ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

Ingegneria Elettronica, Telecomunicazioni e Tecnologie dell'Informazione

Ciclo 36

Settore Concorsuale: 09/F2 - Telecomunicazioni

Settore Scientifico Disciplinare: ING-INF/03 - Telecomunicazioni

Grant-free protocols for massive multiple access

Presentata da: ***Lorenzo Valentini***

Coordinatore Dottorato

Aldo Romani

Supervisore

Marco Chiani

Co-supervisore

Enrico Paolini

Esame finale anno 2024

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
Ingegneria Elettronica, Telecomunicazioni e Tecnologie dell'Informazione

Ciclo XXXVI

Settore Concorsuale: 09/F2 - Telecomunicazioni

Settore Scientifico Disciplinare: ING-INF/03 - Telecomunicazioni

GRANT-FREE PROTOCOLS FOR MASSIVE MULTIPLE ACCESS

Presentato da:
LORENZO VALENTINI

Supervisore:
Prof. Ing.
MARCO CHIANI

Coordinatore Dottorato:
Prof. Ing.
ALDO ROMANI

Co-supervisore:
Prof. Ing.
ENRICO PAOLINI

Bologna, Ottobre 2023

Keywords

Massive multiple access

Coded random access

Grant-free protocols

Massive MIMO

6G

Abstract

In next generation Internet-of-Things, the overhead introduced by grant-based multiple access protocols may engulf the access network as a consequence of the proliferation of connected devices. Grant-free access protocols are therefore gaining an increasing interest to support massive multiple access. In addition to scalability requirements, new demands have emerged for massive multiple access, including latency and reliability. The challenges envisaged for future wireless communication networks, particularly in the context of massive access, include: i) a very large population size of low power devices transmitting short packets; ii) an ever-increasing scalability requirement; iii) a mild fixed maximum latency requirement; iv) a non-trivial requirement on reliability. To this aim, we suggest the joint utilization of grant-free access protocols, massive MIMO at the base station side, framed schemes to let the contention start and end within a frame, and succesive interference cancellation techniques at the base station side. In essence, this approach is encapsulated in the concept of coded random access with massive MIMO processing.

These schemes can be explored from various angles, spanning the protocol stack from the physical (PHY) to the medium access control (MAC) layer. In this thesis, we delve into both of these layers, examining topics ranging from symbol-level signal processing to succesive interference cancellation-based scheduling strategies. In parallel with proposing new schemes, our work includes a theoretical analysis aimed at providing valuable system design guidelines. As a main theoretical outcome, we propose a novel joint PHY and MAC layer design based on density evolution on sparse graphs.

Contents

Acronyms

ACK acknowledgement

AWGN additive white Gaussian noise

BCH Bose–Chaudhuri–Hocquenghem

BN burst node

BS base station

CHB channel hardening-based

CRA coded random access

CRC cyclic redundancy check

CRDSA contention resolution diversity slotted ALOHA

CSA coded slotted ALOHA

CSI channel state information

eMBB enhanced mobile broad-band

IoT Internet-of-Things

i.i.d. independent and identically distributed

IRSA irregular repetition slotted ALOHA

LDPC low-density parity-check

MAC medium access control

MIMO multiple-input multiple-output

ML maximum likelihood

MMA massive multiple access

mMTC massive machine-type communication

MPR multi-packet reception

MRC maximal ratio combining

PAB payload aided-based

PDF probability density function

PGF probability generating function

PHY physical

PLR packet loss rate

PRCE perfect replica channel estimation

QAM quadrature amplitude modulation

QPSK quadrature phase-shift keying

SC spatial coupling

SIC succesive interference cancellation

SN slot node

SNR signal-to-noise ratio

SSC spaced spatial coupling

URLLC ultra-reliable and low-latency communication

Introduction

Recent years have witnessed an increasing interest in wireless Internet-of-Things (IoT) communications. Among the scientific community and the industry there is consensus on the fact that, according to the current trend, cellular IoT will become pervasive in future 6G systems [1, 2]. Typical IoT networks involve a massive set of battery-powered (or harvesting-powered) devices autonomously transmitting "small data", i.e., short packets during short activity periods separated by random idle periods, to a common base station (BS). This regime, where the number of total devices is large and the number of active devices is an unknown subset of the whole set of devices, is often referred to as massive multiple access (MMA) and, sticking to the 5G nomenclature, the corresponding services belong to the class of massive machine-type communication (mMTC) [3–5]. Although expressions such as "massive random multiple access" or simply "massive random access" are also used in literature, to emphasize the random and intermittent devices' activity, in this dissertation, we will only adopt the most common nomenclature "massive multiple access" via its acronym MMA.

Medium access control (MAC) layer solutions, currently available for cellular IoT networks, mostly adopt scheduled and grant-based transmissions. However, coordination of a massive number of IoT devices wishing to access the channel is extremely inefficient, as it requires control signaling that may even outnumber data and increases latency. As a matter of fact, the future massive IoT networks represent a natural venue for *grant-free* and *uncoordinated* communication protocols, where the channel is dynamically shared by a very large population of nodes emitting small data at unpredictable instants, and among which only a low level of coordination (or even no coordination at all) is established [6–9]. Examples of recent grant-free schemes are the one in [10–15]. The idea behind these ac-

cess protocols is to let machine-type devices, generating per-user intermittent and sporadic traffic with random activity periods, access the channel in a grant-free fashion, i.e., without any prior agreement with the BS and without any coordination with the other devices that are active at the same time. This approach greatly simplifies the protocol on the device side, while increasing the computational complexity on the BS. Typically, devices transmit short uplink packets composed of a known preamble (i.e., a pilot) for user activity detection and channel estimation, and a payload containing (channel coded) data. Here, we refer to users and devices interchangeably.

Effective grant-free schemes for MMA applications should encompass both the physical (PHY) layer and the MAC layer, whose joint design is expected to allow optimizing the system performance. In the current stage of research, the goal is to investigate whether approaches that foresee a heavy coordination and a signaling overhead that scales with the number of devices (irrespective of their actual activity) really represent the high road for MMA applications or if grant-free approaches should be pursued. As distinctive features of MMA schemes towards 6G, they will address the potential offered by BSs featuring massive multiple-input multiple-output (MIMO) processing [16, 17] and intelligence [18, 19], i.e., capability of taking advantage from application of artificial intelligence algorithms, for example, for user activity detection.

Cellular system services are currently categorized by 3GPP into the three classes called enhanced mobile broad-band (eMBB), ultra-reliable and low-latency communication (URLLC), and mMTC [4]. These three classes are most often regarded as separate, each one with its own performance metrics and requirements. For example, in mMTC services emphasis is essentially on scalability, with no particularly constraining reliability and latency targets. In contrast, in URLLC services emphasis is on reliability and latency, whereas scalability is usually not an issue. New emerging use cases in the framework of the future IoT, including industrial IoT applications, vehicle-to-infrastructure communications, and smart city, are however calling into question this rigid scheme as they require convergence, for example, between mMTC and URLLC with different trade-off points among scalability, latency, and reliability [20–28]. Next generation MMA systems shall therefore be able to support mMTC services where scalability will still

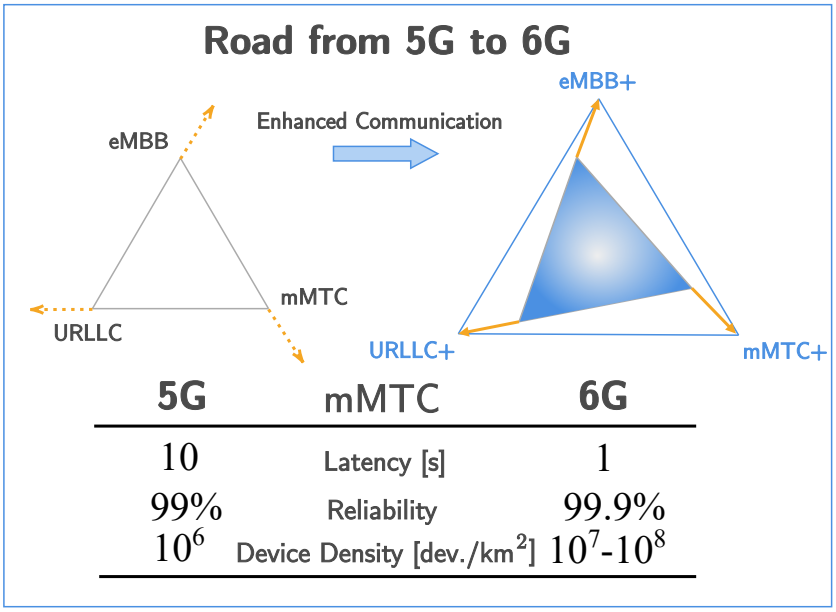

Figure 1: 5G and envisaged 6G key performance indicators, and target values for mMTC services.

be the main performance metric, but also with relatively challenging (although non-URLLC) reliability and latency requirements [3, 6, 7, 29]. Some of the requirements for mMTC are reported in Fig. 1.

In our vision, future challenges in wireless communication networks will arise from: i) a very large population size of low power devices transmitting short packets; ii) an ever-increasing scalability requirement; iii) a mild fixed maximum latency requirement; iv) a non-trivial requirement on reliability. To this aim, we suggest the joint utilization of grant-free access protocols, massive MIMO at the BS side, framed schemes to let the contention start and end within a frame, and succesive interference cancellation (SIC) techniques at the BS side. In short, coded random access (CRA) with massive MIMO processing.

The Recipe for Future Massive Multiple Access

I) We Need Slot Diversity

Slot diversity in grant-free access is a key ingredient that allows the system to reach high reliability. The basic slotted ALOHA scheme (i.e., no slot diversity is exploited) with no multi-packet reception (MPR) capability, operated without retransmissions, achieves a theoretically peak throughput of 0.37 packets/slot with a packet loss rate of 63%; this is due to the generic packet being successfully received if and only if no other transmission has occurred in the same slot. The same performance is observed in framed slotted ALOHA with no retransmissions, where slots are grouped to form fixed-length MAC frames and where each active device picks randomly one slot in the current frame and attempts packet transmission in the chosen slot. Imposing a framed system has the effect to fix a maximum latency for the communication. Focusing on reliability, if we impose the activation of only two users per frame we achieve a packet loss rate of $1/N_{\mathrm{s}}$ considering idealized PHY layer signal processing, where N_{s} is the number of slots per frame. Since the frame could not be too large for latency constraints, it is not possible to achieve good reliabilities. Higher throughput and reliability values have been obtained by introduction of slot diversity. The key idea is that each user that want to transmit an information message, transmits two copies of that information packet in two different time slots of a frame. This scheme concept, named diversity slotted ALOHA, was proposed in [30]. Considering to repeat r times the information packet in a frame composed by N_{s} slots, we can now achieve a packet-loss probability of $1/\binom{N_{\mathrm{s}}}{r}$ when considering only two active users per frame. Although, when increasing the number of active users per frame, the reliability metric drops. This behaviour can be seen in Fig. 2 where we set $N_{\mathrm{s}} = 50$ and $r = 3$. For completeness, we also report the reliability target of 99.9% envisaged for next generation protocols.

II) We Need Coded Random Access

In framed and slotted random access schemes, a major breakthrough was achieved with the introduction of succesive interference cancellation (SIC) techniques in addition to slot diversity. The idea is to transmit multiple copies of the

same packet in different slots of the frame and, whenever any of them is successfully decoded, attempt cancelling the interference generated by the replicas in the corresponding slots. In order to perform this cancellation it is required to retrieve information about the positions of the other replicas. This can be easily done by letting such positions be a function of the packet data payload. Hence, whenever we successfully decode a new packet we are able to reconstruct all the choices that the user has made.

The term SIC usually refers to an iterative procedure that subtracts (cancels in the best case scenario and attenuates in realistic cases) interference contributions from a received signal, aiming to consequently decode new information from the same signal. Through the SIC phase in our CRA scenario, the BS processes all previously decoded packets, cancelling their contribution of interference from the samples of the frame slot where they have been initially detected along with the contribution of interference of their replicas from the received signal samples in the corresponding slots. Then, in each such slot the BS reattempts to recover new packets. Processed packets are removed from the SIC buffer, while newly successfully decoded packets are added to it; this process is iterate until the buffer is empty. Under ideal assumptions regarding signal processing at PHY layer, this technique largely improves the number of simultaneously active uncoordinated devices that the system can serve while meeting a given reliability target. However, under a realistic setting it is important to define strategies to make SIC accurate and effective.

In terrestrial scenario the difficulty of this task is increased even more by the fact that the channel coefficients of any user could vary slot-by-slot. In other words, the channel coherence time is usually lower than the frame time, but larger or equal to the slot time. A possible and common channel assumption in this scenario is the block fading channel model. In such a setting, whenever a user has been successfully decoded in a slot, we cannot reuse the channel estimate obtained in that resource to subtract interference generated by replicas in other slots. Re-estimation of the channel coefficients is therefore required to make the SIC phase properly work.

The first such scheme was contention resolution diversity slotted ALOHA (CRDSA) [31], in which the number of replicas per active user is constant. The

scheme was then extended in [32] by introduction of irregular repetition slotted ALOHA (IRSA), in which users active on the same frame are allowed to employ different packet repetition rates. A very strong bridge was also established with codes on sparse graphs, which helped to significantly improve system design and analysis on simple channels such as the collision channel. A more general strategy, known as coded slotted ALOHA (CSA), has been proposed in [33] by generalizing simple packet repetitions with packet fragmentation and packet-level coding of the obtained fragments; this approach includes CRDSA and IRSA as special cases and allows achieving different trade-offs between throughput and power efficiency. Hereafter we will refer to these schemes with expression coded random access (CRA) (sometimes also referred to as "modern random access" in literature) [31–36]. If we adopt ideal SIC in diversity slotted ALOHA, obtaining then a CRA scheme, we achieve an improved reliability vs. scalability trade-off curve as reported in Fig. 2. Note that SIC does not help in improving the minimum achievable packet loss rate, which is again $1/\binom{N_\mathrm{s}}{r}$, when there are two active users in the frame.

III) We Need Massive MIMO

The possibility to deploy a massive number of antennas at the BS enables multi-packet reception at the receiver, i.e., capability to decode multiple packets per slot. In fact, given channel state information, the receiver can decode multiple interfering users in the same slot relying on channel hardening and favorable propagation (statistical quasi-orthogonality of the propagation vectors) [37]. To obtain channel state information, a simple approach consists of resorting on orthogonal pilot sequences pre-assigned to users; such an approach is however not viable in the MMA context due to the too large size K of the users' population. A possible solution to this issue consists of defining a set of $N_\mathrm{P} \ll K$ orthogonal pilots, and letting each active user pick one pilot randomly from this set in every slot in which the user performs a transmission [11]. The randomly chosen pilot is concatenated with the data payload, obtained by encoding the original message and mapping the codeword bits onto symbols of a complex constellation. Whenever a pilot is chosen by a single active user in a slot (i.e., it is a singleton pilot), the user channel can be estimated very reliably in that slot. This assumes that the

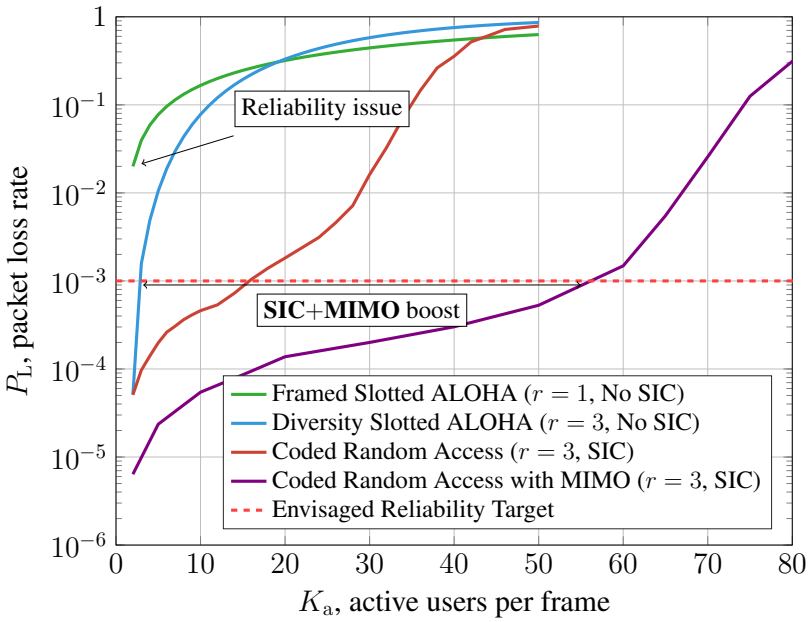

Figure 2: Reliability vs. scalability trade-off when a fixed maximum latency is imposed by the frame length. Number of slots per frame $N_{\mathrm{s}} = 50$.

coherence time is larger than the time slot. In this way, if a user is the only one using a pilot sequence and the total number of active user per slot is considerably lower than the number of antenna elements M, we can decode its data payload. In Fig. 2 we show an example using a set of two orthogonal pilots $N_{\mathrm{P}} = 2$ and $M \gg N_{\mathrm{P}}$ ($M = 64$ in the example, but we can practically obtain the same curve also with smaller M). As expected, the boost that MIMO processing can give, when jointly combined with SIC, is remarkable. In case of two active user we have now a packet loss rate of $1/\left(\binom{N_{\mathrm{s}}}{r} N_{\mathrm{P}}^{r}\right)$. We will investigate more general bounds and performance analysis through the thesis, this example is just to give a general idea to the reader.

To summarize this section, we propose in Fig. 2 a graphical representation of the improvement given by the three aforementioned ingredients for MMA. By including time diversity to framed slotted ALOHA we can achieve low error probability (high reliability). By introducing SIC algorithms we can guarantee such reliabilities also when the number of simultaneously active users per frame increases. Finally, combining SIC algorithm and MIMO processing we

can serve a larger number of simultaneously active users, while improving the maximum achievable reliability. Originally proposed in the context of satellite uplink [38, 39], CRA schemes have more recently been investigated to support massive access in terrestrial networks. Their capability to achieve unprecedented, beyond-5G tradeoff points between scalability, reliability, and latency makes them a candidate mMTC+ in the future 6G networks. This topic will be explored and investigated through this manuscript.

Thesis Organization

This manuscript aims at providing a comprehensive overview on grant-free uncoordinated access protocols for massive multiple access based on coded random access. We survey this class of protocols starting from the basic versions and then discussing several variants. Results are also presented and discussed in a systematic matter, both over simple channel models capturing the bursty and intermittent nature of transmissions but neglecting the underlying processing at PHY layer (including noise), and over PHY layer channel models that include noise and the fading effect. The thesis is structured as follows.

- Chapter 1 presents a system overview. The chapter starts by describing a general framework for MMA schemes and by introducing some nomenclature and notation used throughout the thesis. The main performance metrics considered in MMA applications are then described, followed by channel model overviews that are typically used to analyze MMA schemes. To have a common scheme to compare with other schemes and propose variants of it, we define a baseline protocol combining ideas from the literature to fit in our scenario. Finally, the density evolution tool for CRA parameters optimization is reviewed under the collision channel.

- Chapter 2 presents grant-free uncoordinated MAC protocols based on the CRA paradigm, as well as PHY layer processing, aiming at improving the scalability of the system at a given reliability. It describes several variants of the baseline scheme, spanning from spatial coupling and acknowledgement based proposals. Regarding PHY layer, it shows how to perform SIC

exploiting the payload knowledge for channel estimation and how to effectively schedule SIC operations.

- Chapter 3 is devoted to analytical results on fundamental limits for MMA. Among them we derive a lower bound for each proposed access protocol, as well as for the baseline one. We derive the analytical packet loss probability of a scheme without SIC. We evaluate the impact of PHY layer on the decoding probability to show the limits of a surrogate collision channel. Finally, we present a joint PHY and MAC layer optimization via density evolution.

- Chapter 4 reports several numerical results on selected topics from previous chapters.

Notation

Throughout the manuscript, capital and lowercase bold letters denote matrices and vectors, respectively. Symbols $(\cdot)^T$ and $(\cdot)^H$ are used to indicate transposition and conjugate transposition, respectively, while $\|\cdot\|$ denotes Euclidean norm and $|\cdot|$ denotes cardinality or absolute value, depending on the context. Regarding probability, $\mathbb{E}\{\cdot\}$ denotes expectation, $\mathbb{V}\{\cdot\}$ variance, and $\mathbb{P}\{\mathcal{E}\}$ the probability that the event $\mathcal{E}$ holds. Whenever possible, to keep a clean and compact notation through the thesis, we will denote the probability that a random variable A is equal to value a, $\mathbb{P}\{\mathsf{A}=a\}$ as $\mathbb{P}\{a\}$. Similarly, we write $\mathbb{P}\{a,b|c\}$ to indicate the probability $\mathbb{P}\{\mathsf{A}=a,\mathsf{B}=b\,|\,\mathsf{C}=c\}$.

Chapter 1

System Overview and Assumptions

Massive multiple access for machine-type traffic is a very broad topic, encompassing different layers of the communication stack and featuring different problems in terms of access protocol and PHY layer procedures. As such, the MMA literature is often fragmented as it tends to address the topic from several different perspectives. This chapter starts by offering a general outlook on synchronous MMA systems in Section 1.1, stating explicitly the typical assumptions. In Section 1.2 the performance metrics of interest for MMA applications are presented, and in Section 1.3, we review the most common channel models. Finally, in Section 1.4 we present a baseline protocol used through the thesis for comparisons.

1.1 General MMA Framework

With reference to Fig. 1.1, a typical massive multiple access (MMA) scenario consists of a very large number of transmitters, usually referred to as *users*, and one receiver. *Devices* and also user equipments are synonyms we could find in literature to refer to transmitters. This is due to the fact that users can be thought as IoT devices, wireless sensors, smart meters, etc. The common receiver is a base station (BS) of the radio access network, sometimes referred to as access point in the literature. In such a context, users want to sporadically communicate one information packet to the common BS.

The users' population size is denoted by K, while the BS is equipped with

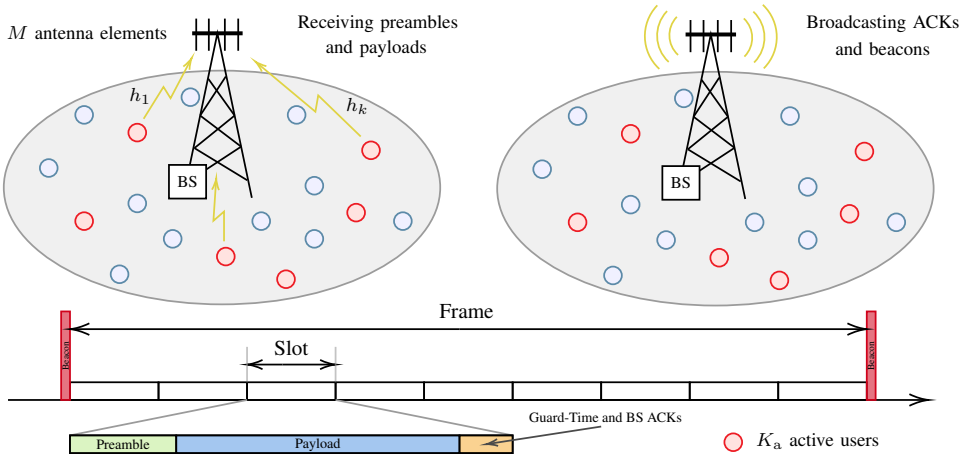

Figure 1.1: Pictorial representation of the considered scenario. There are K_a simultaneously active users, out of K users (K very large), contending for grant-free uplink to a base station with M antennas. The time is framed and slotted, the users act in an uncoordinated fashion and may interfere with each other. The base station can broadcast simple feedback messages such as beacons and acknowledgments.

