curves have the same trend, with mean value equal to 0 for Fig. 5.4a and to $\sqrt{N_{ZC}}$ for Fig. 5.4b, that are the ideal value assumed in the absence of noise and interference by Φ and Φ_P, respectively. It can be clearly observed that the results obtained by the model and by simulations are extremely similar for most of the ϕ values. However, in order to provide with an in-depth look at the accuracy of our model, in Table 5.2 the $F_\Phi(\phi)$ values obtained by simulation, by our analytic model and the error values are reported as an example for $M = 50$.

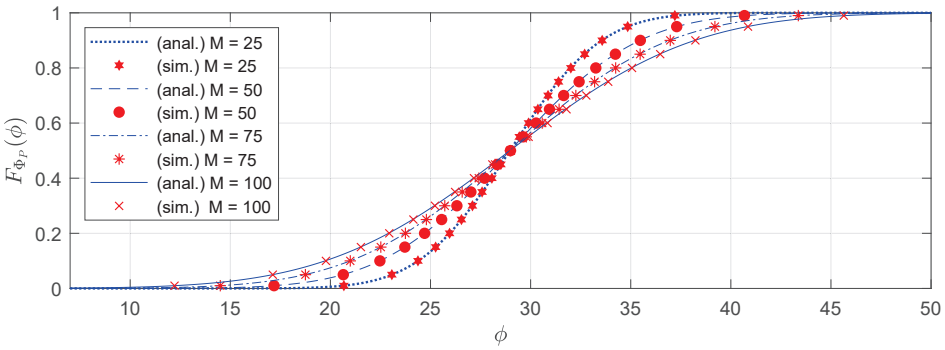
Once we have proven the accuracy of the proposed analytic model, we can use it to evaluate the effectiveness of preamble detection strategies as M and the SNR values change. Starting from the above results, in order to detect with probability ϵ the preambles $p_s \in \mathcal{P}_\mathcal{S}$, we define a proper threshold, T_P^ϵ, such that $Pr\left[\Phi_p \geq T_P^\epsilon\right] = \epsilon$ and, by using (5.19), we obtain:

$$1 - \frac{1}{2}\left[1 + \text{erf}\left(\frac{T_P^\epsilon - |\mathcal{M}_{p_s}|\sqrt{N_{ZC}}}{\sqrt{M + 2\sigma_N^2}}\right)\right] = \epsilon. \tag{5.35}$$

Given the probability ϵ, since the preamble detection probability should



(a) *CDF of* Φ.



(b) *CDF of* Φ_P *(when* $|\mathcal{M}_{p_s}| = 1$*).*

Figure 5.4: *CDFs of* Φ_P *(when* $|\mathcal{M}_{p_s}| = 1$*) and* Φ *obtained by simulation and our model, for several* M *values, with* $SNR = 20dB$.

be respected even in the worst case, i.e., $|\mathcal{M}_{p_s}| = 1$, the threshold T_P^{ϵ} is computed in this case, as:

$$T_P^{\epsilon} = \sqrt{N_{ZC}} - \sqrt{M + 2\sigma_N^2} \quad \text{erf}^{-1}(2\epsilon - 1). \tag{5.36}$$

Let us note that in a generic point $\tau' \neq p_s N_{CS}, \forall p_s \in \mathcal{P}_{\mathcal{S}}$, a false positive preamble could be detected if the interference assumes high values. For this reason, in order to non-detect false positive preambles with probability ϵ, we also define for the other variable Φ a threshold, denoted as T_{IP}^{ϵ}, that is, $Pr\left[\Phi \leq T_{IP}^{\epsilon}\right] = \epsilon$. Then, given the probability ϵ, the threshold T_{IP}^{ϵ} can be calculated, by using (5.18), as $T_{IP}^{\epsilon} = \sqrt{N_{ZC}} - T_P^{\epsilon}$.

In Figs. 5.5a we show the thresholds T_P^{ϵ} and T_{IP}^{ϵ} vs the number of devices M, with different SNR values, when $\epsilon = 0.999$. Let us analyze Fig. 5.5a. $T_{IP}^{0.999}$ is the minimum threshold which guarantees a true negative preamble probability at least equal to 0.999 in a point $\tau' \neq p_s N_{CS}$, $\forall p_s \in \mathcal{P}_{\mathcal{S}}$, while $T_P^{0.999}$ is the maximum threshold which guarantees

Table 5.2: *Simulated vs analytical results of $F_{\Phi}(\phi)$, when $M = 50$ and $SNR = 20$ dB*

Percentile	Simulation	Analytical	Error
0.99	11.7772	11.8003	-0.0231
0.95	8.3655	8.3435	0.0220
0.90	6.5271	6.5006	0.0265
0.85	5.2885	5.2573	0.0312
0.80	4.3001	4.2691	0.0310
0.75	3.4509	3.4213	0.0295
0.70	2.6894	2.6600	0.0294
0.65	1.9822	1.9545	0.0277
0.60	1.3101	1.2851	0.0250
0.55	0.6628	0.6374	0.0254
0.50	0.0200	0	0.0200
0.45	-0.6144	-0.6374	0.0230
0.40	-1.2660	-1.2851	0.0191
0.35	-1.9378	-1.9545	0.0167
0.30	-2.6432	-2.6600	0.0168
0.25	-3.4084	-3.4213	0.0130
0.20	-4.2533	-4.2691	0.0158
0.15	-5.2476	-5.2573	0.0097
0.10	-6.4842	-6.5006	0.0164
0.05	-8.3190	-8.3435	0.0245
0.01	-11.7423	-11.8003	0.0580

a true positive preamble probability at least equal to 0.999 in a point $\tau'' = p_s N_{CS}, \forall p_s \in \mathcal{P}_{\mathcal{S}}$ with $|\mathcal{M}_{p_s}| = 1$.

As expected, for guaranteeing a given true positive preamble probability, the $T_P^{0.999}$ threshold is a monotonically increasing function as the number of devices (and, consequently, the interference) increases. An opposite behavior occurs for the threshold $T_{IP}^{0.999}$ to guarantee a given true negative preamble probability.

In order to ensure a very low detection error probability, the detection strategy should ensure that both true positive preamble and true negative preamble probabilities are high, that is, in a generic instant time τ, it should be $T_{IP}^{\epsilon} \leq C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau] \leq T_P^{\epsilon}$, with ϵ approaching 1. Accordingly, it is possible to derive a valid working zone as long as $T_P^{\epsilon} \geq T_{IP}^{\epsilon}$, that is, up to the point of intersection between the two curves. This working zone corresponds to a maximum number of attempting devices M_{max} which can be effectively managed by the detector. As depicted in Fig.

Table 5.3: *Simulated vs analytical results of $F_\Omega(\omega)$, when $M = 50$ and $SNR = 20$ dB*

Percentile	Simulation	Analytical	Error
0.99	19.7529	19.8029	-0.0500
0.95	14.0663	14.0017	0.0646
0.90	10.9918	10.9091	0.0827
0.85	8.9039	8.8226	0.0814
0.80	7.2377	7.1642	0.0734
0.75	5.7996	5.7415	0.0581
0.70	4.5086	4.4639	0.0447
0.65	3.3115	3.2800	0.0315
0.60	2.1704	2.1566	0.0138
0.55	1.0690	1.0697	-0.0007
0.50	-0.0177	0	-0.0177
0.45	-1.1000	-1.0697	-0.0303
0.40	-2.2035	-2.1566	-0.0469
0.35	-3.3401	-3.2800	-0.0601
0.30	-4.5350	-4.4639	-0.0711
0.25	-5.8231	-5.7415	-0.0816
0.20	-7.2520	-7.1642	-0.0878
0.15	-8.9152	-8.8226	-0.0927
0.10	-10.9853	-10.9091	-0.0761
0.05	-14.0311	-14.0017	-0.0294
0.01	-19.6485	-19.8029	0.1544

5.5a, it corresponds to $M_{max} = 13$ when $SNR = 0$ dB, $M_{max} = 35$ when $SNR = 10$ dB, and $M_{max} = 43$ when $SNR = 20$ dB. We note that, for high SNR values, the contribution of interference dominates with respect to the noise.

Clearly, this approach is also valid for thresholds with different probability requirements. In fact, on basis of a given RA procedure strategy, the effect of a false positive preamble could be less severe than a false negative preamble. For instance, in Fig. 5.5b we show the curves related to a probability of true positive preamble equal to 0.999 and a probability of true negative preamble equal to 0.98. It results $M_{max} = 19$ when $SNR = 0$ dB, $M_{max} = 49$ when $SNR = 10$ dB, and $M_{max} = 61$ when $SNR = 20$ dB.

The above analysis shows that the effective working zones could be small and, generally, depend on the pair of values (M, SNR). This result suggests that an adaptive detection strategy may be needed (e.g., based

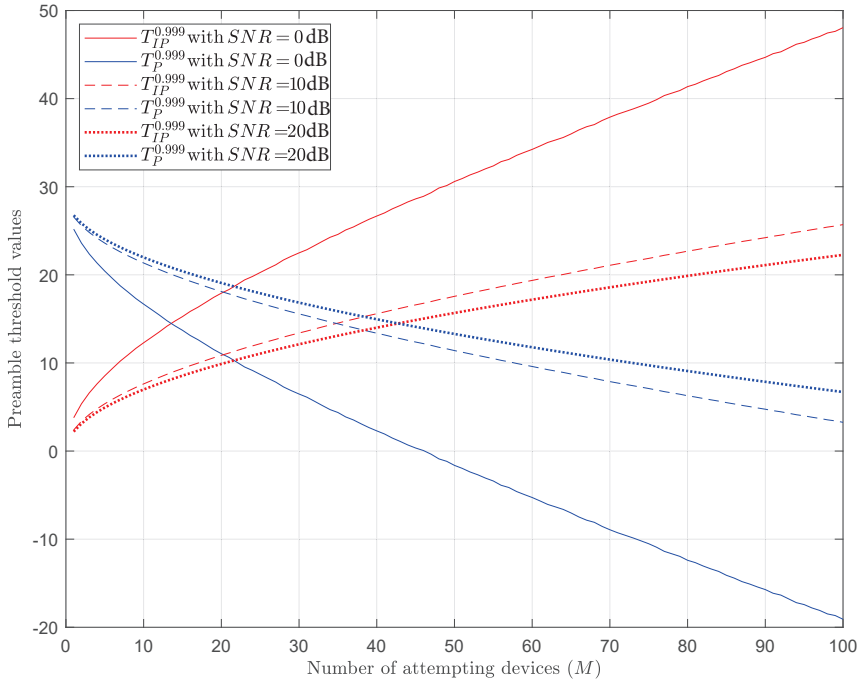
on a dynamic threshold value that adapts to the current traffic load M and SNR conditions). However, if we consider high SNR values (e.g. $SNR > 10$ dB) the threshold values depend substantially only on the value of M (i.e., the contribution of interference dominates with respect to the noise). In this case, a fixed threshold value can be considered as a function of M_{max}, and an appropriate ACB factor can be adopted to limit the maximum number of attempting devices to M_{max}. Therefore, in a scenario where a massive number of access attempts are expected, it may be necessary to carry out adequate cellular planning, and/or transmission power control to maintain a high SNR value on the receiver.

