Feedback-Aided Coded Random Access With Intentional Power Unbalance

Lorenzo Valentini, *Member, IEEE*, Alessandro Mirri, *Graduate Student Member, IEEE*, and Enrico Paolini, *Senior Member, IEEE*

***Abstract*— In this paper, feedback-aided coded random access (CRA) protocols for grant-free massive access, with and without exploitation of the power domain, are investigated. The developed schemes can support non-instantaneous acknowledgment messages, along with their time resources, with very little penalty in terms of the achieved tradeoff between scalability, reliability, and latency. The new CRA-type protocols rely on waiting slots introduced between consecutive transmissions from the same device to pipeline the feedback reception without degrading the throughput. Their performance can be further enhanced by exploitation of the power domain, in particular by introduction of a deterministic power selection scheme designed for transmission of different packet replicas over a short time window. The system performance is investigated assuming a realistic wireless channel model, a massive MIMO base station, and a realistic processing, via simulation and analysis. The achieved results show that the realistic feedback-aided protocols with deterministic power selection can support 12.8 grant-free users per slot guaranteeing a packet loss rate of 10^{-4} while a massive MIMO base station equipped with 128 antennas is employed.**

***Index Terms*— Coded random access, feedback-aided multiple access, grant-free access, Internet of Things, massive MIMO, massive multiple access, power unbalance, waiting slot.**

I. Introduction

THE advent of the Internet of Things (IoT) has fostered an increasing interest towards machine-type communications (MTC), i.e., communication between autonomous connected devices, operating without human supervision [1]. Recently, the expression "massive MTC" (mMTC) has surfaced as a consequence of the explosion of the number of IoT devices in several application domains [2]. In mMTC, a massive number of devices, each transmitting short packets with a low duty cycle and at unpredictable time instants, contend for wireless access to the network through the same base station (BS); in the uplink, this problem is typically referred to as massive multiple access (MMA) [3].

Manuscript received 24 November 2023; revised 25 May 2024; accepted 26 July 2024. Date of publication 2 August 2024; date of current version 16 January 2025. Supported in part by the CNIT/WiLab and the WiLab-Huawei Joint Innovation Center, and in part by the European Union through the Italian National Recovery and Resilience Plan of NextGenerationEU, partnership on "Telecommunications of the Future" under Grant PE00000001- "RESTART." An earlier version of this paper was presented in part at the 2023 IEEE International Conference on Communications (ICC 2023), Rome, Italy, May/June 2023 [DOI: 10.1109/ICC45041.2023.10278984]. The associate editor coordinating the review of this article and approving it for publication was J. Wang. *(Corresponding author: Enrico Paolini.)*

The authors are with the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" and CNIT/WiLab, University of Bologna, 40136 Bologna, Italy (e-mail: lorenzo.valentini13@unibo.it; alessandro.mirri7@unibo.it; e.paolini@unibo.it).

Digital Object Identifier 10.1109/TCOMM.2024.3437480

The MMA problem deviates from a traditional multiple access one. Devices are sporadically active and their transmitted packets are short, but their number is extremely large and their activity pattern is unknown to the BS [4], [5]. Moreover, they act without any coordination with each other. Such a peculiar scenario requires ad-hoc medium access control (MAC) protocols, which are often grant-free to let devices share the channel dynamically and avoid unacceptable overheads. In turn, grant-free access protocols unavoidably lead to interference, which raises the need for advanced signal processing algorithms at physical (PHY) layer to achieve the necessary levels of quality of service.

A class of grant-free MAC access protocols that is gaining interest to efficiently support massive access from uncoordinated devices is coded random access (CRA) [6], [7], [8], based on the idea of combining packet diversity with successive interference cancellation (SIC) at the receiver, thus attaining multi-packet reception (MPR) working across the whole frame. While these protocols have so far been studied in a feedback-free context, very recently the possibility to exploit a feedback channel to enhance their performance and increase energy efficiency has been considered [9], [10], [11], [12]. The feedback channel is usually operated using a time division duplexing approach, according to which active users wait for acknowledgement (ACK) messages in dedicated time windows.[1] In this context, since broadcasting of feedback messages requires a certain amount of time depending on the number of users to be acknowledged, the system latency suffers from a degradation [13]. On the other hand, if a maximum latency requirement is imposed, the system may suffer from throughput degradation.

Another option to further increase the number of simultaneously-active devices supported by a CRA scheme in a grant-free fashion consists of exploiting the power domain to enhance MPR capabilities at the receiver. Exploitation of the power domain for multiple access was originally proposed in [14] outside the CRA context. The power domain multiple access (PDMA) principle may be summarized as intentionally unbalancing the power levels of the received packets at the BS to "capture" multiple overlapping packets by application of an interference cancellation operation to be performed at single slot level. Some recent works have investigated the possibility of nesting PDMA into CRA protocols. For example,

[1] In this paper, the terms "acknowledgement" and "feedback" are used interchangeably.

coded slotted ALOHA (CSA) schemes [7] with power diversity are proposed in [15], where encoded packet fragments from each active device are transmitted with different power levels. Moreover, irregular repetition slotted ALOHA (IRSA) schemes [6] with power diversity are investigated in [16] and [17]. Density evolution analysis and degree distribution optimization in presence of a different number of available power levels are addressed again in [15] and [16] and, in addition, in [18]. We remark that the analyses in these works, although very innovative, are mainly based on the received signal power and on considerations related to the signal-to-interference-plus-noise ratio (SINR). As such, the provided numerical results are obtained without a detailed modeling of the signal at PHY layer so that validity of the results under a more realistic PHY layer modeling is not necessarily guaranteed. A similar issue was pointed out in [19] outside the PDMA context. Here, the authors have proposed an analytical tool for CRA over a non-idealized setting. The proposed method captures both MAC and PHY layer aspects, going beyond the simplifying assumption of collision channel model. In particular, a Rayleigh block fading channel is considered, a realistic PHY layer processing is performed and a receiver equipped with a massive number of antennas is deployed. On the other hand, in the context of PDMA considering a BS with a very large number of antennas, it could happen that nesting a power unbalance (PU) strategy based on a random power selection in CRA over a realistic PHY layer does not necessarily lead to a tangible performance improvement with respect to power balance (PB). This may occur due to the fact that the MPR capabilities offered by massive multiple-input multiple-output (MIMO) are already exploited at the receiver. On the other hand, as discussed in this paper, remarkable performance improvements can be attained tailoring the power diversity strategy to the specific MAC protocol.

In this paper, we pursue the investigation of feedback-aided CRA schemes, with and without exploitation of the power domain. We address the problem of supporting realistic (i.e., non-instantaneous) ACK messages, along with their consumed time resources, without sacrificing the achievable scalability-reliability-latency tradeoff. To this aim we propose new CRA-type protocols, referred to as spaced spatial coupling (SSC), based on the idea of introducing waiting slots between consecutive transmissions from the same device in order to pipeline the ACKs reception without degrading the throughput. We refer to such protocols using the term "spaced" to emphasize the waiting time between successive transmissions. Both a deterministic and a randomized version of SSC protocols are analyzed, under a realistic wireless channel model, a massive MIMO BS, and a realistic PHY layer processing. In case of exploitation of the power domain, the proposed SSC protocols are combined with a specifically tailored PU strategy that brings to a noticeable boost in the number of supported devices for the same level of reliability. We can summarize the paper key contributions of this paper as follows:

- we propose new MAC layer protocols able to effectively pipeline the ACK messages, reducing the intrinsic feedback overhead;
- we develop a power diversity mechanism which synergies effectively with the proposed MAC layer protocols;
- we analytically investigate the performance in the packet loss rate error floor region to provide design guidelines;
- we address the design of both ID-based and pilot-based ACK implementations in our schemes.

The paper is organized as follows. Section II introduces preliminary concepts, the system model, and some background material. Section III presents the proposed ACK design and MAC layer access techniques, also exploiting the power domain. Numerical results are provided in Section IV. Finally, conclusions are drawn in Section V. A subset of the results presented in this paper appeared in the conference version [20]. With respect to [20]: i) ACK design is provided for both pilot-based and ID-based feedback; ii) a novel power diversity based scheme is proposed; iii) the error floor analysis has been deepened; iv) several numerical results have been added.

Notation: Throughout the paper, capital and lowercase bold letters denote matrices and vectors, respectively. The conjugate transposition of a matrix or vector is denoted by $(\cdot)^H$, while $\|\cdot\|$ indicates the Euclidean norm. The operator $|\cdot|$ applied to a set represents the cardinality of that set. To keep a clean and compact notation, we denote the probability that a random variable A takes the value a, $\mathbb{P}\{\mathsf{A}=a\}$, as $P(a)$. Similarly, we write $P(a,b\,|\,c)$ to indicate the probability $\mathbb{P}\{\mathsf{A}=a,\mathsf{B}=b\,|\,\mathsf{C}=c\}$, and $\mathbb{P}\{\mathcal{E}\}$ to indicate the probability that an event $\mathcal{E}$ holds.

II. Preliminaries and Background

We consider an MMA system in which the time is organized in MAC frames, each composed of $N_{\rm s}$ slots. In any frame, $K_{\rm a}$ active devices (or "users") contend for transmission of one information message each, with $K_{\rm a}$ random and unknown to the receiver. Users are both frame- and slot-synchronous owing to the presence of a beacon signal broadcast by the BS at the beginning of each frame; such a beacon is also used for power control. Each packet (or "burst") transmitted by an active user has a transmission time that fits the slot time and arrives at the BS aligned with one slot.

A. Coded Random Access

In the considered setting, each user has an intermittent activity, waking up unpredictably and attempting transmission of one message without any coordination with the other active devices. As such, "collisions" between packets transmitted in the same slot by users that are active on the same frame necessarily occur. Some of these collisions can be resolved at PHY layer by processing each slot individually, even in presence of non-orthogonal users' transmissions. For example, the availability of a massive number of BS antennas and the consequent "favorable propagation" [21], combined either with randomly-chosen orthogonal pilots [22], [23], or with randomly-generated non-orthogonal pilots [24], allows decoding multiple packets in the same slot.

Unresolved collisions at slot level can be handled combining the channel access strategy at MAC layer and signal processing at PHY layer. In CRA, in particular, this is done by performing

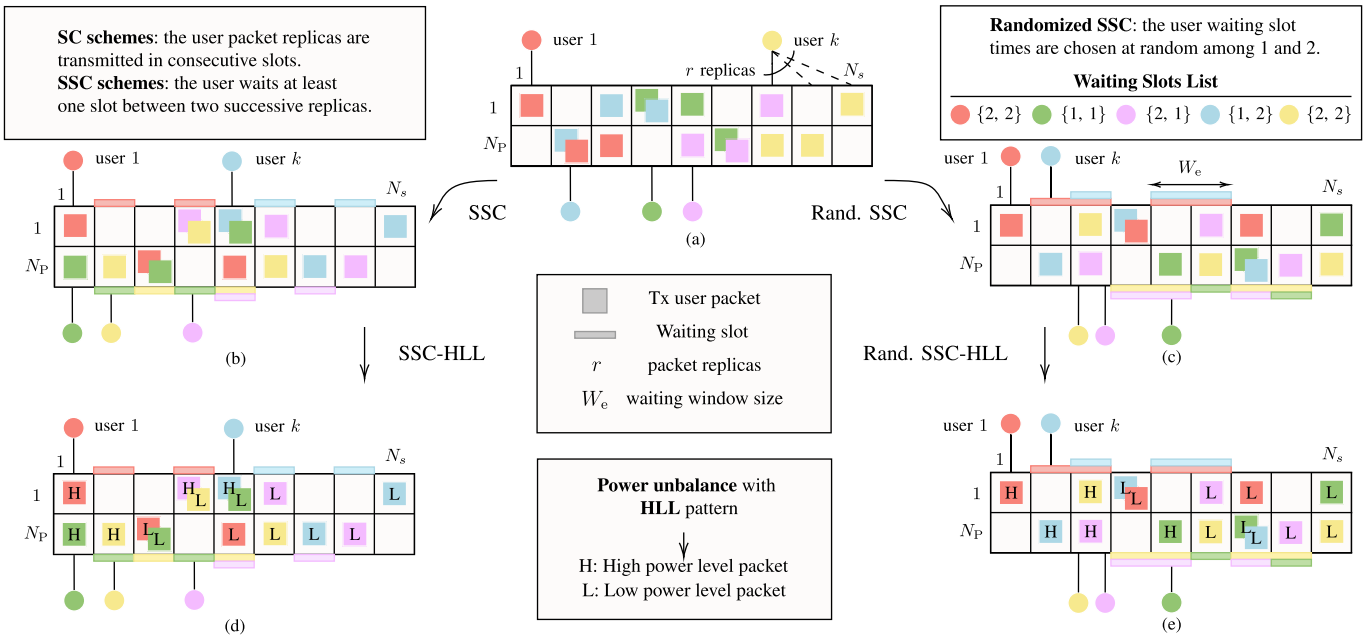

Fig. 1. CRA protocols with N_s slots per frame, N_P orthogonal pilots, and uniform repetition rate $r = 3$. (a) Intra-frame SC. (b) Intra-frame SSC and $W_e = 1$. (c) Randomized intra-frame SSC and $W_e = 2$. (d) Intra-frame SSC-HLL and $W_e = 1$. (e) Randomized intra-frame SSC-HLL and $W_e = 2$. The figure assumes that all r replicas are transmitted by each active user.