M antenna elements. Both K and M have been considered large since we are targeting next generations systems. We consider a synchronous MMA system in which the time is organized in MAC frames, each composed of N_s slots. In particular, active users are both frame-synchronous and slot-synchronous with the BS. Time synchronization in framed schemes can be achieved by a beacon signal that is broadcast by the BS at the beginning of the frame. In general, this signal can be used also for other procedures such as power control operations. In this way, users which desire to communicate await for the beacon signal, and then contend for transmission in the frame starting right after the beacon. Using this approach, the maximum latency (which will be better defined in the next Section) is fixed to two times the frame length.

In this population, a user is said to be *active* when it has a new data packet to transmit towards the BS. In fact, the K users are not all simultaneously active at the same time: the number of users, active for transmission on a given frame, $K_a = K_a(t)$ is actually an integer-valued random process, taking in principle values between 0 and K, whose statistical description depends on the users' *activation model* and it is unknown to the BS. In practice, the value of K_a is usually small compared with the users' population size ($K_a \ll K$). This happens when

devices have a very low duty cycle, being in an idle state for most of the time and becoming sporadically active. In literature, this scenario is usually referred to as a massive MIMO system with K_{a} transmitting antennas and M receiving ones. Although, transmitters act in an uncoordinated fashion since they cannot communicate between each others.

Active users contend to communicate their information to the receiver, where this information is in the form of a *data packet*, sometimes also referred to as *burst*. To communicate it, an active device may use the channel a certain amount of times depending on the adopted PHY and MAC layer protocols. In wireless channels such packets are transmitted via complex symbols (or real, depending on the modulation) at some *symbol rate* $B_{\mathrm{s}}\,[\mathrm{symbols/s}]$[1]. Packets transmitted by all users have the same packet duration, that coincides with the slot duration; it follows that in a framed and slotted scheme, each packet arriving at the BS is aligned with one of the frame slots. Time slots represent an orthogonal resource that users can use for transmission.

When a user becomes active, it attempts transmission of its data packet according to an *access protocol* which, in a broad sense, defines the access and transmission rules encompassing both the PHY layer and the MAC layer. This thesis is focused on access protocols that are both *grant-free* and *uncoordinated*. In grant-free protocols, there is no handshake procedure between an active user and the BS to request and obtain the grant to transmit and schedule the transmission resources. As such, the protocol lets active users share dynamically the communication channel behaving independently and performing grant-free data transmissions without any prior agreement with the BS. Uncoordinated protocols require not only absence of coordination between active users and the BS, but also among active users; the actions taken by an active user are taken individually, without any form of coordination or cooperation with the other devices. Protocols with these characteristics are very suitable to MMA scenarios, as the number of potential users, K, is very large and devices typically activate sporadically and unpredictably to transmit short information packets. In these situations, in fact, the amount of control signalling required by a scheduled and coordinated access

[1] By symbol rate we mean the rate at which symbols are transmitted over the communication channel. This is sometimes referred to as the *baud rate*.

protocol may even outnumber the amount of data.

Grant-free protocols tend to be very simple on the device side, a desirable feature especially when devices are low-cost and energy constrained. However, these protocols also come with their drawbacks. First of all, since the devices' transmissions are uncoordinated and there is no BS scheduling, active devices may interfere each other, yielding to packet collision events, requiring extra care during the signal processing phase. For this reason, grant-free protocols tend to move most of the computational burden to the BS, whose processing becomes more complex than the one required by scheduled protocols. To this challenge, we add that the BS has no prior knowledge of the number of active users contending within a contention window (it might not even know the total population size K) and that the BS has no channel state information (CSI) for any of the active users. Therefore, before being in a position to recover the information bits of any user, the BS requires *users' activity detection* followed by *channel estimation*. The first operation is aimed at detecting active users' transmissions and may be defined in different ways depending on the protocol.

In some schemes activity detection is meant as detection of which users, out of K ones, are currently active: Not only we require to know how many users are simultaneously active, but also their IDs. In some others it is meant simply as preamble detection, without any attempt to identify which users are active and which ones are not. Sometimes it is also not required in order to let the system properly work, but can be useful to lower the BS complexity, for example avoiding to process empty slots or too crowded slots. The second operation aims at estimating CSI to perform demodulation followed by channel decoding on a detected burst. Channel estimation is typically performed on a pilot symbol sequence that is appended to the information data payload. Clearly, in order to make useful the CSI acquired on pilot symbols for demodulation and decoding, it is necessary that the time duration of a burst is less than the channel coherence time. Moreover, the transmissions occur under narrow band assumptions, which is practically equivalent to state that the symbol rate B_{s} is sufficiently lower than the coherence bandwidth.

Remarkably, the grant-free schemes proposed in this section all be contextualized within the described general model, but differ from each other in two

key aspects, namely, the access protocol on the device side and the data recovery procedure at the BS.

- The access protocol defines the actions taken by an active device to communicate its data packet to the BS by exploiting the available channel uses. It includes elements that are typical of the MAC layer, such as rules according to which a device schedules transmissions of its bursts (e.g., random choice of r slots out of N_{s} ones) or packet erasure coding. It also includes aspects related to information transmission at PHY layer, such as channel coding, modulation, preamble structure, choice of the preamble (e.g., random choice of one preamble out of τ orthogonal ones).

- The data recovery procedure defines the actions taken by the BS to recover the data packets transmitted over the current contention window. Similar to the access protocol, the data recovery procedure concerns MAC layer aspects, such as cyclic redundancy check (CRC) validation tests or packet erasure decoding, and PHY aspects. These latter include signal processing algorithms for activity detection, channel estimation, interference cancellation, soft or hard demapping, as well as channel decoding algorithms. A data recovery procedure is typically tailored to a particular access protocol. However, different data recovery procedures with different trade-offs between performance and complexity, may exist for the same access protocol.

Through the thesis, different schemes will be presented following the framework developed in this section and their performance will be compared with respect to the metrics addressed in the following Section.

1.2 Performance Metrics

In this section we summarize the main performance metrics employed to analyze grant-free schemes for MMA through this thesis. As pointed out in Introduction, new mMTC use cases are emerging towards next generation cellular IoT, where scalability is going to be the key metric, while mild reliability and latency constraints are fixed.

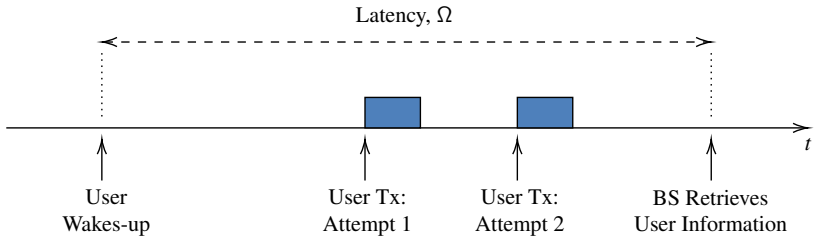

Figure 1.2: Example of a user activity and definition of the latency experienced by the user.

In mMTC applications, reliability is usually expressed as the packet loss rate (PLR) denoted as P_{L} throughout the document. This metric is defined as the probability that the data packet transmitted by the generic user is not retrieved by the BS at the end of the data recovery procedure. According to the frequentist approach of probability, it can be computed as

$$P_{\mathrm{L}} = \frac{\text{Number of unsuccessfully recovered data packets}}{\text{Total number of transmitted data packets}} \tag{1.1}$$

where the numbers in the numerator and denominator are relevant to a sufficiently long observation period. For the sake of generality, we consider the PLR conditioned to the number of active users per frame K_{a} instead of choosing a particular arrival model. In fact, it is always possible to retrieve the packet loss probability of a particular arrival model, for example a Poisson model with parameter λ, starting from the packet loss probability conditioned to the number of active users, as

$$\mathbb{P}\{\text{“packet loss”} \,|\, \lambda\} = \sum_{K_{\mathrm{a}}} \mathbb{P}\{\text{“packet loss”} \,|\, K_{\mathrm{a}}\} \; \mathbb{P}\{K_{\mathrm{a}} \,|\, \lambda\} \tag{1.2}$$

where $\mathbb{P}\{K_{\mathrm{a}} \,|\, \lambda\} = e^{-\lambda}\lambda^{K_{\mathrm{a}}}/(K_{\mathrm{a}})!$ and $\mathbb{P}\{\text{“packet loss”} \,|\, K_{\mathrm{a}}\}$ is the packet loss probability (PLR in Monte Carlo simulations) considered in the manuscript. This way, we avoid to stick with a particular arrival model, making the results more general. For the sake of simplicity, we henceforth drop the nuisances that distinguish between packet loss probability and rate, using the acronym PLR or the symbol P_{L} to indicate both of them.

On the other hand, the latency is defined as the time elapsed from the node wake up to the time instant when the BS successfully decodes the user informa-

tion, as depicted in Fig. 1.2. Concerning latency, we are interested to the maximum one imposed by the protocol. In this way, we can guarantee a given constraint. The maximum (or worst-case) latency of a successfully decoded user k as a random variables Ω_k is therefore defined as

$$\Omega = \max_k \{\Omega_k\} \,. \tag{1.3}$$

In MMA, the number of users potentially active at a given time, K_a, may be become too large, leading to a service outage, i.e., a situation where the latency or reliability constraints are not fulfilled. Mathematically, this happens when $P_\mathrm{L} \geqslant P_\mathrm{L}^*$ or $\Omega \geqslant \Omega^*$, where P_L^* and Ω^* are the maximum tolerable PLR and worst-case latency. As such, a reasonable approach to assess the performance of an MMA scheme and to compare different schemes is to evaluate the largest K_a the system can support while both latency and reliability constraints are satisfied. This is our main key performance indicator as already anticipated in Fig. 2 in the Introduction.

Another important metric is the energy efficiency on device side. Several MMA schemes feature transmission of multiple bursts per active user (this is the case, for example, for CRA-type schemes). In such a situation, we have that the energy spent by each user per data packet, E_dp, is given by

$$\mathsf{E}_\mathrm{dp} = \mathsf{N}_\mathrm{burst}\, \mathsf{E}_\mathrm{burst} \tag{1.4}$$

where $\mathsf{N}_\mathrm{burst}$ and $\mathsf{E}_\mathrm{burst}$ are independent random variables representing the number of bursts transmitted by the active user within the contention window and the energy dissipated at each burst transmission, respectively. Importantly, the parameter $\mathsf{N}_\mathrm{burst}$ depends on the MAC protocol, while the energy $\mathsf{E}_\mathrm{burst}$ depends on the PHY layer. Assuming symbols belonging to a constellation with a constant envelop, such as the quadrature phase-shift keying (QPSK) one, we can write

$$\mathsf{E}_\mathrm{burst} = \mathsf{P}_\mathrm{tx} \frac{N_\mathrm{D}}{B_\mathrm{s}} \tag{1.5}$$

where P_tx is the power per symbol at the transmitter, N_D is the number of symbols per burst, and B_s is the symbol rate. Hence, if two scheme under comparison

adopt the same PHY layer protocol, it is sufficient to compare the quantity $r_{\mathrm{avg}} = \mathbb{E}\{\mathsf{N}_{\mathrm{burst}}\}$ to have an indication about which scheme is better in term of energy consumption.

Finally, with reference to the general framework described in Section 1.1, a common performance metric particularly suitable to MMA, is represented by the sum rate, measured in total information bits per second achieved by the system as

$$\gamma = \frac{u\,\mathbb{E}\{S_{\mathrm{a}}\}}{T_{\mathrm{F}}} \tag{1.6}$$

where u is the number of information bits per active user, T_{F} is the frame time, and $S_{\mathrm{a}} \leq K_{\mathrm{a}}$ is the number of active users in a frame that have been successfully served. The sum rate can be extended to the case of packets with different lengths.

1.3 Channel Models

Massive multiple access protocols are often analyzed from a MAC layer perspective, by adopting "surrogate" channel models in which the PHY layer is mimicked in a simple, often idealized, way. These channel models are often referred to (in a broad sense) as "collision channels" and the data recovery procedures based on them as "collision resolution" approaches [40]. This class of channels can be referred to as "MAC layer" channel models since they capture essential phenomena, enabling protocol analysis, design, and optimization. As a common feature, such channels capture the bursty nature of users' transmissions but ignore the presence of noise as well as of PHY layer algorithms whose noisy operations may considerably degrade performance. Moreover, they treat bursts as "atomic units" neglecting the fact that they are composed of symbols. For a correct usage of these channel models for MAC layer analysis and design it is therefore fundamental a clear understanding of the PHY layer assumptions behind them. In the following, we introduce both some MAC layer channels and a complete PHY-MAC channel of common use.

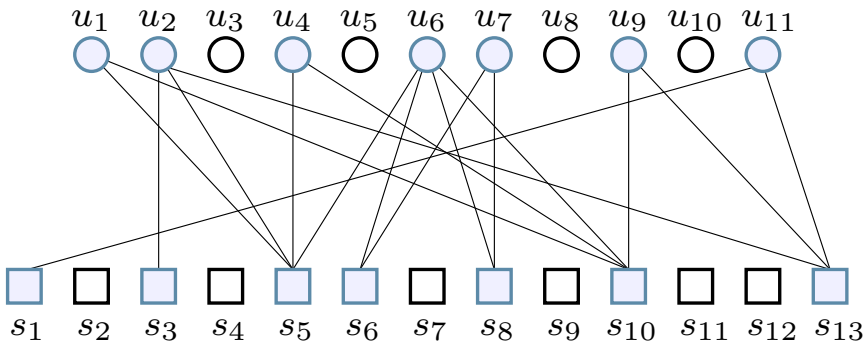

Figure 1.3: Graphical representation of IRSA access with $K = 11$ users ($K_{\mathrm{a}} = 7$ contending) and $N_{\mathrm{s}} = 13$ slots. Light-blue circles are contending users and blank circles idle users. Blank squares empty slots and light-blue squares are slots where at least one transmission occurred.

1.3.1 Synchronous Collision Channel

The synchronous collision channel model assumes that, whenever a burst, transmitted by some user in a slot, is not interfered by any burst from another user, the corresponding information bits are always correctly recovered. On the other hand, if multiple bursts arrive in the same slot, the corresponding observation at the receiver is the sum (over an appropriate field, e.g., the complex or real one) of the arriving bursts and no information can be extracted by the receiver from it. In this latter case, an error is detected by the receiver. Moreover, the receiver always discriminates with no errors between an empty slot, a slot whose observation corresponds to a single burst, and a slot whose observation corresponds to the sum of multiple bursts. Such a model dates back to [41, 42]. This channel model is useful for MAC layer analysis and design in situations characterized by effective power control (no capture effect is possible as bursts arrive at the receiver with the same power), large enough signal-to-noise ratio (SNR) (as packets corrupted only by noise are always correctly decoded), effective error detection capability at the receiver. A pictorial representation of this protocol using a bipartite graph is given in Fig. 1.3.

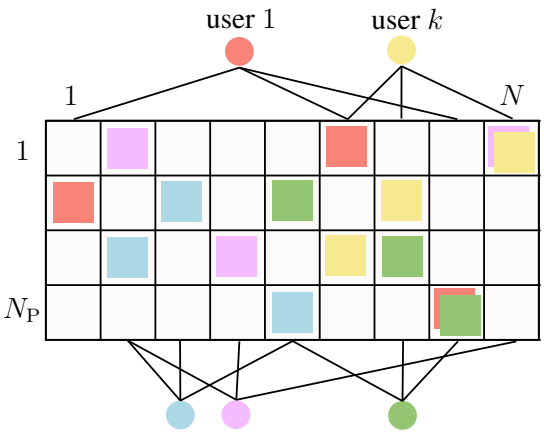

Figure 1.4: Pictorial example of conventional repetition-based CRA protocol with random pilot selection and repetition rate $r = 3$. There are $N = 9$ slots and $N_{\mathrm{P}} = 4$ orthogonal resources per slot. An active user chooses a slot r-tuple uniformly at random as well as a pilot uniformly at random in each chosen slot. For instance, user 1 chooses slots 1 with pilot 2, slot 6 with pilot 1, and slot 8 with pilot 4. Circles represent users and squares represent packets.

1.3.2 Synchronous τ-Fold Collision Channel

The τ-fold synchronous collision channel model is similar to the previous one, the difference being represented by the presence of τ orthogonal resources per slot, τ being a positive integer. A burst, received in a slot-resource pair in which no other burst is received, is always considered as correctly received (the corresponding information bits are successfully extracted). In contrast, when multiple bursts from different users arrive in the same slot-resource pair, their sum is observed in the slot-resource pair, with no possibility to extract any information. Again, the receiver always discriminates with no errors between an empty slot-resource pair, a slot-resource pair in which a single burst was received, and a slot-resource pair where multiple bursts were received, interfering each other. From an equivalent perspective, this channel may be seen as composed by τ parallel orthogonal synchronous collision channels. The synchronous τ-fold collision channel model is reasonable in situations where, besides featuring effective power control, large enough SNR, effective error detection, and effective activity detection, the system allows some form of user orthogonality in some domain. As an example, the τ parallel channels may correspond to τ different orthogonal subcarriers in a

multicarrier system. As another example, they may correspond to τ orthogonal preambles employed by the users to allow channel estimation at a massive MIMO BS. A pictorial representation of this channel is given in Fig. 1.4, where $\tau = N_{\mathrm{P}}$.

1.3.3 PHY-MAC Layer Model

To incorporate the PHY layer in the channel model, we adopt a block Rayleigh fading channel with additive white Gaussian noise (AWGN). In this channel model, the vector $\boldsymbol{h}_k = (h_{k,1}, \ldots, h_{k,M})^T \in \mathbb{C}^{M\times 1}$, where M is the number of antenna elements at the BS, is the vector of channel coefficients of the k-th user transmitting on the frame. Such a channel vector varies from slot to slot due to coherence time which is assumed equal to the slot duration T_{s}. This implies statistical independence of the channel coefficients of the same user across different slots. Moreover, the elements of $\boldsymbol{h}_k$ are modeled as zero-mean, circularly symmetric, complex Gaussian independent and identically distributed (i.i.d.) random variables, i.e., $h_{k,i} \sim \mathcal{CN}(0, \sigma_{\mathrm{h}}^2)$ for all $k \in \mathcal{A}$ and $i \in \{1, \ldots, M\}$. When coherence times are large, it is possible to subdivide slots into sub-slots as done recently in [43] (where compressed sensing and SIC across sub-slots are used), with the advantage that the user channel remains the same in all sub-slots. In this thesis we consider relatively small coherence times and for this reason we stick with the framed and slotted structure. We do not consider shadowing effects owing to the assumption of perfect power control.[2]

The packet is composed by two parts: i) a preamble of length N_{P} symbols, drawn from a set of known ones, used for channel estimation; ii) a data payload of length N_{D} symbols, containing the information bits. Hence, the signal received in a slot may be expressed as $[\boldsymbol{P}, \boldsymbol{Y}] \in \mathbb{C}^{M\times(N_{\mathrm{P}}+N_{\mathrm{D}})}$ where

$$\begin{aligned} \boldsymbol{P} &= \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{s}(k) + \boldsymbol{Z}_p \\ \boldsymbol{Y} &= \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{x}(k) + \boldsymbol{Z}. \end{aligned} \tag{1.7}$$

In (1.7), $\mathcal{A}$ is the set of users transmitting a packet (or burst) in the consid-

[2]Power control operations can be performed exploiting the beacon signal used also for synchronization.

ered slot $\boldsymbol{s}(k) \in \mathbb{C}^{1\times N_{\mathrm{P}}}$ and $\boldsymbol{x}(k) \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ are the orthogonal pilot sequence picked by user k in the current slot and the user's payload, respectively. Finally, $\boldsymbol{Z}_p \in \mathbb{C}^{M\times N_{\mathrm{P}}}$ and $\boldsymbol{Z} \in \mathbb{C}^{M\times N_{\mathrm{D}}}$ are matrices whose elements are Gaussian noise samples. The elements of both $\boldsymbol{Z}_p$ and $\boldsymbol{Z}$ are i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2)$. Due to power control, we adopt the normalization $\sigma_{\mathrm{h}}^2 = 1$ for all users' channel coefficients. Note that, we are tacitly assuming that transmissions from different users are synchronous at symbol level. This assumption allows a mathematical formulation of the problems and is essentially made in the whole MMA literature addressing signal processing at PHY layer.

1.4 Baseline Access Protocol

By "baseline" CRA protocol, we refer to a repetition-based CRA access protocol [31–33], in which each active user chooses r different slot indexes in the set $\{1, \ldots, N_{\mathrm{s}}\}$, uniformly at random without replacement. In general, according to the IRSA approach, the number of replicas (bursts) r can be drawn at random using the probability generating function (PGF) $\Lambda(x) = \sum_r \Lambda_r x^r$, where Λ_r represents the probability to pick r replicas. Whenever the number of replicas is fixed (e.g., $\Lambda(x) = x^r$), we will just indicate the value of r. In this baseline protocol, users pick uniformly at random a preamble according to a set of N_{P} available ones [11]. Due to orthogonality of the preambles, whenever a user chooses a unique preamble in a slot, it is possible to retrieve its channel vector without any interference. Ideally, we can achieve the performance of a τ-fold collision channel (see Section 1.3.2). Note that, $N_{\mathrm{P}} \ll K$ and therefore a unique pre-assignment cannot be done. For this reason, users are let to pick at random.

The BS, equipped with M antennas, process the whole frame slot-by-slot in order to retrieved users. It firstly attempts the channel estimation and then performs data payload estimations in each slot and for all possible preambles. If a packet is successfully retrieved (this can be guaranteed by a CRC), its interference is subtracted in all slots it has used for transmission, through a SIC algorithm. Due to the large amount of antenna elements, the adopted SIC is based on channel hardening assumptions as done in [11]. Finally, the BS reattempts channel and payload estimation in slots where SIC is performed, to recover new users that

were previously interfered by the deleted packet. This is repeated until no new users are found.

The details of this protocol are here reported in the next two subsections.

1.4.1 Baseline Protocol: User Side

As anticipated, the protocol has low-complexity on user side. Hereafter a detailed summary is proposed.

Repetition-based CSA with random pilot selection

1. After wake-up, the active user picks a repetition degree r according to the PGF $\Lambda(x)$.
2. It generates r different slot indexes according to a pre-defined slot selection rule.
3. For each such slot, the user chooses a preamble according to a predefined set of N_{P} orthogonal pilot sequences.
4. It appends a CRC message to the information bits, constructing a payload of k bit.
5. This bits are then encoded into n bits via a channel encoder and modulated according to a complex constellation, obtaining a payload of N_{D} symbols.
6. The device waits for the start of the next frame, signaled by the BS beacon and, by means of this signal, performs preliminary operations (e.g., synchronization, power control, etc.).
7. In each of the r pre-selected slots, the device transmits a packet composed of the corresponding pilot symbols concatenated with the data payload ones.

1.4.2 Baseline Protocol: Base Station Side

As shown in Section 1.3.3, the BS receives a signal in the form $[\boldsymbol{P}, \boldsymbol{Y}]$ in each slot. The BS processing acts on these matrices to decode all transmitting users. In particular, the processing is split into two phases: i) initialization phase; ii) SIC phase.

Initialization Phase

The purpose of this initial step is to retrieve users which have transmitted in a singleton resource. A singleton resource is defined as a slot-pilot pair which was chosen by a unique users. The initialization is performed slot-by-slot in a real-time fashion. This phase can be again split into two sub-phases.

1. Estimation of a channel coefficient for each pilot. In this step, the BS attempts maximum likelihood (ML) channel estimation for all possible pilots by computing $\boldsymbol{\phi}_j \in \mathbb{C}^{M\times 1}$, for all $j \in \{1, \ldots, N_{\mathrm{P}}\}$, as

$$\boldsymbol{\phi}_j = \frac{\boldsymbol{P}\, \boldsymbol{s}_j^H}{\|\boldsymbol{s}_j\|^2} \tag{1.8}$$

where $\boldsymbol{s}_j \in \mathbb{C}^{1\times N_{\mathrm{P}}}$ is the j-th pilot sequence.