5.5.2 Tag Detection Analysis and Thresholds

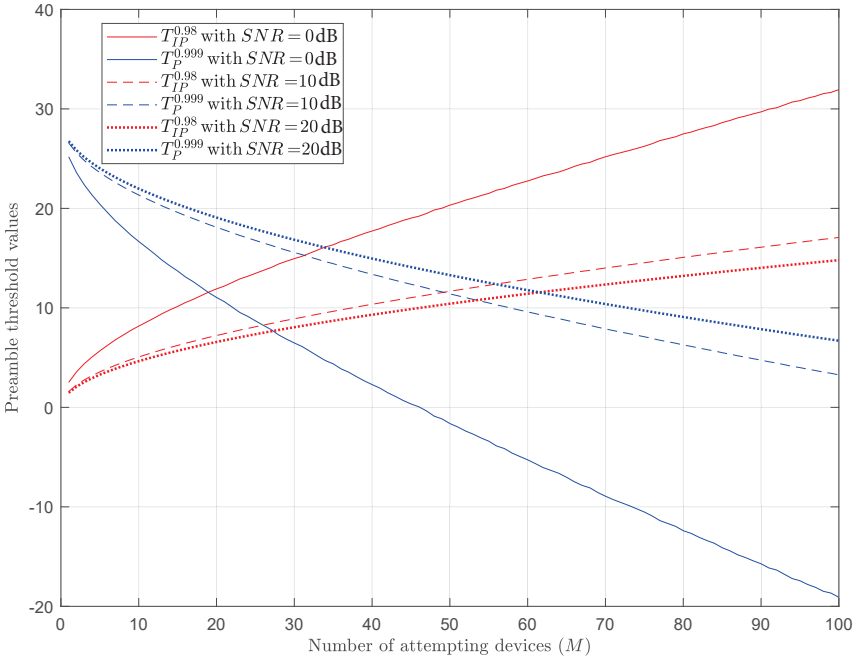
The analysis carried out here is similar to that of the previous subsection. In fact, we begin by comparing the analytical CDFs $F_{\Omega}(\omega)$, derived in (5.33), and $F_{\Omega_T}(\omega)$, derived in (5.34), with the numerical results obtained by simulation. Let us remember that the random variable Ω_T represents the correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_{k_{p_s}}}[\tau'']\right\}$ in a point $\tau'' = t_i N_{TS}$, in which the preamble $t_i \in \mathcal{T}_{p_s}$ has been transmitted by at least one device, while $F_{\Omega}(\omega)$ represents the correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_{k_{p_s}}}[\tau']\right\}$ in a point $\tau' \neq t_i N_{CS}$, $\forall t_i \in \mathcal{T}_{p_s}$, in which no tag has been transmitted.

In Fig. 5.6a we illustrate the comparison between the analytical and the simulation results of $F_{\Omega}(\omega)$ with $M = 25, 50, 75, 100$ with $SNR = 20$ dB, whereas in Fig. 5.6b we show $F_{\Omega_T}(\omega)$. Similarly to the previously analyzed preamble CDFs, both curves have the same trend, with mean value equal to 0 for Fig. 5.6a and to $\sqrt{N_{ZC}}$ for Fig. 5.6b, that are the ideal values assumed by Ω and Ω_T, respectively. Also, since the results obtained by our analytical model and by simulation are very similar, in Table 5.3 we show, as example, the values of $F_{\Omega}(\omega)$ for $M = 50$.

Having proved the accuracy of the proposed analytic model related to the tag detection, we use it to evaluate the effectiveness of tag detection strategies by deriving proper thresholds at the aim of discerning the tags from the interference. Therefore, we define the threshold T_T^{ϵ}, with the aim of detecting a tag $t_i \in \mathcal{T}_{p_s}$ with a probability equal to ϵ, and a threshold, T_{IT}^{ϵ}, with the aim of non-detecting an interference point as a



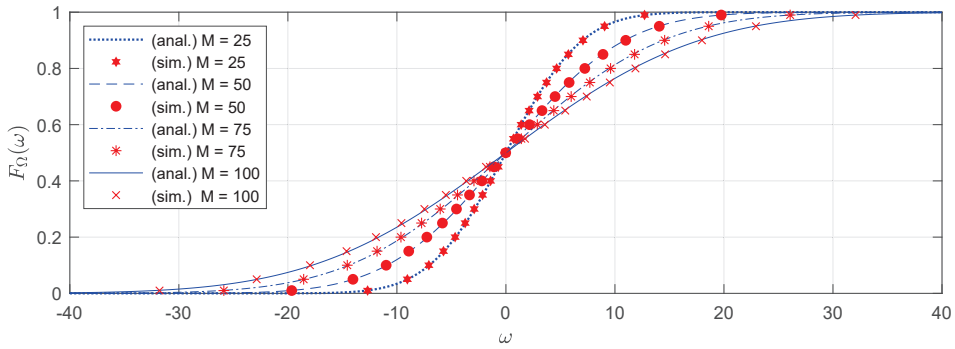
(a) $T_P^{0.999}$ *and* $T_{IP}^{0.999}$



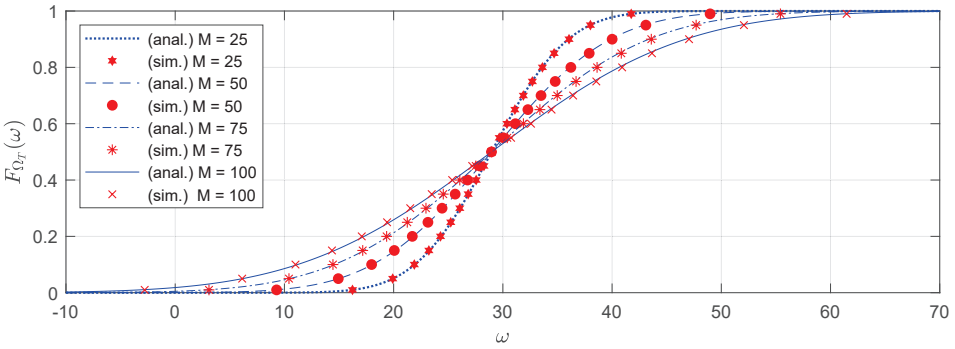
(b) $T_P^{0.999}$ *and* $T_{IP}^{0.98}$

Figure 5.5: *Preamble thresholds vs the number of attempting devices* M*, with different* SNR *values.*

tag with a probability equal to ϵ. In a similar way to what described for the preambles, given a probability ϵ, the threshold T_{IT}^{ϵ} can be calculated



(a) *CDF of* Ω.



(b) *CDF of* Ω_T.

Figure 5.6: *CDFs of* Ω *and* Ω_T *obtained by simulation and by our model, for several* M *values, with* $SNR = 20$ *dB.*

as:

$$T_{IT}^{\epsilon} = \sqrt{\frac{(M-1)^2}{N_P} + 2M - 1 + 2\sigma_N^2} \quad \text{erf}^{-1}(2\epsilon - 1), \tag{5.37}$$

whereas the threshold T_T^{ϵ} can be calculated as $T_T^{\epsilon} = \sqrt{N_{ZC}} - T_{IT}^{\epsilon}$.

In Figs. 5.7a we show the thresholds T_T^{ϵ} and T_{IT}^{ϵ}, when $\epsilon = 0.999$ vs the number of devices M, with different SNR values. In order that $T_{IT}^{\epsilon} \leq C_{y_{r,k_{\mathcal{P}_\mathcal{S}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}},z_r}[\tau] \leq T_T^{\epsilon}$ (i.e., $T_{IT}^{\epsilon} \leq T_T^{\epsilon}$), we get $M_{max} = 10$ when $SNR = 0$ dB, $M_{max} = 18$ when $SNR = 10$ dB, and $M_{max} = 19$ when $SNR = 20$ dB. Also for the tag detector, on basis of the access procedure strategy, the effect of a false positive tag could be less severe than a false negative tag. For instance, in Fig. 5.7b we show the curves related to a probability of true positive tag equal to 0.999 and a probability of true negative tag equal to 0.98. It results $M_{max} = 14$ when $SNR = 0$ dB, $M_{max} = 24$ when $SNR = 10$ dB, and $M_{max} = 26$ when $SNR = 20$ dB.

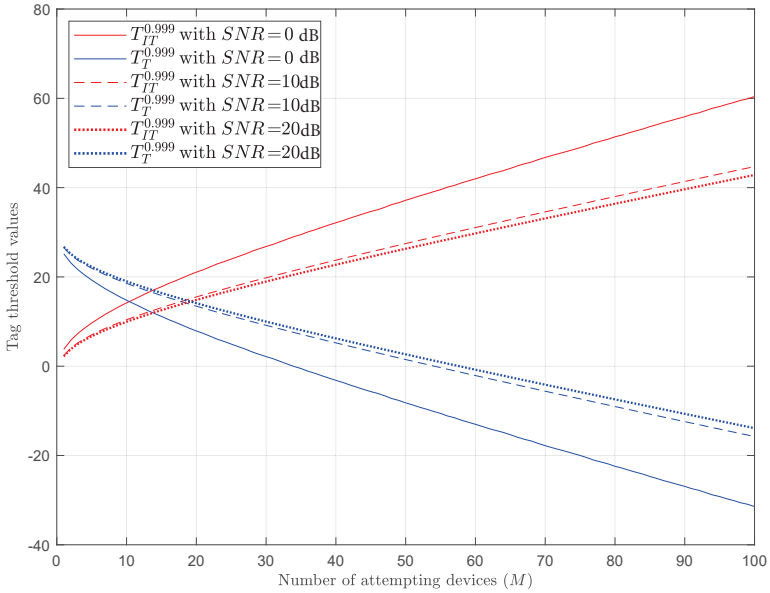
Compared to preamble detection, the tag detection has a more limited working zone, due to more severe interference. This means that, with the same preamble and tag detection probability, the working zone is constrained by the tag detection. So, a possibility to increase the working zone is to adopt a less stringent tag detection probability than the preamble probability. All the considerations made at the end of the previous subsection are valid also here for the tag detection. However, we highlight that the working areas valid for tag detection are unfortunately very small, and in general a straightforward strategy is not suitable for a massive MTC scenario. The overall working zones analysis suggests to investigate new optimized strategies for tagged preamble sequence detection.