SIC across different slots of the frame. A simple form of CRA consists of letting active users transmit multiple copies (or "replicas") of their data payload in the same frame; whenever any such copy is decoded in a slot, the interference generated by the other replicas is subtracted from the corresponding slots. The number of copies r transmitted by an active device in the frame, called the user repetition rate, can be the same for all users [25] or it can be a random variable sampled independently by each user at each transmission [6]. More involved strategies, based on packet fragmentation and erasure coding, are possible [26], [27].

B. CRA: Intra-Frame Spatial Coupling With Feedback

The device-side access protocol considered in this paper, from wake up to transmission, may be summarized as follows. The active user chooses r different slots in the frame. For each such slot, the user picks uniformly at random one pilot sequence of length N_P symbols, out of N_P available orthogonal ones, and concatenates it with a payload of length N_D symbols, obtained by encoding the information message with a channel encoder and by mapping the encoded bits onto a complex constellation. The device then waits for the start of the next frame, signaled by the BS beacon, and finally transmits r packets in the r pre-selected slots. Note that the r packets are characterized by the same payload, while the pilot may differ in each of them. Replica transmissions are interrupted if an ACK message is received from the BS, as explained in the remainder of the section.

A conventional way to choose the r slot indexes by an active device consists of picking them uniformly at random, without replacement, in the set $\{1, \ldots, N_s\}$. A different strategy consists of choosing one slot index n, uniformly at random, in $\{1, \ldots, N_s - (r-1)\}$ and of transmitting the r packet replicas in slots n, $n+1$, $\ldots$, $n+r-1$ (Fig. 1a). This slot selection technique has been called intra-frame spatial coupling (SC) [10] because, together with the iterative PHY layer BS processing described in Section II-D, it is able to trigger an "interference cancellation wave" propagating from the edges of the frame towards the center of it, where decoded packets foster interference cancellation in adjacent slots in which new packets can be successfully decoded. This phenomenon is analogous to SC in LDPC coding [28].

The benefits of intra-frame SC are augmented when a feedback channel is available for transmission of ACK messages from the BS to the users [10]. An ACK consists of a message broadcast by the BS, at the end of each slot, to notify users whose messages have been decoded in the current slot. Users informed of correct reception of their messages abort transmissions of all not yet transmitted replicas, leading both to a reduction of the interference in future slots and to transmit energy savings. This mechanism provides remarkable enhancements to the system performance in terms, for example, of packet loss rate (PLR) versus the number of active users. Note that interrupting transmissions of unnecessary replicas (since the user's message has already been decoded) is equivalent to these replicas being transmitted and ideally cancelled.

C. Channel Model

We consider a Rayleigh block fading channel model where the channel coherence time equals the slot time. Perfect power control is assumed. The BS has M antennas, each with independent fading coefficient per user. The received signal in a slot is expressed as $[\boldsymbol{P}, \boldsymbol{Y}] \in \mathbb{C}^{M \times (N_P + N_D)}$ where

$$\boldsymbol{P} = \sum_{k \in \mathcal{A}} \boldsymbol{h}_k \boldsymbol{s}(k) + \boldsymbol{Z}_p \quad \text{and} \quad \boldsymbol{Y} = \sum_{k \in \mathcal{A}} \boldsymbol{h}_k \boldsymbol{x}(k) + \boldsymbol{Z}. \quad (1)$$

In (1), $\mathcal{A}$ is the set of users transmitting a burst in the slot; $\boldsymbol{h}_k = (h_{k,1}, \ldots, h_{k,M})^T \in \mathbb{C}^{M\times 1}$ is the channel coefficient vector of user $k \in \mathcal{A}$, whose elements are independent and identically distributed (i.i.d.) random variables with distribution $\mathcal{CN}(0, \sigma_{\mathrm{h}}^2)$ for all $k \in \mathcal{A}$. Owing to perfect power control, without loss of generality we assume $\sigma_{\mathrm{h}}^2 = 1$. Moreover, $\boldsymbol{s}(k) \in \mathbb{C}^{1\times N_{\mathrm{P}}}$ and $\boldsymbol{x}(k) \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ are the pilot and the data payload of user $k \in \mathcal{A}$ in the slot. 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 of Gaussian noise samples.

D. PHY Layer Processing at the Receiver

The processing of each frame is split into two phases. In phase 1, all slots are processed in order. In each slot, for all $j \in \{1, \ldots, N_{\mathrm{P}}\}$ the BS attempts channel estimation $\boldsymbol{\phi}_j \in \mathbb{C}^{M\times 1}$ and then payload estimation $\widehat{\boldsymbol{x}}_j$ via $\boldsymbol{\phi}_j$, according to

$$\boldsymbol{\phi}_j = \frac{\boldsymbol{P}\,\boldsymbol{s}_j^H}{\|\boldsymbol{s}_j\|^2} = \sum_{k\in\mathcal{A}^j} \boldsymbol{h}_k + \boldsymbol{z}_j \quad \text{and} \quad \widehat{\boldsymbol{x}}_j = \frac{\boldsymbol{\phi}_j^H \boldsymbol{Y}}{\|\boldsymbol{\phi}_j\|^2}. \tag{2}$$

In (2), $\mathcal{A}^j$ is the set of active users employing pilot j in the current slot, $\boldsymbol{s}_j \in \mathbb{C}^{1\times N_{\mathrm{P}}}$ is the j-th pilot, and $\boldsymbol{z}_j \in \mathbb{C}^{M\times 1}$ is a noise vector. Demapping and channel decoding are performed on $\widehat{\boldsymbol{x}}_j$. If a user $\ell \in \mathcal{A}$ is the only one picking pilot j in the currently processed slot ($\mathcal{A}^j = \{\ell\}$) then $\boldsymbol{\phi}_j$ provides an accurate estimate of this user's channel vector. We refer to these users as "singleton" users. Upon successful channel decoding performed on $\widehat{\boldsymbol{x}}_j$, the decoded message is stored in a buffer awaiting for phase 2. Note that, even in the case $\mathcal{A}^j = \{\ell\}$, if $|\mathcal{A}|$ is too large it may happen that the user payload is not successfully decoded, despite accuracy of the channel estimate $\boldsymbol{\phi}_j$ (see, e.g., [23])

In phase 2, an SIC algorithm is performed across slots. Its task is to subtract the interference due to copies of a decoded packet in the corresponding slots. Then, reprocessing these slots it may be possible to retrieve previously undecodable messages. Assuming payload aided based (PAB) subtractions [23], after correct decoding of the message of some user ℓ, $\boldsymbol{P}$ and $\boldsymbol{Y}$ are updated in all slots picked by user ℓ as

$$\boldsymbol{P}^{(i+1)} = \boldsymbol{P}^{(i)} - \boldsymbol{h}\,\boldsymbol{s}(\ell) \tag{3}$$

$$\boldsymbol{Y}^{(i+1)} = \boldsymbol{Y}^{(i)} - \boldsymbol{h}\,\boldsymbol{x}(\ell) \tag{4}$$

where $\boldsymbol{P}^{(0)} = \boldsymbol{P}$, $\boldsymbol{Y}^{(0)} = \boldsymbol{Y}$, and $\boldsymbol{h}$ is the channel estimate for user ℓ in the slot under processing.[2] In the slot where the user message has been decoded, $\boldsymbol{h}$ in (3) and (4) is set equal to $\boldsymbol{\phi}$ in (2), while in the other slots channel estimation relies on the user data payload (the same in all r replicas), as

$$\widehat{\boldsymbol{h}}_\ell = \frac{\boldsymbol{Y}^{(i)}\boldsymbol{x}(\ell)^H}{\|\boldsymbol{x}(\ell)\|^2}. \tag{5}$$

Under PAB SIC, every time the matrices $\boldsymbol{P}$ and $\boldsymbol{Y}$ are updated, (2) is re-computed in the current slot for each pilot, to check if any other user can be correctly decoded after interference subtraction. Whenever a new user is successfully decoded its message is buffered; phase 2 iterates until the buffer is empty.

[2]In practice, there are different ways to let the BS know the set of slots and the corresponding pilots chosen by a user. A simple way is to make them a function of the random message. Note also that, in case ACK messages are used, the BS disables subtraction of interference of future replicas of a decoded packet, since they will not be transmitted.

III. Proposed MAC Layer Access Techniques

This section provides an in-depth discussion of the ACK message design for the described MMA feedback-based schemes (Section III-A and Section III-B) and highlights how, for a fixed user payload size and symbol rate, increasing the ACK message time deteriorates the performance under a maximum latency constraint. To cope with this issue, in Section III-C new protocols able to support low-latency notifications to many active devices without degrading the overall performance are introduced. Next, the effective exploitation of the degrees of freedom offered by the power domain within the proposed protocols is addressed in Section III-D. Finally, analytical performance bounds are derived in Section III-E.

A. Acknowledgements in Massive Multiple Access

Concerning feedback, we define a misdetection event as the event in which a user, whose packet has been successfully decoded in phase 1, does not receive the corresponding notification by the BS. Similarly, a false alarm event occurs when a user, whose transmission was unsuccessful, is erroneously notified by the BS. False ACKs can heavily impair performance: users receiving a false ACK interrupt transmissions of their replicas although their message has not yet been decoded, which leaves them completely in the care of phase 2. On the other hand, a missing ACK is not as critical: the user has been decoded, so the interference it generates in future slots is subtracted, though imperfectly, by SIC. In fact, the worst case of a systematic misdetection event simply leads back to a CRA scheme without ACKs.

In feedback-aided massive access, the structure of the ACK messages is an important issue, since the problem of broadcasting them to the users becomes critical as the population size gets very large. Conventional ACK messages are "ID-based", meaning that the ACK message contains the IDs of the devices whose packet has been successfully decoded. In principle, however, the MAC protocol described in Section II admits another form of ACK messages, potentially more efficient than ID-based ones, referred to as "pilot-based" [10]. These two different feedback strategies are discussed in the remainder of this subsection. We also remark that several techniques to reduce the number of feedback bits, possibly involving enhanced compression techniques, have been proposed in the literature. For example, [29] deals with the problem of broadcasting a minimum-length feedback message for collision-free scheduling in massive random access. Here, a scheduled approach is considered and a feedback message is used to notify the active users about which orthogonal slot they have to transmit in, guaranteeing absence of collisions. After an initial user's detection phase, the BS sends the feedback message to the users, and only at the end of the scheduling procedure the users transmit their payload information.

1) Pilot-Based Feedback: Under the condition that only packets from singleton users can be successfully decoded during phase 1, the number of users to be notified at the end of each slot cannot exceed the number of available orthogonal pilots N_{P}. Thus, N_{P} bits per ACK message can be used, where bit j is set to 1 if a packet has been decoded when processing pilot j and is set to 0 otherwise. The above-mentioned condition about singleton packets is always satisfied in presence of perfect power control and a large number of BS antennas.