2. Payload Estimation. During this process, the BS computes the quantities $\boldsymbol{f}_j \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ and $g_j \in \mathbb{R}$ as

$$\boldsymbol{f}_j = \boldsymbol{\phi}_j^H\, \boldsymbol{Y} \quad \text{and} \quad g_j = \|\boldsymbol{\phi}_j\|^2 . \tag{1.9}$$

Then, the BS attempts estimation of the payload using conventional maximal ratio combining (MRC) as

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{f}_j}{g_j} = \frac{\boldsymbol{\phi}_j^H\, \boldsymbol{Y}}{\|\boldsymbol{\phi}_j\|^2} . \tag{1.10}$$

Exploiting the pilot sequences orthogonality in (1.8), we have that

$$\boldsymbol{\phi}_j = \frac{\boldsymbol{P}\, \boldsymbol{s}_j^H}{\|\boldsymbol{s}_j\|^2} = \sum_{k \in \mathcal{A}^j} \boldsymbol{h}_k + \boldsymbol{z}_j \tag{1.11}$$

where $\mathcal{A}^j$ is the set of active devices employing pilot j in the current slot and $\boldsymbol{z}_j \in \mathbb{C}^{M \times 1}$ is a noise vector with i.i.d. $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2/N_{\mathrm{P}})$ entries. Note that in absence of noise, when pilot j is picked by a single user in the current slot (i.e., pilot j in that slot is a singleton resource), $\boldsymbol{\phi}_j$ equals the vector of channel coefficients for that user. Orthogonal sequences are able to guarantee an accurate channel estimation of singleton users (users which has picked a singleton resource).

The quantities $\boldsymbol{f}_j \in \mathbb{C}^{1 \times N_{\mathrm{D}}}$ and $g_j \in \mathbb{R}$ are used to the next phase (SIC phase). This is the reason why they are separately computed. Analyzing more in details (1.9), we have that

$$\boldsymbol{f}_j = \boldsymbol{\phi}_j^H\, \boldsymbol{Y} = \sum_{k \in \mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \sum_{k \in \mathcal{A}^j} \sum_{m \in \mathcal{A} \setminus \{k\}} \boldsymbol{h}_k^H\, \boldsymbol{h}_m \boldsymbol{x}(m) + \tilde{\boldsymbol{z}}_j \tag{1.12}$$

and

$$g_j = \|\boldsymbol{\phi}_j\|^2 = \sum_{k \in \mathcal{A}^j} \left(\|\boldsymbol{h}_k\|^2 + \sum_{m \in \mathcal{A}^j \setminus \{k\}} \boldsymbol{h}_m^H\, \boldsymbol{h}_k \right) + \tilde{n}_j \tag{1.13}$$

where $\tilde{\boldsymbol{z}}_j \in \mathbb{C}^{1 \times N_{\mathrm{D}}}$ and $\tilde{n}_j$ are noise terms. Under the hypothesis that all cross-terms (i.e., terms involving a product $\boldsymbol{h}_k^H\, \boldsymbol{h}_m$ with $k \neq m$) can be neglected, (1.12) and (1.13) can be approximated as

$$\boldsymbol{f}_j \approx \sum_{k \in \mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \tilde{\boldsymbol{z}}_j \tag{1.14}$$

$$g_j \approx \sum_{k \in \mathcal{A}^j} \|\boldsymbol{h}_k\|^2 + \tilde{n}_j\,. \tag{1.15}$$

This approximation holds due to channel hardening and favorable propagation [37], when $M \gg |\mathcal{A}|$. Therefore, under the same hypothesis, when a single user ℓ employs pilot j in the current slot, (1.10) can be taken as an estimate $\hat{\boldsymbol{x}}(\ell)$, of the user's payload. Symbol demapping and channel decoding is then performed

on the estimated payload $\hat{\boldsymbol{x}}$: If channel decoding returns a valid codeword and a CRC test is passed, then a message decoding success is declared and the message is stored in a buffer for the SIC phase. The same processing is executed for each slot.[3]

Successive Interference Cancellation Phase

The second step of the processing, named SIC phase, is triggered at the end of the frame. This second phase, in which an iterative subtraction of interfering terms is performed to attempt decoding of messages not yet recovered at the end of the initialization, is addressed in the following using the SIC proposed in [11]. In particular, this low-complexity SIC technique heavily rely on channel hardening, and for this reason we name it channel hardening-based (CHB) SIC. It relies on the assumption, whose range of validity is analyzed and discussed later (see Section 3.3), that in a massive MIMO setting (1.12) and (1.13) can be approximated as in (1.14) and (1.15). In other words, the algorithm relies on assuming that the cross-terms in (1.12) and (1.13) (i.e., terms featuring a product $\boldsymbol{h}_k^H \boldsymbol{h}_m$ with $k \neq m$) can be neglected with respect to the main terms.

Assume that we have initially computed $\boldsymbol{f}_j$ and g_j, $j = 1, \ldots, N_{\mathrm{P}}$, in all slots and that user ℓ payload is successfully decoded in a slot and therefore its packet stored in the SIC buffer. The BS can recover all the random choices made by the user, such as slot and pilot selections, if we simply imposed that those choices are function of the information bits. For example, this can be done by letting the information payload be the seed of a random number generator. In this way the BS acquire knowledge about where each packet has to be deleted.

The approximations in (1.14) and (1.15), lead naturally to the SIC procedure where we update $\boldsymbol{f}_j$ and g_j as

$$\boldsymbol{f}_j \leftarrow \boldsymbol{f}_j - \|\boldsymbol{h}_\ell\|^2 \, \boldsymbol{x}(\ell) \quad \text{and} \quad g_j \leftarrow g_j - \|\boldsymbol{h}_\ell\|^2 \tag{1.16}$$

in all slots where replicas of the ℓ-th user's payload are present. As such, this SIC algorithm subtracts only the main interfering term from (1.12) and (1.13). The

[3] Activity detection can be adopted to improve the BS computational complexity, avoiding to process empty or too crowded resources.

update requires knowledge of $\|\boldsymbol{h}_\ell\|^2$ in the replica slots where, due to the block fading assumption, the channel coefficients are different. For this issue, in [11] the authors invoke temporal stability of $\|\boldsymbol{h}_\ell\|^2$ through the whole frame. Here, we simply use the expectation $\mathbb{E}\{\|\boldsymbol{h}_\ell\|^2\} = M$ to perform SIC which is more accurate under block Rayleigh fading assumptions with $\sigma_{\mathrm{h}}^2 = 1$. Hence, the SIC procedure can be described by the updates

$$\boldsymbol{f}_j \leftarrow \boldsymbol{f}_j - M\,\boldsymbol{x}(\ell) \quad \text{and} \quad g_j \leftarrow g_j - M. \tag{1.17}$$

Finally, we want to foreshadow that the approximations (1.14) and (1.15) are not very accurate when the cardinality of $\mathcal{A}$ is large. In fact, since we have

$$\begin{aligned} \mathbb{E}\big\{\boldsymbol{h}_k^H\,\boldsymbol{h}_m\big\} &= 0 \\ \mathbb{V}\big\{\boldsymbol{h}_k^H\,\boldsymbol{h}_m\big\} &= M \end{aligned} \tag{1.18}$$

for $m \neq k$, the corresponding interfering terms in (1.12) and (1.13) may prevent from decoding a user packet even if it is the only one with a specific pilot. In the following we analyze this phenomenon by evaluating the probability that a user, being the only one with a specific pilot in a slot, is nevertheless not decoded.

1.5 Density Evolution over the Collision Channel

In this last section of preliminaries and background, we want to review the density evolution tool used to optimize $\Lambda(x)$ distributions over the collision channel described in Section 1.3.1. This optimization was used both for IRSA and CSA protocols [32, 33].

Let us assume that K_{a} user nodes, here referred to as burst nodes (BNs), are connected with N_{s} slot nodes, hereafter referred to as slot nodes (SNs), where K_{a} and N_{s} are, the number of contending users and the number of slot in each synchronized frame, respectively. A pictorial representation of the status of the frame is given in Figure 1.3 in Section 1.3.1. A BN has r edges representing the r replicas sent by the corresponding user. In IRSA schemes the repetition degree r is a random variable chosen according to a probability distribution with PGF

$\Lambda(x)$ [32]. Then, we can define λ_r as the probability that an edge is connected to a degree-r BN; this is given by

$$\lambda_r = \frac{\Lambda_r \, r}{\sum_h \Lambda_h \, h} \, . \tag{1.19}$$

On the other hand, each SN has c edges representing the number of users which have selected the corresponding slot to transmit a packet replica. Similarly, we define ρ_c as the probability that an edge is connected to a SN of degree c; this is given by

$$\rho_c = \frac{\Psi_c \, c}{\sum_h \Psi_h \, h} \tag{1.20}$$

where Ψ_c is probability that c users performed a transmission in the slot.

The BN degree distribution, $\Lambda(x) = \sum_r \Lambda_r \, x^r$, is the design parameter, being actually an input parameter of the density evolution procedure aimed at computing the asymptotic load threshold; this is indeed not the case for the SN degree distribution $\Psi(x) = \sum_c \Psi_c \, x^c$, that is fully defined by the system load G and by the average burst repetition rate, $\sum_r r \, \Lambda_r = \Lambda^{'}(1)$. In particular, for a large users' population size K, it is licit to assume that the number of transmissions in a slot follows a Poisson distribution. Specifically, we can write

$$\Psi_c = \frac{(G\Lambda^{'}(1))^c}{c!} \exp(-G\Lambda^{'}(1)). \tag{1.21}$$

The collision channel assumptions can be summarized as:

Assumption 1: If the number of arrivals in a slot is larger than one, then the receiver is unable to successfully decoded none of these packets.

Assumption 2: If there is only one arrival in a slot, then the packet is successfully decoded with zero error probability.

Under a collision channel model, the receiver attempts recovery of all packets using the usual iterative SIC-based procedure. Each SIC iteration comprises two steps described as follows, under the assumption that each replica carries information about the number and the position of other replicas:

1. In every resolvable resource, all packets are correctly received.

2. For each such packet, the interference of every replica of the packet is subtracted from the corresponding resource, possibly leading to new resolvable resources.

In this scenario, let $q_\ell^{(r)}$ be the probability that an edge connected to a degree-r BN is unknown at the end of iteration ℓ. Let us also define $p_\ell^{(c)}$ as the probability that an edge, connected to a degree-c SN, is unknown at the end of iteration ℓ. Then, exploiting the edge-oriented distributions λ_r and ρ_c we can define the average probabilities q_ℓ and p_ℓ as

$$q_\ell = \sum_r \lambda_r \, q_\ell^{(r)} \tag{1.22}$$

and

$$p_\ell = \sum_c \rho_c \, p_\ell^{(c)} \, . \tag{1.23}$$

Next, consider a degree-r BN. An edge is revealed whenever at least one of the other edges connected to the same BN has been revealed. This is true due to the repetition code assumption, which makes it possible to retrieve all the replicas from a single successfully decoded packet. Thus, the probability that a packet has not been retrieved by the "layer MAC repetition code" is

$$q_\ell^{(r)} = p_{\ell-1}^{r-1} \, . \tag{1.24}$$

Similarly, consider a degree-c SN. An edge is revealed whenever all the other edges have been revealed due to the collision channel assumptions. Hence, the probability that a packet in a slot is not cancelled by SIC is

$$p_\ell^{(c)} = 1 - (1 - q_\ell)^{c-1} \, . \tag{1.25}$$

As an example, an iteration of this procedure is depicted in Figure 1.5. For IRSA over the collision channel we therefore end up with the recursive equations

$$q_\ell = \sum_r \lambda_r \, p_{\ell-1}^{r-1} \tag{1.26}$$

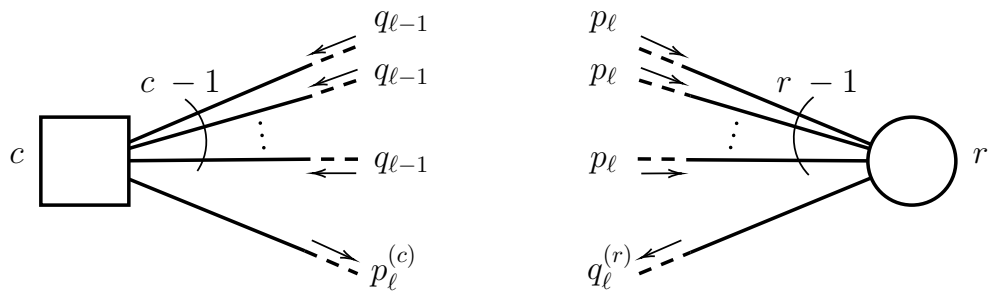

Figure 1.5: Representation of the density evolution procedure for successive interference cancellation over bipartite graphs.

and

$$p_\ell = \sum_c \rho_c \left(1 - (1 - q_\ell)^{c-1}\right) \tag{1.27}$$

where λ_r and ρ_c are degree distributions defined from an edge perspective.

Imposing as initial condition $q_0 = p_0 = 1$ (there are no revealed edges at the beginning of the process) we can define the load threshold as

$$G^\star = \sup\{G > 0 : p_\ell \to 0 \text{ as } \ell \to \infty\}\,. \tag{1.28}$$

At each $\Lambda(x)$ is therefore associated a load threshold value $G^\star$. This value is used in literature to optimize degre distributions [32, 33, 44–46]. In other works, optimization is carried out targeting different goals, such as the total power consumption [47].

As density evolution [48,49], this analysis assumes statistical independence of messages along the edges of the graph. Thus, the accuracy of (1.26) and (1.27) is subject to the absence of loops in the graph (recall that loops introduce correlation in the evolution of the erasure probabilities). This condition is met in the limit where $K_\mathrm{a} \to \infty$, $N_\mathrm{s} \to \infty$, and $K_\mathrm{a}/N_\mathrm{s} = a$ is constant.

Chapter 2

Grant-free Protocols based on Coded Random Access

To support new use cases, MMA protocols and signal processing algorithms shall be designed to address not only node density, but also latency and reliability. Both grant-based and grant-free multiple access schemes are nowadays widely used, although grant-free ones have recently attracted more interest for MMA applications. As pointed out in Introduction, this is mainly due to their ability to reduce control signalling, a very beneficial feature when the number of devices connected to the same BS becomes very large and when active devices contend for transmission of short packets. In a nutshell, the main difference between these access protocol classes lies on the presence (for grant-based) or the absence (for grant-free) of a connection establishment procedure between the BS and the machine-type device prior to data transmission.

This chapter describes several proposal we have made to improve the baseline scheme in Section 1.4 in the context of synchronous grant-free and uncoordinated access. We tackle this challenge from both a MAC and PHY layer perspective.

2.1 MAC Layer Improvements

2.1.1 CRA with Intra-Frame Spatial Coupling

In the baseline protocol (Section 1.4), each active user chooses its r different slot indexes in the set $\{1, \ldots, N_\mathrm{s}\}$ uniformly at random and without replacement. In contrast, in this access scheme, an active user only picks one slot index randomly in the set $\{1, \ldots, N_\mathrm{s} - (r-1)\}$. Denoting by n the drawn index, the user then transmits its r packet replicas in slots n, $n+1$, $\ldots$, $n+r-1$. In each such slot, the packet payload is the same, while the pilot is still chosen randomly from a set of N_P orthogonal ones. This is exemplified in Fig. 2.1 for a repetition rate $r = 3$. Compared with the baseline scheme, the introduced access strategy yields a lower degree of randomization in the choice of the slots by each user, which might jeopardize the iterative interference cancellation process. In fact, if multiple users pick the same index n, they necessarily transmit all replicas in the same r slots, increasing the probability that all transmissions from the same user experience a pilot collision.

The proposed access protocol, however, also potentially brings substantial performance advantages in terms of packet loss probability due to its capability to trigger an effect similar to the well-know spatial coupling (SC) one in the framework of low-density parity-check (LDPC) coding [50, 51]. This is due to the fact that the physical load in the first and in the last $r-1$ slots of the frame (where the physical load in a slot is defined as the number of packet replicas arriving in it) is on average lower than the physical load in the other slots. More specifically, given that there are K_a active devices, the average physical load in slot n, denoted by $G_\mathrm{phy}(n|K_\mathrm{a})$, is given by

$$G_\mathrm{phy}(n|K_\mathrm{a}) = \begin{cases} n\,\gamma & \text{if } n \in \{1, \ldots, r-1\} \\ r\,\gamma & \text{if } n \in \{r, \ldots, N_\mathrm{s}-r+1\} \\ (N_\mathrm{s}+1-n)\,\gamma & \text{if } n \in \{N_\mathrm{s}-r+2, \ldots, N_\mathrm{s}\} \end{cases} \tag{2.1}$$

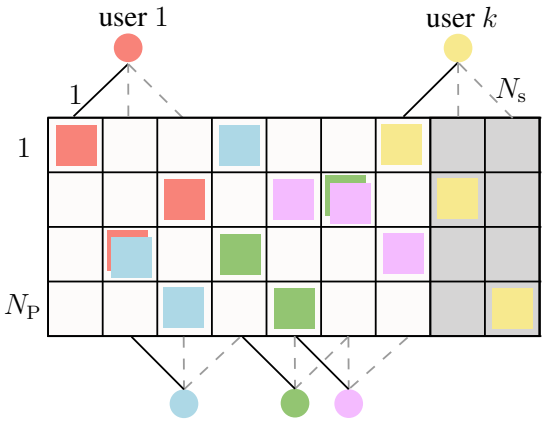

Figure 2.1: Pictorial representation of repetition-based CSA with intra-frame spatial coupling and random pilot selection. Repetition rate $r = 3$. Only the first replica slot is randomly selected, after that, the following $r - 1$ ones must be adjacent.

where

$$\gamma = \frac{K_a}{N_s - r + 1} \,. \tag{2.2}$$

Owing to the reduced load on the frame terminations, i.e., on the $r - 1$ initial and the $r - 1$ final slots of the frame, termination slots are characterized by a lower pilot collision probability; as a consequence, they exhibit a higher probability of successful packet decoding. Thus, applying a SIC algorithm, the reduced load on the frame terminations is potentially able to trigger an "interference cancellation wave" propagating from the edges of the frame towards the center of it, where previously decoded packets foster interference cancellation in adjacent slots in which new packets can be successfully decoded.

We anticipate that, this protocol can be used effectively with the feedback message strategy addressed in the following Section. 2.1.3. As it will be shown in Chapter 4, this access scheme can provide substantial enhancements to the system performance, in terms of packet loss probability versus the number of simultaneously active users K_a, when a feedback channel is available. In fact, feedback strategies fit very well with intra-frame SC strategy, favoring propagation of the forward interference cancellation wave.

This scheme was proposed in [52], and extended in [53].

2.1.2 CRA with Randomized Intra-Frame Spatial Coupling

The lower degree of randomization in the selection of the slots offered by the access protocol described in Section 2.1.1 can be mitigated introducing a window of W slots for each transmitting user, as depicted in Fig. 2.2. In this variant of the scheme, hereafter called CRA with randomized intra-frame SC, each active device initially chooses one offset slot index n at random in the set $\{1, 2, \ldots, N_{\mathrm{s}} - W + 1\}$. Then, it randomly picks the r slots in which to perform transmissions of its r packet replicas in the set $\{n, \ldots, n + W - 1\}$, uniformly and without replacement. The window size W can range from $W = r$ to $W = N_{\mathrm{s}}$, where $W = r$ corresponds to the scheme of Section 2.1.1 and $W = N_{\mathrm{s}}$ is equivalent to the baseline scheme of Section 1.4. Given that the number of active users is K_{a}, the average physical load in slot n, $G_{\mathrm{phy}}(n|K_{\mathrm{a}})$, may in this case be expressed as

$$G_{\mathrm{phy}}(n|K_{\mathrm{a}}) = c(n)\,\frac{r}{W}\,\frac{K_{\mathrm{a}}}{N_{\mathrm{s}} - W + 1}\,. \tag{2.3}$$

where

$$c(n) = \min(n,\, N_{\mathrm{s}} - W + 1) - \max(1,\, n - W + 1) + 1\,. \tag{2.4}$$

As expected, (2.3) recovers (2.1) for $W = r$ and is constant for all n when $W = N_{\mathrm{s}}$. Despite W can assume values in the range $[r, N_{\mathrm{s}}]$, in order to keep the advantages achieved using SC, W has to be chosen close to r value as it will be shown in Chapter 4. As for the previous schemes, also this protocol can be used with or without the feedback message strategy that will be addressed in Section. 2.1.3, exhibiting performance enhancements when a feedback channel is available. We also point out that in the Section 3.1 we will draw a lower bound on the performance of the presented access protocols, that turns to be tight in the low traffic regime, discussing the effect of the window size W on this lower bound. In particular, we will show how this protocol can decrease the error floor region of the performance curve, enabling the system to reach higher reliabilities.

This scheme was proposed in [53].

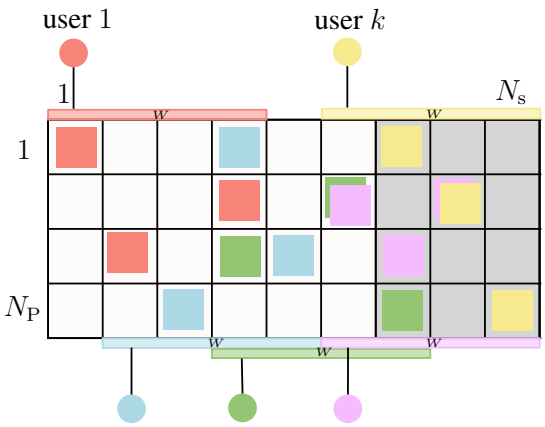

Figure 2.2: CRDSA Randomized Intra-Frame SC protocol. N_s slots, N_P orthogonal pilot sequences, repetition rate $r = 3$ and window size $W = 4$. Note that in this variant, replicas must select a slot within the window size W.

2.1.3 CRA with ACK Messages

We address in this section the role of acknowledgement (ACK) messages, or equivalently feedback messages, and how they can enhance the overall system performance. The use of ACK message that exploit the BS broadcast capabilities is in fact a clever way to improve scalability and efficiency of CRA protocols. Note that the use of ACK messages can be combined with any of the described strategies for the selection of the r slots, such as CRDSA, IRSA and even with CSA. Availability of feedback messages is not usually an issue, even in the current 5G systems.

Here we assume that a feedback message is transmitted by the BS to the users at the end of each slot. Other strategies are in principle possible when considering different frame structures [45]. Concerning the structure of ACK messages, under perfect power control we assume that they simply carry the indexes of the pilots associated with packet replicas that have been successfully decoded in the current slot. This ACK structure works when a packet replica is never correctly decoded when the corresponding pilot has been chosen by at least another active user in the same slot, which is always verified in practice if power control is enabled. Whenever this could not guaranteed other ACK messages construction can be adopted, from the naive user's IDs concatenation to more sophisticated and efficient ACK

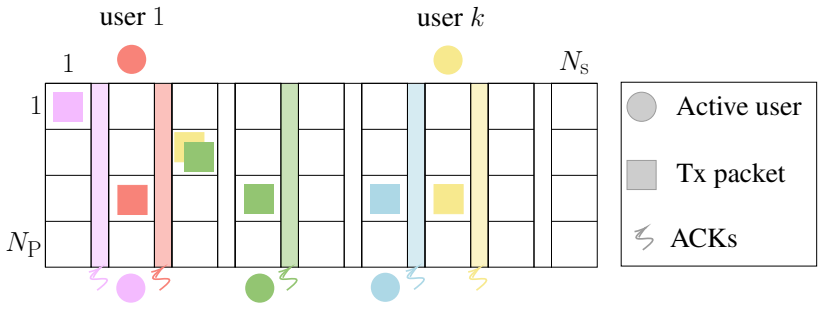

Figure 2.3: Introduction of ACK times between each slot of the frame. This provides: energy saving (less replicas transmitted), ideal interference cancellations (less packets received), but also an overhead increment.

design [54, 55].