5.5.3 Assessment of the Analytical Thresholds Accuracy

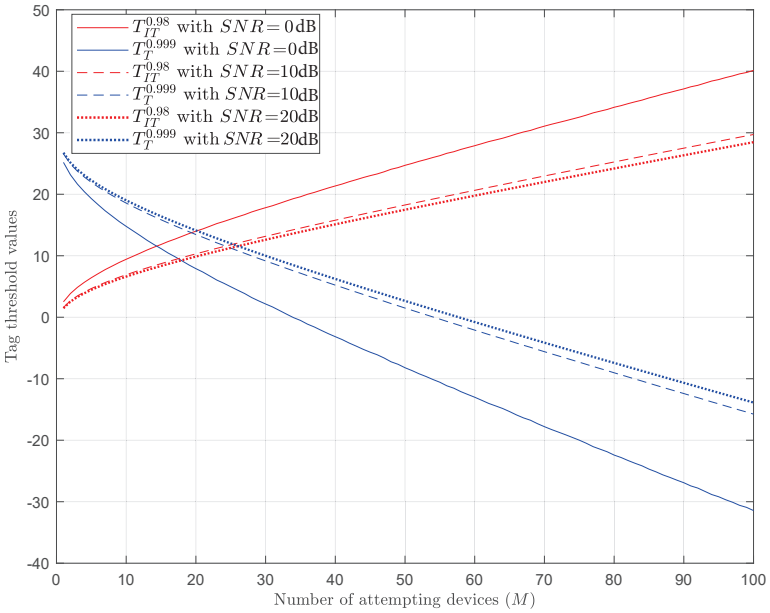
For the sake of completeness, in this subsection, we evaluate by simulation the accuracy of the thresholds derived from the analytical model in subsection 5.5.1 and 5.5.2. In this way, the accuracy of working zones is also verified.

We denote with $\bar{P}_{PD}(\epsilon)$ the average percentage of detected preambles above the threshold T_P^{ϵ}, and with $\bar{P}_{DT}(\epsilon)$ the average percentage of detected tags above the threshold T_T^{ϵ}. In Fig. 5.8 we show $\bar{P}_{DP}(\epsilon)$ and $\bar{P}_{DT}(\epsilon)$, for $\epsilon \in \{0.90, 0.95, 0.99\}$ with $SNR = 10$ dB. As regards $\bar{P}_{DP}(\epsilon)$, the simulation results demonstrate that the average percentage of detected preambles above the T_P^{ϵ} threshold value derived by the analytical model is very close to the expected ϵ value. Similar results were obtained with other SNR values. As for $\bar{P}_{DT}(\epsilon)$, we can see that the accuracy is slightly lower compared to the model used for the preambles. This result was expected because the model adopted for the tags introduces some approximations. In particular, the variances σ_I^2 and σ_L^2 have been approximated with their average value. However, the variation of the simulation results with respect to the theoretical expected values is very small and, consequently, the rightness of the working zones identified in the previous subsections is confirmed.

Similar considerations are valid for the percentage of interference/noise points below the thresholds T_{IP}^{ϵ} and T_{IT}^{ϵ}, but the related graphs are not reported for space reasons.



(a) $T_T^{0.999}$ *and* $T_{IT}^{0.999}$

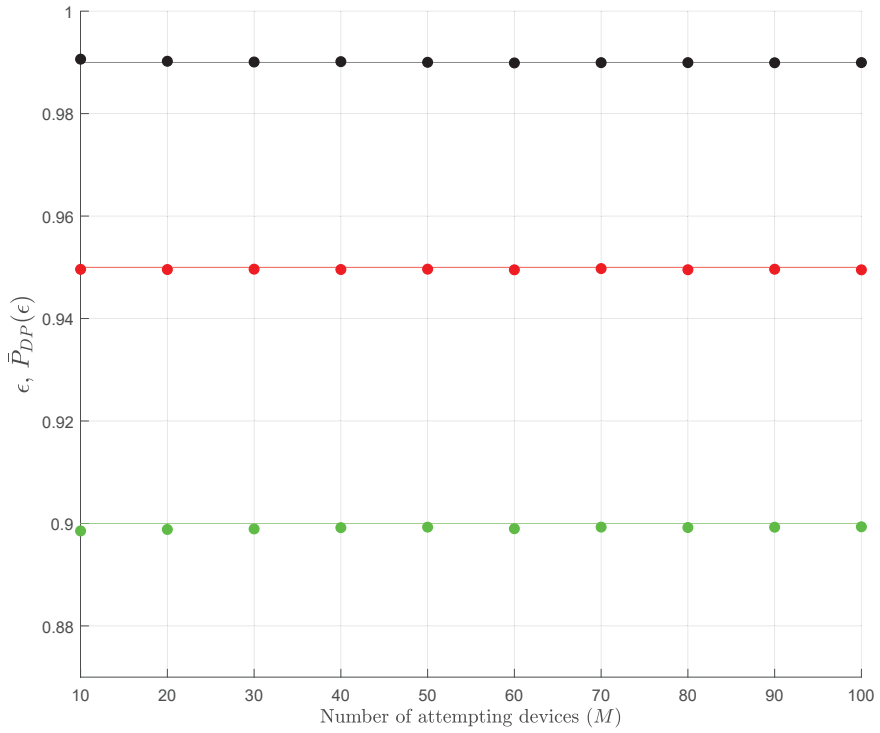


(b) $T_T^{0.999}$ *and* $T_{IT}^{0.98}$

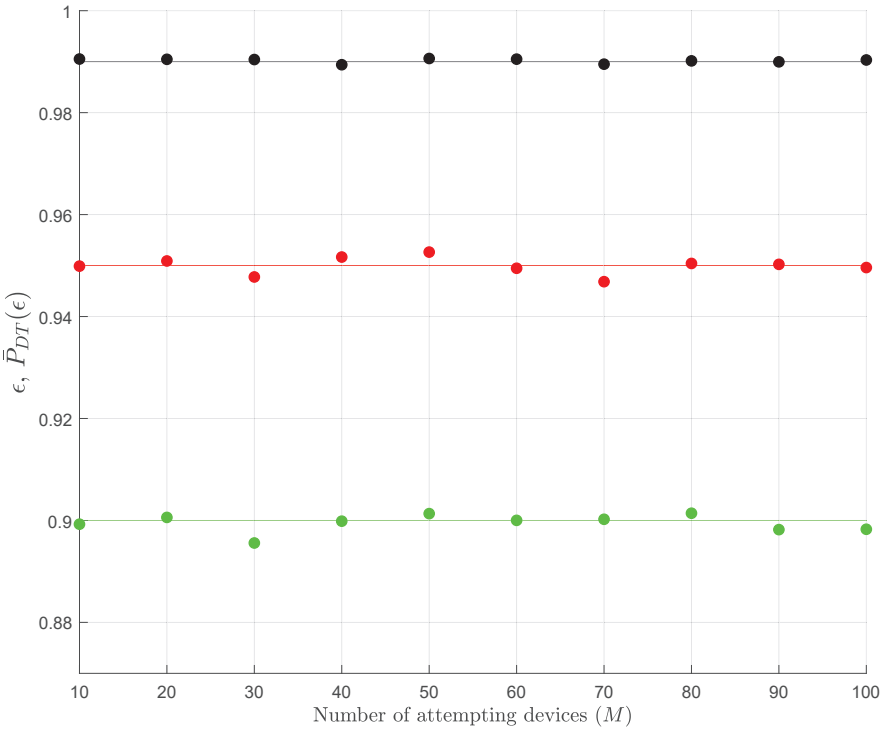
Figure 5.7: *Tag thresholds vs the number of attempting devices M, with different SNR values.*

5.5.4 Impact of failed preamble-tag detections on the signaling overhead and energy consumption

In this subsection, we analyze the performance of the detector inside the gNB receiver, when a simple preamble-tag detection strategy and the two-step RA procedure [14] are applied. In particular, we evaluate how



(a) $\bar{P}_{DP}(\epsilon)$



(b) $\bar{P}_{DT}(\epsilon)$

Figure 5.8: *Simulation results for* $\epsilon \in \{0.90; 0.95; 0.99\}$*, with* $SNR = 10$ *dB.*

much the erroneous detections impact on the signaling overhead and on the extra energy consumption experienced by the MTC devices, under different preamble and tag detection probabilities, ϵ_P and ϵ_T, respectively.

Given a RA cycle, for each preamble $p \in \mathcal{P}$, the detector may either detect the preamble (i.e., $p \in \mathcal{P}_\mathcal{D}$), or not. In the first case, thanks to the tags, the preamble can be classified either as successfully transmitted (i.e., $p \in \mathcal{P}_{\mathcal{SD}}$), or as collided (i.e., $p \in \mathcal{P}_{\mathcal{CD}}$). Obviously, $\mathcal{P}_{\mathcal{SD}} \cup \mathcal{P}_{\mathcal{CD}} = \mathcal{P}_\mathcal{D}$ and $\mathcal{P}_{\mathcal{SD}} \cap \mathcal{P}_{\mathcal{CD}} = \emptyset$. Let us define the following disjoint failure events for the detector.

Event A: The MTC device has selected a preamble $p_s \in \mathcal{P}_\mathcal{S}$ that was successfully transmitted (i.e., $|\mathcal{M}_{p_s}| = 1$), but it was either not detected or classified as collided (i.e., $p_s \in (\mathcal{P} - \mathcal{P}_\mathcal{D}) \cup \mathcal{P}_{\mathcal{CD}}$).

Event B: The MTC device has selected a preamble $p_s \in \mathcal{P}_\mathcal{S}$ that was collided (i.e., $|\mathcal{M}_{p_s}| \geq 2$), and it was classified by the detector as successfully transmitted (i.e., $p_s \in \mathcal{P}_{\mathcal{SD}}$).

Event C: The MTC device has selected a preamble $p_s \in \mathcal{P}_\mathcal{S}$ that was a collided preamble and it was not detected by the detector (i.e., $p_s \notin \mathcal{P}_\mathcal{D}$).

Event D: No MTC device has selected the preamble p (i.e., $p \in \mathcal{P} - \mathcal{P}_\mathcal{S}$), but it was detected as sent by the detector, (i.e. $p \in \mathcal{P}_\mathcal{D}$).