Again under the assumption that only packets from singleton users can be decoded during phase 1, pilot-based feedback represents the most efficient strategy to notify users in CRA protocols without incurring into false alarm or misdetection events. Remarkably, the number of bits per pilot-based ACK can be reduced admitting a controlled amount of misdetection events. Specifically, if a maximum number of users $U^{\star}_{\mathrm{ACK}}$ to be notified per slot is set, then the required number of bits per ACK is the one that allows encoding all possible words of N_{P} bits with Hamming weight up to $U^{\star}_{\mathrm{ACK}}$. Clearly, misdetection events occur whenever the number of successfully decoded packets in a slot exceeds $U^{\star}_{\mathrm{ACK}}$. In the following, we refer to this strategy as "compressed" pilot-based feedback.

As mentioned above, pilot-based feedback is efficient in terms of required bits per ACK, however it relies on the assumption that no capture effect (decoding of multiple packets per slot-pilot pair owing to power unbalance) can be exploited. When such an assumption does not hold, pilot-based feedback yields false alarms. This may happen not only due to power control imperfectness, but also when PU is intentionally introduced to enhance the receiver MPR capabilities as addressed later in Section III-D. A false ACK is generated whenever only one of the users colliding in the same slot-pilot pair is successfully decoded.

2) ID-Based Feedback: ID-based ACK messages require more information bits, causing an increment of the time resources needed by ACK transmission. A naive approach to acknowledge U_{ACK} users without incurring into false ACKs consists of broadcasting the concatenation of their IDs. This technique is simple but inefficient in terms of ACK length, making it problematic in MMA applications. A simple solution investigated in literature consists of compressing the user IDs via a hash function and of concatenating the ID-hashes instead of the IDs [13], [30]. Users are therefore associated with their ID-hashes, yielding a shortening of the ACK message size at the cost of possible hash collisions. This leads again to false alarm events, whose probability is however controllable. Considering equiprobable hashes of b bits, the probability that a device receives a false ACK in a slot where U_{ACK} users are notified is given by

$$P_{\mathrm{FA}}(U_{\mathrm{ACK}}, b) = \frac{U_{\mathrm{ACK}}}{2^b}\,. \tag{6}$$

In ID-based feedback with ID-hashes, the number of bits per ACK can be further reduced allowing also a controlled misdetection probability $P_{\mathrm{MD}}(U_{\mathrm{ACK}})$. This is addressed with more detail in Section III-B.

B. Acknowledgement Design

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 cyclic redundancy check (CRC) and by a channel code with rate $\mathcal{R}_{\mathrm{a}}$, the ACK packet size in symbols is

$$N_{\mathrm{ACK}} = \left\lceil N_{\mathrm{P,ACK}} + \frac{n_{\mathrm{b}} + n_{\mathrm{CRC}}}{\mathcal{R}_{\mathrm{a}} \log_2(\mathsf{M}_{\mathrm{ACK}})} \right\rceil \tag{7}$$

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 per ACK message. For the sake of notation simplicity, we define the user packet size in symbols as $N_{\mathrm{pkt}} = N_{\mathrm{P}} + N_{\mathrm{D}}$, where N_{P} and N_{D} are the number of pilot and payload symbols, respectively. The time requested for the ACK and uplink packet transmission is

$$T_{\mathrm{ACK}} = \frac{N_{\mathrm{ACK}}}{B_{\mathrm{s}}} \quad \text{and} \quad T_{\mathrm{pkt}} = \frac{N_{\mathrm{pkt}}}{B_{\mathrm{s}}} \tag{8}$$

where B_{s} is the symbol rate, assumed to be the same for uplink and feedback. We denote by α the ratio of the ACK message time to the packet time, i.e.,

$$\alpha = \frac{T_{\mathrm{ACK}}}{T_{\mathrm{pkt}}} = \frac{N_{\mathrm{ACK}}}{N_{\mathrm{pkt}}}\,. \tag{9}$$

The parameter α can be used to measure the time resources employed by the above-presented feedback implementations. Due to time resources spent for ACK message transmission at the end of each slot, the adoption of feedback necessarily reduces the number of slots per frame, for fixed frame time T_{F} and packet time T_{pkt}. As we will show, this yields a performance degradation of the CRA scheme, an issue that will be addressed in Section III-C. To be precise, letting $T_{\mathrm{s}} = T_{\mathrm{pkt}} + T_{\mathrm{ACK}}$ be the slot time, the number of slots per frame can be computed as

$$N_{\mathrm{s}} = \left\lfloor \frac{\Omega\, B_{\mathrm{s}}}{2\,(N_{\mathrm{P}} + N_{\mathrm{D}})\,(1+\alpha)} \right\rfloor \tag{10}$$

where $\Omega = 2T_{\mathrm{F}}$ is the maximum latency constraint. The maximum latency is experienced by a user that wakes up right after the BS beacon that initializes a frame (making it await for the next beacon after a time T_{F}) and that is decoded in the last slot of the subsequent frame or in its SIC phase.

1) Pilot-Based Acknowledgement Design: As described in Section III-A, pilot-based feedback aims at acknowledging in which pilots the BS has successfully decoded an uplink message. Therefore, with reference to (7), we have $n_{\mathrm{b}} = N_{\mathrm{P}}$ in uncompressed pilot-based feedback. However, as pointed out above, the number of ACK information bits n_{b} can be further reduced by adopting a compressed version of this feedback strategy, at the price of a nonzero (but controllable) misdetection probability. Specifically, we can fix a maximum tolerable misdetection probability $P^{\star}_{\mathrm{MD}}$ and calculate the maximum number of users $U^{\star}_{\mathrm{ACK}}$ which can be acknowledged in

a slot. Then, imposing that up to $U^{\star}_{\mathrm{ACK}}$ users can be notified by the BS, we calculate n_{b} as

$$n_{\mathrm{b}} = \left\lceil \log_2 \left(\sum_{i=0}^{U^{\star}_{\mathrm{ACK}}} \binom{N_{\mathrm{P}}}{i} \right) \right\rceil \leq N_{\mathrm{P}} \tag{11}$$

where $\sum_{i=0}^{U^{\star}_{\mathrm{ACK}}} \binom{N_{\mathrm{P}}}{i}$ is the total number of possible sequences of N_{P} bits with up to $U^{\star}_{\mathrm{ACK}}$ bits equal to "1". The procedure to obtain $U^{\star}_{\mathrm{ACK}}$ from $P^{\star}_{\mathrm{MD}}$ is detailed below.

2) ID-Based Acknowledgement Design: As mentioned above, ID-based ACKs can be efficiently deployed accepting non-zero false alarm and misdetection probabilities. Similar to compressed pilot-based ACKs, imposing a target misdetection probability $P^{\star}_{\mathrm{MD}}$ allows us finding the maximum number of users that can be notified per single slot, $U^{\star}_{\mathrm{ACK}}$. Then, fixing a target false alarm probability $P^{\star}_{\mathrm{FA}}$, from (6) we are able to obtain the number of ID-hash bits $b = \lceil \log_2(U^{\star}_{\mathrm{ACK}}/P^{\star}_{\mathrm{FA}}) \rceil$. Thus, we compute the number of required information bits n_{b} as

$$n_{\mathrm{b}} = b\, U^{\star}_{\mathrm{ACK}}. \tag{12}$$

Note that, whenever the ACK time T_{ACK} is imposed, it must be checked that n_{b} obtained from (12) fits the requirement.

3) Limiting ACK Messages via Controlled Misdetection Probability: For a given number of active users in a frame, K_{a}, the misdetection probability is $P(U_{\mathrm{ACK}} > U^{\star}_{\mathrm{ACK}} | K_{\mathrm{a}})$. Intuitively, the misdetection probability increases with K_{a} in the low traffic regime, reaches a maximum, and then decreases as K_{a} keeps growing due to system congestion. To release the analysis from the dependence on K_{a}, we define

$$P_{\mathrm{MD}}(U^{\star}_{\mathrm{ACK}}) = \max_{K_{\mathrm{a}}} \{ P(U_{\mathrm{ACK}} > U^{\star}_{\mathrm{ACK}} | K_{\mathrm{a}}) \} \tag{13}$$

and we design $U^{\star}_{\mathrm{ACK}}$ for this worst case scenario.

The probability $P_{\mathrm{MD}}(U^{\star}_{\mathrm{ACK}})$ depends on all PHY and MAC layer parameter choices, making its exact derivation analytically intractable. To address this problem, we introduce an idealized setting where the number of active user per slot is not influenced by previous ACKs. Since $P(U_{\mathrm{ACK}} > U^{\star}_{\mathrm{ACK}} | K_{\mathrm{a}})$ depends on the number of active user per slot K_{s}, we have a mismatch with the idealized setting. In particular, if in the realistic setting we have that K_{a} produces an average K_{s}, then, in the idealized setting we have that the same K_{s} is achieved for a smaller K_{a}, denoted in the following as $K_{\mathrm{a,eq}}$. This assumption makes the problem tractable as shown in Appendix A. Finally, we approximate P_{MD} as

$$P_{\mathrm{MD}}(U^{\star}_{\mathrm{ACK}}) \approx \max_{K_{\mathrm{a,eq}}} \{ 1 - P(U^{\star}_{\mathrm{ACK}} \,|\, K_{\mathrm{a,eq}}) \} \tag{14}$$

where $P(U^{\star}_{\mathrm{ACK}} \,|\, K_{\mathrm{a,eq}})$ is the probability to decode $U^{\star}_{\mathrm{ACK}}$ users which have picked a unique resource (i.e., slot-pilot pair), in a generic slot, given the total number of active user within the frame $K_{\mathrm{a,eq}}$ in the idealized setting (see (26) in Appendix A).

Let us now discuss the two main causes which make (14) an approximation. The first approximation is due to the fact that the mathematical problem we solve in Appendix A does not account for load decreasing due to past ACK messages. Consequently, users in a collision state with only acknowledged ones, could become singleton. Hence, this effect lead us to underestimate the number of ACKs (or decoded singleton) considered in the analysis. However, we cannot consider (14) as an upper bound because we are counting, as new decoded users, also singletons which could have been already decoded in past slots (i.e., users already acknowledged which should not re-transmit). In contrast with the previous effect, this leads to overestimate the number of counted ACKs. Nonetheless, it seems that from several simulation carried out, the effect which makes (14) an upper bound is dominant, making it useful for system design. This will be illustrated in Section IV.

C. Spaced Spatial Coupling Protocols

The proposed SSC protocols are variants of the intra-frame SC ones reviewed in Section II-B. They 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.[3] In a first SSC version, an active user picks one slot index n randomly in the set $\{1, \ldots, N_{\mathrm{s}} - (r-1)(W_{\mathrm{e}}+1)\}$, where W_{e} is the *waiting window* size, and transmits its r packet replicas in slots n, $n + W_{\mathrm{e}} + 1$, $\ldots$, $n + (r-1)(W_{\mathrm{e}}+1)$. 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 to be informed about success of its transmission in the previous slot. (Note that other active users may use these waiting slots for transmission of their replicas.) The access protocol is exemplified in Fig. 1b for $W_{\mathrm{e}} = 1$.

The proposed access protocol with spaced repetitions brings tangible advantages in terms of ACK scheduling. In absence of spacing and with half-duplex devices, it is necessary to allocate time resources at the end of each slot to accommodate reception of the ACK message by the users active in the slot. In presence of a maximum latency constraint Ω, this boils down to reducing the number of slots per frame N_{s}, which deteriorates the performance of CRA systems, as it will be shown in Section IV-B. In particular, as from (10), it is evident that increasing $T_{\mathrm{ACK}} = \alpha\, T_{\mathrm{pkt}}$ the number of slots per frame turns to be smaller. In contrast, no extra time for ACKs has to be accounted in SSC, since this protocol guarantees an entire slot for ACK messages reception by the active users. Hence, the number of slots per frame for SSC can be computed as

$$N_{\mathrm{s}} = \left\lfloor \frac{\Omega\, B_{\mathrm{s}}}{2\,(N_{\mathrm{P}} + N_{\mathrm{D}})} \right\rfloor \tag{15}$$

in the hypothesis that $T_{\mathrm{ACK}} = T_{\mathrm{pkt}}$ is a sufficient time for the feedback (a condition equivalent to $\alpha \leq 1$).