When a device performs transmission of a packet replica without running into a pilot collision and the replica is successfully decoded, the user becomes aware of successful transmission from the ACK message broadcast by the BS at the end of the slot. The device immediately stops transmissions of the remaining replicas, which yields a two-fold benefit: i) the device consumes less transmission energy; ii) the device does not generate unnecessary interference in subsequent slots. An example is depicted in Fig. 2.3.

Very remarkably, aborting the transmission of the not yet sent replicas can be equivalently interpreted as an *ideal* cancellation of the interference that these replicas would generate in the slots where they would be transmitted in absence of ACK messages. Although this provides no performance advantages over simple surrogate channels, such as the collision channel, where interference cancellation is always assumed as ideal, when modeling the system including a realistic wireless channel model, noise, and accurate physical layer processing, these "ideal cancellations" can boost the system performance. Clearly, when the BS broadcasts an ACK message with a list of pilot indexes for which successful decoding occurred in that slot, the BS deactivates interference cancellation of all corresponding packet replicas in future slots.

With reference to the baseline access protocol presented in Section 1.4.1, point 7 specializes as

7. In each of the r pre-selected slots, the device transmits a packet composed of the corresponding pilot symbols concatenated with the data payload ones.

when no BS feedback is used, while it specializes as

7. In each of the r pre-selected slots, the device transmits a packet composed of the corresponding pilot symbols concatenated with the data payload ones, unless the user has received an ACK on successful transmission of a previous replica.

when the BS feedback is exploited.

Regardless of the feedback scheme (pilot- or ID-based), since the ACK messages are transmitted over a feedback channel that is interference-free and therefore noise-limited, a shorter preamble for channel estimation and a higher order constellation may be used for transmission of ACK messages compared to the uplink. Letting the ACK message be protected by a CRC and by a channel code with rate $\mathcal{R}_{\mathrm{a}}$, the ACK packet size in symbols is

$$N_{\mathrm{ACK}} = \left\lceil N_{\mathrm{P,ACK}} + \frac{n_{\mathrm{b}} + n_{\mathrm{CRC}}}{\mathcal{R}_{\mathrm{a}} \log_2(\mathsf{M}_{\mathrm{ACK}})} \right\rceil \tag{2.5}$$

where $N_{\mathrm{P,ACK}}$ is the ACK preamble length, n_{CRC} is the number of CRC bits, $\mathsf{M}_{\mathrm{ACK}}$ is the constellation order, and n_{b} is the number of information bits transmitted per message in the ACK time. In case of pilot-based ACKs, $n_{\mathrm{b}} = N_{\mathrm{P}}$.

The introduction of ACKs at the end of each slot produces an overhead increment. When the maximum latency is fixed (i.e., the frame time is fixed), we suffer a decrement of the total number of slots in order to have sufficient space for all ACK times. This problem can be solved with our next proposal.

The exploitation of ACK messages to improve intra-frame SC was proposed in [52], and extended in [53].

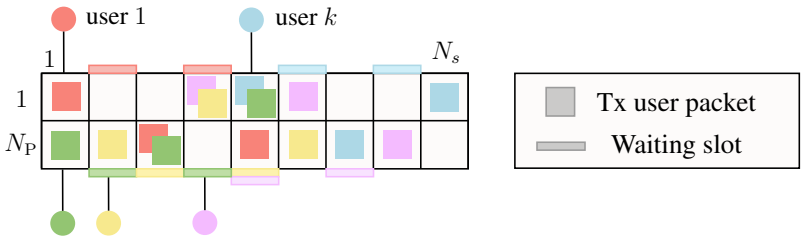

Figure 2.4: Intra-frame SSC protocol with $W_e = 1$, N_s slots per frame, N_P orthogonal pilots, and uniform repetition rate $r = 3$.

2.1.4 CRA with Spaced Spatial Coupling

The proposed spaced spatial coupling (SSC) protocols are variants of the intra-frame SC one presented in Section 2.1.1. Although, the same concepts of "spaced replicas" can be applied to other access strategy such us the one adopted for the baseline protocol. Spaced schemes are characterized by the fact that any two subsequent packet replicas transmitted by the same user are "spaced" by a certain number of waiting slots. In a first version of SSC protocols, an active user picks one slot index n randomly in the set $\{1, \ldots, N_s - (r-1)(W_e+1)\}$, where W_e is the *waiting window* size, and transmits its r packet replicas in slots n, $n+W_e+1$, $\ldots$, $n+(r-1)(W_e+1)$. The parameter W_e indicates the number of waiting slots that must occur between transmissions of two successive replicas by the same user. During the first such waiting slot, the user listens for the ACK message from the BS, which are now sent in a full-duplex way during the slots of the frame, to be informed about success of its transmission in the previous slot. In this manner, we can efficiently pipeline the feedback in order to avoid the overhead increment given by ACK times. The access protocol is exemplified in Fig. 2.4 for $W_e = 1$. Note that in SSC, the parameter W_e should not be too large, otherwise central slots in the frame suffer from higher traffic and are likely to prematurely stop the SIC algorithm.

A second version of the SSC protocol features a randomization in the number of waiting slots between two successive replicas from the same user. For this reason, we name it randomized SSC. In this case, for fixed W_e the number of waiting slots after each transmission is chosen uniformly at random by a user in the set $\{1, \ldots, W_e\}$. The randomization in the number of waiting slots is intro-

duced to achieve error floor reductions, which will be studied in Chapter 4 and clearly shown in Chapter 4. In particular, these scheme will be used in Chapter 4 to address the problem of overwhelming overheads introduced by ACK times.

This scheme was proposed in [56].

2.2 PHY Layer Enhancing Techniques

In Section 1.3.3 we addressed the physical channel model and defined analytically the corresponding $[\boldsymbol{P}, \boldsymbol{Y}]$ symbols matrix received in each slot. In Section 1.4.2, we described the CHB SIC. Here another SIC technique and a SIC scheduler are presented, aiming at improving the performance from a signal processing point of view.

2.2.1 Payload-Aided-Based Interference Cancellation

Motivated by the fact that CHB SIC does not improve the probability that a singleton is retrieved after a cancellation of a user transmitting with a different pilot in the same slot, we propose a different SIC strategy. First of all, we remind here that CHB cancellations act only on specific slot-pilot pair. More specifically, considering that a user has to be cancelled in a slot where it had picked the pilot j, CHB SIC only updates the vector $\boldsymbol{f}_j$ and the scalar g_j (see Section 1.4.2). This simple SIC mechanism has an intrinsic problem. For instance, if we have a singleton on pilot i, but due to the fact that the slot is too crowded (we have many users in the same slot picking pilot $j \neq i$), the payload estimation in (1.10) could be inaccurate, leading to an unsuccessful channel decoding. In this example, even if we cancel all interfering users in the slots, we are not able to retrieve the user in pilot i because we have never updated $\boldsymbol{f}_i$ and g_i, which again will lead to a failure decoding.

Our proposal is to use the user's payload the BS has retrieved to perform channel estimation in slots where no accurate channel estimation is available. For this reason we name this SIC algorithm payload aided-based (PAB). Assume one of the replicas sent by a user, say user ℓ, is successfully decoded in a slot, in correspondence of some pilot $\boldsymbol{s}_j$. The BS available information consists of: i) the

exact[1] user's payload $\boldsymbol{x}(\ell)$, which is common to all replicas; $ii)$ the indexes of the slots where the other replicas have been transmitted; $iii)$ the indexes of the pilots used in each such replica; $iv)$ the estimate $\boldsymbol{\phi}_j$ of the channel coefficients in the generator slot computed as per (1.11). The interference subtraction operation in the "generator" slots, i.e., where the user has been successfully decoded, is performed as

$$\begin{aligned} \boldsymbol{P}^{(i+1)} &= \boldsymbol{P}^{(i)} - \boldsymbol{\phi}_j \boldsymbol{s}_j \\ \boldsymbol{Y}^{(i+1)} &= \boldsymbol{Y}^{(i)} - \boldsymbol{\phi}_j \boldsymbol{x}(\ell) \end{aligned} \tag{2.6}$$

where we let $\boldsymbol{P}^{(0)} = \boldsymbol{P}$ and $\boldsymbol{Y}^{(0)} = \boldsymbol{Y}$. As from (2.6), in the generator slot we do not recompute the channel estimate since the estimation provided by $\boldsymbol{\phi}_j$ is impaired only by noise. Note that we are now updating $\boldsymbol{P}$ and $\boldsymbol{Y}$, avoiding in this way the CHB SIC problem.

In the other replica slots, we exploit knowledge of the payload to estimate the channel coefficients as

$$\hat{\boldsymbol{h}}_\ell^{(i)} = \frac{\boldsymbol{Y}^{(i)} \boldsymbol{x}(\ell)^H}{\|\boldsymbol{x}(\ell)\|^2} = \boldsymbol{h}_\ell + \tilde{\boldsymbol{h}}_\ell \,. \tag{2.7}$$

Then, using this PAB channel estimate, in the replica slots we can perform subtraction of interference, similar to (2.6), as

$$\begin{aligned} \boldsymbol{P}^{(i+1)} &= \boldsymbol{P}^{(i)} - \hat{\boldsymbol{h}}_\ell^{(i)} \boldsymbol{s}(\ell) \\ \boldsymbol{Y}^{(i+1)} &= \boldsymbol{Y}^{(i)} - \hat{\boldsymbol{h}}_\ell^{(i)} \boldsymbol{x}(\ell) \,. \end{aligned} \tag{2.8}$$

In this SIC algorithm, hereafter referred to as PAB, each time an update of the matrices $\boldsymbol{P}$ and $\boldsymbol{Y}$ has been carried out we re-compute (1.11) and (1.10) for each pilot in the current slot, to check if any other user can be successfully decoded after interference subtraction. We point out that exploiting the preamble (instead of the payload) to perform channel estimation in slots where we wish to subtract interference may heavily deteriorate the estimation quality due to preamble collisions.

In the particular case in which we perform the first subtraction operation in a

[1]Note that this is not the payload estimation, but the exact reconstruction after channel decoding and CRC verification.

slot using (2.8), we have

$$\hat{\boldsymbol{h}}_{\ell}^{(0)} = \boldsymbol{h}_{\ell} + \sum_{k \in \mathcal{A} \setminus \{\ell\}} \boldsymbol{h}_{k} \frac{\boldsymbol{x}(k)\, \boldsymbol{x}(\ell)^{H}}{\|\boldsymbol{x}(\ell)\|^{2}} + \boldsymbol{z}_{\mathrm{h}} \tag{2.9}$$

where $\boldsymbol{z}_{\mathrm{h}}$ is the residual noise term. In this specific case, we can derive the statistical properties of the estimation error $\tilde{\boldsymbol{h}}_{\ell}$, given that the payload symbols are independent among users, as

$$\begin{aligned} \mathbb{E}\left\{\tilde{h}_{\ell,n}\right\} &= 0 \\ \mathbb{V}\left\{\tilde{h}_{\ell,n}\right\} &= \frac{|\mathcal{A}| - 1 + \sigma_{n}^{2}}{N_{\mathrm{D}}} \end{aligned} \tag{2.10}$$

where $n = 1, \ldots, M$. We observe that, as expected, the accuracy of the channel coefficients estimate improves as the number of payload symbols increases. On the other hand, the channel estimate deteriorates as the number of users transmitting in the slot increases. Among all possible $\tilde{\boldsymbol{h}}_{\ell}$ obtained running the SIC algorithm, this represents the worst case in terms of estimation accuracy.

This SIC algorithm was proposed in [57], and extended in [58].

2.2.2 Scheduling of Interference Cancellation Operations

In this section we propose a BS processing technique that is able to improve the overall performance in different MAC and PHY layer configurations. The key idea is to introduce a priority scheduler for interference subtraction operations based on the accuracy of the corresponding channel estimates.

Let us initially focus our attention on PAB processing. Recalling (1.11), we see that pilot-based channel estimation is impaired by noise only in case of a singleton user on the j-th pilot (namely, when $|\mathcal{A}^{j}| = 1$). On the other hand, the samples corresponding to replicas of a successfully decoded packet are subtracted from the received matrices $\boldsymbol{P}$ and $\boldsymbol{Y}$ using the payload-based estimation of the channel coefficient according to (2.8). Since the payloads are not orthogonal with each other, payload-based estimation is impaired by both noise and interference. We can therefore categorize interference subtraction operations based on the ac-

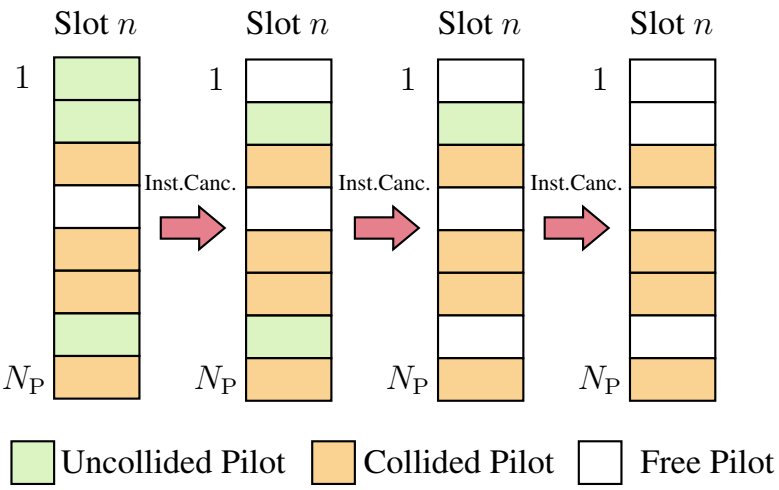

Figure 2.5: Pictorial representation of the Instantaneous Cancellation technique. In the example have been used $N_\mathrm{P} = 8$ orthogonal pilots per slot. In green are represented pilots chosen by one user (singleton), in orange the pilots used by two or more users, and in white the unused pilots.

curacy of the channel estimation on which they rely and schedule "high quality" subtractions first. Since in each SIC iteration channel estimations are performed on the current $\boldsymbol{P}$ and $\boldsymbol{Y}$ matrices, as per (1.11) and (2.7), it is expected that giving priority to those subtractions that deteriorate these matrices less (in terms of interference residue after the subtraction is performed) helps to increase the number of successful channel decoding operations, hence to trigger new SIC iterations.

Based on the above discussion, interference subtraction operations relying on singleton user pilot-based channel estimation tend to be of higher quality than the ones relying on payload-aided channel estimation and should be scheduled first. Channel estimates are affected by several parameters, namely, the noise variance and the the pilot length for pilot-based ones and the noise variance, the payload length, and the number of users active in the slot for payload-based ones, and the number of SIC iterations done. In particular, a payload-aided channel estimation, on a $\boldsymbol{Y}$ that has not been yet modified, is less accurate compared to a pilot-based one if

$$\frac{N_\mathrm{P}}{N_\mathrm{D}} \frac{|\mathcal{A}| - 1 + \sigma_\mathrm{n}^2}{\sigma_\mathrm{n}^2} > 1\,. \tag{2.11}$$

This priority SIC scheduling can be implemented adopting the following technique, that we name "instantaneous cancellation". Consider that the BS is receiving frame symbols in real-time. In conventional schemes, after the reception of each slot symbol block, the BS attempts packet decoding for each pilot. In this procedure, all decoded packets are buffered, waiting for SIC phase. When SIC phase starts, the subtractions are scheduled first-to-last (or last-to-first) decoded user. Under instantaneous cancellation scheduling, we instead perform subtractions of singleton users slot by slot, and retry the decoding step for each pilot whenever a user is successfully decoded.

This provides a second benefit which is exemplified in Fig. 2.5. In this example, we are processing a generic slot n when the total number of pilots is $N_{\mathrm{P}} = 8$. There are three singleton users in pilot $p \in \{1, 2, 7\}$, the pilot 4 is unused, while the other pilots have been chosen by more than one user. Starting from pilot one, the decoder finds a user in the first pilot. It performs instantaneous cancellation and retries the decoding phase from pilot one[2]. When the decoder attempts to decode the user in pilot 2, it fails. This behaviour is justified by the curves in Fig. 3.3, for $|\mathcal{A}^j| = 1$, which state that a singleton user could not been correctly decoded due to interference and noise. Then, the decoder finds a packet using pilot 7 and it subtracts the corresponding symbols in the slot. Due to the fact that the decoder retries from pilot one and in the slot there is less interference compared to the previous decoding step, it is possible that the packet using pilot 2 is found. In contrast, using the conventional scheduling of interference cancellation operations, the user in pilot 2 cannot be found and, even if that user is found in another slot, the subtraction in the slot n would be impaired by both noise and interference.

This algorithm fits effectively also with feedback-aided CSA protocols (Section 2.1.3), because a larger number of ACKs messages is more likely to be triggered. In general, the instantaneous cancellation technique can be seen as a pre-SIC processing which is performed slot by slot. Hence, it can be employed both by CHB and PAB processing schemes.

This SIC scheduling technique was proposed in [58].

[2]Considering that only a singleton user can be successfully decoded, it is possible to optimize this procedure avoiding to search for packets in pilots where a user has already been found.

Chapter 3

Analytical Design Tools

In this chapter we present some theoretical analysis carried out in the context of MMA. The analysis in the following sections span from MAC to PHY layer error evaluations.

3.1 Error Floor Analysis

We define the error floor region of a PLR performance curve as the curve region in which we have a low PLR and its value slowly vary when varying another parameter. On the contrary, waterfall region is defined as the curve region transitioning from high to low PLR. For the sake of clarity, we report an example in Fig. 3.1. This kind of shape is typical of PLR in MMA and channel codes where is plotted the bit error rate against the signal-to-noise ratio.

Estimation of error floors provides a useful design guideline. For example, we can tune parameters to have a lower error floor PLR compared to a target PLR. In general, to estimate error floors it is sufficient to extract the main source of errors, and then analytically derive the corresponding probability. In our setup, due to perfect power control, all packets are received with the same energy. For this reason, we can retrieve packets only if they are "alone"in a resource (slot-pilot pair). Hence, we have an unresolvable MAC layer error whenever two users pick the same slots, and in those slots they pick exactly the same pilots. Having K_{a} active users, the probability that at least two users pick the same resources can be

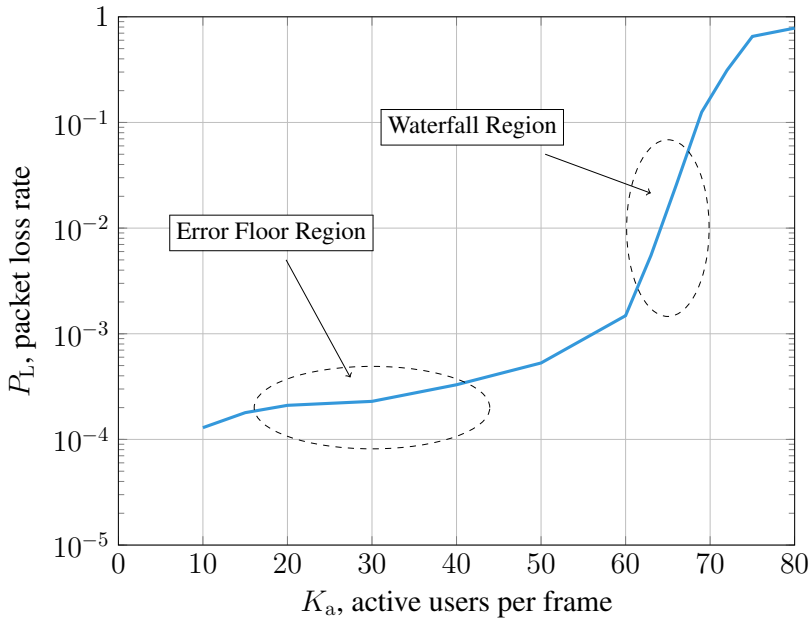

Figure 3.1: Definitions of error floor and waterfall regions.

mapped into a birthday problem.

Let us recall that the probability that at least two people (users) have the same birthday (pick the same resources) is

$$P_{\text{coll}} = 1 - \prod_{i=0}^{N-1} \frac{C-i}{C}, \tag{3.1}$$

where C is the total number of possible birthday dates (available resource choices) and N is the total number of people (number of active users $N = K_{\mathrm{a}}$). This assumes equiprobability of the available dates (all the resources can be picked with the same probability). It can be easily proven that in case of different probabilities, (3.1) represents a lower bound [59].

In general, the PLR given K_{a} can be written as

$$P_{\mathrm{L}} = \sum_{k=0}^{K_{\mathrm{a}}} \mathbb{P}\{k \text{ collisions}\} \ \frac{k}{K_{\mathrm{a}}} \tag{3.2}$$

where $\mathbb{P}\{k \text{ errors}\}$ is the probability to have k errors in a frame transmission.

Restricting our view to the sole error source of unresolvable collisions, and considering the best case scenario in which only one unresolvable collision per frame could occur and only among two users, we can lower bound (3.2) with

$$P_{\mathrm{L}} = \sum_{k=0}^{K_{\mathrm{a}}} \mathbb{P}\{k \text{ collisions}\} \, \frac{k}{K_{\mathrm{a}}} \geq \frac{2}{K_{\mathrm{a}}} \left[1 - \prod_{i=0}^{K_{\mathrm{a}}-1} \frac{C-i}{C} \right] . \tag{3.3}$$

Note that (3.3) is valid also when equiprobability of the birthday dates does not hold.

In a repetition based scheme (i.e., $\Lambda(x) = x^r$), the r replicas are placed by the active user r in different slots and, for each such slot, one pilot is chosen randomly out of the N_{P} available ones. Hence, a bijection is established between the r replicas and an r-tuple of the available resources. For the baseline scheme we have that the total number of possible r-tuple is $C_{\mathrm{B}} = \binom{N_{\mathrm{s}}}{r} N_{\mathrm{P}}^r$. This is due to the fact that we have to pick r slots at random without replacement and for each slot we can choose a pilot out of N_{P} available ones. Due to (3.3), we have that protocols with higher C exhibits lower error floor regions at a given K_{a}.

To compute the value of C for the intra-frame SC protocol described in Section 2.1.1 we directly compute $C_{\mathrm{RSC}}(W)$ for the randomized intra-frame SC protocol described in Section 2.1.2 with parameter W. This can be done because setting $W = r$ is equivalent to have a scheme without randomization. In particular, for any offset slot n there are $\binom{W}{r}$ admissible slot r-tuples in the corresponding window. Unresolvable collisions could occur not only if two users pick the same offset slot n, but also for different n. For example, taking $r = 3$ and $W = 4$; the slots triplet $(2, 3, 4)$ is included in the windows starting at $n = 1$ and $n = 2$. To correctly enumerate the admissible slot r-tuples, we slide the window of size W from the first possible offset slot $(n = 1)$ to the last one $(n = N - W + 1)$, counting each time only the r-tuples that cannot be included in the subsequent window positions (to avoid counting the same combination more times). For all windows starting at some offset slot $n < N - W + 1$, the unique r-tuples that could not be selected for $n' > n$ are the ones that include slot n nd their number is $\binom{W-1}{r-1}$. Only for the last window position, i.e., at offset slot $n = N - W + 1$, we need to

count all $\binom{W}{r}$ possible r-tuples. Hence, we have

$$C_{\mathrm{RSC}}(W) = \left[\binom{W-1}{r-1}(N_{\mathrm{s}} - W) + \binom{W}{r}\right] N_{\mathrm{P}}^{r} \tag{3.4}$$

which also yields

$$C_{\mathrm{RSC}}(W = r) = C_{\mathrm{SC}} = (N_{\mathrm{s}} - r + 1)\, N_{\mathrm{P}}^{r} \tag{3.5}$$

for the intra-frame SC scheme, and we can check that for the baseline scheme we have

$$C_{\mathrm{RSC}}(W = N_{\mathrm{s}}) = C_{\mathrm{B}} = \binom{N_{\mathrm{s}}}{r} N_{\mathrm{P}}^{r}\,. \tag{3.6}$$

It is possible to prove that the derived lower bound on the packet loss probability is monotonically decreasing with W [53]. However, increasing W may jeopardize the benefits brought SC for large number of simultaneously active users, as it will be shown in Chapter 4 by numerical analysis.