As regards the MTC device experiencing event A, it is not affected by any additional signaling overhead. However, it is affected by an additional energy consumption, denoted as E_A, due to the preamble transmission and the permanence in the RX active state waiting for the RAR message. As for the MTC device experiencing Event B, it is subject to a signaling overhead of O_B bytes related to the transmission of the data packet including the device ID, and an additional energy consumption, E_B, due to the data packet transmission and the waiting for the ACK message. As regards Events C and D, they do not involve any increase in both signaling overhead and energy consumption from the point of view of the MTC device.

Let us denote with $Pr\{A\}$ and $Pr\{B\}$ the probability that an attemping MTC device is experiencing Event A and Event B, respectively. The additional average energy consumption per access attempt is $E_{add} = Pr\{A\}E_A + Pr\{B\}E_B$, whereas the additional signaling overhead is $O_{add} = Pr\{B\}O_B$ bytes. The values of $Pr\{A\}$ and $Pr\{B\}$ depend on

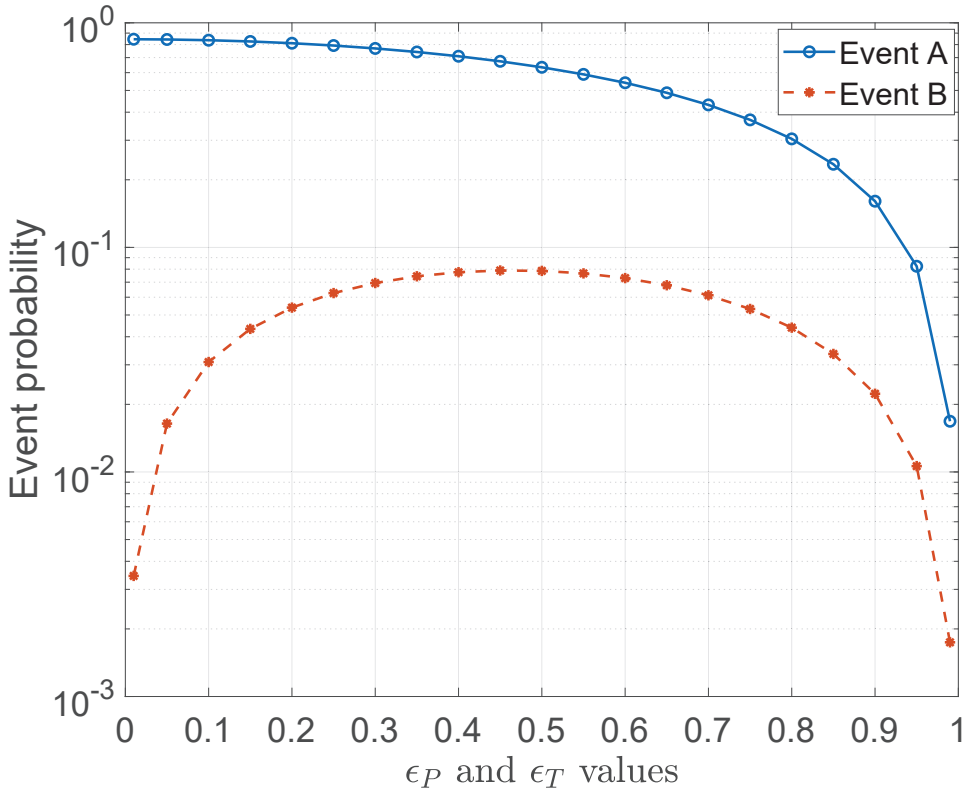


Figure 5.9: *Variation of $Pr\{A\}$ and $Pr\{B\}$ with respect to ϵ_P and ϵ_T, with $SNR = 20$ dB, for $M = 10$.*

the detection strategy adopted and on ϵ_P and ϵ_T values.

As regards the preamble detection strategy, we note that the detector does not know a priori whether a point τ of $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_r}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]\right\}$ is an interference point (i.e., $\tau \neq p_s N_{CS}$) or not. So, for any point τ, a simple strategy is to apply a unique threshold $T_P^{\epsilon_P}$ in order to detect the transmitted preambles. The same consideration applies to the tag detection, and we assume the unique threshold $T_T^{\epsilon_T}$ for each point of $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_{k_{p_s}}}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]\right\}$.

Then, in Figs. 5.9 and 5.10 we show the variation of $Pr\{A\}$ and $Pr\{B\}$ with respect to ϵ_P and ϵ_T, with $SNR = 20$ dB, for $M = 10$ and $M = 30$, respectively. For low values of ϵ_P and ϵ_T, the $Pr\{A\}$ value is high and $Pr\{B\}$ is low because there is a high probability that the preamble is not detected. As ϵ_P and ϵ_T increase to about 0.5, $Pr\{B\}$ increases because the probability of detecting a preamble increases, but the probability of correctly detecting the tags remains low (less than 0.5). Obviously, for $\epsilon_P = \epsilon_T > 0.5$, $Pr\{A\}$ and $Pr\{B\}$ decrease. Comparing

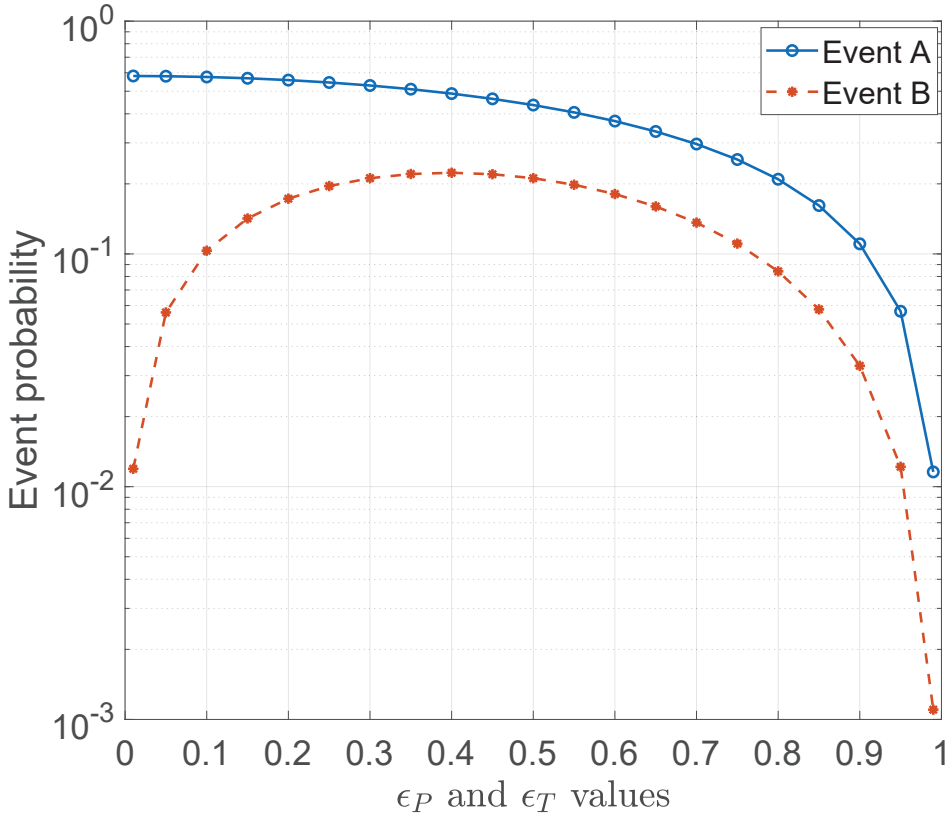


Figure 5.10: *Variation of $Pr\{A\}$ and $Pr\{B\}$ with respect to ϵ_P and ϵ_T, with $SNR = 20$ dB, for $M = 30$.*

Fig. 5.9 with 5.10, it can be seen that for $M = 10$ the $Pr\{A\}$ curve is higher than the one for $M = 30$, whereas the reverse holds true for $Pr\{B\}$, because for $M = 10$ the probability that a MTC device transmits a preamble with success is higher and the collision probability is lower.

Finally, as a case study, to quantify the signaling overhead and energy consumption of the MTC device, we consider the two-step RA procedure and the energy model proposed in [14]. The MTC device that experiences Event A will consume 3 mJ for transmitting the preamble, and 1.277 mJ for permaning in the RX active state during the RAR message window, i.e., $E_A = 4.277$ mJ. As regards the MTC device experiencing Event B, it consumes the same amount of energy to transmit the data packet and to wait for the relative ACK reception, i.e., $E_B = E_A$. In addition, there will be a signaling overhead of $O_B = 9$ bytes, related to the additional information piggybacked with the data. As example, if $\epsilon_P = \epsilon_T = 0.95$ and $M = 10$, it results $E_{add} = 0.3978$ mJ and $O_{add} = 0.095$ bytes, whereas with $M = 30$, it follows $E_{add} = 0.2947$ mJ

and $O_{add} = 0.11$ bytes.

5.6 Appendix

5.6.1 Calculation of $\mathbb{E}\{|\mathcal{M}_{p_s}|\}$

$|\mathcal{M}_{p_s}|$ is the number of devices that have selected the preamble p_s, given that the preamble p_s has been selected by at least one device. Then, the probability that $|\mathcal{M}_{p_s}| = h$ results:

$$Pr\{|\mathcal{M}_{p_s}| = h\} = \binom{M-1}{h-1} \left(\frac{1}{N_P}\right)^{h-1} \left(1 - \frac{1}{N_P}\right)^{M-h}, \quad (5.38)$$

with $h = \{1, \ldots, M\}$. The expected value of $|\mathcal{M}_{p_s}|$ is defined as:

$$\mathbb{E}\{|\mathcal{M}_{p_s}|\} = \sum_{h=1}^{M} hPr\left\{|\mathcal{M}_{p_s}| = h\right\}. \quad (5.39)$$

Let us put $j = h - 1$, $n = M - 1$, $p = \frac{1}{N_P}$, and $q = 1 - p$. Then:

$$\begin{aligned} \mathbb{E}\{|\mathcal{M}_{p_s}|\} &= \sum_{j=0}^{n} (j+1) \binom{n}{j} p^j q^{n-j} = \\ &= \sum_{j=0}^{n} j \binom{n}{j} p^j q^{n-j} + \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j}. \end{aligned} \quad (5.40)$$

By exploiting the binomial formula, it follows:

$$\mathbb{E}\{|\mathcal{M}_{p_s}|\} = \sum_{j=0}^{n} j \binom{n}{j} p^j q^{n-j} + (p+q)^n. \quad (5.41)$$

Since the first addend of the second member is the average value of a binomial distribution with parameters n and p, and the second addend is equal to 1, it follows:

$$\mathbb{E}\{|\mathcal{M}_{p_s}|\} = np + 1 = \frac{M-1}{N_P} + 1. \quad (5.42)$$

5.6.2 Lindeberg's condition

In this section, we verify that the sequence of N independent random variables $\{L_p, 2L_p, \ldots, NL_p\}$ satisfies the Lindeberg's condition.