A second version of the SSC protocol features a randomization in the number of waiting slots between two successive replicas from the same user. 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 $\{1, \ldots, W_{\mathrm{e}}\}$. This is exemplified in Fig. 1c for $W_{\mathrm{e}} = 2$. Here, the generic user k

[3]The idea of spacing can be also applied on a baseline scheme adopting inter-slot ACKs, where the slots are picked uniformly at random and no waiting slots are considered.

transmits its $r = 3$ replicas waiting one slot between the first and the second replica and two slots between the second and the third transmission. The randomization in the number of waiting slots is introduced as it allows achieving a remarkable performance improvement in the error floor region of the PLR curves, as addressed in Section III-E.

D. Enhancing SSC by Exploiting the Power Domain

A possible strategy to increase the scalability of CRA schemes consists of resolving packet collisions by exploitation of the power domain, as first proposed in [14]. Thus, aiming at improving the receiver MPR capability in MMA applications we can further scale the system exploiting power diversity besides massive MIMO. The key idea is to harness the capture effect, varying the transmitting power levels (again in a power control setting) to obtain a discrete power unbalance (PU), i.e., set of receiving power levels. An accurate design of the power levels can make both the payload estimation (phase 1) and the SIC processing (phase 2) even more effective, causing a consistent boost in the overall system performance.

Adopting PDMA with discrete PU in a grant-free setting, each device may randomly select different received power levels for different packet replicas, out of a set of available ones. Let $\{P_q\}_{q=1}^{Q}$ be the set of available power levels, $P_1 > P_2 > \cdots > P_Q$, with $Q \leqslant r$ (where r is again the number of packet replicas per user). Then, through power control, each user adjusts its transmit power to reach the receiver with the desired power level. The expressions of the received pilot and payload complex matrices $\boldsymbol{P}$ and $\boldsymbol{Y}$, previously defined in Section II-C, can be reformulated to incorporate PU as

$$\boldsymbol{P} = \sum_{k \in \mathcal{A}} \sqrt{P(k)} \boldsymbol{h}_k \boldsymbol{s}(k) + \boldsymbol{Z}_p \tag{16}$$

and

$$\boldsymbol{Y} = \sum_{k \in \mathcal{A}} \sqrt{P(k)} \boldsymbol{h}_k \boldsymbol{x}(k) + \boldsymbol{Z} \tag{17}$$

where $P(k) \in \{P_q\}_{q=1}^{Q}$ is the received power level chosen by user $k \in \mathcal{A}$ for its replica in the current slot. Without loss of generality we apply the usual normalization $\sigma_{\mathrm{h}}^2 = 1$. Hereafter we consider the cases $Q = 2$ and $Q = 3$. For $Q = 2$, we denote the two power levels as σ_{H}^2 and σ_{L}^2 with $\sigma_{\mathrm{H}}^2 > \sigma_{\mathrm{L}}^2$, where "H" stands for high and "L" for low. Therefore, each element of $\sqrt{P(k)}\boldsymbol{h}_k$ has distribution $\mathcal{CN}(0, \sigma_{\mathrm{H}}^2)$ if $P(k) = \sigma_{\mathrm{H}}^2$ and $\mathcal{CN}(0, \sigma_{\mathrm{L}}^2)$ if $P(k) = \sigma_{\mathrm{L}}^2$. For $Q = 3$, we introduce an intermediate power level σ_{M}^2, with $\sigma_{\mathrm{H}}^2 > \sigma_{\mathrm{M}}^2 > \sigma_{\mathrm{L}}^2$, where "M" stands for middle. Again, each element of $\sqrt{P(k)}\boldsymbol{h}_k$ has distribution $\mathcal{CN}(0, \sigma_{\mathrm{M}}^2)$ if $P(k) = \sigma_{\mathrm{M}}^2$. However, we will show by numerical analysis that $Q = 2$ is optimal in our settings, and for this reason we will focus more on that case during our treatment.

The use of discrete PU on the transmission side requires some variations in the signal processing at the receiver. The main difference with respect to Section II-D lies in the SIC phase, upon subtraction of a packet in the slot where it has been decoded. In fact, if we decode a packet received with power σ_{H}^2 interfered in the same resource by another packet received with power σ_{L}^2 (or σ_{M}^2), the vector $\boldsymbol{\phi}$ contains the estimate of the channel vectors sum, as per (2). Hence, the subtraction operation (3) using $\boldsymbol{\phi}$ would cancel also the preamble of the interfering packet, making its future decoding not possible in that slot. To avoid this, we rely on a low-complexity energy detector with the purpose of deciding among: i) using $\boldsymbol{\phi}$ as channel estimate since a singleton user has been detected in that resource; ii) re-estimating the channel via (5) since multiple packets have been detected in that resource. Note that the same energy detector can also be used to optimize the number of operations the BS has to perform. For example, it can be used to avoid decoding attempts in empty or "too-crowded" resources.

Unfortunately, we have found by numerical analysis that the adoption of discrete PU with a random power selection (i.e., where each user chooses a power level randomly for each replica) does not provide a substantial improvement in CRA schemes with massive MIMO. Motivated by this empirical observation, we propose a new deterministic PU scheme which synergies very well with ACK-aided SSC protocols. In the considered scheme, each user adopts P_i for replica i, if $i < Q$, and adopts P_Q for all replicas i with $i \geqslant Q$. We remark that, to be fair in the comparisons between different strategies, the power levels $\{P_q\}_{q=1}^{Q}$ must be designed subject to the normalization constraint over the average power level, i.e., $\left[\sum_{q=1}^{Q-1} P_q + (r + 1 - Q)P_Q\right]/r = 1$. Considering $Q = 2$, the proposed power selection scheme, dubbed "HLx", still exploits the two power levels σ_{H}^2 and σ_{L}^2. Accordingly, the high-power level is always employed for the first replica transmitted by an active device, while the low-power level is applied to all subsequent $r - 1$ replicas. For example, if the number of replicas is $r = 3$ then the deterministic power level pattern is high-low-low (HLL). This choice of the power levels fits well with the use of an ACK mechanism at the end of each slot. In fact, the systematic use of a high power level for the first replica and of a low power level on the subsequent ones favors correct decoding of the first replica, rendering the ACK process more effective in avoiding transmission of unnecessary successive replicas of the same packet, with a sensible reduction of the load on the frame. The SSC-HLL scheme is pictorially represented in Fig. 1d (for non-randomized SSC) and Fig. 1e (for randomized SSC). Furthermore, we extend the proposed power selection strategy to the case $Q = 3$. We refer to this scheme as "HMLx". Again, with reference to $r = 3$, the deterministic pattern is high-middle-low (HML). Accordingly, the first replica employs an high-power level, the second replica employs an intermediate power level, and the last one employs a low-power level.

As mentioned above, we adopt an energy detector to switch between two SIC techniques when discrete PU is enabled. In particular we use the detection rule

$$\frac{\|\boldsymbol{\phi}\|^2}{M} \lessgtr \gamma \tag{18}$$

where γ is the switching threshold. Following a likelihood ratio test approach, a simple design of γ for large M and $Q = 2$ (see Appendix B), is given by

$$\gamma = \frac{2\sigma_{\mathrm{H}}^2 \left(\sigma_{\mathrm{L}}^2 + \sigma_{\mathrm{H}}^2\right)}{\sigma_{\mathrm{L}}^2 + 2\sigma_{\mathrm{H}}^2}. \tag{19}$$

Regarding ACK message design in presence of discrete PU with two power levels, ID-based feedback is not a concern, while it is still possible to exploit a pilot-based feedback as addressed next. Introducing two separate ACK messages, both of length N_{P} bits (i.e., $n_{\mathrm{b}} = 2\,N_{\mathrm{P}}$ bits), we can acknowledge both high-power and low-power users. This way, if two active users are present in the same resource with different power levels, we can separately acknowledge them. As usual in pilot-based ACKs, two users with the same power cannot be decoded in the same resource. For example, let us consider $N_{\mathrm{P}} = 4$ pilots and a slot in which we decode: one user with high power in pilot 1; one user with low power in pilot 2; two users with high and low power in pilot 4. The ACK message for high power users is "1001" and for low power ones is "0101", resulting in a final message "10010101". Similarly to what proposed in Section III-A, we can adopt a compressed version of this pilot-based strategy.

E. Error Floor Analysis of SSC

We can analytically derive PLR lower bounds for SSC schemes by extending the approach developed in [10] for SC ones and based on the "birthday paradox" [31], [32]. The lower bound is tight in the small K_{a} regime, and therefore can be used for error floor estimation, providing useful system design guidelines. For example, system configurations leading to poor performance in the error floor region (PLR values above the target one) can be preemptively discarded. Let "resource" be a slot-pilot pair. Each user employs a resource r-tuple $\{(n_1, p_1), \ldots, (n_r, p_r)\} \subseteq \mathcal{C}$, where n_i is the i-th chosen slot, p_i is the pilot chosen in slot n_i, and $\mathcal{C}$ is the set of all resource r-tuples. The size of $\mathcal{C}$ depends on the adopted CRA-type protocol. With perfect power control, an error event which cannot be resolved, even in noiseless conditions, occurs when at least two users choose the same resource r-tuple.

Under (non-randomized) 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. Since in each slot there are N_{P} available pilots, the size of $\mathcal{C}$ in the SSC case is

$$|\mathcal{C}_{\mathrm{SSC}}\,(W_{\mathrm{e}})| = [N_{\mathrm{s}} - (r-1)(W_{\mathrm{e}}+1)]\,N_{\mathrm{P}}^{r}\,. \tag{20}$$

Regarding randomized SSC, the first replica location is again drawn from $\{1, \ldots, N_{\mathrm{s}} - (r-1)(W_{\mathrm{e}}+1)\}$, while multiple options are available for the subsequent $r-1$ slots. The main intent of randomization is to increase the size of $\mathcal{C}$ by adding more admissible resource combinations, while spacing any two replicas from the same user by at least one slot. Once the first slot has been picked, an active user may choose between W_{e} slots for the second replica. Given the second chosen slot, the user has again W_{e} choices for the third one, and so on. This way, the number of resource combinations increases by a factor of W_{e}^{r-1}, yielding

$$|\mathcal{C}_{\mathrm{RSSC}}\,(W_{\mathrm{e}})| = [N_{\mathrm{s}} - (r-1)(W_{\mathrm{e}}+1)]\,W_{\mathrm{e}}^{r-1}\,N_{\mathrm{P}}^{r}\,. \tag{21}$$

Next, given that there are K_{a} active users, the probability that at least two of them choose the same resource r-tuple, under the assumption of a *uniform* probability distribution on the $|\mathcal{C}|$ admissible resource r-tuples, is

$$P_{\mathrm{u}} = 1 - \prod_{i=0}^{K_{\mathrm{a}}-1} \frac{|\mathcal{C}| - i}{|\mathcal{C}|}\,. \tag{22}$$

The true probability distribution on the resource r-tuples is non-uniform. Letting P_{c} be the probability that at least two active users (out of K_{a}) choose the same resource r-tuple under the actual probability distribution, owing to the birthday paradox (according to which the uniform probability distribution represents the worst case) we have $P_{\mathrm{c}} \geqslant P_{\mathrm{u}}$. Hence, denoting by C_k the event that k active users experience a resource r-tuple "collision", we can lower bound the packet loss probability P_{L} as

$$\begin{aligned} P_{\mathrm{L}} &\geqslant \frac{1}{K_{\mathrm{a}}} \sum_{k=2}^{K_{\mathrm{a}}} k\,P(C_k) \geqslant \frac{2}{K_{\mathrm{a}}} \sum_{k=2}^{K_{\mathrm{a}}} P(C_k) \\ &= \frac{2}{K_{\mathrm{a}}} P_{\mathrm{c}} \geqslant \frac{2}{K_{\mathrm{a}}} \left[1 - \prod_{i=0}^{K_{\mathrm{a}}-1} \frac{|\mathcal{C}| - i}{|\mathcal{C}|}\right] \end{aligned} \tag{23}$$

where the first inequality is due to the fact that resource r-tuple collisions are not the only sources of error (e.g., in a realistic setting a packet may not be successfully decoded even if it is the only one using a certain pilot in the slot where it arrives), the last inequality to $P_{\mathrm{c}} \geqslant P_{\mathrm{u}}$ and (22), and where $|\mathcal{C}|$ is given by (20) and (21) for SSC and randomized SSC, respectively.