Under SSC, the r replicas from the same user are evenly spaced, any two subsequent ones being separated by exactly W_{e} slots. Having drawn the first slot from $\{1, \ldots, N_{\mathrm{s}} - (r-1)(W_{\mathrm{e}}+1)\}$, there is only one option for placement of the remaining replicas. This is the same of the SC case, but with different number of possible slots choices. Then, we have

$$C_{\mathrm{SSC}} = [N_{\mathrm{s}} - (r-1)(W_{\mathrm{e}}+1)]\, N_{\mathrm{P}}^{r}\,. \tag{3.7}$$

Analysis Output: This analysis provides useful system design guidelines. For example, using this tool it is possible to discard schemes having an error floor higher than the target PLR without running simulations. Moreover, it can be used to finely tune the randomized spatial coupling window W in order to set the floor below a specific target. In addition, since the error floor analysis is generalized for all the proposed schemes, it can be used to derive extremely low error floors in the baseline case where simulations would take extremely long time (some examples in Fig. 3.2). Since the schemes are usually design to work in the waterfall region, this tool cannot provide a straightforward performance comparisons. This

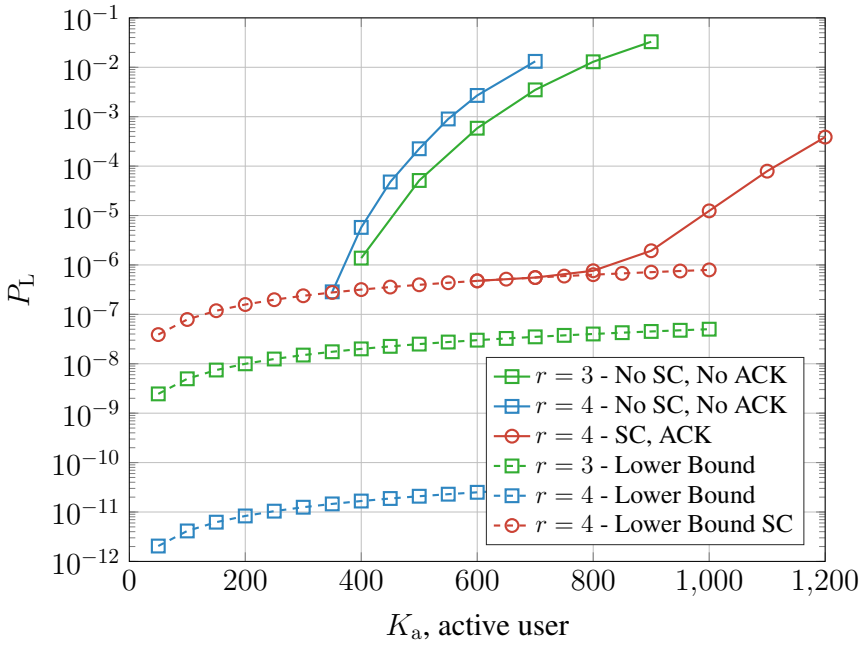

Figure 3.2: Estimation of the error floor region using lower bounds for the baseline scheme and the intra-frame SC. The number of slots per frame is set to $N_{\mathrm{s}} = 78$ and the available pilots to $N_{\mathrm{P}} = 64$.

show that the performance of the investigated access schemes in the low load regime (i.e., small number of simultaneously active users K_{a}) depends essentially on the access protocol rather than on the receiver processing and can be analyzed via simple combinatorial analysis, the performance analysis in this regime is presented here. Finally, while at high load values errors in payload decoding may fail due to several causes related to PHY layer procedures, such as imperfect interference cancellation, inaccurate channel state information acquisition, channel code decoding errors, or unresolvable collisions, in the low load regime the few error events are caused essentially by the unresolvable interference between two users choosing exactly the same resources, i.e., the same slot-pilot pairs. It turns out that the derived lower bound on the PLR that turns very tight for small K_{a}, i.e., in the error floor region.

This analysis was proposed and adopted in [52, 53, 56].

3.2 Performance Analysis without SIC

In this section we derive the average number of successfully decoded users, assuming a collision channel over resources model, when no SIC is performed.

Let us consider the following problem. There are K_{a} active devices, each of which transmits r replicas of its packet into a frame composed of N_{s} slots. The device can put no more than one replica in each slot, and in each slot it can choose between N_{P} possible orthogonal pilots. Therefore we can describe the frame as a grid of $N_{\mathrm{s}} \cdot N_{\mathrm{P}}$ resources. Defining as *uncollided* a user, any replica of which has arrived alone in a resource, under a collision channel model the number of successful users in the current frame equals the number of uncollided ones. We can write the total number of uncollided users as

$$\mathsf{X} = \mathsf{X}_1 + \mathsf{X}_2 + \cdots + \mathsf{X}_{K_{\mathrm{a}}} \tag{3.8}$$

where

$$\mathsf{X}_i = \begin{cases} 1 & \text{if at least one replica of user } i \text{ is uncollided} \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.9}$$

The average number of uncollided users can therefore be written as

$$\mathbb{E}\{\mathsf{X}\} = \sum_{i=0}^{K_{\mathrm{a}}} \mathbb{E}\{\mathsf{X}_i\} = K_{\mathrm{a}} \cdot \mathbb{P}\{\mathsf{X}_i = 1\}\ . \tag{3.10}$$

Denoting by $\mathcal{U}$ the event that the generic replica transmitted by an active user arrives alone in a resource, we have

$$\mathbb{P}\{\mathsf{X}_i = 1\} = 1 - (1 - \mathbb{P}\{\mathcal{U}\})^r\,. \tag{3.11}$$

Next, let us focus on a single replica from an active device. Let the considered replica be interfered by J replicas transmitted by other devices that have chosen

the same slot. By law of total probability we can write

$$\mathbb{P}\{\mathcal{U}\} = \sum_j \mathbb{P}\{\mathcal{U}, j\} = \sum_j \mathbb{P}\{\mathcal{U}|j\}\ \mathbb{P}\{j\} \tag{3.12}$$

where it is immediate to see that

$$\mathbb{P}\{\mathcal{U}|j\} = \left(\frac{N_\mathrm{P}-1}{N_\mathrm{P}}\right)^j . \tag{3.13}$$

To derive $\mathbb{P}\{j\}$, we firstly write the probability that none of the r replicas is transmitted in a specific slot as

$$\frac{(N_\mathrm{s}-1)\dots(N_\mathrm{s}-r)}{N_\mathrm{s}\dots(N_\mathrm{s}-r-1)} = 1-\frac{r}{N_\mathrm{s}} . \tag{3.14}$$

Consequentially, we can derive $\mathbb{P}\{j\}$ as

$$\mathbb{P}\{j\} = \binom{K_\mathrm{a}-1}{j}\left(\frac{r}{N_\mathrm{s}}\right)^j\left(1-\frac{r}{N_\mathrm{s}}\right)^{K_\mathrm{a}-1-j} \tag{3.15}$$

and conclude that

$$\begin{aligned}\mathbb{P}\{\mathcal{U}\} &= \sum_{j=0}^{K_\mathrm{a}-1}\binom{K_\mathrm{a}-1}{j}\left(\frac{r}{N_\mathrm{s}}\frac{N_\mathrm{P}-1}{N_\mathrm{P}}\right)^j\left(1-\frac{r}{N_\mathrm{s}}\right)^{K_\mathrm{a}-1-j}\\ &= \left(1-\frac{r}{N_\mathrm{s}N_\mathrm{P}}\right)^{K_\mathrm{a}-1} .\end{aligned} \tag{3.16}$$

Finally, in absence of SIC the packet loss probability is

$$P_{\mathrm{L,noSIC}} = 1-\frac{\mathbb{E}\{X\}}{K_\mathrm{a}} = \left(1-\left(1-\frac{r}{N_\mathrm{s}N_\mathrm{P}}\right)^{K_\mathrm{a}-1}\right)^r . \tag{3.17}$$

Analysis Output: This analysis can be used as a benchmark to evaluate the effectiveness of the proposed SIC strategy. Moreover, in case of framed schemes without SIC mechanisms it can be used to optimize the parameters to achieve a certain PLR at a specific traffic regime, K_a.

This analysis was proposed in [58].

3.3 Analysis of CHB Interference Cancellation

Hereafter we provide a theoretical analysis of the interference effects to understand their impact in a realistic setting. Such an analysis is conducted for the CHB SIC presented in Section 1.4.2. We remind here that $\mathcal{A}$ is the set of users transmitting simultaneously in the considered slot, while $\mathcal{A}^j \subset \mathcal{A}$ is the set of users transmitting using the pilot j in the considered slot.

Let us focus on the following scenario. Assume $|\mathcal{A}^j| - 1$ users from the set $\mathcal{A}^j$ have been successfully decoded in other slots. Then, in the current slot, we can apply CHB interference subtraction which, as previously discussed, mitigates but does not eliminate completely the interference. At this point, there is only one undecoded user adopting the j-th pilot (singleton) in the slot. We want to understand how much is likely to decode that user. Note that, if $|\mathcal{A}^j| = 1$, CHB cancellations play no role in this experiment and we are evaluating the probability to find a singleton (which was singleton from the beginning).

To analyze the probability that this user is successfully decoded, we focus on the interfering and noisy terms in (1.12). Then, from (1.12) we can write

$$\boldsymbol{f}_j = \sum_{k \in \mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \, \boldsymbol{x}(k) + \boldsymbol{I}_j \tag{3.18}$$

where

$$\begin{aligned} \boldsymbol{I}_j = & \sum_{k \in \mathcal{A}^j} \sum_{m \in \mathcal{A} \setminus \{k\}} \boldsymbol{h}_k^H \, \boldsymbol{h}_m \, \boldsymbol{x}(m) + \sum_{m \in \mathcal{A}} \boldsymbol{z}_j^H \, \boldsymbol{h}_m \, \boldsymbol{x}(m) \\ & + \sum_{k \in \mathcal{A}^j} \boldsymbol{h}_k^H \, \boldsymbol{Z} + \sum_{m \in \mathcal{A}} \boldsymbol{z}_j^H \, \boldsymbol{Z} \, . \end{aligned} \tag{3.19}$$

Let us define $\boldsymbol{\xi}_1(k,m) = \boldsymbol{h}_k^H \, \boldsymbol{h}_m \, \boldsymbol{x}(m)$. Since $\boldsymbol{h}_k$ and $\boldsymbol{h}_m$ are length-M vectors whose entries are modeled as i.i.d. $\mathcal{CN}(0,1)$ random variables and $\boldsymbol{x}$ is a length-N_{D} payload vector with i.i.d. entries, it follows that each entry $\xi_1(k,m)$, when $k \neq m$, fulfills

$$\mathbb{E}\{\xi_1(k,m)\} = 0 \, , \quad \mathbb{V}\{\xi_1(k,m)\} = M \, . \tag{3.20}$$

The second group of terms in (3.19) can be represented by $\boldsymbol{\xi}_2(m) = \boldsymbol{z}_j^H \, \boldsymbol{h}_m \, \boldsymbol{x}(m)$ where $\boldsymbol{z}_j$ is a noise vector with i.i.d. $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2/N_{\mathrm{P}})$ entries. Therefore each entry $\xi_2(m)$ fulfills

$$\mathbb{E}\{\xi_2(m)\} = 0\,, \quad \mathbb{V}\{\xi_2(m)\} = \frac{M}{N_{\mathrm{P}}}\,\sigma_{\mathrm{n}}^2\,. \tag{3.21}$$

Similarly, the third group of terms in (3.19) can be represented by $\boldsymbol{\xi}_3(k) = \boldsymbol{h}_k^H \, \boldsymbol{Z}$ where $\boldsymbol{Z}$ is a matrix whose elements are i.i.d. $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2)$. Then, each entry $\xi_3(k)$ fulfills

$$\mathbb{E}\{\xi_3(k)\} = 0\,, \quad \mathbb{V}\{\xi_3(k)\} = M\,\sigma_{\mathrm{n}}^2\,. \tag{3.22}$$

Finally the last term $\boldsymbol{\xi}_4 = \boldsymbol{z}_j^H \, \boldsymbol{Z}$ has entries characterized by

$$\mathbb{E}\{\xi_4\} = 0\,, \quad \mathbb{V}\{\xi_4\} = \frac{M}{N_{\mathrm{P}}}\,\sigma_{\mathrm{n}}^4\,. \tag{3.23}$$

We now make the approximation which considers entry independence between $\boldsymbol{\xi}_1(k,m)$, $\boldsymbol{\xi}_2(m)$, $\boldsymbol{\xi}_3(k)$, and $\boldsymbol{\xi}_4$. Under this independence assumption, each element I_j of $\boldsymbol{I}_j$ fulfills

$$\begin{aligned} &\mathbb{E}\{I_j\} = 0 \\ &\mathbb{V}\{I_j\} = M\left(|\mathcal{A}^j|\left(|\mathcal{A}| - 1 + \sigma_{\mathrm{n}}^4\right) + \frac{\sigma_{\mathrm{n}}^2}{N_{\mathrm{P}}}\left(|\mathcal{A}| + \sigma_{\mathrm{n}}^2\right)\right). \end{aligned} \tag{3.24}$$

At this point, consider the case where $|\mathcal{A}^j| - 1$ users using pilot j are decoded in other slots. Performing interference cancellation based on CHB, new residual interfering terms arise. We recast (3.18) as

$$\begin{aligned} \boldsymbol{f}_j &= \|\boldsymbol{h}_\ell\|^2 \, \boldsymbol{x}(\ell) + \sum_{k \in \mathcal{A}^j \setminus \{\ell\}} \left(\|\boldsymbol{h}_k\|^2 - M\right) \boldsymbol{x}(k) + \boldsymbol{I}_j \\ &= \|\boldsymbol{h}_\ell\|^2 \, \boldsymbol{x}(\ell) + \tilde{\boldsymbol{I}}_j \end{aligned} \tag{3.25}$$

where the subscript ℓ denotes the only remaining user employing pilot j in the slot under analysis. Since $\mathbb{E}\{\|\boldsymbol{h}_k\|^2\} = M$ and $\mathbb{V}\{\|\boldsymbol{h}_k\|^2\} = M$, we can incorporate

these terms in our approximation, leading to

$$\begin{aligned}\mathbb{E}\left\{\tilde{I}_j\right\} &= 0 \\ \mathbb{V}\left\{\tilde{I}_j\right\} &= M\left(|\mathcal{A}^j|\left(|\mathcal{A}|+\sigma_{\mathrm{n}}^4\right)-1+\frac{\sigma_{\mathrm{n}}^2}{N_{\mathrm{P}}}\left(|\mathcal{A}|+\sigma_{\mathrm{n}}^2\right)\right).\end{aligned} \tag{3.26}$$

Due to summation of a large amount of terms we can approximate $\tilde{I}_j$ as a circularly symmetric complex Gaussian distribution with mean and variance reported in (3.26). Then, dividing by M we can estimate the payload of user ℓ as

$$\hat{\boldsymbol{x}}(\ell) = \frac{\|\boldsymbol{h}_\ell\|^2}{M}\,\boldsymbol{x}(\ell) + \frac{\tilde{\boldsymbol{I}}_j}{M}\,. \tag{3.27}$$

For a realistic analysis we also consider modulation and channel coding. Employing an M-quadrature amplitude modulation (QAM) constellation and hard-decision decoding, the symbol error probability given $w = \frac{2}{\sigma_h^2}\|\boldsymbol{h}_\ell\|^2$ can be written as [60]

$$P_{\mathrm{e}|w} = A_{\mathsf{M}}\,\mathrm{erfc}\left(\sqrt{\frac{C_{\mathsf{M}}\,w^2}{\mathbb{V}\left\{\tilde{I}_j\right\}}}\right) - \frac{A_{\mathsf{M}}^2}{4}\,\mathrm{erfc}^2\left(\sqrt{\frac{C_{\mathsf{M}}\,w^2}{\mathbb{V}\left\{\tilde{I}_j\right\}}}\right) \tag{3.28}$$

where $A_{\mathsf{M}} = 2 - 2/\sqrt{\mathsf{M}}$ and $C_{\mathsf{M}} = 3/(8\mathsf{M}-8)$. Finally, we assume an error correcting code with bounded-distance decoding, able to correct up to t errors, and constellation Gray mapping. We can express the probability that decoding of a user packet is unsuccessful given w as

$$P_{\mathrm{fail}|w} \approx 1 - \sum_{d=0}^{t}\binom{N_{\mathrm{D}}}{d}P_{\mathrm{e}|w}^d\left(1-P_{\mathrm{e}|w}\right)^{N_{\mathrm{D}}-d} \tag{3.29}$$

where N_D is the number of payload symbols. Equality in (3.29) would hold if, whenever a symbol is erroneous, only one of its bits is received in error. In general this is not true, but exploiting Gray mapping this is a well-fitting approximation.

Hence, the probability to have a decoding failure of a user packet in a slot, where its $|\mathcal{A}^j|-1$ pilot-interferers are subtracted and a total of $|\mathcal{A}|$ users were

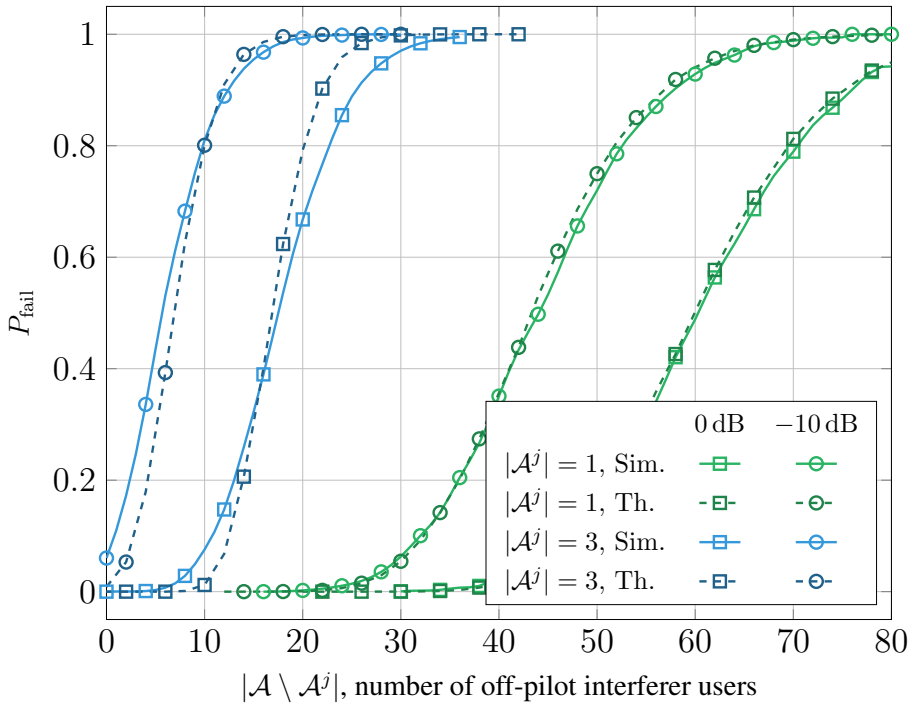

Figure 3.3: Probability to unsuccessfully decode a singleton user after $|\mathcal{A}^j| - 1$ CHB iterations. Comparison between the analytical approximation and the simulation for $N_{\text{D}} = 256$, $t = 10$, $M = 256$, QPSK constellation, and $\sigma_{\text{n}}^2 \in \{1, 10\}$.

initially allocated in the slot, is

$$P_{\text{fail}} = \int_0^{\infty} P_{\text{fail}|w} \frac{1}{2^M \, \Gamma(M)} \, w^{M-1} \, e^{-w/2} \, dw \tag{3.30}$$

due to the fact that w is distributed according to a chi-squared distribution with $2M$ degrees of freedom ($\sigma_h^2 = 1$). We observe that, to increase the resilience of singleton users to interference in terms of packet error probability, we can increase either the number of BS antennas M or the code error correction capability t for fixed N_{D} (which decreases the error correcting code rate). Note that, a simple approximation can be made, observing that for a large number of antennas M the probability density function (PDF) narrowed around the mean value and therefore we have $\mathbb{E}\{f(w)\} \simeq f(\mathbb{E}\{w\})$.

Analysis Output: We report in Fig. 3.3 the analytical approximations derived in (3.30) in comparison with Monte Carlo simulations for $N_{\text{D}} = 256$, $t = 10$,

$M = 256$, QPSK constellation, and two noise levels $\sigma_{\mathrm{n}}^2 \in \{1, 10\}$. Despite the approximations, the analytical results provide a good estimate of the simulated curves also in the presence of noise. In particular, when $|\mathcal{A}^j| = 1$, no interference subtractions are performed and the user experiences the most favorable interference conditions. The $|\mathcal{A}^j| = 1$ curve in Fig. 3.3 reveals the actual performance of MRC payload estimation in (1.10) when interferers, using different orthogonal preambles, are captured in the model. Indeed, this is a major non-ideality, degrading the general performance of MAC protocols when a realistic channel model is accounted. On the other hand, when $|\mathcal{A}^j| > 1$, the estimation deteriorates even more, revealing the non-ideality of the SIC procedure. Moreover, we point out that, whenever a device using pilot j in the current slot is successfully decoded and CHB is performed, the interference on pilots different from j is not mitigated. This is the most critical point we have identified in the CHB approach and in the next section we propose a technique that is able to overcome this problem. Finally, we point out that this analysis can be used to account for PHY layer effects in higher level analysis. This will be done in the next section.

This analysis was proposed in [57] and extended in [58].

3.4 Density Evolution over MIMO Fading Channels

In this section we describe a novel asymptotic threshold analysis able to capture realistic aspects of the PHY processing.

3.4.1 Assumptions and Channel Model

The specifications of the system, to which our threshold analysis applies, can be summarized as:

- Block fading channel with power control (variance of the fading coefficient equal to one for all users).
- The receiver has M antennas, each with independent fading coefficient per user.

- Each user picks, for each replica, an orthogonal pilot uniformly at random from a set with cardinality N_{P} for channel estimation purposes.
- Grey mapped QPSK modulation with hard decision.
- The payload is composed by N_{D} symbols and protected using a channel code able to correct up to t errors per codeword.
- CHB SIC processing and IRSA distribution.

In this scenario we refer to a slot-pilot pair as a resource. Then, if a contending device transmits d packet replicas, it chooses d resources to schedule its transmissions that must differ in the slot, since no device can send multiple packets in a single slot, but not necessarily in the pilot. Moreover, the receiver attempts recovery in all resources and, whenever decoding of some messages succeeds in a resource, the contribution of interference of the decoded user is subtracted across slots in a SIC fashion.

Example 3.1. In Figure 3.4 we provide a pictorial representation of a frame as an $N_{\mathrm{P}} \times N_{\mathrm{s}}$ grid in which each row corresponds to a pilot, each column to a slot, and each cell to a *resource*. In the specific example we have $N_{\mathrm{P}} = 4$ pilots and $N_{\mathrm{s}} = 9$ slots; moreover, there are $K_{\mathrm{c}} = 5$ contending users, corresponding to the circles, all exploiting repetition rate $d = 3$. Note that in the example, only two packet replicas do not experience a resource collision (i.e., pilot collision in a slot), one in slot 2 and one in slot 7. Furthermore, extending the simple collision channel model to a "pilot-based collision channel" where we consider collisions on the resources, the messages of active users i_1 and i_5 are decoded at the first iteration in slots 2 and 7, respectively. Interference subtraction allows cleaning the packet replica of user i_2 in slot 5; hence, the message of user i_2 is decoded at the second iteration. A further stage of interference subtraction allows decoding messages of users i_3 and i_4 in the third iteration.