Let $s_N^2 = \sum_{h=1}^{N} \sigma_{hL_p}^2 = \sum_{h=1}^{N} \frac{h^2}{2}$. The Lindeberg's condition is the following:

$$\lim_{N\to\infty} \frac{1}{s_N^2} \sum_{h=1}^{N} \mathbb{E}[(hL_p - \mu_{hL_p})^2 \cdot \mathbf{1}\{|hL_p - \mu_{hL_p}| > \varepsilon s_N\}] = 0 \quad (5.43)$$

for all $\varepsilon > 0$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

Let us analyze the term $\mathbb{E}[(hL_p - \mu_{hL_p})^2 \cdot \mathbf{1}\{|hL_p - \mu_{hL_p}| > \varepsilon s_N\}]$. In our case, it becomes:

$$\mathbb{E}\left[(hL_p)^2 \cdot \mathbf{1}\{|hL_p| > \varepsilon s_N\}\right] h^2 \mathbb{E}\left[L_p^2 \cdot \mathbf{1}\{h|L_p| > \varepsilon s_N\}\right], \quad (5.44)$$

where

$$\mathbf{1}\{h|L_p| > \varepsilon s_N\} = \begin{cases} 1 & \text{if } h|L_p| > \varepsilon s_N \\ 0 & \text{otherwise.} \end{cases} \quad (5.45)$$

Let us analyze the indicator function:

$$h|L_p| > \varepsilon s_N \Rightarrow h|L_p| > \varepsilon \sqrt{\sum_{h=1}^{N} \frac{h^2}{2}} \Rightarrow h^2 L_p^2 > \frac{\varepsilon^2}{2} \sum_{h=1}^{N} h^2. \quad (5.46)$$

By exploiting the Faulhaber's formula [94], it follows:

$$h^2 L_p^2 > \frac{\varepsilon^2}{2} \frac{N(N+1)(2N+1)}{6}. \quad (5.47)$$

Since $0 \leq L_p^2 \leq 1$ and $1 \leq h \leq N$, when N approaches infinity, the second member goes to infinity faster than the first one. Therefore, for N large enough, it follows $\mathbb{E}\left[L_p^2 \cdot \mathbf{1}\{h|L_p| > \varepsilon s_N\}\right] = 0$, and the Lindeberg's condition (5.43) is verified.

5.6.3 Calculation of the average value of σ_L^2

$$\mathbb{E}\{\sigma_L^2\} = \mathbb{E}\left\{\frac{1}{2} \sum_{h=1}^{M} h^2 |\mathcal{P}_h|\right\} = \frac{1}{2} \sum_{h=1}^{M} h^2 \mathbb{E}\left\{|\mathcal{P}_h|\right\}, \quad (5.48)$$

where $\mathbb{E}\{|\mathcal{P}_h|\}$ is the average number of preambles that have been selected by h out of M devices. We define the random variables X_h^j, with $h \in \{1, \ldots, M\}$, and $j \in \{1, \ldots, N_P\}$ as:

$$X_h^j = \begin{cases} 1 & \text{if preamble } j \text{ has been selected by } h \text{ devices} \\ 0 & \text{otherwise.} \end{cases} \tag{5.49}$$

So, the number of preambles selected by h devices can be written as:

$$|\mathcal{P}_h| = \sum_{j=1}^{N_P} X_h^j, \tag{5.50}$$

and the mean value results:

$$\mathbb{E}\{|\mathcal{P}_h|\} = \mathbb{E}\left\{\sum_{j=1}^{N_P} X_h^j\right\} = \sum_{h=1}^{N_P} \mathbb{E}\{X_h^j\}. \tag{5.51}$$

Since X_h^j are N_P random variables identically distributed, the mean value of X_h^j, for each j, is equal to:

$$\begin{aligned} \mathbb{E}\{X_h^j\} &= \Pr(X_h^j = h) = \\ &= \binom{M}{h} \left(\frac{1}{N_P}\right)^h \left[1 - \left(\frac{1}{N_P}\right)\right]^{M-h}. \end{aligned} \tag{5.52}$$

So:

$$\begin{aligned} \mathbb{E}\{|\mathcal{P}_h|\} &= N_P \mathbb{E}\{X_h^j\} = \\ &= N_P \binom{M}{h} \left(\frac{1}{N_P}\right)^h \left[1 - \left(\frac{1}{N_P}\right)\right]^{M-h}. \end{aligned} \tag{5.53}$$

Equation (5.48) becomes:

$$\mathbb{E}\{\sigma_L^2\} = \frac{1}{2} N_P \sum_{h=1}^{M} h^2 \binom{M}{h} \left(\frac{1}{N_P}\right)^h \left[1 - \left(\frac{1}{N_P}\right)\right]^{M-h}. \tag{5.54}$$

Let us put $p = \frac{1}{N_P}$, $q = 1 - p$, and substitute h^2 with $h(h-1) + h$.

Then:

$$\mathbb{E}\{\sigma_L^2\} = \left[\sum_{h=2}^{M} h(h-1)\binom{M}{h}p^h q^{M-h} + \sum_{h=0}^{M} h\binom{M}{h}p^h q^{M-h}\right]\frac{1}{2p} = \\ = \left[M(M-1)p^2\sum_{h=2}^{M}\frac{(M-2)!}{(h-2)!(M-h)!}p^{h-2}q^{M-h} + Mp\right]\frac{1}{2p}. \tag{5.55}$$

Let us put $k = h-2$ and $N = M-2$ in the summation, it follows:

$$\mathbb{E}\{\sigma_L^2\} = \left[M(M-1)p^2\sum_{k=0}^{N}\binom{N}{k}p^k q^{N-k} + Mp\right]\frac{1}{2p} = \\ = \left[M(M-1)p^2(p+q)^N + Mp\right]\frac{1}{2p}. \tag{5.56}$$

Since $(p+q)^N = 1$, it results:

$$\mathbb{E}\{\sigma_L^2\} = \left[M(M-1)\left(\frac{1}{N_P}\right)^2 + \frac{M}{N_P}\right]\frac{N_P}{2} = \\ = \frac{1}{2}\left[\frac{M(M-1)}{N_P} + M\right]. \tag{5.57}$$

5.6.4 Multi-path fading, propagation delay, and Doppler spread Analysis

In the development of the analytical model, we assumed a single path, ideal channel conditions, and the propagation delay was neglected. In this section, we add some discussions on the implications of the delay spread due to multi-path fading, the propagation delay, and the Doppler spread, due to the relative motion between the MTC device and the gNB. Clearly, this signal degradation reduces the orthogonality of the tagged preamble sequences. To overcome this issue, we introduced, as reported in Section 4.2, one cyclic shift both for the preambles, N_{CS}, and for the tags, N_{TS}, aiming to guarantee the orthogonality of the sequences regardless of all the above phenomena. Particularly, both the cyclic shifts should be properly dimensioned based on the considered mMTC scenario. In the following we focus on N_{CS}, but similar considerations are

valid for N_{TS}. As reported by 3GPP in [95], the recommended deployment scenario for mMTC is denoted as "Urban coverage for massive connection", and it is characterized by the carrier frequency (f) equal to 700 MHz, and the movement speed (v) fixed to 3 km/h, identical for all the MTC devices. First of all, we observe that this test scenario implies a negligible frequency offset due to Doppler spread, calculated as $\frac{vf}{c}$, where c is the speed of light. In fact, it is equal to 1.944 Hz and results very lower than the PRACH sub-carrier spacing, Δf =1.25 kHz. So, it is known that:

$$N_{CS} \geq \left\lceil (2\delta_{p,max} + \tau_{DS,max}) \frac{N_{ZC}}{T_{SEQ}} \right\rceil , \tag{5.58}$$

where $\delta_{p,max} = R/c$ is the maximum propagation delay, R is the cell radius, $\tau_{DS,max}$ is the maximum delay spread, $T_{SEQ} = 1/\Delta f$ is the sequence duration.

Let us start neglecting the propagation delay, i.e., $\delta_{p,max} = 0$. Then, Eq. (5.58) depends only on $\tau_{DS,max}$, that is, a measure of the multipath richness of the communication channel. At this aim of estimating $\tau_{DS,max}$, we consider the Tapped Delay Line (TDL) models for wireless channels provided by the 3GPP in [96]; these models include all fading phenomena considered in (5.9). In particular, we utilize the TDL-C model for Non-Line-Of-Sight (NLOS), and the TDL-E model for Line-Of-Sight (LOS). Each TDL model is scaled in delay so that the model achieves a desired Root Mean Square (RMS) delay spread, $\tau_{DS_{RMS}}$, equal to 93 ns for the short-delay profile, and 363 ns for the normal-delay profile [96]. From the TDL-C model, the maximum delay spread is $\tau_{DS,max} = 8.6523 \cdot \tau_{DS_{RMS}}$, so by using (5.58), the minimum N_{CS} value is 1 for the short-delay profile, and 4 for the normal-delay profile. Conversely, in the TDL-E model, $\tau_{DS,max} = 20.6519 \cdot \tau_{DS_{RMS}}$, then N_{CS} should be at least equal to 3 and 8, for the short-delay profile and the normal-delay profile, respectively. Since the minimum value provided by the standard is $N_{CS} = 13$, we can estimate the maximum cell size, by taking into account the propagation delay $\delta_{p,max}$. In the TDL-C model, the maximum cell radius R is 1.78 km for the short-delay profile, and 1.38 km for the normal-delay profile. Conversely, in the TDL-E model, R is 1.57 km and 0.73 km, for the short-delay profile and the normal-delay profile, respectively. Clearly, in the case of larger cells, a greater