It is easy to recognize that, in case of non-randomized SSC, the lower bound on the packet loss probability increases monotonically with the size of the waiting window W_{e}. In fact, with reference to (20) we note that the size of the set of all resource r-tuples, $|\mathcal{C}_{\mathrm{SSC}}|$, decreases with W_{e} causing a concurrent increment of the packet loss probability P_{L}. Therefore, the best choice of the waiting window size to reduce the error floor (as predicted by the lower bound, that becomes tight in the low traffic regime) consists of $W_{\mathrm{e}} = 1$, i.e., it is recommended to impose a waiting window of minimum size.

In contrast with its non-randomized version, for the randomized SSC protocol the lower bound on the packet loss probability P_{L} is not monotonically increasing with the maximum waiting window size W_{e}. We can derive the value of W_{e}, denoted by $\widehat{W}_{\mathrm{e}}$, that makes the lower bound minimum for given number of slots N_{s}, number of pilots N_{P}, and number of packet replicas r. Treating the quantity W_{e} as a real number, we can compute the derivative of $|\mathcal{C}_{\mathrm{RSSC}}\,(W_{\mathrm{e}})|$ in (21) with respect to W_{e} as

$$\frac{\partial |\mathcal{C}_{\mathrm{RSSC}}(W_{\mathrm{e}})|}{\partial W_{\mathrm{e}}} = (r-1)\,W_{\mathrm{e}}^{r-2}\,N_{\mathrm{P}}^{r}\{[N_{\mathrm{s}} - (r-1)] - r\,W_{\mathrm{e}}\}\,. \tag{24}$$

This results shows that $|\mathcal{C}_{\mathrm{RSSC}}\,(W_{\mathrm{e}})|$ increases from zero to $w_{\mathrm{e}} = (N_{\mathrm{s}} - (r-1))/r$ and then it decreases. Hence, we can compute $\widehat{W}_{\mathrm{e}}$ as

$$\widehat{W}_{\mathrm{e}} = \begin{cases} \lfloor w_{\mathrm{e}} \rfloor & \text{if } |\mathcal{C}_{\mathrm{RSSC}}\,(\lfloor w_{\mathrm{e}} \rfloor)| \geqslant |\mathcal{C}_{\mathrm{RSSC}}\,(\lceil w_{\mathrm{e}} \rceil)| \\ \lceil w_{\mathrm{e}} \rceil & \text{if } |\mathcal{C}_{\mathrm{RSSC}}\,(\lfloor w_{\mathrm{e}} \rfloor)| < |\mathcal{C}_{\mathrm{RSSC}}\,(\lceil w_{\mathrm{e}} \rceil)|\,. \end{cases} \tag{25}$$

For example, assuming $N_s = 78$ slots, $N_P = 64$ pilots and $r = 3$ replicas, applying (24) and subsequently (25) we obtain $\widehat{W}_e = 25$. We can also perform an analysis of the effect of the number of slots N_s and the number of packet replicas r on the lower bound. Specifically, we observe that the minimum window size $\widehat{W}_e$ increases when the number of slots N_s grows, while it decreases for a growing repetition rate r. Moreover, as expected, the value of the minimizing $\widehat{W}_e$ is always less than $[N_s - (r-1)]/(r-1)$, which is the maximum value of W_e allowed by the randomized SSC protocol.

We finally point out that the developed lower bound on the packet loss probability keeps holding, with no modification, also for the SSC-HLx scheme in which the SSC protocol is combined with power diversity. The reason is that, when at least two users choose the same resource r-tuple, due to the deterministic power pattern all different replicas collide with the same power level (the high-level and the low-level ones collide in the same resources for all the interfering users).

IV. Performance Evaluation

A. Simulation Setup

We provide Monte Carlo simulation results for the proposed CRA schemes, in which each user transmits information messages encoded 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. Then, the BCH codeword is padded with a final zero bit and the encoded bits are mapped onto a quadrature phase-shift keying (QPSK) constellation with Gray mapping, yielding a data payload of $N_D = 256$ symbols. We adopt a bounded distance decoder which corrects up to t errors. Simulation results are given for a symbol rate $B_s = 1\,\text{Msps}$, $N_P = 64$ pilots, and $r = 3$ replicas per active user if not otherwise stated. An orthogonal pilot set is constructed using Hadamard matrices. The noise variance is set to $\sigma_n^2 = 1$ (i.e., signal-to-noise ratio equal to 0 dB). All MAC protocols assume the possibility to exploit ACKs to preemptively stop the transmissions. The numerical results assume a perfect feedback channel for ACKs (i.e., all ACK messages are always successfully received) and a perfect power control mechanism ($\sigma_h^2 = 1$). In case of power unbalance (PU), with reference to (16) and (17) we adopt the two power levels $\sigma_H^2 = 1.6$ and $\sigma_L^2 = 0.7$ for the HLx power selection scheme ($Q = 2$). Instead, for the HMLx one ($Q = 3$) we employ $\sigma_H^2 = 1.3$, $\sigma_M^2 = 1$ and $\sigma_L^2 = 0.7$ for the high, middle, and low level, respectively. These choices derive from numerical optimization. The key performance indicator we consider is the packet loss rate (PLR), P_L, when a maximum latency constraint $\Omega = 50\,\text{ms}$ is imposed. In this framed system the maximum latency is two times the frame time [10]. The PLR is plotted against the number of simultaneously active users per frame, K_a, representing the system scalability parameter.

B. Numerical Results

1) Misdetection Probability Analysis: In Fig. 2, the misdetection probability, both simulated and analytically evaluated through the method developed in Appendix A, is reported for the randomized SSC protocol with $W_e = 2$, assuming $M = 128$ BS antennas. The simulated curves are obtained for $K_a \in \{700, 1025, 1200\}$. These numbers are opportunely chosen to represent different traffic conditions in the frame. The red curve characterizes a low traffic regime, which causes the transmission of a relatively small number of ACKs to the users. The azure one shows a high throughput regime where users experience a low error probability, with a large number of transmitted ACK messages. The green curve refers to a congested system where the overall number of ACKs is again small due to frequent decoding failures. For example, for $K_a = 1025$ we observe that, fixing a misdetection probability of 0.1% (i.e., $P_{MD}^\star = 10^{-3}$), approximately $U_{ACK}^\star = 21$ users are notified in a slot by the BS. Note that $U_{ACK}^\star$ is considerably lower than $N_P = 64$, which represents the maximum possible number of ACKs per slot.

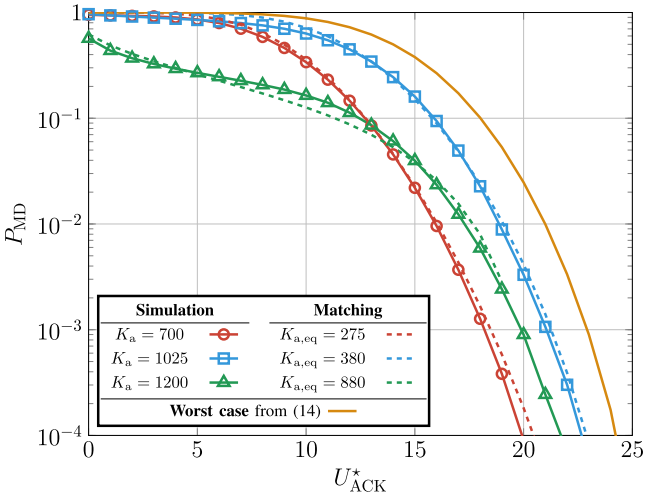

Fig. 2. Misdetection probability of randomized SSC protocol with $W_e = 2$, employing $M = 128$ antennas. Solid-marked: Simulation. Dashed: Matching. Solid-ocher: Worst case scenario derived from (14).

In dashed lines the theoretical curves based on the analysis conducted in Appendix A are plotted. As explained in Section III-B, the matching is obtained considering $K_{a,eq} < K_a$ due to the fact that the idealized setting described therein does not capture the traffic reduction caused by ACKs. Despite this, the analytical approximation fits very well the numerical simulation in all considered traffic regime. Lastly, the ocher curve shows the analytical worst case scenario derived using the approximation formulated in (14).

In the following we take the azure curve of Fig. 2 as a reference for ACK design. Based on exhaustive simulations, in fact, this curve is the one achieving the largest $U_{ACK}^\star$ for given P_{MD} and for the considered system parameters; therefore it represents the most challenging situation for ACK design. Note that, without running exhaustive simulations, the analytical ocher curve may also be used for ACK design.

2) ACK Design and Analysis: Table I reports ACK design results based on Section III-A, Section III-B, and the above misdetection probability analysis. All adopted parameters are reported in the north-west side of the table. First, we present a pilot-based ACK design for $N_P \in \{64, 128\}$. In this uncompressed feedback scheme, neither the false alarm nor

TABLE I

COMPARISON BETWEEN DIFFERENT ACK FEEDBACK TECHNIQUES WITH AN INTRA-FRAME SC PROTOCOL. TOP TABLE: PILOT-BASED FEEDBACK. MIDDLE TABLE: COMPRESSED PILOT-BASED FEEDBACK WITH FIXED MISDETETECTION PROBABILITY. BOTTOM TABLE: ID-BASED FEEDBACK WITH FIXED NUMBER OF PILOTS

System Parameters			
Maximum latency, Ω [ms]	50	Symbol rate, B_s [Msps]	1
Number of antennas, M	128	Packet repetition rate, r	3
High power level, σ_H^2	1.6	Low power level, σ_L^2	0.7
Noise variance, σ_n^2	1.0	Channel variance, σ_h^2	1.0

Packet Parameters		**ACK Parameters**	
Payload symbols, N_D	256	Preamble length, $N_{P,ACK}$	4
Code rate, $\mathcal{R}_c$	0.82	Code rate, $\mathcal{R}_a$	1/3
QPSK modulation, M	4	Modulation order, M_{ACK}	16
		CRC bits, n_{CRC}	32

Pilot-based	$N_P = 64$	$N_P = 128$
ACK bits, n_b	64	128
ACK symbols, N_{ACK}	76	124
Ratio T_{ACK}/T_{pkt}, α	0.2375	0.3229
Effective slots, N_s	63	49

C. Pilot-based, $P_{MD} = 5 \cdot 10^{-2}$	$N_P = 64$	$N_P = 128$
ACK bits, n_b	53	77
ACK symbols, N_{ACK}	68	86
Ratio T_{ACK}/T_{pkt}, α	0.2125	0.2239
Effective slots, N_s	64	53

ID-based, $N_P = 64$	$P_{FA} = 10^{-3}$, $P_{MD} = 5 \cdot 10^{-2}$	$P_{FA} = 10^{-3}$, $P_{MD} = 10^{-2}$	$P_{FA} = 10^{-4}$, $P_{MD} = 10^{-2}$	$P_{FA} = 10^{-6}$, $P_{MD} = 5 \cdot 10^{-3}$
Hash bits, b	10	10	14	20
ACK bits, n_b	180	190	266	400
ACK symbols, N_{ACK}	163	171	228	328
Ratio T_{ACK}/T_{pkt}, α	0.5094	0.5344	0.7125	**1.0250**
Effective slots, N_s	51	50	45	38

the misdetection probability need to be considered. For $N_P = 64$ pilots, applying (7) we have $N_{ACK} = 76$ symbols, descending in a ratio $\alpha = T_{ACK}/T_{pkt} = 0.2375$. For example, adopting an SC protocol, the number of available transmission slots is reduced to $N_s = 63$, while for an SSC protocol we have $N_s = 78$ as per (15) since $\alpha \leq 1$. Allowing a low misdetection probability in the ACK design, the compressed pilot-based feedback permits to further reduce the feedback message length. For example, imposing a $P^\star_{MD} = 5 \cdot 10^{-2}$, from the azure curve in Fig. 2 we get $U^\star_{ACK} = 18$, $n_b = 53$ bits (from (11)) and $\alpha = 0.2125$. It translates into a saving of 1 slot into the whole transmission frame ($N_s = 64$). This recovered slot can be used, for example, to ensure an additional guard and processing time to the BS.