Since SIC and packet decoding in realistic systems are not well-approximated by collision channel assumptions, we introduce a novel PHY layer-aware channel model. Similarly to collision channel, this realistic channel assumes that, when two or more users choose the same resource, it is not possible to successfully

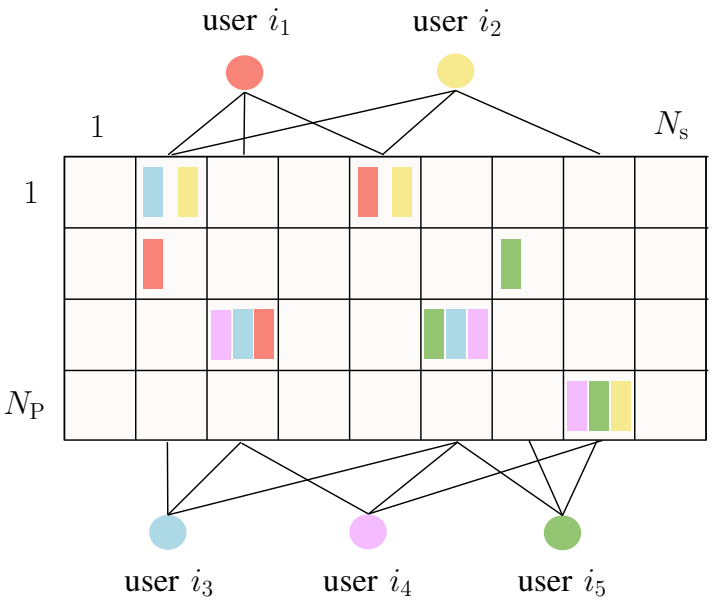

Figure 3.4: Grid-based representation of the pilot-based access protocol using $N_P = 4$ pilots, $N_s = 9$ slots, $d = 3$ repetition code, and $K_c = 5$ contending users. Assuming a collision channel model over resources, through SIC iterations it is possible to successfully decode messages from all users.

decode the corresponding packet replicas (note that no capture effect based on energy diversity is possible owing to power control). On the other hand, when only a user chooses a resource, the packet is successfully decoded according to a probability depending on the total number of interfering users in that slot and the PHY layer parameters. We adopt the same graphical representation mentioned in Section 1.5 to describe the realistic channel model: a bipartite graph with K_a BNs connected to N_s SNs.

Example 3.2. The left side of Figure 3.5 highlights a particular slot. The top-right side of the figure shows an example using $N_P = 8$ pilots (p_0, p_1, ..., p_7) showing the pilots choice of the users that have transmitted a replica inside the s_4 slot. Finally, the bottom-right side of the describes a possible configuration of the s_4 slot after some SIC iterations when PHY layer is considered. Adopting the standard collision channel, s_4 is a collision slot and, at the SIC-initialization step, none of its packets can be recovered. On the other hand, adopting a collision channel over the resources, the users u_0 and u_1 are resolvable due to the fact

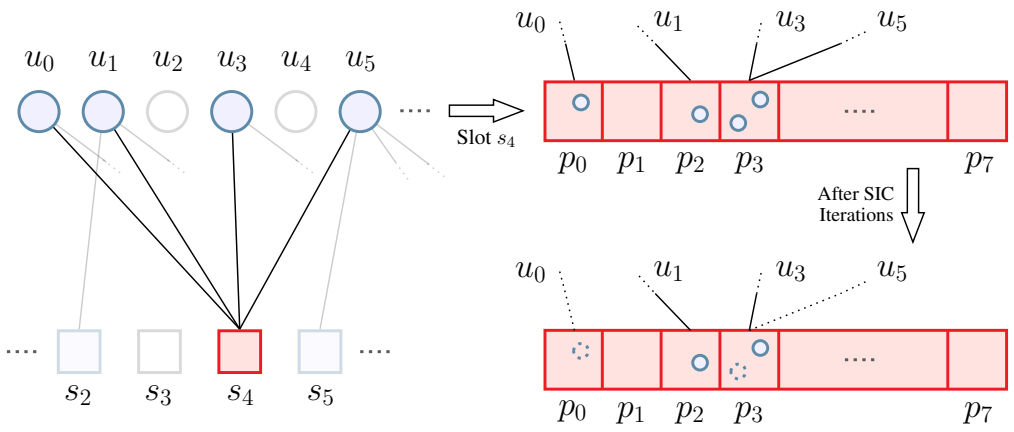

Figure 3.5: An example of SIC over block fading channel with massive MIMO.

that they transmit using pilots p_0 and p_2, respectively. Note that, unresolvable collisions of u_3 and u_5 could be resolved after SIC iterations. Nevertheless, in realistic scenarios, assuming collision channel among the resources could be way too optimistic. In fact, due to payload estimation failures it is possible that the packet from user u_1 cannot be decoded despite it chooses an uncollided pilot. The same could happen to users like u_3 which was initially collided by user u_5 and through SIC iterations across slots remains the only one using pilot p_3. In this example, only user u_0 is successfully decoded as a result of slot s_4 processing.

3.4.2 Density Evolution taking into account the Physical Layer

To derive ad-hoc density evolution equations to account for PHY layer it is necessary to find the probability update function as in in (1.24) and (1.25). To be fair, since we are using the same MAC layer protocol, the expression in (1.24) remains the same, while (1.25) changes due to the introduction of the PHY layer. As a reminder, we have to compute the probability $p_\ell^{(c)}$ that a packet replica, arriving in a slot where c users have transmitted ($c-1$ interfering users), is not successfully decoded in that slot accounting for PHY layer. To this aim, let us define the failure event $\mathcal{F}$ = "the replica corresponding to an edge is not decoded" and the random variable C describing the total number of active users in a slot. From a user's point of view, we also define the random variable I as the number of interfering users which have chosen the same pilot as the given user (pilot-colliding users). Then,

the probability that a user has exactly i pilot-colliders, given c total users in the slot and N_{P} available pilots, is

$$\mathbb{P}\{i|c\} = \binom{c-1}{i} \left(\frac{1}{N_{\mathrm{P}}}\right)^{i} \left(1-\frac{1}{N_{\mathrm{P}}}\right)^{c-1-i} . \tag{3.31}$$

Moreover, considering that from previous interference cancellations users can be subtracted with probability $1-q_{\ell-1}$, we define the random variable S as the number of pilot-colliding users subtracted. The probability that exactly s subtractions are performed, given i pilot-colliding users and $q_{\ell-1}$ pilots, is

$$\mathbb{P}\{s|i\} = \binom{i}{s} \left(1-q_{\ell-1}\right)^{s} \left(q_{\ell-1}\right)^{i-s} . \tag{3.32}$$

Noting that $\mathbb{P}\{s|i,c\} = \mathbb{P}\{s|i\}$, we can write

$$\begin{aligned} p_{\ell}^{(c)} &= \mathbb{P}\{\mathcal{F}|c\} \\ &= \sum_{i}\sum_{s} \mathbb{P}\{\mathcal{F},i,s|c\} \\ &= \sum_{i}\sum_{s} \mathbb{P}\{\mathcal{F}|i,s,c\}\ \mathbb{P}\{s|i\}\ \mathbb{P}\{i|c\} . \end{aligned} \tag{3.33}$$

Since is not possible to successfully decode a pilot-collided replica, we have

$$\mathbb{P}\{\mathcal{F}|i,s,c\} = 1, \quad s \neq i\,. \tag{3.34}$$

On the other hand, when $s = i$ and a realistic channel is considered, the probability to successfully decoded a user is not always zero. Using the P_{fail} approximation highlighted in Section 3.3, we can write

$$\mathbb{P}\{\mathcal{F}|i,s,c\} = \begin{cases} P_{\text{fail}}((i+1)c-1), & s=i \\ 1, & s\neq i \end{cases} \tag{3.35}$$

where

$$P_{\text{fail}}(n) = 1 - \sum_{d=0}^{t} \binom{N_{\text{D}}}{d} P_{\text{e}}^{d}(n) \left(1 - P_{\text{e}}(n)\right)^{N_{\text{D}}-d} \tag{3.36}$$

$$P_{\text{e}}(n) = \text{erfc}\left(\sqrt{\frac{M}{2n}}\right) - \frac{1}{4}\text{erfc}^2\left(\sqrt{\frac{M}{2n}}\right) . \tag{3.37}$$

From (3.36) and (3.37), it is possible to note that P_{fail} depends on the number of available pilots N_{P}, the number of antennas M, the error correction capability of the PHY error correcting code t, and the number of interfering users. Finally, we can write

$$\begin{aligned} p_{\ell}^{(c)} &= \sum_{i=0}^{c-1} \sum_{s=0}^{i} \mathbb{P}\{\mathcal{F}|i,s,c\} \; \mathbb{P}\{s|i\} \; \mathbb{P}\{i|c\} \\ &= \sum_{i=0}^{c-1} \sum_{s=0}^{i-1} \mathbb{P}\{s|i\} \; \mathbb{P}\{i|c\} + \sum_{i=0}^{c-1} P_{\text{fail}}((i+1)c-1) \; \left(1 - q_{\ell-1}\right)^{i} \; \mathbb{P}\{i|c\} \\ &= \sum_{i=0}^{c-1} \left[1 + \left(1 - q_{\ell-1}\right)^{i} \left[P_{\text{fail}}((i+1)c-1) - 1\right]\right] \; \mathbb{P}\{i|c\} \; . \end{aligned} \tag{3.38}$$

All the other density evolution equations remain the same. Substituting (3.38) in (1.25) we can find the asymptotic thresholds.

Analysis Output: In this section we have extended the asymptotic load threshold of IRSA to a wireless MIMO fading channel, for a specific setting and PHY layer signal processing. A main outcome of our analysis is that the IRSA distributions that are optimum over the simple collision channel model turn suboptimum in this new and more realistic setting. This will be shown clearly in Chapter 4. As such, when designing multiple access protocols for fading MIMO channels, employing IRSA schemes designed for surrogate channels, such as the collision one, is likely to yield suboptimum performance and to jeopardize the overall system performance. Indeed, a suitable modeling of the PHY layer is of utmost importance to correctly determine the theoretical limits and to optimize the protocol design parameters. The analysis developed in this chapter can be used as a building block for accurate IRSA design over wireless channels. More specifically, the

developed threshold analysis tool can be exploited within an optimization procedure, e.g., an evolutionary optimization algorithm [61], to design IRSA distributions characterized by optimum waterfall performance. Suitable constraints to the optimization procedure should be imposed in order to achieve a good compromise between waterfall and error floor performance.

Chapter 4

Numerical Results

In this chapter we provide simulation results to validate both the scheme proposals and the analytical results. To estimate probability we make use of Monte Carlo simulations where several instances of the considered random scenario are generated to have statistically meaningful estimate. As a rule of thumb we let the simulations go until 100 error events occur. In this way, to estimate a point on a curve with 10^{-4} probability we run about 10^6 simulations. A summary of the setup is reported below. In each following section, all parameters are set as in this brief summary if not otherwise stated.

We provide simulation results is a setting where each user encodes its messages with an $(n = 511, k = 421, t = 10)$ binary Bose–Chaudhuri–Hocquenghem (BCH) code. Part of the k information bits are used to validate the decoded packets via a CRC. After padding the BCH codeword with a final zero bit, the encoded bits are mapped onto a QPSK constellation with Gray mapping, yielding a payload of $N_{\mathrm{D}} = 256$ symbols. Simulation results are given for $B_{\mathrm{s}} = 1$ Mbps symbol rate, $N_{\mathrm{P}} = 64$ pilots, $r = 3$ replicas per active user, and $M = 256$ antennas. Pilots are constructed using Hadamard matrices and noise variance is set to $\sigma_{\mathrm{n}}^2 = 1$ as the channel vector variance σ_{h}^2. When ACKs are adopted, numerical results will assume a perfect feedback channel (i.e., all ACK messages are always successfully received). As a reminder the main metric is the PLR against the number of active users per frame K_{a} (see Section 1.2), under a maximum latency constraint $\Omega = 50$ ms. In this way, we can show the reliability-scalability trade-off, given

the maximum latency. For a given maximum latency Ω, and neglecting guard and processing times for the sake of simplicity, the number of slots per frame N_{s} is set equal to

$$N_{\mathrm{s}} = \left\lfloor \frac{\Omega\, B_{\mathrm{s}}}{2\,(N_{\mathrm{P}} + N_{\mathrm{D}})} \right\rfloor . \tag{4.1}$$

The factor 2 is due to the fact that a user could wake up right after a beacon, wait for the next one, and successfully decoded at the end the next frame.

Adopting pilot-based ACK messages (i.e., the BS notifies in which pilot a user has been decoded), only N_{P} information bits are required to be broadcast as an ACK message. Since this message is transmitted over a feedback channel that is not interfered, the BS may use a shorter preamble for channel estimation purposes and a more compact constellation such M-QAM. In practice, the ACK message is protected by a CRC and by a specific channel code. Then, the ACK packet size in symbols is

$$N_{\mathrm{ACK}} = N_{\mathrm{P,ACK}} + \frac{(N_{\mathrm{P}} + N_{\mathrm{CRC}})}{\mathcal{R}_{\mathrm{a}} \log_2(\mathsf{M})} \tag{4.2}$$

where $N_{\mathrm{P,ACK}}$ is the ACK preamble length, N_{CRC} is the number of CRC bits protecting the ACK message, and $\mathcal{R}_{\mathrm{a}}$ is the code rate. For example, considering $N_{\mathrm{P,ACK}} = 4$, $N_{\mathrm{CRC}} = 16$, $\mathsf{M} = 256$, and $\mathcal{R}_{\mathrm{a}} = 2/3$, we end up with $N_{\mathrm{ACK}} = 19$ when $N_{\mathrm{P}} = 64$ and $N_{\mathrm{ACK}} = 31$ when $N_{\mathrm{P}} = 128$. A simple way to include ACK time is substituting $N_{\mathrm{P}} + N_{\mathrm{D}}$ with $N_{\mathrm{P}} + N_{\mathrm{D}} + N_{\mathrm{ACK}}$ in equation (4.1). In order not to stick to a particular ACK implementation, we consider ideal instantaneous ACKs ($N_{\mathrm{ACK}} = 0$). This is due to the fact that, in Section 4.1.4, we will show how SSC protocols can solve this issue.

4.1 MAC Layer Protocol Evaluation

4.1.1 Spatial Coupling and Acknowledgement Benefits

Fig. 4.1 compares several MAC protocols in terms of PLR versus the number of active users per frame K_{a}. The considered protocols are framed slotted ALOHA,

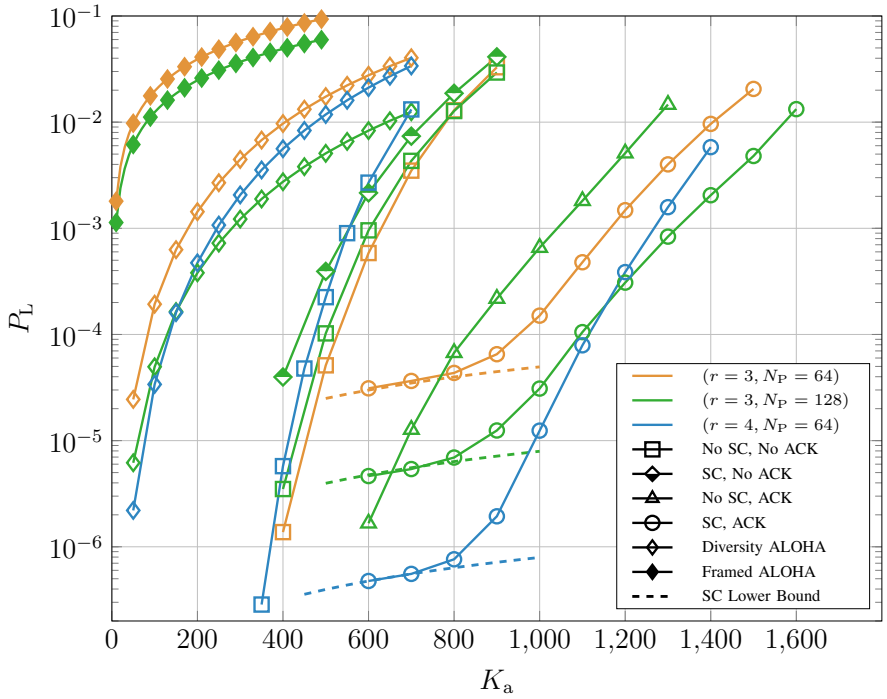

Figure 4.1: PLR comparison between schemes with different combinations of acknowledgments and intra-frame SC activation, for the baseline with $r \in \{3, 4\}$ and $(N_{\mathrm{P}}, N_{\mathrm{s}}) \in \{(64, 78), (128, 62)\}$. Framed ALOHA uses $r = 1$ despite of the curve color. Error floors derived in Section 3.1 are reported in dashed line.

diversity slotted ALOHA, the baseline scheme (reported in figure as "No SC, No ACK"), and CRA with intra-frame SC introduced in Section 2.1.1, with and without feedback (see Section 2.1.3). For the framed slotted ALOHA and diversity slotted ALOHA we have verified and then used the theoretical performance derived in (3.17). Simulations are run for different choices of the number of available pilots N_{P} and repetition rate r. The number of slot vary accordingly to (4.1) when N_{P} is changed.

From the plot we can observe, as expected, that framed ALOHA is not well-suited for access schemes when a non-trivial reliability is required. In fact, transmitting a single packet within the frame, the PLR is limited by the probability that two users pick the same pair slot-pilot. An improvement is achieved by diversity ALOHA where the user transmits r replicas in the frame. The difference between this access scheme and the baseline with r replicas, is the absence of a SIC phase.

Diversity ALOHA is able to achieve a better performance compared to framed ALOHA due to the fact that the scheme is now limited by the probability that all r replicas are collided. However, a high quality of service is achievable only for a limited number of active users per frame. In order to improve scalability having constraints in latency and reliability, well-designed SIC algorithms are required.

Let us focus at first on the $N_{\mathrm{P}} = 128$ case. As we can see, use of SC without ACK messages tends to worsen performance with respect to the baseline. A closer inspection reveals that this effect is associated with failures in SIC physical layer processing due to an increased number of active devices choosing the same r slots, even with different pilots, which makes the cross terms in (1.12) and (1.13) not negligible. This observation is supported by the fact that, when ACK messages from the BS are enabled, the proposed SC scheme exhibits the most pronounced performance boost. In fact, intra-frame SC gives rise to a lower number of resource collisions in the first slots, which stops a higher number of replica transmissions in subsequent slots and reduces interference in them. Note that, here we are assuming instantaneous feedback. For this reason, adopting ACKs does not decrease the number of slots. In Section 4.1.4 we will accurately investigate this problem, showing that both the baseline with ACKs and the intra-frame SC with ACKs performance curves translate on the left due to ACK time overhead. Despite of this translation, intra-frame SC with ACKs outperforms the baseline with ACKs (both translate about the same amount). For example, at $P_{\mathrm{L}} = 10^{-3}$, the baseline scheme supports $K_{\mathrm{a}} = 600$ active users per frame, which are pushed to more than $K_{\mathrm{a}} = 1300$ active users per frame combining of intra-frame SC and ACK messages. Accounting for ACK delay assumptions, leading to $N_{\mathrm{s}} = 60$ when $N_{\mathrm{ACK}} = 31$ ($N_{\mathrm{P}} = 128$), the performance of the intra-frame SC and ACK scheme degrades from $K_{\mathrm{a}} = 1300$ to approximately $K_{\mathrm{a}} = 1250$. Nevertheless, the improvement with respect to the baseline remains remarkable. Notably, the intra-frame SC protocol with ACK messages performs better in the $N_{\mathrm{P}} = 128$ case than is the $N_{\mathrm{P}} = 64$ one, despite the fact that the number of slots N_{s} decreases according to (4.1). In fact, with $N_{\mathrm{P}} = 128$, resource collisions in the first slots are less likely, making the ACK-based procedure more effective when the SC scheduling is adopted. On the contrary, increasing N_{P} in the baseline scheme worsens the system performance, as the cross-term interference increases due to

the reduction in the number of slots and the non-ideality of the SIC processing.

In Fig. 4.1 we also report the PLR bound (3.3), applied to the SC schemes. As we can see, the bound is tight in the error floor region, which makes it useful for design purposes. Moreover, despite the fact that (3.3) has been derived without considering the wireless channel effects, it is remarkable that it well-fits the behavior of the schemes also under realistic physical layer processing. The different behavior of the PLR in the waterfall and floor regions highlights the different trade-offs achieved by the two schemes (with and without SC). However, for well-designed system parameters, the error floor of SC schemes is below the PLR targets in MMA scenarios (e.g., $P_{\mathrm{L}} = 10^{-4}$, considered as a tightening requirement in MMA applications).

4.1.2 Randomized Spatial Coupling Optimization

In Fig. 4.2 we focus on the randomized SC protocol of Section 2.1.2, assuming ACK messages enabled, number of pilots $N_{\mathrm{P}} = 64$, and repetition rate $r = 3$. The total number of slots in the frame is $N_{\mathrm{s}} = 78$ in accordance to (4.1). For comparison we also report the baseline scheme with ACKs, represented in figure with $W = N_{\mathrm{s}} = 78$. As expected from the discussion about the dependence of the lower bound (3.3) on W, the schemes using a larger window size W exhibit a lower error floor. In this regard, we observe that (3.3) remains tight also in the case where $r < W < N$ despite the fact that the "birthday dates"are not uniformly chosen by users. As another important observation, for small $W > r$ window size, the benefits in error floor region come at no loss in terms of waterfall performance, where a small improvement is even observed. Then, depending on the target PLR P_{L}^{*}, we can note that there exists an optimal value of W. This is shown explicitly in Fig. 4.3, where we plot the maximum number of served users versus the window size W for given P_{L}^{*}, rate r, and number of available pilots N_{P}.

4.1.3 Energy Saving due to Acknowledgements

Another fundamental metric in MMA protocols is the energy efficiency. In Fig. 4.4 we plot the average number of transmitted packet replicas per active user (proportional to the average transmit energy per active user) versus the number of active

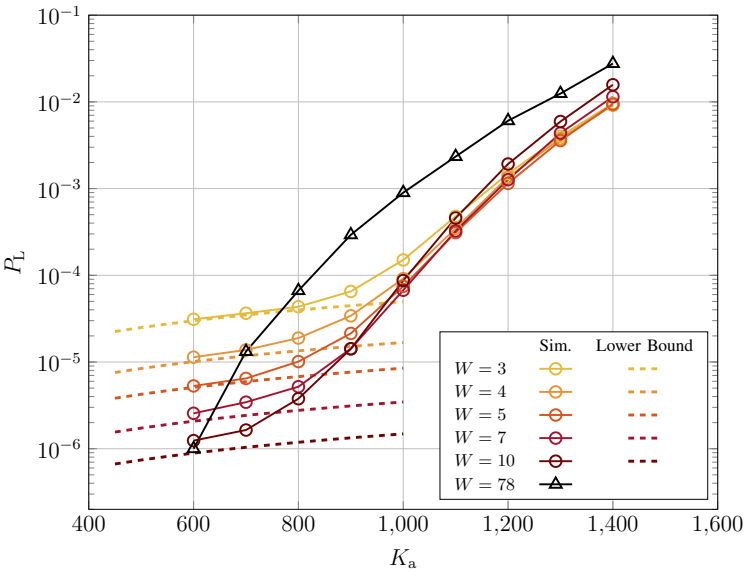

Figure 4.2: Packet loss rate comparison between schemes adopting randomized intra-frame SC and ACK messages, for CRA with $r = 3$, $N_{\mathrm{P}} = 64$, $N = 78$, and $W \in \{3, 4, 5, 7, 10, 78\}$. The baseline scheme is represented by the case $W = N = 78$, while the standard intra-frame SC by the case $W = r = 3$. Error floors derived in Section 3.1 are reported in dashed line.