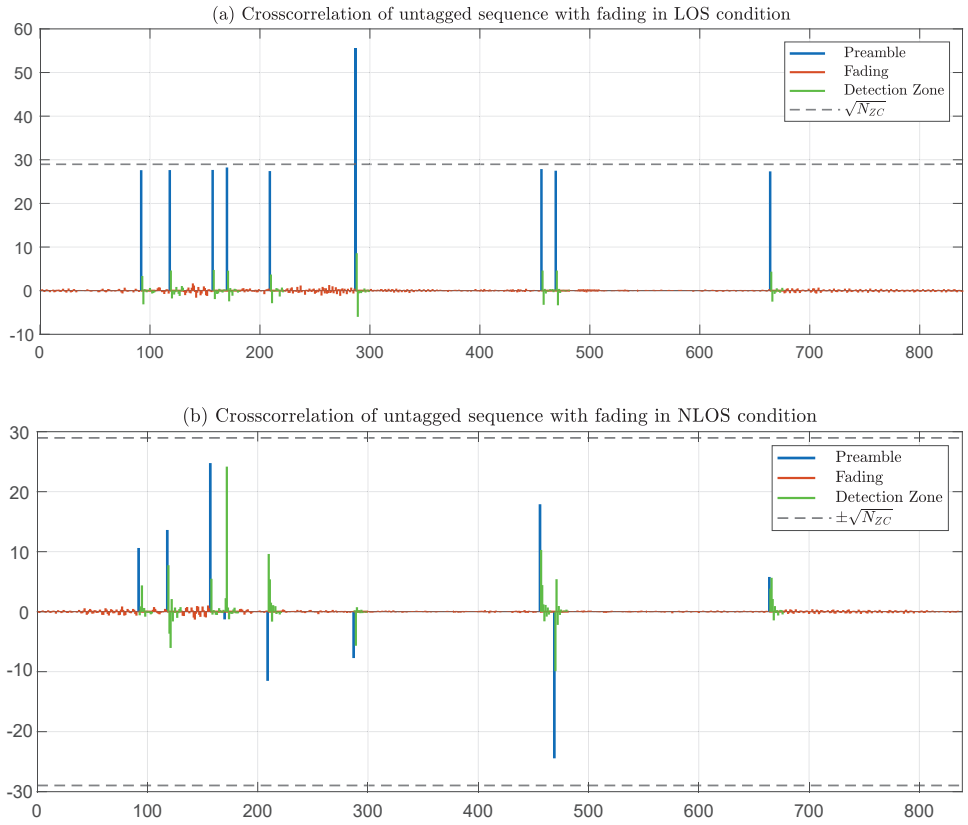


Figure 5.11: *Example of* $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_r}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]\right\}$ *when* $M = 10$ *with fading in LOS (a) and NLOS (b) conditions.*

value of N_{CS} needs to be set. also, we note that our analytical model is derived for a generic N_{CS}, so it is still valid. As regards N_{TS}, the same dimensioning is applied, but there is no a minimum value set by the standard.

Now, we want to evaluate the overall effects of the multi-path transmissions, non-ideal channel gains ($h_{i,g} \neq 1$) and $d_{i,g} \neq 0$ on the preamble detection. Firstly, for each device $i \in \mathcal{M}$ transmitting an untagged preamble sequence $x_r^{p_i}[n]$, we calculated the output sequence across an independent TDL-E and TDL-C modeled channel, with normal-delay profile, $N_{CS} = 13$, and $N[n] = 0$. Then, we calculated the sum of the above sequences and evaluate the cross-correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_\mathcal{S}}},z_r}^{\mathcal{P}_\mathcal{S},\mathcal{T}_\mathcal{S}}[\tau]\right\}$, reported in Figs.5.11a and 5.11b. As regards the LOS (see Fig. 5.11a), due to the multiple paths, in each point $\tau = p_i N_{CS}$, $\forall p_i \in \mathcal{P}_\mathcal{S}$, the correlation value is slightly lower than the expected value $|\mathcal{M}_{p_i}|\sqrt{N_{ZC}}$, and

different from 0 in the other points. Specifically, significant secondary peaks occur in the ranges $\{p_i N_{CS}+1,\ldots,p_i N_{CS}+N_{CS}-1\}, \forall p_i \in \mathcal{P}_{\mathcal{S}}$. Therefore, given $p_i \in \mathcal{P}_{\mathcal{S}}$, we introduce the "Detection Zone for preamble p_i", $DZ_{p_i} = \{p_i N_{CS},\ldots,p_i N_{CS}+N_{CS}-1\}$, as the interval of values containing the main and the secondary peaks of the correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau]\right\}$. Outside these ranges, i.e., $\tau \notin DZ_{p_i}, \forall p_i \in \mathcal{P}_{\mathcal{S}}$, the values assumed by $\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau]\right\}$ approaches zero. Conversely, as regards the NLOS (see Fig. 5.11b), the sign of the main peaks of the correlation $\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau]\right\}$ can be either positive or negative. In addition, $\left|\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau]\right\}\right|$ can be significantly lower than $|\mathcal{M}_{p_i}|\sqrt{N_{ZC}}$. Also in this case, $\forall p_i \in \mathcal{P}_{\mathcal{S}}$, several peaks occur in the range DZ_{p_i}, while outside these detection zones, the values assumed approaches zero. Also, we note that the main peak of $\left|\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau]\right\}\right|$ may be in a point $\tau \neq p_i N_{CS}$, but $\tau \in DZ_{p_i}$; this is due to the absence of a direct path which dominates in terms of power compared to the other secondary paths and arrives with an additional delay equal to 0.

Having introduced the detection zones, we are now able to appropriately adapt the Φ_P and Φ random variables in our model. At this aim, the range $\{0,\ldots,(N_{ZC}-1)\}$ needs to be divided into N_P detection zones $DZ_p, \forall p \in \mathcal{P}$, and for each zone, the scheduler should calculate:

$$\tau'' = \underset{\forall \tau \in DZ_p}{\arg\max}\left|\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau]\right\}\right|. \tag{5.59}$$

The random variable Φ_P is re-defined as $\Phi_P = \left|\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau'']\right\}\right|$, where τ'' is calculated in (5.59), $\forall p \in \mathcal{P}_{\mathcal{S}}$. We also need to re-define the random variable Φ as $\Phi = \left|\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}}^{\mathcal{P}_{\mathcal{S}},\mathcal{T}_{\mathcal{S}}},z_r}[\tau'']\right\}\right|$, where τ'' is calculated in (5.59), $\forall p \in \mathcal{P}-\mathcal{P}_{\mathcal{S}}$. Then, the T_P^{ϵ} and T_{IP}^{ϵ} values and the related working zones need to be recalculated.

As already pointed out, in LOS the effect of fast fading does not severely change the values assumed by both the adapted Φ_P and Φ vari-

ables compared to the ideal values (i.e., with the absence of noise, fading, and interference). We therefore expect a very slight variation in the threshold values and the related working zone. Instead, in NLOS, only the variation of the Φ values is negligible compared to the ideal ones, while Φ_P is subject to large variations. For this reason, we expect T_P^{ϵ} to assume much lower values compared to the ones of the analysis previously made and, therefore, the working zones will be reduced accordingly.

Now, we take into account the propagation delays effects, in absence of multi-path propagation and noise. Typically, each device $i \in \mathcal{M}$ experiences a propagation delay different from the other ones. Since the N_{CS} value in (5.58) is properly dimensioned taking into account the $\delta_{p,max}$ value, the Detection Zones strategy is valid for the preamble transmissions of any device. In addition, we note that, given a preamble $p_s \in \mathcal{P}_{\mathcal{S}}$, the gNB receiver could detect in $\Re\left\{C_{y_{r,k_{\mathcal{P}_{\mathcal{S}}}},z_r}^{\mathcal{P}_{\mathcal{S}},\tau_{\mathcal{S}}}[\tau]\right\}$, for $\tau \in DZ_{p_s}$, $|\mathcal{M}_{p_s}|$ different peaks with amplitude $\sqrt{N_{ZC}}$ belonging to the same detection zone DZ_{p_s}, instead of a single peak of amplitude $|\mathcal{M}_{p_s}|\sqrt{N_{ZC}}$ in $\tau = p_s N_{CS}$. Consequently, the value assumed by the adapted random variable Φ_P would be equal to $\sqrt{N_{ZC}}$. However, in our model the preamble threshold T_P^{ϵ} has been derived in the worst case (i.e., $|\mathcal{M}_{p_s}| = 1$), therefore, the preamble p_s will still be detected with the related ϵ probability and the eventual collision will be detect by the tag analysis. Finally, we also add the multi-path effects. Since each gth path of the MTC device i will suffer the same additional propagation delay, the main and secondary peaks of the cross-correlation will be rigidly shifted forward. Nevertheless, since the N_{CS} value in (5.58) is properly dimensioned taking into account both $\delta_{p,max}$ and $\tau_{DS,max}$, the peaks remain within the Detection Zone, allowing the detector to properly work.

5.7 Conclusion

In this chapter, we presented a rigorous methodology for modeling and analyzing the signal processed by the gNB receiver, when tagged preamble sequences are transmitted at the first step of the RA procedure. More specifically, we have derived the expression of the probability distributions of the related variables, in closed form, in presence of additive

white Gaussian noise and interference due to other preambles and tags. Under several traffic conditions, we have assessed the accuracy of proposed models compared to simulation results. We used our analytical approach to derive the threshold values to detect both the preamble and the tag with a given probability, and to determine the working zones where a basic preamble detection strategy can well operate. The high accuracy obtained allows the researcher to investigate and elaborate optimized detection strategies or to evaluate the implications of different ACB factor values without carrying out a large number of simulations, which are typically highly time-consuming in a massive scenario. Furthermore, we proposed also a qualitative evaluation of the detection performance in the presence of a multi-path fading, considering both the LOS and NLOS visibility conditions.

We emphasize that the detection strategy for the tag-preamble pairs adopted by the gNB receiver plays a fundamental role on the performance of the contention-based access procedures. For example, by considering a simple detection strategy and the 2-phase RA procedure, our models allowed us to estimate the relationship between the tag-preamble pair detection probabilities and both the signaling overhead and the energy consumption per access attempt from the MTC device's perspective, under different traffic loads. In conclusion, our analytical approach is a powerful tool that can be easily adopted to investigate and to propose new and effective strategies of tagged preamble detection and, accordingly, to evaluate innovative and efficient RA strategies tailored for the mMTC scenario.

CHAPTER 6

Conclusions and Perspectives

6.1 Conclusions

In this thesis, we addressed the problem of allocating radio resources in 5G cellular networks for eMBB and mMTC usage scenarios. Specifically, we presented four research activity.

In the first one, we focused on the radio resource scheduling problem in a sub-6GHz band eMBB scenario and proposed a channel and QoS aware scheduling scheme. We conducted a simulation study under different traffic types and channel conditions, and the results show that the proposed scheduling scheme always outperforms other benchmark algorithms, by exhibiting a higher degree of fairness and larger amount of satisfied GBR DRBs.

In the second research activity, we considered a D2D-enabled MMB and proposed a TDMA-based centralized access control scheme which jointly manages D2D communications and transmissions in both the access and the backhaul networks. Extensive simulations demonstrated that our access scheme outperforms the considered reference scheme in terms of throughput, end-to-end packet delay, and fairness. Furthermore, we proposed a Radio Network Planning for the presented architecture

composed of a coverage and a capacity planning.