Concerning ID-based feedback, the goal is to find the value of α corresponding to given $P^\star_{FA}$ and $P^\star_{MD}$. For example, assuming a target $P^\star_{FA} = 10^{-4}$ and $P^\star_{MD} = 10^{-2}$, from (12), (7), and (9) we find $\alpha = 0.7125$. From (10) the number of available slots of a non-spaced access protocol is reduced to $N_s = 45$. It is worth noting that, since $\alpha < 1$, the proposed SSC scheme with $W_e = 1$ permits to solve the frame length contraction issue. Imposing more severe requirements, for example $P^\star_{FA} = 10^{-6}$ and $P^\star_{MD} = 5 \cdot 10^{-3}$, we end up with $\alpha > 1$. In this situation, even the SSC protocol needs a reduction of the number of slots N_s.[4]

3) Benefits and Costs of the Feedback: Fig. 3 elaborates on the cost of feedback, in terms of reduction of the largest supported K_a (scalability penalty) for a given PLR (reliability). In particular, the intra-frame SC protocol (Section II-B) and the proposed SSC protocols with our without power unbalance (Section III-C and Section III-D) are considered. We fairly compare them fixing the maximum latency Ω and assuming $M = 128$ BS antennas. The dashed azure curve in Fig. 3 corresponds to the ideal intra-frame SC system with

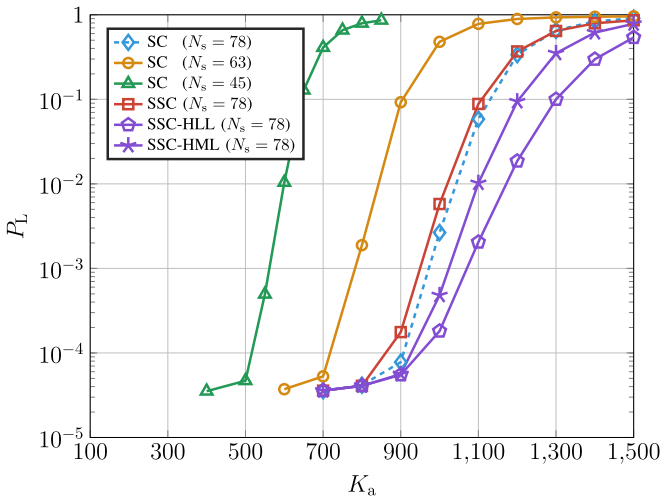

Fig. 3. Packet loss rates of CRA-type protocols (SC, SSC and SSC-HLL) for different ACK time resources consuming, and $M = 128$ antennas. The SSC curves assume $W_e = 1$.

an instantaneous feedback, not consuming any time resources ($\alpha = 0$); for this ideal scheme we obtain $N_s = 78$ slots per frame from (10). Considering then a realistic system (i.e., with non-instantaneous feedback) having for example $\alpha = 0.2375$ (obtained for pilot-based ACKs, see Table I), we obtain $N_s = 63$. As shown in Fig. 3, this causes a tangible performance degradation with respect to the idealized scheme. Moreover, in a scenario where pilot-based ACKs are not practicable we have to adopt ID-based ACK messages. For example, with reference again to Table I, with $\alpha = 0.7125$ we obtain $N_s = 45$ slots, leading to a dramatic performance degradation compared to the ideal case. The SSC protocol represents an elegant and effective solution to this problem. As shown in the figure, for $W_e = 1$, the presence of a waiting window between successive replica transmissions from the same device guarantees $N_s = 78$ slots per frame as per (15), whenever $\alpha < 1$. We emphasize that this simple approach achieves the same performance as the idealized intra-frame

[4]We mention that a possible solution to this issue, not deepened in this paper, could involve an SSC scheme in which additional small sub-slots for the ACKs, able to ensure the residual time $(\alpha - 1)T_{pkt}$, are introduced between the packets transmission slots.

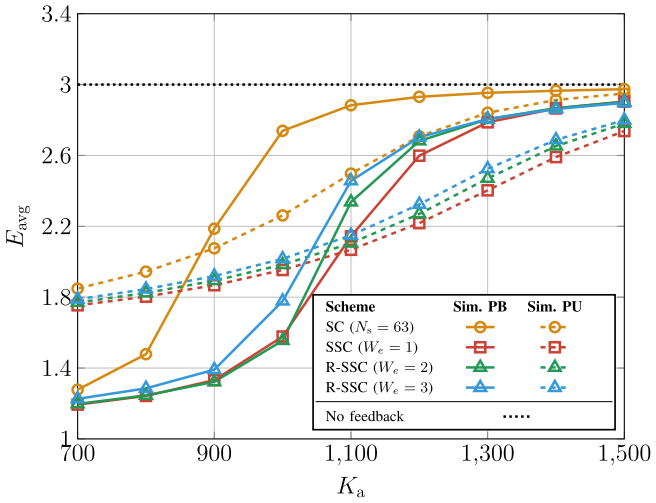

Fig. 4. Average energy cost per user per information message for different CRA-type protocols (SC, SSC, R-SSC) exploiting both power balance and power unbalance, with $M = 128$ antennas.

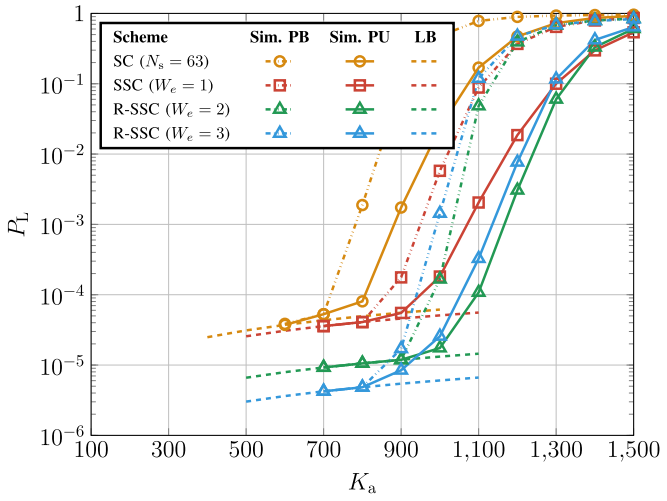

Fig. 5. Packet loss rates achieved by CRA-type protocols (SC, SSC and R-SSC) exploiting both power balance and power unbalance, with $M = 128$ antennas. Lower bound predictions computed by (23).

SC system. Moreover, adopting a power diversity approach we can improve performance even further. In this example, using the SSC-HLL scheme, we are able to achieve a 10% scalability boost at a target PLR of 10^{-3}. Instead, adopting the SSC-HML one, a 5% boost at the same target PLR is achieved.

Next, we discuss how ACKs can lead to energy savings. For the sake of fairness in comparing different MAC protocols, we assume that each packet replica has a cost in energy equal to its power level. Then, considering all actual packet transmissions in the frame (weighted by their corresponding power levels) and dividing this value by K_{a}, we evaluate the average energy cost, E_{avg}. Such a metric is related to the average energy consumption per user per information message. Fig. 4 shows the trend of E_{avg} versus K_{a} for feedback-aided CRA-type protocols (SC, SSC, and R-SSC) with $r = 3$, adopting both power balance and power unbalance. Considering a system without feedback and letting the energy spent per replica be normalized to 1, the average energy cost is always $E_{\text{avg}} = 3$, since each active user always transmits $r = 3$ replicas per frame, independently of the system load. Regarding the schemes with power balance, we can observe how the feedback mechanism guarantees conspicuous energy savings, especially in the low traffic regime. For example, considering a target load $K_{\text{a}}^{\star} = 700$, we note that the average energy cost remains considerably below 2, meaning that the majority of the active users are decoded after transmission of their first replica. We also see that, when the system load increases, the energy metric E_{avg} converges to its maximum value of 3, because a considerable amount of users needs to transmit nearly all replicas to have a chance of being decoded. Next, the energy consumption exhibits a slightly different trend when power unbalance is exploited. Specifically, for low traffic values, the average energy cost is higher with respect to the power balance case, since the first replica is transmitted with a higher power, according to the HLx deterministic scheme. On the other hand, the PLR performance of the HLx scheme is better.

4) Impact of Waiting Time Randomization: Fig. 5 elaborates on the results of Fig. 3 to show the effect of randomization in the number of SSC waiting slots, again assuming $M = 128$ BS antennas. The dash-dotted curves, relative to power balance (PB) and corresponding to SC with $N_{\text{s}} = 63$ and to SSC with $W_{\text{e}} = 1$ are the same appearing in Fig. 3. In addition, the performance of randomized SSC schemes with $W_{\text{e}} = 2$ and $W_{\text{e}} = 3$, denoted as R-SSC in the figure legend, is reported. Notably, we observe that a randomization in the number of waiting slots not only provides a better error floor performance, as expected from the analysis in Section III-E (dashed lines), but also a better performance in the waterfall region. With the considered simulation parameters, we found that the best waterfall performance for (non-randomized) SSC is achieved for a waiting window size $W_{\text{e}} = 1$, while for randomized SSC it is achieved for $W_{\text{e}} = 2$; these are the minimum window sizes for the two protocol versions. For a target PLR $P_{\text{L}}^{\star} = 10^{-3}$, randomized SSC with $W_{\text{e}} = 2$ improves the system scalability, with a gain as high as 15% over the SC protocol, while supporting a T_{ACK} that is 10 times larger than the SC one. Applying the analysis in Section III-E, we have $\widehat{W}_{\text{e}} = 25$ for the chosen system parameters. As predicted by such an analysis, randomized SSC schemes with progressively higher waiting window sizes tend to perform better in the error floor region as long as $W_{\text{e}} \leq \widehat{W}_{\text{e}}$. However, we also observed that the improved error floor performance comes at the cost of a deterioration in the waterfall performance as shown in Fig. 5 for $W_{\text{e}} = 3$ in comparison to $W_{\text{e}} = 2$, this latter value offering the best tradeoff between the two regions of the performance curve.

5) Impact of Power Unbalance: As already observed upon commenting Fig. 3, intra-frame SSC with discrete PU is able to outperform both idealized SC ($\alpha = 0$) and SSC with PB. In particular, we consider the HLL power pattern ($Q = 2$ and $r = 3$) previously introduced is Section III-D, where the first replica is transmitted with high power ($\sigma_{\text{H}}^2 = 1.6$) and the others with low power ($\sigma_{\text{L}}^2 = 0.7$). We also consider the HML scheme ($Q = 3$ and $r = 3$), where the first

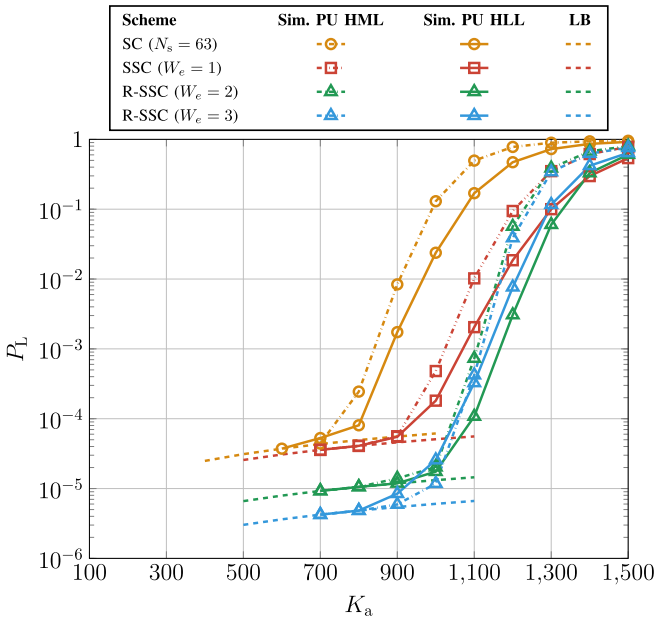

Fig. 6. Packet loss rates achieved by CRA-type protocols (SC, SSC and R-SSC) exploiting power unbalance with HLL and HML power selection schemes and employing $M = 128$ antennas. Lower bound predictions computed by (23).

replica is transmitted with high power ($\sigma^2_{\mathrm{H}} = 1.3$), the second replica with intermediate power ($\sigma^2_{\mathrm{M}} = 1$) and the last one with low power ($\sigma^2_{\mathrm{L}} = 0.7$). The performance improvement that was pointed out in Fig. 3 for the HLL scheme can be systematically observed also in Fig. 5 as the consequence of the fact that the considered pattern has been applied to all spatial coupling schemes. In this example, we observe a uniform 10% scalability boost at $P^{\star}_{\mathrm{L}} = 10^{-3}$, given by PU. On the other hand, observing the error floor region, we have no difference between PB and PU, as anticipated in Section III-E.