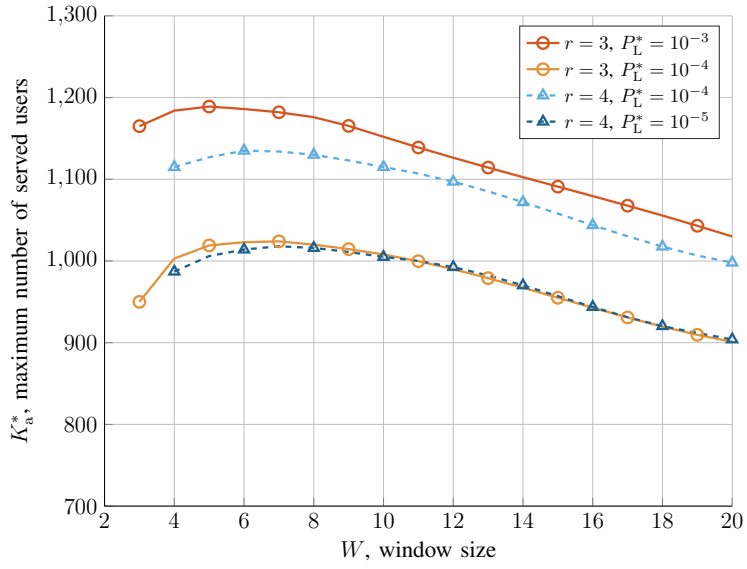

Figure 4.3: Window size W analysis at $P_{\mathrm{L}}^{*} \in \{10^{-3}, 10^{-4}, 10^{-5}\}$ for schemes adopting randomized intra-frame SC and ACK messages, using $N = 78$, $N_{\mathrm{P}} = 64$, $M = 256$, $N_{\mathrm{D}} = 256$, $r \in \{3, 4\}$. In the y-axis is reported the number of served user per frame given the target PLR.

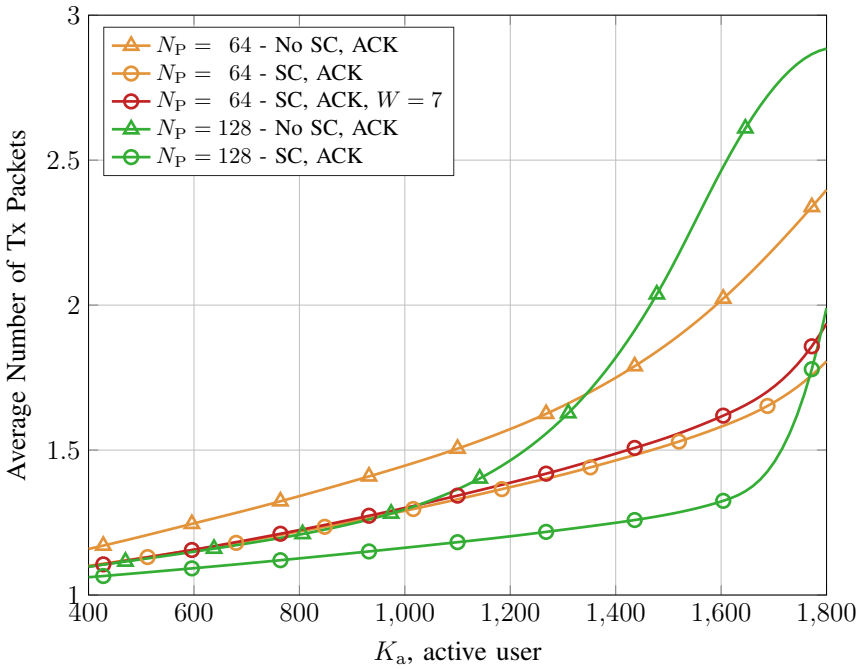

Figure 4.4: Average number of transmitted packets by the users. The comparison is carried out between schemes characterized by different combinations of acknowledgments and intra-frame SC, for CRA with $r = 3$, $(N_P, N) \in \{(64, 78), (128, 62)\}$. Randomized SC is reported using $W = 7$ in red line.

users per frame, for the different access schemes. Clearly, this value equals r for all schemes not exploiting ACK messages from the BS. For example, with reference to Fig. 4.1 and Fig. 4.4, considering the intra-frame SC protocol with ACK messages enabled, $r = 3$, and $N_P = 64$, each user transmits on the average less than 1.4 replicas per frame at $P_L^* = 10^{-3}$. We see from the figure that exploiting ACK messages provides substantial savings, with an average number of transmitted packet replicas below 1.5, over a large range of K_a. Concerning randomized SC schemes, for small window sizes W (which are the values of interest, as pointed out in Fig. 4.3), the average number of transmitted packet replicas does not increase significantly. It can be verified that, when W is significantly larger than r, the energy efficiency worsens compared to the $W = r$ case.

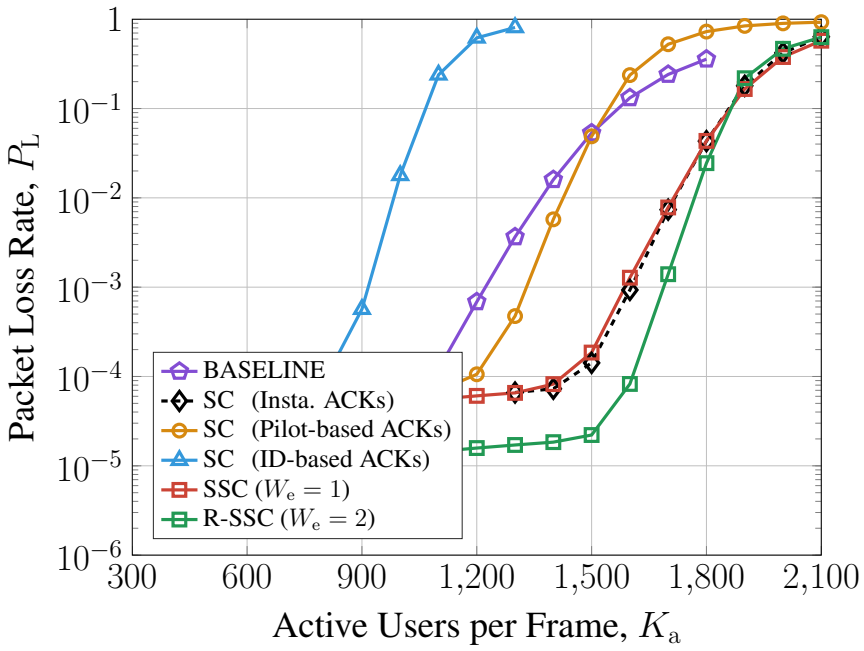

Figure 4.5: Degradation of the PLR due to ACK times. For pilot-based ACKs $N_{\mathrm{ACK}} = 76$, while for ID-based ACKs $N_{\mathrm{ACK}} = 228$. SSC is able to overcome the degradation due to ACK overhead.

4.1.4 Solve ACK Overheads with Spaced Spatial Coupling

Fig. 4.5 shows the PLR P_{L} versus K_{a} for CRA systems based on intra-frame SC (Section 2.1.1) or on the proposed SSC protocol (Section 2.1.4). Here, the dashed black curve corresponds to an ideal intra-frame SC system with an instantaneous feedback not consuming any time resources ($N_{\mathrm{ACK}} = 0$). Using the usual parameter adopted in these numerical sections, through (4.1), we obtain a number of slots per frame $N_{\mathrm{s}} = 78$. Fig. 4.5 also shows the performance of the same system when a realistic, non-instantaneous feedback is considered. Assuming $N_{\mathrm{P,ACK}} = 4$, $N_{\mathrm{CRC}} = 32$, $\mathsf{M} = 16$, and $\mathcal{R}_{\mathrm{a}} = 1/3$, we obtain $N_{\mathrm{ACK}} = 76$ for pilot-based ACKs. Hence, the number of slots per frame reduces to $N_{\mathrm{s}} = 71$, causing a visible performance loss with respect to the idealized system with instantaneous feedback. Note that the BS may also use a more compact constellation such M-QAM for transmission of ACK messages, if affordable in terms of link budget. In contrast, an ID-based ACK technique, in which hashes of decoded users are

concatenated to form an ACK message, tends to be larger in terms of symbols.[1] For example, for an hash size of 14 bits, a maximum of 19 notified users per slot by ACKs and the same transmission parameters of the pilot-based case, we end up with $N_{\mathrm{ACK}} = 228$. This leads to $N_{\mathrm{s}} = 45$ slots per frame and to a catastrophic performance degradation in terms of PLR. This degradation could make the intra-frame SC scheme worse than the baseline scheme with ACKs in terms of reliability. In particular, the baseline scheme is assuming instantaneous feedback which is not fair. However, letting the user listen for ACKs in the subsequent slot of a transmission: $i)$ if they don't have to transmit and they have been successfully decoded, they will receive the ACK; $ii)$ if they have to transmit and they have been successfully decoded, they will miss the ACK. The sporadic occurrence of event $ii)$ make the performance of this scheme practically equal to the baseline with instantaneous feedback. Regarding schemes with SC, SSC protocols represent an elegant solution to this problem. As shown in the figure, the presence of a wait window (used for receiving ACKs) between successive replicas from the same device guarantees $N_{\mathrm{s}} = 78$ slots per frame, with a negligible performance degradation with respect to the idealized SC case. This holds whenever the ACK message could fit in a slot time. Finally, we show the impact of randomization in SSC using $W_{\mathrm{e}} = 2$. As the randomization in intra-frame SC shown in Section 4.1.2, we can obtain some improvements both in the waterfall and the error floor region.

4.1.5 Spaced Spatial Coupling varying the Antennas

In Fig. 4.6 we compare the performance of SC, SSC with $W_{\mathrm{e}} = 1$, and randomized SSC with $W_{\mathrm{e}} = 2$ protocols, for different numbers M of BS antennas. Specifically, we consider $M = 64, 128, 256$, while the other simulation parameters remaining unchanged. For intra-frame SC we consider here that the ACK time is one tenth of the packet size. We observing how the trend is the same for all values of M, the randomized SSC scheme always achieving the best performance. As expected, increasing the number of BS antennas improves system scalability for a given target PLR; this is due to better PHY layer characteristics,

[1] More details in [56].

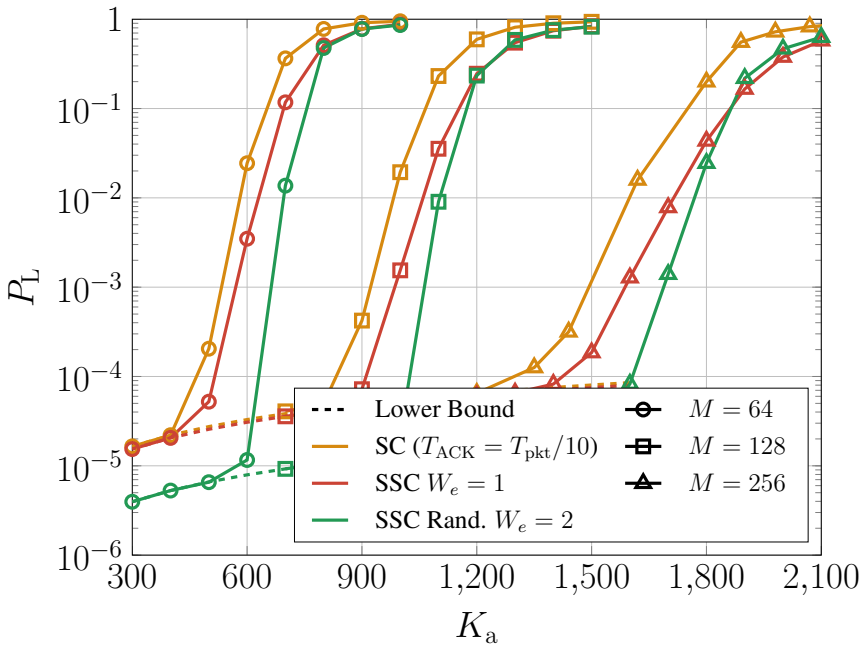

Figure 4.6: Packet loss rates achieved by CRA-type protocols (SC and SSC) for 64, 128, and 256 BS antennas. Dashed: Error floor analytical predictions (right-hand side of (3.3)).

such as singleton channel hardening and favorable propagation. In general, the approach featuring a random spacing of replicas yields a good tradeoff between scalability and reliability at both low and high traffic loads. As usual, the lower bounds derived in Section 3.1 are very tight in the error floor region.

4.2 PHY Layer Processing Evaluation

In this section, we present numerical results about several PHY layer processing strategies. Moreover, we compare the techniques discussed in previous sections with some representative benchmarks, using also different MAC protocols. Besides the benchmark derived in Section 3.2, we consider a setting in between the ideal collision channel and the realistic channel we usually assume. In particular, payload estimation is performed as in (1.10); upon successful message decoding in a slot, PAB processing is applied under the assumption that the subtractions are perfect (i.e., ideal SIC). In this setting, referred to as perfect replica channel estimation (PRCE), the performance is therefore limited by payload estimation (1.10) only. This establishes a second upper bound on the number of simultaneously active users; this upper bound is generally tighter than the logical performance with SIC one.

4.2.1 Payload Aided Based SIC

In Fig. 4.7 we report the PLR varying the symbol payload size N_{D} while keeping the rate of the channel code (a BCH code) as much constant as possible, for the CHB (baseline scheme), PAB, and PRCE interference cancellation. To be precise, for $N_{\mathrm{D}} \in \{128, 256, 512\}$ the corresponding BCH codes are $(255, 207, 6)$, $(511, 421, 10)$, and $(1023, 843, 18)$. In this particular example, we adopt the baseline MAC fixing $N_{\mathrm{P}} = 64$ leading to $N_{\mathrm{s}} \in \{130, 78, 43\}$ in accordance with (4.1). As expected, the CHB processing curves degrade when N_{D} increases due to the fact that the number of slots per frame N_{s} is decreasing. The same behavior can be observed for PRCE. In the case of PAB processing, instead, the trend is not so obvious. In fact, its performance tends to degrade when N_{s} decreases as for the other schemes, however, a gain in term of SIC quality is also expected from (2.10). In Fig. 4.7 we can see the gap between the PRCE and the PAB reduces, highlighting the effectiveness of the proposed technique in a complete scenario which accounts for both the PHY and MAC layers. In this particular example, these two effects counterbalance each other resulting in approximately 1000 active users per frame at $P_{\mathrm{L}} = 10^{-4}$, for all N_{D} under examination using PAB.

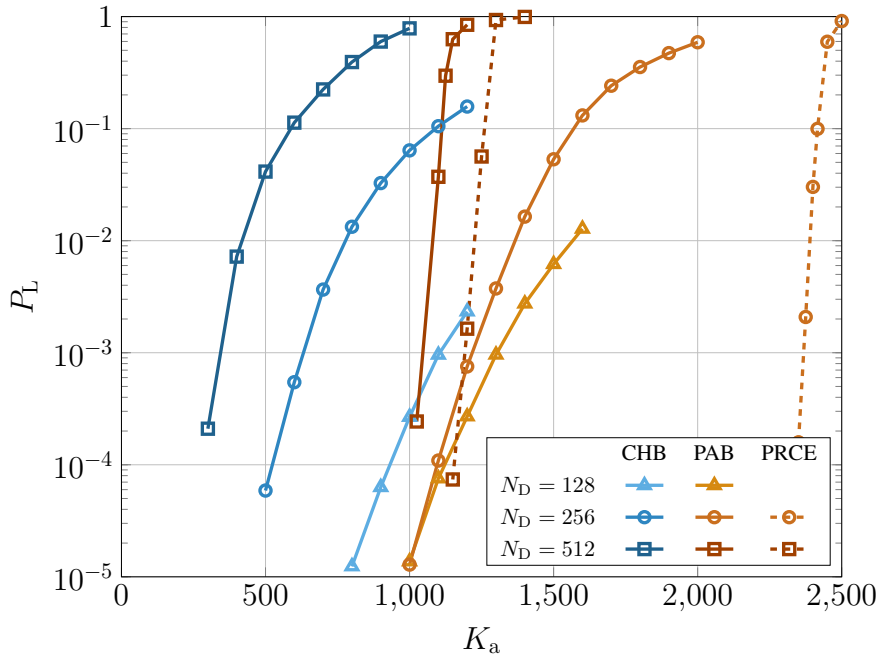

Figure 4.7: Packet loss rate values of schemes characterized by different SIC techniques and payload sizes $N_{\mathrm{D}} = \{128, 256, 512\}$. Baseline MAC with $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = \{130, 78, 43\}$, and $M = 256$ antennas. Comparison between the CHB, the proposed PAB and the ideal SIC case (PRCE). For the sake of completeness, the PRCE curve at $N_{\mathrm{D}} = 128$ intersect $P_{\mathrm{L}}^{*} = 10^{-3}$ around $K_{\mathrm{a}} = 4500$.

4.2.2 Impact of SIC scheduling

In Fig. 4.8 we plot a comparison between the CHB and PAB SIC techniques, using the baseline MAC protocol. We also apply instantaneous cancellation presented in Section 2.2.2, and plot the relative performance for both methods. The number of payload symbols is set to $N_{\mathrm{D}} = 256$, leading to a $(511, 421, 10)$ BCH code when an information payload of about 50 Bytes is considered. The PAB processing exhibits an improvement compared to the CHB. This is motivated by the fact that PAB subtractions have a beneficial effect on all users transmitting in a slot, while CHB ones influence only the users employing a particular pilot. Enabling instantaneous cancellation we obtain a remarkable performance boost in both SIC algorithms. Targeting for example a PLR $P_{\mathrm{L}} = 10^{-3}$, we see that the logical performance without SIC achieves up to 180 users per frame, the CHB processing increases this number to 650, and PAB with instantaneous cancellation achieves a

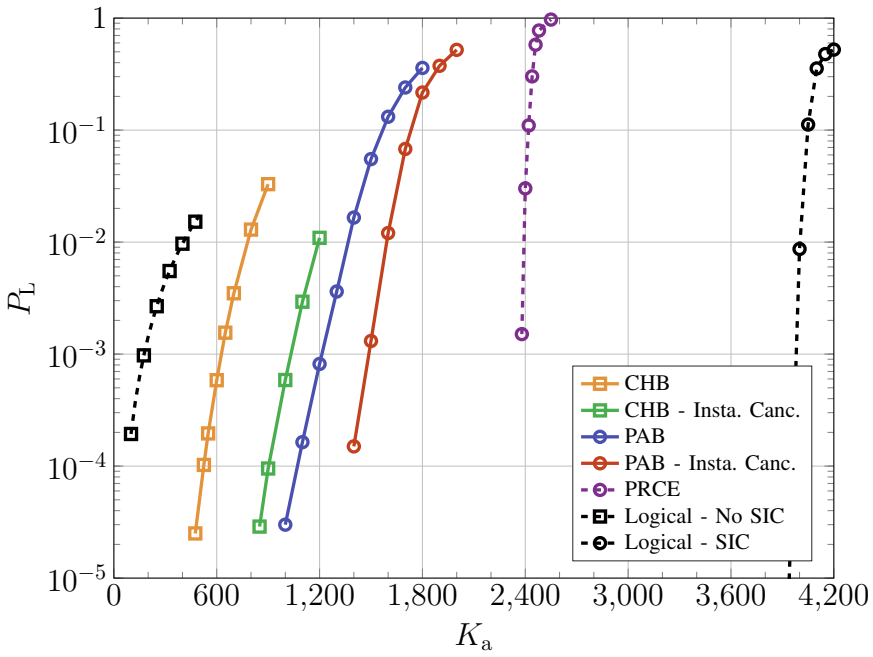

Figure 4.8: Packet loss rate comparison between different PHY layer schemes, when a baseline MAC protocol based on CSA using repetition code with $r = 3$ is employed. Maximum latency $\Omega = 50$ ms, $M = 256$ antennas, $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = 78$, and $N_{\mathrm{D}} = 256$.

K_{a} of approximately 1500. This $8\times$ increase in scalability motivates the interest on grant-free CRA schemes under a realistic PHY layer processing.

With reference to the same figure, we also point out the performance gap between a system performing realistic SIC and two idealized schemes, the PRCE and the logical one using SIC. The PAB and PRCE curves rely on the same payload estimation, and for this reason their performance gap depends on channel estimation imperfections. At the same time, there is a remarkable gap between the PRCE curve and the logical one using SIC as a result of payload estimation non-idealities. Comparing the performance of actual schemes with these benchmarks reveals how neglecting the PHY layer processing in real scenarios may lead to wrong conclusions and suboptimum optimizations. In Fig. 4.9 we also report the performance of the same PHY layer processing techniques of Fig. 4.8, when the intra-frame SC and ACKs ($N_{\mathrm{ACK}} = 0$) are adopted. Despite the MAC protocol change, the proposed PHY layer processing techniques provide again a considerable performance improvement.

Let us now discuss how the PRCE performance (i.e., same processing as PAB but with ideal SIC) can be approached using the proposed techniques. As anticipated when discussing Fig. 4.7, one possibility to reduce the gap between PAB and PRCE is to increase N_{D}. However, since we are considering a scenario where maximum latency is constrained, the degrading effect cause by N_{s} reduction is dominant. Hence, reaching PRCE in this way could not give an overall boost in performance. Another case in which PRCE curve can be reached is depicted in Fig. 4.9. So far we have considered block fading channel where the coherence time T_{c} is equal to the slot time T_{s}. However, if the time slot is sufficiently small it is possible that, in some scenarios, the coherence time is several times T_{s}. Exploiting the characteristic of intra-frame SC, we can therefore have the same user channel coefficients among all the replicas ($T_{\mathrm{c}} \geqslant r\, T_{\mathrm{s}}$). Hence, when noise is sufficiently small, we can subtract interference of all replicas using the channel estimates of singleton users, approaching ideal cancellation performance of PRCE. Despite we are not using the payload information, we report this scheme as PAB with $T_{\mathrm{c}} = r\, T_{\mathrm{s}}$ because it adopts iterative subtractions in (2.6).

In Fig. 4.8 and Fig. 4.9 we remark the notable gap between PRCE and the logical curve using SIC. This gap is essentially due to the fact that singleton replicas (either the ones that arrived alone in a resource or those becoming singleton ones during the SIC process) are not decoded with probability one and, thus, it is strictly related to Fig. 3.3. The analytical derivation developed in Section 3.3, and in particular the expression of P_{fail} in (3.30), suggests possible solutions to narrow this gap: for example, we can increase the number of antennas M, or increase the error correction capability t of the channel code (at the cost, however, of reducing the code rate and therefor the sum rate presented next). Some of these solutions are intuitively obvious, but the conducted analysis allows precisely quantifying the effect of a variation of each system parameter. Another important factor which should be considered is the noise level. Nevertheless, since we have used $\sigma_n^2 = 1$ in the numerical evaluation, having a smaller noise level does not improve significantly the performance. This is due to the fact that the system is interference-limited.

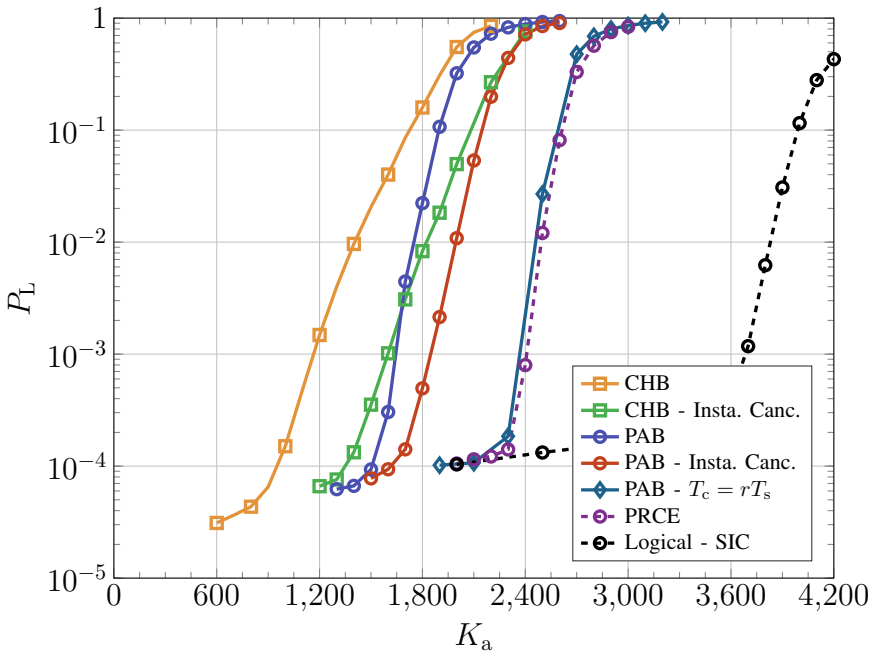

Figure 4.9: Packet loss rate comparison between different PHY layer schemes, when intra-frame spatial coupling and ACKs are enabled. CSA using repetition code with $r = 3$ is employed, maximum latency $\Omega = 50$ ms, $M = 256$ antennas, $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = 78$, and $N_{\mathrm{D}} = 256$.