As regards the third research activity, we proposed two new access control schemes tailored for mMTC scenarios that optimize the radio resource allocation between PRACH and PUSCH according to the traffic load condition in every RA cycle. Moreover, the EDURD exploits also the unused PUSCH resources in a contention-based mode. Simulation results showed that the proposed dynamic dimensioning control schemes allow to achieve significant improvements in terms of both system throughput and MTC device energy consumption compared to any traditional static dimensionings.

Finally, in the last research activity, we presented a rigorous methodology for modeling and analyzing the signal processed by the gNB receiver, when the MTC devices transmit tagged preamble sequences in the PRACH. Our analytical model considered the presence of interference, noise, and multi-path fading, and its accuracy was evaluated by means of a large number of simulations.

6.2 Future works

The work done in this Dissertation in the different areas of 5G cellular networks can be considered as a step towards Beyond 5G (B5G) and 6G cellular networks. In fact, although 5G wireless systems are not fully deployed yet, B5G and 6G wireless systems are gaining more importance. These system are expected to require artificial intelligence as an essential component of their technology. In this context, during my PhD course, I attended, *inter alia*, the 2019 International School on Network and Computer Sciences with thematic "Machine Learning methodologies and applications" that was held in Lipari, the "Huawei Workshop on Intelligent IoT for 6G", and followed the "Machine Learning" course from the Department of Mathematics and Computer Sciences of the University of Catania.

The vision of B5G or 5.5G is the large use of machine learning techniques as tools to optimize the performance of the wireless communication, to optimize communications building blocks or network function blocks. The main principle for this approach is to keep the communication links or network as it is. Inspired by the great success of the typical

AI technologies, especially Machine Learning (ML) and Deep Learning (DL) in areas like computer vision, automatic speech recognition, and natural language processing, many researchers are attempting to introduce AI into mobile network systems with the capability to optimize a variety of wireless network problems [97]. Machine learning is capable of optimizing various complex mathematical problems including the problems that cannot be modeled using mathematical equations. In this context, inspired by the work [98], we are now evaluating to adopt the AI to a predictive resource allocation in the mMTC scenario to further improve the management of a huge number of access requests. Specifically, we are working on a Dense Neural Network (DNN)-based methods to estimate the traffic load, to be applied in ACB schemes and/or dynamic radio resource dimensionings. The studied methods are based on the informations related to the PRACH, and new advanced technique to exploit also the PUSCH resources, if the information available in the PRACH are not sufficient for an accurate load estimation.

On the other hand, in the 6G vision, the traditional communications should be designed from the scratch, meaning the AI and the ML techniques are not used only as an optimization tool but as a communication function, or an information processing block. The 6G network is seen as a revolutionary step since the communication will not just happen on the bit level, but on the intelligence level by means of deep neural networks. In this context, we are working on the adoption of DNNs both at the receiver and the transmitter side in a SCMA-based mMTC scenario. The main disadvantage of the SCMA technique is the complexity at the receiver side the decoding the received signal is based on the iterative MPA algorithm. Specifically, we are working on Generative Adversarial Networks (GANs), in order to generate optimal SCMA codewords on the receiver side and optimal reconstruction of the SCMA modulated signals corrupted by additive white Gaussian noise at the decoder side.

Bibliography

[1] 3GPP, "System Architecture for the 5G System (5GS); Stage 2 (Release 16)," Technical Specification (TS) 23.501, 3rd Generation Partnership Project (3GPP), 09 2019. Version 16.2.0.

[2] 3GPP, "E-UTRA; User Equipment (UE) radio transmission and reception," TS 36.101, 3GPP, 01 2019. Version 15.5.0.

[3] ITU-T, "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," Recommendation M.2083, International Telecommunication Union, Geneva, Sep. 2015.

[4] 3GPP, "Study on new radio access technology Physical layer aspects," TR 38.802, 3GPP, 09 2017. Version 14.2.0.

[5] 3GPP, "Study on new radio access technology," TR 38.912, 3rd Generation Partnership Project (3GPP), 07 2018. Version 15.0.0.

[6] 3GPP, "NR; Physical layer procedures for data," TS 38.214, 3rd Generation Partnership Project (3GPP), 09 2019. Version 15.7.0.

[7] A. Aijaz, "Towards 5G-enabled tactile internet: Radio resource allocation for haptic communications," in *2016 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 145–150, April 2016.

[8] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, pp. 1617–1655, thirdquarter 2016.

[9] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, pp. 1801–1819, Fourthquarter 2014.

[10] W. ur Rehman, J. Han, C. Yang, M. Ahmed, and X. Tao, "On scheduling algorithm for device-to-device communication in 60 GHz networks," in *2014 IEEE Wireless Commun. and Netw. Conf. (WCNC)*, pp. 2474–2479, April 2014.

[11] I. K. Son, S. Mao, M. X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *2012 Proceedings IEEE INFOCOM*, pp. 2149–2157, March 2012.

[12] Y. Niu, Y. Li, D. Jin, L. Su, and D. Wu, "A two stage approach for channel transmission rate aware scheduling in directional mmWave WPANs," *Wireless Commun. and Mobile Comput.*, vol. 16, no. 3, pp. 313–329, 2016.

[13] 3GPP, "Service accessibility," TS 22.011, 3rd Generation Partnership Project (3GPP), 06 2010. Version 9.4.0.

[14] L. Miuccio, D. Panno, and S. Riolo, "Joint Control of Random Access and Dynamic Uplink Resource Dimensioning for massive MTC in 5G NR based on SCMA," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[15] B. Fekade, T. Maksymyuk, M. Kyryk, and M. Jo, "Probabilistic recovery of incomplete sensed data in IoT," *IEEE Internet of Things Journal*, vol. 5, pp. 2282–2292, Aug 2018.

[16] D. Panno and S. Riolo, "A new joint scheduling scheme for GBR and non-GBR services in 5G RAN," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, Sep. 2018.

[17] D. Panno and S. Riolo, "An enhanced joint scheduling scheme for GBR and non-GBR services in 5G RAN," *Wireless Networks*, Jan 2020.

[18] S. Riolo, D. Panno, and A. Di Maria, "A new centralized access control for mmWave D2D communications," in *13th IEEE Intl. Conf. on Wireless and Mobile Comput. Netw. and Commun. (WiMob)*, (Rome, Italy), pp. 546–553, Oct. 2017.

[19] D. Panno and S. Riolo, "On mmWave radio network planning based on a centralized access control," in *2018 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, pp. 1–8, June 2018.

[20] D. Panno and S. Riolo, "A new centralized access control scheme for D2D-enabled mmWave networks," *IEEE Access*, vol. 7, pp. 80697–80716, 2019.

[21] L. Miuccio, D. Panno, and S. Riolo, "Dynamic uplink resource dimensioning for massive MTC in 5G networks based on SCMA," in *European Wireless 2019; 25th European Wireless Conference*, pp. 1–6, May 2019.

[22] L. Miuccio, D. Panno, and S. Riolo, "Joint congestion control and resource allocation for massive MTC in 5G networks based on SCMA," in *2019 15th International Conference on Telecommunications (ConTEL)*, pp. 1–8, July 2019.

[23] L. Miuccio, D. Panno, and S. Riolo, "Joint control of random access and dynamic uplink resource dimensioning for massive MTC in 5G NR based on SCMA," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5042–5063, 2020.

[24] L. Miuccio, D. Panno, and S. Riolo, "A New Contention-Based PUSCH Resource Allocation in 5G NR for mMTC Scenarios," *IEEE Communications Letters*, pp. 1–5, 2020.

[25] S. Riolo, D. Panno, and L. Miuccio, "Modeling and Analysis of Tagged Preamble Transmissions in Random Access Procedure for

mMTC Scenarios," *IEEE Transactions on Wireless Communications, In Press*, pp. 1–17, 2021.

[26] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G New Radio numerologies and their impact on the end-to-end latency," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1–6, Sep. 2018.

[27] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2016.

[28] D. C. Abrahão and F. H. T. Vieira, "Resource allocation algorithm for LTE networks using fuzzy based adaptive priority and effective bandwidth estimation," *Wireless Networks*, vol. 24, pp. 423–437, Feb 2018.

[29] M. Kawser, H. Farid, A. Hasin, A. Sadik, and I. Razu, "Performance comparison between Round Robin and Proportional Fair scheduling methods for LTE," *International Journal of Information and Electronics Engineering*, vol. 2, no. 5, pp. 678–681, 2012.

[30] N. Guan, Y. Zhou, L. Tian, G. Sun, and J. Shi, "QoS guaranteed resource block allocation algorithm for LTE systems," in *7th IEEE Intl. Conf. on Wireless and Mobile Comput. Netw. and Comm.*, pp. 307–312, Oct 2011.

[31] A. Akhtar and H. Arslan, "Downlink resource allocation and packet scheduling in multi-numerology wireless systems," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 362–367, April 2018.

[32] A. Morgado, K. M. S. Huq, S. Mumtaz, and J. Rodriguez, "A survey of 5G technologies: regulatory, standardization and industrial perspectives," *Digital Communications and Networks*, vol. 4, no. 2, pp. 87 – 97, 2018.

[33] R. Jain, D. Chiu, and W. Hawe, *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System*. Eastern Research Laboratory, Digital Equipment Corporation, 1984.

[34] H. Li, Q. Guo, L. Fang, and D. Huang, "Fairness and capacity analysis of opportunistic feedback protocol with proportional fair or maximum throughput scheduling," in *2012 International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–5, Oct 2012.

[35] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G evolution HSPA and LTE for mobile broadband, Second Edition*. Elsevier, 2008.

[36] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, pp. 150–154, Feb 2001.

[37] A. Banerjea and S. Keshav, "Queueing delays in rate controlled ATM networks," in *IEEE INFOCOM '93 The Conference on Computer Communications, Proceedings*, pp. 547–556 vol.2, March 1993.

[38] D. Wu, J. Wang, Y. Cai, and M. Guizani, "Millimeter-wave multimedia communications: challenges, methodology, and applications," *IEEE Communications Magazine*, vol. 53, pp. 232–238, January 2015.

[39] A. Mastrosimone and D. Panno, "Moving network based on mmWave technology: a promising solution for 5G vehicular users," *Wireless Networks, 2017*, Mar 2017.

[40] S. K. Yoo, S. L. Cotton, R. W. Heath, and Y. J. Chun, "Measurements of the 60 GHz UE to eNB channel for small cell deployments," *IEEE Wireless Commun. Lett.*, vol. 6, pp. 178–181, April 2017.

[41] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, pp. 2052–2069, Oct 2015.

[42] "IEEE standard for high data rate wireless multi-media networks," *IEEE Std 802.15.3-2016 (Revision of IEEE Std 802.15.3-2003*, pp. 1–510, July 2016.