Comparing, instead, the HML and HLL power selection schemes we note that, owing to the imposed power normalization, the first replica of the HLL deterministic pattern is received with a higher power level with respect to the HML one. Instead, the second replica of the HLL scheme experiences a lower power level compared to the second one of HML. Ultimately, the third replica is received with the same power in both schemes. Consequently, the HLL scheme tends to favor decoding of the first replica, while HML favors more the second replica. Since we are using acknowledgements, we expect that HLL outperforms HML in terms of overall packet loss rate. To substantiate this claim, we report in Fig. 6 the comparison between these two power control strategies. For all the analyzed spatial coupling protocols, the HLL scheme shows a better waterfall performance than the HML protocol. As mentioned above, we can attribute this trend to the effectiveness of the feedback mechanism. In fact, as also pointed out in in Section III-D, increasing the power level of the first replica favors its correct decoding, improving the effectiveness of the ACK phase.

6) Results for Different Degree Distributions: In principle, CRA-type protocols allow variable packet repetition rates, meaning that different users might transmit different number of replicas in the frame. Typically, the choice of the packet repetition rate is subject to probability distributions. Let us define the degree distribution through the polynomial $\Lambda(x) = \sum_r \Lambda_r x^r$, where Λ_r is the probability that a generic user transmits r packet replicas over the frame.

So far, we have provided results for a concentrated degree distribution profile with packet repetition rate $r = 3$, meaning that each active user deterministically transmits 3 packet replicas over the frame ($\Lambda(x) = x^3$). This choice follows from the analysis performed in [19], where it was shown that the repetition rate $r = 3$ achieves a good PLR performance tradeoff in a Rayleigh block fading channel setting, where a non-ideal PHY layer processing and a multiple antenna receiver are additionally considered. However, the setting of [19] differs from the one proposed in this paper, since neither advanced MAC layer protocols nor exploitation of power domain are considered in [19]. Hence, we have extended the numerical results simulating more degree distributions, both regular and irregular, aiming to understand the best profile for our scenario.

We report the performance curves in Fig. 7. The simulations are performed for an SSC protocol with waiting window size $W_{\mathrm{e}} = 1$, exploting discrete PU. The BS is equipped with $M = 128$ antennas. The analyzed distributions are: *i)* $\Lambda(x) = x^2$ (regular, with constant repetition degree 2), adopting an HL power profile, with $\sigma^2_{\mathrm{H}} = 1.3$ (high power level) and $\sigma^2_{\mathrm{L}} = 0.7$ (low power level); *ii)* $\Lambda(x) = x^3$ (regular, with constant repetition degree 3), adopting an HLL power profile, with $\sigma^2_{\mathrm{H}} = 1.6$ and $\sigma^2_{\mathrm{L}} = 0.7$; *iii)* $\Lambda(x) = 0.5465\ x^2 + 0.1623\ x^3 + 0.2912\ x^6$ (irregular, with average repetition degree 3.33 [6]), adopting an HLx power profile, with $\sigma^2_{\mathrm{H}} = 1.6$ and $\sigma^2_{\mathrm{L}} = 0.7$; *iv)* $\Lambda(x) = 0.5\ x^2 + 0.5\ x^4$ (irregular, with average repetition degree 3), adopting an HLx power profile, with $\sigma^2_{\mathrm{H}} = 1.6$ and $\sigma^2_{\mathrm{L}} = 0.7$.

The distribution $\Lambda(x) = x^2$ outperforms all of the others in the waterfall region, but is affected by an high error floor (i.e., around $P_{\mathrm{L}} = 10^{-2}$). Out of the two irregular distributions, the more "concentrated" one shows the better waterfall performance; nevertheless, both exhibit a worse trend with respect to $\Lambda(x) = x^2$. The regular distribution $\Lambda(x) = x^3$ achieves the best compromise between the waterfall and the error floor performance. In fact, it only experiences a 10% scalability degradation with respect to $\Lambda(x) = x^2$ at a target PLR $P^{\star}_{\mathrm{L}} = 10^{-2}$. Otherwise, for a low-load system (e.g., $K^{\star}_{\mathrm{a}} = 900$), it can reach a PLR inferior to 10^{-4} while $\Lambda(x) = x^2$ saturates to a PLR slightly lower than 10^{-2}. Despite the above-mentioned setting differences, the analysis performed in [19] remains consistent in the analyzed system. Following, the best degree distribution profile consists of a concentrated one with packet repetition rate $r = 3$.

7) Results for Different Numbers of Pilots: In Fig. 8 the performance is shown for SC with $\alpha = 0.1$, SSC with $W_{\mathrm{e}} = 1$, and randomized SSC with $W_{\mathrm{e}} = 2$ protocols, varying the number N_{P} of orthogonal pilots. Specifically, we consider $N_{\mathrm{P}} \in \{32, 64, 128\}$, while the number of BS antennas is $M = 256$. Coherently with the number of pilots N_{P} also the number of available slots N_{s} changes accordingly to (10) and (15); this leads us to $(N_{\mathrm{P}}, N_{\mathrm{s}}) \in \{(32, 78), (64, 71), (128, 59)\}$ for SC with $\alpha = 0.1$ and to $(N_{\mathrm{P}}, N_{\mathrm{s}}) \in \{(32, 86), (64, 78), (128, 65)\}$ for SSC and

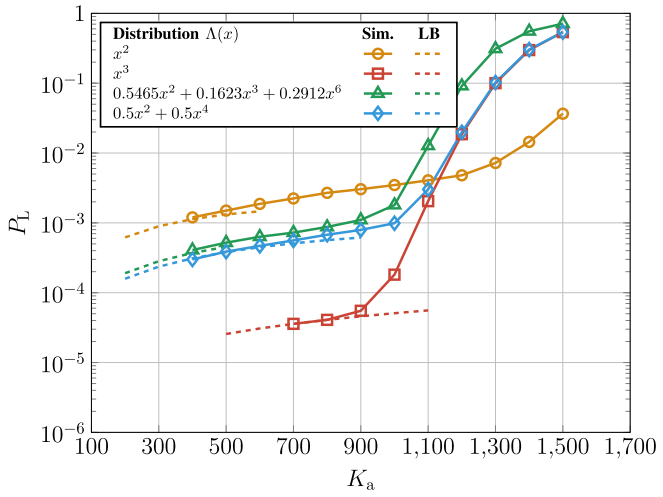

Fig. 7. Packet loss rate achieved by SSC-type protocols for different degree distribution profiles, exploiting power unbalance, with $M = 128$ antennas. Lower bound predictions computed by (23), with slight variations for the irregular distributions.

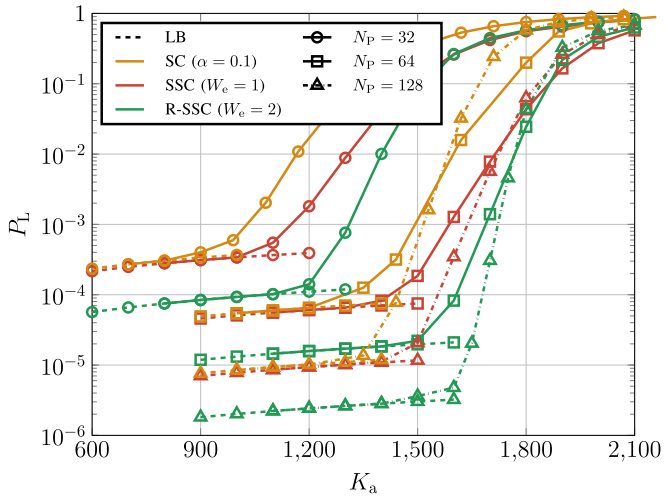

Fig. 8. Packet loss rates achieved by CRA-type protocols (SC, SSC and R-SSC) with $N_P \in \{32, 64, 128\}$, employing $M = 256$ antennas. Lower bound predictions computed by (23).

randomized SSC. Strictly speaking, the comparison between schemes using different N_P is not completely fair, due to the different coherence time assumptions. However, we find this comparison useful to highlight two behaviors. Firstly, increasing the number of pilots N_P, we observe that the scalability performance boost saturates. This is mainly caused by the fact that: *i)* N_P is approaching N_D, resulting in an unacceptable decrease of N_s; *ii)* N_P is approaching M which generates a sort of saturation in terms of multi-packet reception. It seems that N_P should not be larger than M, and this can be true as a general guideline. However, N_P can also be non-trivially increased above M, for example, to increase the probability to have singletons. Secondly, we can decrease the error floor, for example, to place it under a target PLR. This is due to the fact that a higher number of pilots translates into a higher number of available resources, causing less probable unresolvable collisions. Anyway, after a certain value of N_P, a saturation phenomenon arises. Finally, we highlight that an increase of N_P imposes stricter requirements in terms of ACK message symbols, according to (7), which reflects into more severe ACK time constraints.

V. Conclusion

Our study shows that inter-slot ACK messages can be employed without any performance loss by an effective pipeline mechanism, naturally arising in our proposed spaced spatial coupling (SSC) protocol. The possibility to enable such a feedback strategy brings advantages both in term of energy efficiency and interference reduction paving the way to MMA for low power devices. Alongside "spaced" protocols, we introduce a novel algorithm based on intentional power unbalance, specifically designed for the feedback-aided schemes we propose. This algorithm prioritizes the first replica, enhancing the probability of a successful decoding and subsequently triggering an ACK. Indeed, we show how SSC can reach practically the same performance of a spatial coupling scheme with instantaneous feedback, and the tailored power unbalance proposal obtain around a 10% scalability boost. The results are validated using different ACK strategies and analytical lower bound on the packet loss rate are provided.