4.2.3 Sum Rate Evaluation

In Fig. 4.10 we show the sum rate in terms of information bits per channel use, defined as

$$\gamma = (1 - P_{\mathrm{L}})\, K_{\mathrm{a}}\, \frac{N_{\mathrm{D}}\, \log_2(\mathsf{M})\, R_{\mathrm{c}} - N_{\mathrm{extra}}}{N_{\mathrm{s}}\,(N_{\mathrm{P}} + N_{\mathrm{D}})} \tag{4.3}$$

where $N_{\mathrm{extra}} = 33$, $R_{\mathrm{c}} = 421/511$, $\mathsf{M} = 4$, and other parameters are the same used in Fig. 4.8 and Fig. 4.9. The parameter N_{extra} accounts for payload bits which are not used for information data as CRC and zero padding bits. In particular, we report the sum rates of some schemes using intra-frame spatial coupling packet scheduling with ACKs. In this plot we observe that there exists an optimal K_{a} which maximizes the sum rate γ. However, the values of K_{a} yielding the largest γ may correspond to values of reliability not fulfilling the requirements of next generation MMA systems. On the other hand, the maximum value of the sum rate

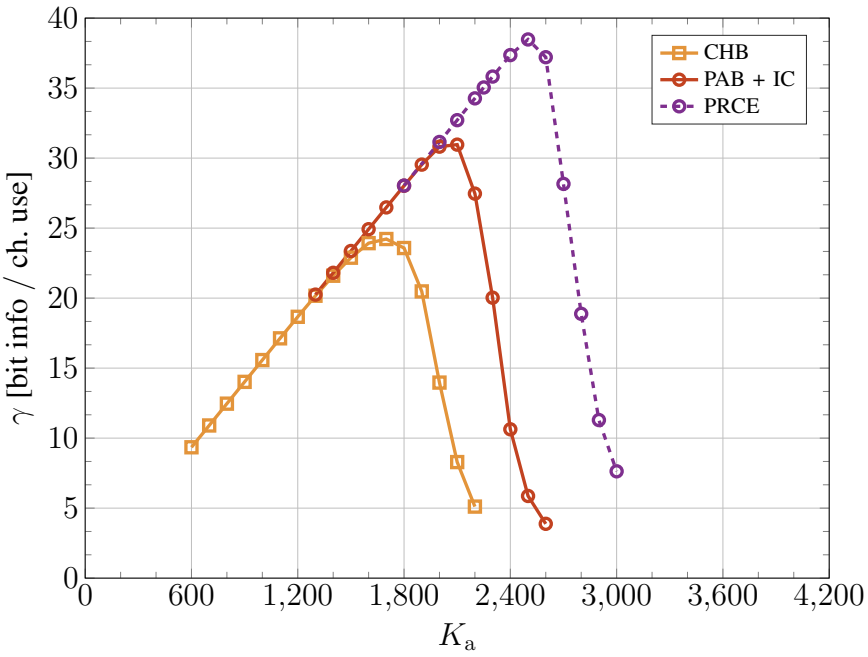

Figure 4.10: Sum rates in information bits per channel use of different PHY layer schemes, when intra-frame spatial coupling and ACKs are enabled. CSA using repetition code with $r = 3$ is employed, maximum latency $\Omega = 50$ ms, $M = 256$ antennas, $N_{\mathrm{P}} = 64$, $N_{\mathrm{D}} = 256$, $N_{\mathrm{s}} = 78$, and $N_{\mathrm{extra}} = 33$.

in information bits per second $\gamma_b = \gamma\, B_{\mathrm{s}}$ can be useful to design the backhaul communication network.

4.3 Joint PHY and MAC Layer Design

We start by presenting numerical results that illustrate the accuracy of the proposed threshold analysis. To this aim, we ran Monte Carlo simulations for some IRSA distributions $\Lambda(x)$ over both the collision channel with orthogonal resources and the MIMO block fading channel with actual signal processing, and performed threshold analysis for the same distributions over these channels. In practice, we declared a value of G as achievable (i.e., $G < G^*$) when density evolution recursion yielded $Q_\ell < 10^{-4}$ after a sufficiently large number of iterations.

In Fig. 4.11 and Fig. 4.12 we report simulation results (in terms of packet loss rate versus the number of active users over the frame) in solid lines, while thresh-

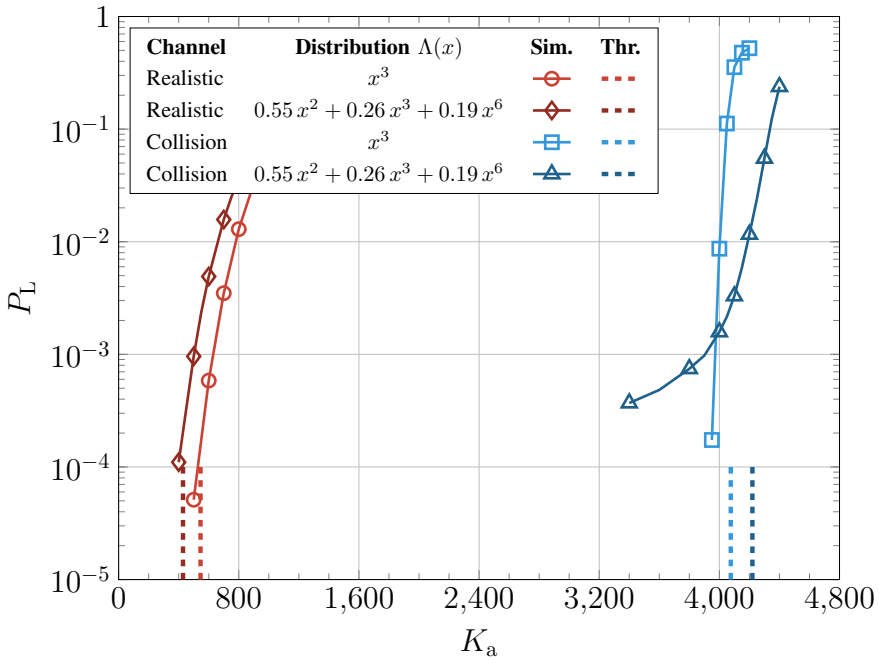

Figure 4.11: Packet loss rate comparison between the IRSA distribution with $\Lambda'(1) = 3$ and maximum repetition degree 6 being optimal over the collision channel (with or without orthogonal resources) and the concentrated distribution $\Lambda(x) = x^3$. Channels: Collision channel with N_{P} orthogonal resources and MIMO block fading channel with realistic signal processing. Parameters: $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = 78$, $M = 256$. Dashed lines: Values of $G^* N_{\mathrm{s}}$.

olds are marked by dashed vertical lines. In these figures, the "thresholds" are defined as $K_{\mathrm{a}}^* = N_{\mathrm{s}} G^*$, which represents an approximation of the number of simultaneously active users the scheme can support. Fig. 4.11 shows that the IRSA distribution with average packet repetition rate $\Lambda'(1) = 3$ and maximum repetition rate 6, having the largest threshold over the collision channel model [33], becomes sub-optimal when the realistic channel and signal processing is considered. In fact, its threshold is outperformed by that of the distribution with a constant repetition rate $\Lambda(x) = x^3$ (that is, CRDSA with repetition rate 3). Very remarkably, as predicted by our threshold analysis, this result is in perfect agreement with the Monte Carlo simulation. Fig. 4.12 shows similar results for other distributions, which again reveal the effectiveness and reliability of the proposed analysis over massive MIMO block fading channels and realistic PHY layer processing. Note that all concentrated (CRDSA) distributions considered in Fig. 4.12 exhibit the

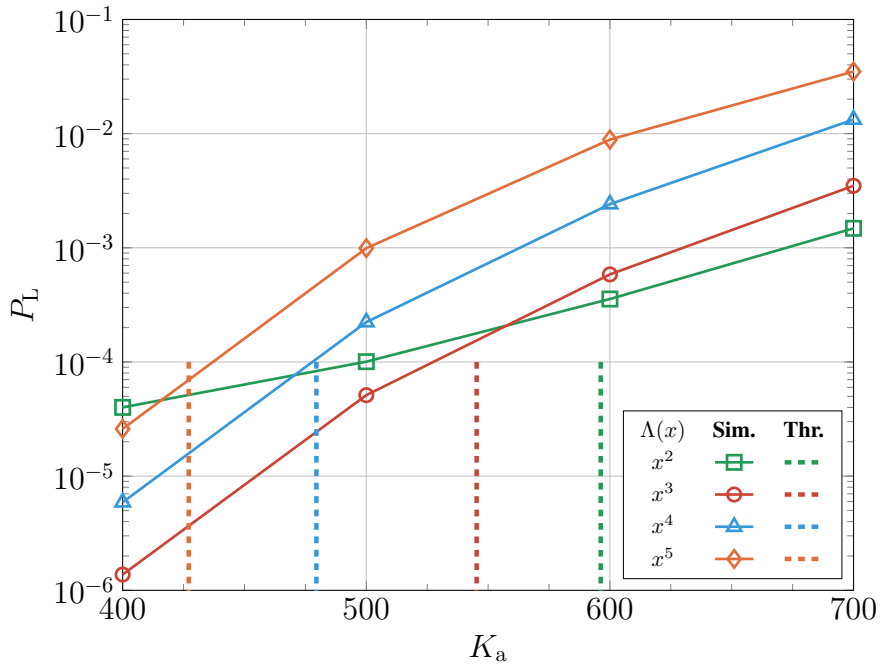

Figure 4.12: Packet loss rate comparison between concentrated distributions characterized by different repetition rates $r \in \{2, 3, 4, 5\}$ over realistic channel. Solid: Monte Carlo simulations. Parameters: $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = 78$, and $M = 256$. Dashed lines: $N_{\mathrm{s}}G^*$.

best threshold constrained to the corresponding integer $\Lambda'(1)$. Looking again at Fig. 4.12 we can see that, as expected, the proposed density evolution analysis is unable to capture error floor phenomena such as the one affecting the distribution $\Lambda(x) = x^2$. In Table 4.1 we list the thresholds estimated through density evolution for some $\Lambda(x)$ distributions with the previous choice of the system parameters.

Lastly, we show in Table 4.2 the results of another analysis we carried out using the proposed tool. For a constrained average repetition rate $\Lambda'(1) = 3$, we let the number of BS antennas M vary, searching for the optimum distribution (in terms of G^*) for each considered M. For all values of M, differential evolution optimization returned the same distribution $\Lambda(x) = x^3$. It is interesting to observe that, while the asymptotic threshold G^* increases monotonically with M, the ratio G^*/M (which represents a sort of efficiency per antenna) is not monotonically increasing but exhibits a maximum value. We attribute the decrease of G^*/M for large M to the constant number of orthogonal pilots $N_{\mathrm{P}} = 64$.

Table 4.1: Asymptotic thresholds obtained through density evolution under realistic channel assumptions.

IRSA Distribution	$\Lambda'(1)$	G^*
$\Lambda(x) = x^2$	2	7.64
$\Lambda(x) = x^3$	3	6.99
$\Lambda(x) = x^4$	4	6.15
$\Lambda(x) = x^5$	5	5.48
$\Lambda(x) = 0.55\,x^2 + 0.26\,x^3 + 0.19\,x^6$	3	5.49
$\Lambda(x) = 0.50\,x^2 + 0.50\,x^3$	2.5	6.64
$\Lambda(x) = 0.51\,x^2 + 0.27\,x^3 + 0.22\,x^8$	3.6	4.63
$\Lambda(x) = 0.55\,x^2 + 0.16\,x^3 + 0.29\,x^6$	3.3	4.97

Table 4.2: Optimum asymptotic thresholds constrained to $\Lambda'(1) = 3.0$ versus the number of antennas M. Optimum distribution $\Lambda(x) = x^3$ in all cases, $N_{\mathrm{P}} = 64$, $t = 10$, $N_{\mathrm{D}} = 256$.

M	G^*	G^*/M
8	0.1356	0.0169
16	0.4409	0.0276
32	1.0562	0.0330
64	2.0778	0.0325
128	3.8167	0.0298
256	6.9909	0.0273

Conclusion

This thesis has presented new schemes for synchronous MMA along with the corresponding signal processing algorithms to be performed in the receiver and the corresponding performance analysis. The output of the research contains advances with respect to current schemes, in several aspects. By exploiting a bridge with state-of-the-art codes on sparse graphs, the presented access schemes support large numbers of active devices while ensuring reliability and latency values that are beyond current mMTC ones. The signal processing is innovative and, leveraging on the dimensions offered by the randomly-chosen orthogonal pilots and on the multiple BS antennas, is effective in achieving multi-packet reception at slot level. The schemes are completely grant-free and uncoordinated, hence they require a minimum amount of control signalling only to let devices synchronize with the BS at frame and slot level. They achieve high performance gains and high energy efficiency when aided by a very simple acknowledgment mechanism implemented on a feedback channel. The thesis also offers an interesting information-theoretic perspective based on density evolution for CRA under realistic channel assumptions and massive MIMO processing. Several possible directions of investigation may be taken to further and extend the obtained results. A few of them are sketched in the following.

In this manuscript it is assumed that all arrivals at the BS are characterized by the same power, owing to the presence of a power control mechanism. In this respect, the introduction of a properly-designed power unbalance mechanism, where active devices intentionally transmit different replicas with different powers, may enhance multi-packet reception and therefore the overall system performance. Interesting research questions concern the design of the power levels, the strategy to assign different power levels to different replicas, and the design of uncoordi-

nated resource (i.e., slot-pilot-power) selection strategies. Asynchronous scenario could also be a valid candidate for further investigation since they decrease even more the complexity at the user side. On the other hand, cell-free architecture are becoming an hot topic due to energy and performance fairness they can provide among users. Evaluating the presented schemes in such an architecture could bring remarkable advantages in terms of energy efficiency at a small cost in performance. Finally, it is worth noting that the theoretical tool we have presented, based on density evolution, can be applied to characterize the behaviors of both the MAC and PHY layers in high-level analyses.

List of Figures

Bibliography

[1] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, 2021, to appear.

[2] S. Verma, S. Kaur, M. A. Khan, and P. S. Sehdev, "Toward green communication in 6G-enabled massive Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5408–5415, 2021.

[3] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[4] S. Henry, A. Alsohaily, and E. S. Sousa, "5G is real: Evaluating the compliance of the 3GPP 5G New Radio system with the ITU IMT-2020 requirements," *IEEE Access*, vol. 8, pp. 42 828–42 840, Mar. 2020.

[5] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Sep. 2017.

[6] G. Chisci, H. Elsawy, A. Conti, M.-S. Alouini, and M. Z. Win, "Uncoordinated massive wireless networks: Spatiotemporal models and multiaccess strategies," *IEEE/ACM Trans. Netw.*, vol. 27, no. 3, pp. 918–931, Jun. 2019.

[7] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. De Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[8] D. Zucchetto and A. Zanella, "Uncoordinated access schemes for the IoT: Approaches, regulations, and performance," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 48–54, Sep. 2017.

[9] E. Paolini, Č. Stefanović, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.

[10] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

[11] J. H. Sørensen, E. De Carvalho, Č. Stefanović, and P. Popovski, "Coded pilot random access for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8035–8046, Dec. 2018.

[12] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-free massive random access with a massive MIMO receiver," in *2019 53rd Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, USA, Nov. 2019, pp. 23–30.

[13] H. Han, Y. Li, W. Zhai, and L. Qian, "A grant-free random access scheme for M2M communication in massive MIMO systems," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3602–3613, Apr. 2020.

[14] J. Choi, J. Ding, N. P. Le, and Z. Ding, "Grant-free random access in machine-type communication: Approaches and challenges," arXiv:2012.10550 [cs.IT], Dec. 2020.

[15] A. T. Abebe and C. G. Kang, "MIMO-based reliable grant-free massive access with QoS differentiation for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 773–787, Mar. 2021.

[16] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[17] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[18] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu, "Artificial-intelligence-enabled intelligent 6G networks," *IEEE Network*, vol. 34, no. 6, pp. 272–280, Nov./Dec. 2020.

[19] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.

[20] A. A. Zaidi, R. Baldemair, V. Moles-Cases, N. He, K. Werner, and A. Cedergren, "OFDM numerology design for 5G New Radio to support IoT, eMBB, and MBSFN," *IEEE Commun. Mag.*, vol. 2, no. 2, pp. 78–83, Jul. 2018.

[21] C. Kalalas and J. Alonso-Zarate, "Massive connectivity in 5G and beyond: Technical enablers for the energy and automotive verticals," in *Proc. 2020 2nd 6G Wireless Summit*, Levi, Finland, Mar. 2020.

[22] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131 796–131 813, Jul. 2020.

[23] B. M. Lee and H. Yang, "Massive MIMO with massive connectivity for industrial Internet of Things," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5187–5196, Jun. 2020.

[24] J. Gao, W. Zhuang, M. Li, X. Shen, and X. Li, "MAC for machine-type communications in industrial IoT—Part I: Protocol design and analysis," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9945–9957, Jun. 2021.

[25] Y. L. Lee, D. Qin, L.-C. Wang, and G. H. Sim, "6G massive radio access networks: Key applications, requirements and challenges," *IEEE Open J. Veh. Technol.*, vol. 2, pp. 54–66, 2021.

[26] R. P. Torres and J. R. Pérez, "A lower bound for the coherence block length in mobile radio channels," *Electronics*, vol. 10, no. 4, 2021.

[27] S. Moloudi, M. Mozaffari, S. N. K. Veedu, K. Kittichokechai, Y.-P. E. Wang, J. Bergman, and A. Höglund, "Coverage evaluation for 5G reduced capability new radio (NR-RedCap)," *IEEE Access*, vol. 9, pp. 45 055–45 067, Mar. 2021.

[28] T. Jiang, J. Zhang, P. Tang, L. Tian, Y. Zheng, J. Dou, H. Asplund, L. Raschkowski, R. D'Errico, and T. Jämsä, "3GPP standardized 5G channel model for IIoT scenarios: A survey," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8799–8815, Jun. 2021.

[29] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[30] G. Choudhury and S. Rappaport, "Diversity ALOHA – a random access scheme for satellite communications," *IEEE Trans. Commun.*, vol. 31, no. 3, pp. 450–457, Mar. 1983.

[31] E. Casini, R. De Gaudenzi, and O. del Rio Herrero, "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access scheme for satellite access packet networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1408–1419, Apr. 2007.

[32] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 477–487, Feb. 2011.

[33] E. Paolini, G. Liva, and M. Chiani, "Coded slotted ALOHA: A graph-based method for uncoordinated multiple access," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815–6832, Dec. 2015.

[34] F. Clazzer, C. Kissling, and M. Marchese, "Enhancing contention resolution ALOHA using combining techniques," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2576–2587, Jun. 2018.

[35] M. Berioli, G. Cocco, G. Liva, and A. Munari, "Modern random access protocols," *Foundations and Trends in Networking*, vol. 10, no. 4, pp. 317–446, 2016.

[36] A. Munari, "Modern random access: An age of information perspective on irregular repetition slotted ALOHA," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3572–3585, Jun. 2021.

[37] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[38] "ETSI TR 101 545-4 V1.1.1 (2014-04) Digital Video Broadcasting (DVB); Second Generation DVB. Interactive Satellite System (DVBRCS2); Part 4: Guidelines for Implementation and Use of EN 301 545," 2014.

[39] "ETSI EN 301 545-2 V1.3.1 (2020-07) Digital Video Broadcasting (DVB); Second Generation DVB. Interactive Satellite System (DVBRCS2); Part 2: Lower Layers for Satellite standard," 2020.

[40] R. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 2, pp. 124–142, Mar. 1985.

[41] L. G. Roberts, "ALOHA packet systems with and without slots and capture," *ARPANET System Note 8 (NIC11290)*, Jun. 1972.

[42] L. Kleinrock and S. Lam, "Packet switching in a multiaccess broadcast channel: Performance evaluation," *IEEE Trans. Commun.*, vol. 23, no. 4, pp. 410–423, Apr. 1975.

[43] J. Liu and X. Wang, "Unsourced multiple access based on sparse tanner graph-efficient decoding, analysis, and optimization," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1509–1521, May 2022.

[44] J. Haghighat and T. M. Duman, "Energy efficiency analysis of a feedback-aided irsa scheme," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 2886–2891.

[45] ——, "An energy-efficient feedback-aided irregular repetition slotted aloha scheme and its asymptotic performance analysis," *IEEE Transactions on Wireless Communications*, 2023.

[46] ——, "Analysis of coded slotted aloha with energy harvesting nodes for perfect and imperfect packet recovery scenarios," *IEEE Transactions on Wireless Communications*, 2023.

[47] K.-H. Ngo, G. Durisi, and A. G. i Amat, "Age of information in prioritized random access," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 2021, pp. 1502–1506.

[48] T. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[49] T. Richardson, M. Shokrollahi, and R. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 619–637, Feb. 2001.

[50] S. Kudekar, T. J. Richardson, and R. L. Urbanke, "Threshold saturation via spatial coupling: Why convolutional LDPC ensembles perform so well over the BEC," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 803–834, Feb. 2011.

[51] G. Liva, E. Paolini, M. Lentmaier, and M. Chiani, "Spatially-coupled random access on graphs," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 478–482.

[52] L. Valentini, A. Faedi, M. Chiani, and E. Paolini, "Coded random access for 6G: Intra-frame spatial coupling with ACKs," in *Proc. 2021 IEEE Global Commun. Conf. Workshops*, Madrid, Spain, Dec. 2021.

[53] L. Valentini, M. Chiani, and E. Paolini, "Massive grant-free access with massive MIMO and spatially coupled replicas," *IEEE Transactions on Communications*, vol. 70, no. 11, 2022.

[54] A. E. Kalør, R. Kotaba, and P. Popovski, "Common message acknowledgments: Massive ARQ protocols for wireless access," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5258–5270, Aug. 2022.

[55] S. Scalise, C. Parraga Niebla, R. De Gaudenzi, O. Del Rio Herrero, D. Finocchiaro, and A. Arcidiacono, "S-MIM: A novel radio interface for efficient messaging services over satellite," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 119–125, Mar. 2013.

[56] L. Valentini, A. Mirri, M. Chiani, and E. Paolini, "Feedback-aided coded random access via replica spacing," in *Proc. 2023 IEEE Int. Conf. Commun.*, Rome, Italy, May/Jun. 2023.

[57] L. Valentini, A. Faedi, M. Chiani, and E. Paolini, "Impact of interference subtraction on grant-free multiple access with massive MIMO," in *Proc. 2022 IEEE Int. Conf. Commun.*, Seoul, South Korea, May 2022.

[58] L. Valentini, M. Chiani, and E. Paolini, "Interference cancellation algorithms for grant-free multiple access with massive MIMO," *IEEE Transactions on Communications*, vol. 71, no. 8, 2023.

[59] D. M. Bloom, "A birthday problem," *Amer. Math. Monthly*, vol. 80, no. 10, pp. 1141–1142, 1973.

[60] A. Conti, M. Win, and M. Chiani, "Invertible bounds for M-QAM in Rayleigh fading," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 1994–2000, Sep. 2005.

[61] T. Back, D. B. Fogel, and Z. Michalewicz, Eds., *Handbook of Evolutionary Computation*. Bristol, UK, UK: IOP Publishing Ltd., 1997.