[43] "ISO/IEC/IEEE International Standard for Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks–Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band (adoption of IEEE Std 802.11ad-2012)," *ISO/IEC/IEEE 8802-11:2012/Amd.3:2014(E)*, pp. 1–634, March 2014.

[44] S. Scott-Hayward and E. Garcia-Palacios, "Multimedia resource allocation in mmwave 5G networks," *IEEE Communications Magazine*, vol. 53, pp. 240–247, January 2015.

[45] L. X. Cai, L. Cai, X. Shen, and J. W. Mark, "REX: A randomized EXclusive region based scheduling scheme for mmWave WPANs with directional antenna," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 113–121, January 2010.

[46] L. Wang, S. Liu, M. Chen, G. Gui, and H. Sari, "Sidelobe interference reduced scheduling algorithm for mmWave device-to-device communication networks," *Peer-to-Peer Networking and Applications*, vol. 12, pp. 228–240, Jan 2019.

[47] Y. Niu, L. Su, C. Gao, Y. Li, D. Jin, and Z. Han, "Exploiting device-to-device communications to enhance spatial reuse for popular content downloading in directional mmWave small cells," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 5538–5550, July 2016.

[48] C. Gao, Y. Li, H. Fu, Y. Niu, D. Jin, S. Chen, and H. Zhu, "Evaluating the impact of user behavior on D2D communications

in millimeter-wave small cells," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 6362–6377, July 2017.

[49] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "D2D-Aware device caching in mmWave-cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 35, pp. 2025–2037, Sept 2017.

[50] N. Wei, X. Lin, and Z. Zhang, "Optimal relay probing in millimeter-wave cellular systems with device-to-device relaying," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 10218–10222, Dec 2016.

[51] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, pp. 101–107, June 2011.

[52] H. Deng and A. Sayeed, "Mm-Wave MIMO channel modeling and user localization using sparse beamspace signatures," in *2014 IEEE 15th Intl. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, pp. 130–134, June 2014.

[53] S. Singh, F. Ziliotto, U. Madhow, E. Belding, and M. Rodwell, "Blockage and directivity in 60 GHz wireless personal area networks: from cross-layer model to multihop MAC design," *IEEE J. Sel. Areas Commun.*, vol. 27, pp. 1400–1413, October 2009.

[54] 3GPP, "Study on channel model for frequency spectrum above 6 GHz," Technical Report (TR) 38.901, 3rd Generation Partnership Project (3GPP), 06 2018. Version 15.0.0.

[55] K. Haneda, H. Asplund, D. Steer, C. Li, T. Balercia, A. Ghosh, T. Thomas, T. Nakamurai, Y. Kakishima, T. Imai, H. Papadopoulas, T. S. Rappaport, G. R. MacCartney, M. K. Samimi, O. Koymen, E. Mellios, A. F. Molisch, S. S. Ghassamzadeh, and A. Ghosh, "Indoor 5G 3GPP-like channel models for office and shopping mall environments," in *2016 IEEE International Conference on Communications Workshops (ICC)*, pp. 694–699, May 2016.

[56] F. Thomson Leighton, "A graph coloring algorithm for large scheduling problems," *Journal of Research of the National Bureau of Standards*, vol. 84, 11 1979.

[57] D. Brélaz, "New methods to color the vertices of a graph," *Commun. ACM*, vol. 22, pp. 251–256, Apr. 1979.

[58] L. Ericsson, "More than 50 billion connected devices," *White Paper*, vol. 14, p. 124, 2011.

[59] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 12–18, 2014.

[60] J. Zhou, R. Qingyang Hu, and Y. Qian, "Scalable distributed communication architectures to support advanced metering infrastructure in smart grid," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, pp. 1632–1642, Sep. 2012.

[61] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, pp. 4–16, First 2014.

[62] L. Ferdouse, A. Anpalagan, and S. Misra, "Congestion and overload control techniques in massive M2M systems: a survey," *Trans. on Emerging Telecomm. Technol.*, vol. 28, no. 2, p. e2936, 2017. e2936 ett.2936.

[63] M. Moltafet, N. Mokari, M. R. Javan, H. Saeedi, and H. Pishro-Nik, "A new multiple access technique for 5G: Power Domain Sparse Code Multiple Access (PSMA)," *IEEE Access*, vol. 6, pp. 747–759, 2018.

[64] M. Aldababsa, M. Toka, S. Gökçeli, G. Kurt, and O. Kucur, "A tutorial on nonorthogonal multiple access for 5G and beyond," *Wireless Communications and Mobile Computing*, vol. 2018, p. 24, 2018.

[65] V. P. Klimentyev and A. B. Sergienko, "Detection of SCMA signal with channel estimation error," in *2016 18th Conference of Open*

Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), pp. 106–112, April 2016.

[66] M. Tavares, D. Samardzija, H. Viswanathan, H. Huang, and C. Kahn, "A 5G lightweight connectionless protocol for massive Cellular Internet of Things," in *2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6, March 2017.

[67] H. S. Dhillon, H. Huang, and H. Viswanathan, "Wide-area wireless communication challenges for the Internet of Things," *IEEE Communications Magazine*, vol. 55, pp. 168–174, February 2017.

[68] H. S. Jang, S. M. Kim, H. Park, and D. K. Sung, "Message-embedded random access for cellular M2M communications," *IEEE Communications Letters*, vol. 20, no. 5, pp. 902–905, 2016.

[69] H. S. Jang, S. M. Kim, H. Park, and D. K. Sung, "A preamble collision resolution scheme via tagged preambles for cellular IoT/M2M communications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1825–1829, 2018.

[70] H. S. Jang, S. M. Kim, H. Park, and D. K. Sung, "An early preamble collision detection scheme based on tagged preambles for cellular M2M random access," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 5974–5984, July 2017.

[71] L. Tello-Oquendo, J. Vidal, V. Pla, and L. Guijarro, "Dynamic access class barring parameter tuning in LTE-A networks with massive M2M traffic," in *2018 17th Med-Hoc-Net Workshop*, pp. 1–8, June 2018.

[72] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Reinforcement learning-based ACB in LTE-A networks for handling massive M2M and H2H communications," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2018.

[73] Y. Gu, Q. Cui, Q. Ye, and W. Zhuang, "Game-theoretic optimization for machine-type communications under QoS guarantee," *IEEE Internet of Things Journal*, vol. 6, pp. 790–800, Feb 2019.

[74] N. Zhang, G. Kang, J. Wang, Y. Guo, and F. Labeau, "Resource allocation in a new random access for M2M communications," *IEEE Communications Letters*, vol. 19, pp. 843–846, May 2015.

[75] 3GPP, "Physical Channels and Modulation," TS 36.211, 3rd Generation Partnership Project (3GPP), 9 2019. Version 15.7.0.

[76] S. Moon, H. Lee, and J. Lee, "SARA: Sparse code multiple access-applied random access for IoT devices," *IEEE Internet of Things Journal*, vol. 5, pp. 3160–3174, Aug 2018.

[77] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 9847–9861, Dec 2016.

[78] 3GPP, "NR; Physical layer procedures for control," Technical Specification (TS) 38.213, 3rd Generation Partnership Project (3GPP), 09 2019. Version 15.7.0.

[79] J. Lee and J. Lee, "Prediction-based energy saving mechanism in 3GPP NB-IoT networks," in *Sensors*, 2017.

[80] GSMA, "5G Implementation Guideline," July 2019. Version 2.0.

[81] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th Annual Intl. Sym. on PIMRC*, pp. 332–336, Sep. 2013.

[82] M. Moltafet, N. M. Yamchi, M. R. Javan, and P. Azmi, "Comparison study between PD-NOMA and SCMA," *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 1830–1834, Feb 2018.

[83] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pp. 1–5, Sep. 2014.

[84] M. Vameghestahbanati, I. D. Marsland, R. H. Gohary, and H. Yanikomeroglu, "Multidimensional constellations for uplink SCMA systems - A comparative study," *CoRR*, vol. abs/1804.05814, 2018.

[85] C. Kahn and H. Viswanathan, "Connectionless access for mobile cellular networks," *IEEE Communications Magazine*, vol. 53, pp. 26–31, Sep. 2015.

[86] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. JÃ¤ntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Communications Magazine*, vol. 53, pp. 10–17, Sep. 2015.

[87] M. Centenaro, L. Vangelista, S. Saur, A. Weber, and V. Braun, "Comparison of collision-free and contention-based radio access protocols for the Internet of Things," *IEEE Transactions on Communications*, vol. 65, pp. 3832–3846, Sep. 2017.

[88] 3GPP, "Study on RAN Improvements for Machine-type Communications," TS 37.868, 3rd Generation Partnership Project (3GPP), 10 2011. Version 11.1.0.

[89] H. He, Q. Du, H. Song, W. Li, Y. Wang, and P. Ren, "Traffic-aware ACB scheme for massive access in machine-to-machine networks," in *2015 IEEE Intl. Conf. on Commun.*, pp. 617–622, June 2015.

[90] L. Zhe, A. Meizhen, B. Linhou, and Z. Enyong, "Correlation analysis on telemetry data of manned spacecraft," in *2018 Chinese Control And Decision Conference (CCDC)*, pp. 377–380, June 2018.

[91] C. Wei, G. Bianchi, and R. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1940–1953, 2015.

[92] I. Leyva-Mayorga *et al.*, "On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7785–7799, 2017.

[93] J. W. Lindeberg, "Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 15, pp. 211–225, 1922.

[94] K. J. McGown and H. R. Parks, "The generalization of Faulhaber's formula to sums of non-integral powers," *Journal of Mathematical Analysis and Applications*, vol. 330, no. 1, pp. 571 – 575, 2007.

[95] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies," TR 38.913, 3GPP, 07 2020. Version 16.0.0.

[96] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," TR 38.901, 3GPP, 01 2020. Version 16.1.0.

[97] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, and C. S. Hong, "6G wireless systems: A vision, architectural elements, and future directions," *IEEE Access*, vol. 8, pp. 147029–147044, 2020.

[98] S. Ali, W. Saad, N. Rajatheva, K. Chang, D. Steinbach, B. Sliwa, C. Wietfeld, K. Mei, H. Shiri, H.-J. Zepernick, T. M. C. Chu, I. Ahmad, J. Huusko, J. Suutala, S. Bhadauria, V. Bhatia, R. Mitra, S. Amuru, R. Abbas, B. Shao, M. Capobianco, G. Yu, M. Claes, T. Karvonen, M. Chen, M. Girnyk, and H. Malik, "6G White Paper on Machine Learning in wireless communication networks," 2020.