Appendix A
Probability Distribution of the Number of Transmitted ACK Messages

In this appendix we derive the probability to decode U_{ACK} users which have picked a unique resource (i.e., pair slot-pilot), also referred as singleton, in a generic slot, given that the total number of active user within the frame K_a. Thus, we consider K_a active devices in the frame, K_s of which are active in the slot under analysis. Using the law of total probability, we can write the probability that U_{ACK} ACKs are sent out at the end of the slot, given K_a, as

$$P(U_{ACK} \mid K_a) = \sum_{K_s} P(K_s \mid K_a) P(U_{ACK} \mid K_s). \tag{26}$$

The probability to have K_s active users in the slot, given K_a total active users in the frame, is

$$P(K_s \mid K_a) = \binom{K_a}{K_s} p^{K_s} (1-p)^{K_a - K_s}. \tag{27}$$

The parameter p in (27) depends on the specific CRA protocol, as

$$p = \begin{cases} \dfrac{r}{N_s} & \text{for baseline CRA} \\ \dfrac{r}{N_s - (W_e+1)(r-1)} & \text{for SSC} \end{cases} \tag{28}$$

where the "baseline CRA" is the CRA access protocol where slots for transmissions of replicas are chosen uniformly at random without replacement. Note that, for the sake of simplicity, we are neglecting termination effects in the derivation of p when addressing SSC protocols. Hence, (28) is exact when considering a "central" slot $n \in \{W_e(r-1)+r, \ldots, N_s - (W_e+1)(r-1)\}$. Therefore, the probability to decode (and

consequently acknowledge) U_{ACK} singleton users, given K_{s}, can be expressed as

$$P(U_{\mathrm{ACK}} \,|\, K_{\mathrm{s}}) = \sum_{\delta} P(U_{\mathrm{ACK}} \,|\, \delta, K_{\mathrm{s}}) P(\delta \,|\, K_{\mathrm{s}}) \tag{29}$$

where δ is the number of singleton users in the slot. For example, assuming a t-error correcting code with hard-decision decoding and QPSK symbols with Gray labelling, and applying the analytical approximation derived in [33], we have

$$P(U_{\mathrm{ACK}} \,|\, \delta, K_{\mathrm{s}}) = \binom{\delta}{U_{\mathrm{ACK}}} \Big[1 - P_{\mathrm{f}}(K_{\mathrm{s}} - 1)\Big]^{U_{\mathrm{ACK}}} \times \Big[P_{\mathrm{f}}(K_{\mathrm{s}} - 1)\Big]^{\delta - U_{\mathrm{ACK}}} \tag{30}$$

where

$$P_{\mathrm{f}}(n) = 1 - \sum_{d=0}^{t} \binom{N_{\mathrm{D}}}{d} P_{\mathrm{e}}^{d}(n) \left(1 - P_{\mathrm{e}}(n)\right)^{N_{\mathrm{D}} - d} \tag{31}$$

and

$$P_{\mathrm{e}}(n) = \mathrm{erfc}\left(\sqrt{\frac{M}{2n}}\right) - \frac{1}{4}\mathrm{erfc}^2\left(\sqrt{\frac{M}{2n}}\right). \tag{32}$$

Finally, the probability to have δ singleton users, given K_{s}, can be obtained recalling the following probability result: the probability to have m cells with exactly k balls inserting B total balls in a total of C cells is [34]

$$g(m; k, B, C) = \frac{(-1)^m}{C^B} \sum_{j} (-1)^j \binom{j}{m} \binom{C}{j} \times (C - j)^{B - jk} \frac{B!}{(k!)^j (B - jk)!} \tag{33}$$

where $j \in \{j : j \geq m,\ j \leq C,\ kj \leq B\}$. Thus, we conclude that

$$P(\delta \,|\, K_{\mathrm{s}}) = g(\delta; 1, K_{\mathrm{s}}, N_{\mathrm{P}}). \tag{34}$$

APPENDIX B
ACTIVITY DETECTION THRESHOLD

In this appendix, we derive the activity detection threshold γ in (19). The objective is to discern if a resource is used by a single user or by multiple users in a discrete PU scenario with two power levels σ_{H}^2 and σ_{L}^2, with $\sigma_{\mathrm{H}}^2 > \sigma_{\mathrm{L}}^2$. The worst case is represented by the situation where we need to discriminate between a single user employing the high power level and two users that employ the high and the low power level, respectively. If a single user is transmitting with the high power level, then $\|\boldsymbol{\phi}\|^2/M$ is chi-squared distributed with $2M$ degrees of freedom. Since M is large we have that $\|\boldsymbol{\phi}\|^2/M$ can be well-approximated by a normal distribution $\mathcal{N}\left(\sigma_{\mathrm{H}}^2, \sigma_{\mathrm{H}}^4/M\right)$. Analogously, in case two users transmit simultaneously with different power levels then the distribution of $\|\boldsymbol{\phi}\|^2/M$ can be approximated as $\mathcal{N}\left(\sigma_{\mathrm{H}}^2 + \sigma_{\mathrm{L}}^2, (\sigma_{\mathrm{H}}^2 + \sigma_{\mathrm{L}}^2)^2/M\right)$. We can now use a likelihood ratio test approach. Let us define $y = \|\boldsymbol{\phi}\|^2/M - \sigma_{\mathrm{H}}^2$. Then, we can write the likelihood ratio as

$$\Lambda(y) = \frac{\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{H}}^2 + \sigma_{\mathrm{L}}^2} \exp\left(-\frac{M}{2}\left[\left(\frac{y - \sigma_{\mathrm{L}}^2}{\sigma_{\mathrm{L}}^2 + \sigma_{\mathrm{H}}^2}\right)^2 - \left(\frac{y}{\sigma_{\mathrm{H}}^2}\right)^2\right]\right). \tag{35}$$

The equivalent log-likelihood ratio test is given by

$$\ln \Lambda(y) \lessgtr \ln \eta \tag{36}$$

where η is the threshold of the likelihood ratio test chosen accordingly to a certain criterion (e.g., Bayes, minimax, Neyman-Pearson, etc.). Solving (36) for y and adding σ_{H}^2, we have that the threshold γ to be used in (18) is

$$\gamma = \sigma_{\mathrm{H}}^2 + \frac{\sigma_{\mathrm{L}}^2 \left(\sqrt{1 + (K^2 - 1)\left(1 + \frac{2}{M}\frac{K}{K-1}\ln(K\eta)\right)} - 1\right)}{K^2 - 1} \tag{37}$$

where $K = 1 + \frac{\sigma_{\mathrm{L}}^2}{\sigma_{\mathrm{H}}^2}$. Moreover, for large M we can approximate γ as

$$\gamma \approx \sigma_{\mathrm{H}}^2 + \sigma_{\mathrm{L}}^2 \frac{K - 1}{K^2 - 1} \tag{38}$$

which justifies (19). Note that (38) is also the intersection of the two Gaussian distributions under examination.

REFERENCES

[1] J. Sachs et al., "Machine-type communications," in *5G Mobile and Wireless Communications Technology*, A. Osseiran, J. Monserrat, and P. Marsch, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2016, ch. 4.

[2] C. Bockelmann et al., "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28969–28992, 2018.

[3] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148–156, Aug. 2020.

[4] N. H. Mahmood et al., "White paper on critical and massive machine type communication towards 6G," *6G Res. Vis.*, vol. 11, no. 11, pp. 1–36, Jun. 2020.

[5] 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.

[6] 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.

[7] 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.

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

[9] J. Haghighat and T. M. Duman, "Energy efficiency analysis of a feedback-aided IRSA scheme," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, Jun. 2022, pp. 2886–2891.

[10] L. Valentini, M. Chiani, and E. Paolini, "Massive grant-free access with massive MIMO and spatially coupled replicas," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7337–7350, Nov. 2022.

[11] J. Haghighat and T. M. Duman, "An energy-efficient feedback-aided irregular repetition slotted ALOHA scheme and its asymptotic performance analysis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9808–9820, Dec. 2023.

[12] M. Bashir, E. Nassaji, D. Truhachev, A. Bayesteh, and M. Vameghestahbanati, "Unsourced random access with threshold-based feedback," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7072–7086, Dec. 2023.

[13] 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.

[14] G. Mazzini, "Power division multiple access," in *Proc. IEEE Int. Conf. Universal Pers. Commun. Conf.*, vol. 1, Florence, Italy, Oct. 1998, pp. 543–546.

[15] J. Su, G. Ren, and B. Zhao, "NOMA-based coded slotted ALOHA for machine-type communications," *IEEE Commun. Lett.*, vol. 25, no. 7, pp. 2435–2439, Jul. 2021.

[16] X. Shao, Z. Sun, M. Yang, S. Gu, and Q. Guo, "NOMA-based irregular repetition slotted ALOHA for satellite networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 624–627, Apr. 2019.

[17] F. Babich, G. Buttazzoni, F. Vatta, and M. Comisso, "Energy-constrained uncoordinated multiple access for next-generation networks," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1808–1819, 2020.
[18] Y. Ma, Z. Yuan, W. Li, and Z. Li, "Novel solutions to NOMA-based modern random access for 6G-enabled IoT," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15382–15395, Oct. 2021.
[19] L. Valentini, M. Chiani, and E. Paolini, "A joint PHY and MAC layer design for coded random access with massive MIMO," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Rio di Janeiro, Brazil, Dec. 2022, pp. 2505–2510.
[20] L. Valentini, A. Mirri, M. Chiani, and E. Paolini, "Feedback-aided coded random access via replica spacing," in *Proc. ICC - IEEE Int. Conf. Commun.*, Rome, Italy, Jun. 2023, pp. 4786–4791.
[21] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.
[22] J. H. Sørensen, E. de Carvalho, C. Stefanovic, and P. Popovski, "Coded pilot random access for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8035–8046, Dec. 2018.
[23] L. Valentini, M. Chiani, and E. Paolini, "Interference cancellation algorithms for grant-free multiple access with massive MIMO," *IEEE Trans. Commun.*, vol. 71, no. 8, pp. 4665–4677, Aug. 2023.
[24] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, Jun. 2018.
[25] E. Casini, R. De Gaudenzi, and O. R. 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.
[26] 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.
[27] Z. Sun, Y. Xie, J. Yuan, and T. Yang, "Coded slotted ALOHA for erasure channels: Design and throughput analysis," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4817–4830, Nov. 2017.
[28] 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.
[29] J. Kang and W. Yu, "Minimum feedback for collision-free scheduling in massive random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 12, pp. 8094–8108, Dec. 2021.
[30] S. Scalise, C. P. 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.
[31] D. M. Bloom, "A birthday problem," *Amer. Math. Monthly*, vol. 80, no. 10, pp. 1141–1142, 1973.
[32] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[33] L. Valentini, A. Faedi, M. Chiani, and E. Paolini, "Impact of interference subtraction on grant-free multiple access with massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022, pp. 1318–1323.
[34] W. Feller, *An Introduction to Probability Theory and its Applications: Volume I*. Hoboken, NJ, USA: Wiley, 1968.

Lorenzo Valentini (Member, IEEE) received the B.S. degree (summa cum laude) in electronics engineering and the M.S. degree (summa cum laude) in electronics and telecommunications engineering from the University of Bologna, Italy, in 2017 and 2019, respectively, and the Ph.D. degree in electronics, telecommunications, and information technologies engineering, in 2024. Currently, he is a Research Fellow with the Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi." His research interests include communication theory, multiple access protocols, and quantum error correction.

Alessandro Mirri (Graduate Student Member, IEEE) received the B.S. degree (summa cum laude) in electronics engineering and the M.S. degree (summa cum laude) in electronics and telecommunications engineering from the University of Bologna, Italy, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree in electronics, telecommunications, and information technologies engineering with the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi." His research interests include communication theory and multiple access protocols.

Enrico Paolini (Senior Member, IEEE) received the Dr.Ing. degree (summa cum laude) in telecommunications engineering and the Ph.D. degree in electrical engineering from the University of Bologna, Italy, in 2003 and 2007, respectively. During the Ph.D. degree, he was a Visiting Research Scholar with the Department of Electrical Engineering, University of Hawai'i at Mānoa, Honolulu, HI, USA. He was a Visiting Scientist with the Institute of Communications and Navigation, German Aerospace Center, in 2012 and 2014, under DLR-DAAD fellowships. He is currently an Associate Professor with the Department of Electrical, Electronic, and Information Engineering, University of Bologna. His research interests include digital communication systems, error-correcting codes, massive multiple access protocols, and detection and tracking in radar systems. He served as the Co-Chair for the ICC 2014, ICC 2015, and ICC 2016 Workshop on Massive Uncoordinated Access Protocols (MASSAP), the VTC 2019-Fall Workshop on Small Data Networks, the 2018 IEEE European School of Information Theory (ESIT), and the 2020 IEEE Information Theory Workshop (ITW 2020). He served as the TPC Co-Chair for the IEEE GLOBECOM 2022 Communication Theory Symposium and the IEEE GLOBECOM 2019 Communication Theory Symposium. He is the Past Chair of the ITSoc Italy Section Chapter and the Vice-Chair of the IEEE ComSoc Radio Communications Committee. He was an Editor of IEEE COMMUNICATIONS LETTERS from 2012 to 2015 and IEEE TRANSACTIONS ON COMMUNICATIONS (in coding and information theory) from 2015 to 2020.

Open Access funding provided by 'Alma Mater Studiorum - Università di Bologna' within the CRUI CARE Agreement