Massive Grant-Free Access With Massive MIMO and Spatially Coupled Replicas

Lorenzo Valentini, *Graduate Student Member, IEEE*, Marco Chiani, *Fellow, IEEE*, and Enrico Paolini, *Senior Member, IEEE*

***Abstract*— Massive multiple access schemes, capable of serving a large number of uncoordinated devices while fulfilling reliability and latency constraints, are proposed. The schemes belong to the class of grant-free coded random access protocols and are tailored to massive multiple input multiple output (MIMO) base station processing. High reliability is obtained owing to an intra-frame spatial coupling effect, triggered by a simple device access protocol combined with acknowledgements (ACKs) from the base station. To provide system design guidelines, analytical bounds on error floor and latency are also derived. The proposed schemes are particularly interesting to address the challenges of massive machine-type communications in the framework of next generation massive multiple access systems.**

***Index Terms*— Grant-free multiple access, coded random access, energy efficiency, future internet of things, massive MIMO, massive multiple access, spatial coupling, successive interference cancellation, 6G.**

I. Introduction

CELLULAR system services are currently categorized by 3GPP into the three classes called enhanced mobile broad-band (eMBB), ultra-reliable and low-latency communication (URLLC), and massive machine-type communication (mMTC) [1], [2]. These three classes are most often regarded as separate, each one with its own performance metrics and requirements. For example, in mMTC services, emphasis is essentially on scalability (defined as the capability to serve as many machine-type devices as possible), with no particularly tightening reliability and latency constraints. In contrast, URLLC services emphasis is on reliability and latency, whereas scalability is usually not as issue.

New emerging use cases in the framework of the future Internet of Things (IoT), including some industrial IoT applications, vehicle-to-infrastructure communications, and smart city, are however calling into question this rigid scheme as they require convergence, for example, between mMTC and URLLC with different trade-off points between scalability, latency, and reliability [3], [4], [5], [6], [7]. Next generation massive multiple access (MMA) systems shall therefore be able to support mMTC services where scalability will still be the main performance metric, but also with relatively tightening (although non-URLLC) reliability and latency requirements.

To support new use cases, next generation MMA protocols should be designed to achieve very high scalability in presence of latency and reliability constraints [8], [9], [10], [11]. In this respect, grant-free multiple access schemes (as opposed to grant-based ones) have gained an increasing interest, as they can drastically reduce control signalling for connection establishment, which is beneficial in terms of both scalability and latency. Examples of grant-free schemes are the ones recently proposed in [12], [13], [14], [15], [16], and [17]. As a particular class of grant-free access schemes, uncoordinated protocols based on the coded random access (CRA) paradigm [18], [19], [20], [21], [22], [23], [24], also known as "modern random access", ensure high reliability and are currently regarded as a candidate for 6G owing to their ability to bridge random access with iterative decoding of codes on sparse graphs [25]. Additional features envisaged for next MMA schemes for the future IoT are the extreme simplicity and very high energy efficiency on the device side, as well as the capability of the protocol to exploit massive MIMO base stations or cell-free architectures [12], [13], [14], [16], [26], [27], [28], [29], [30], [31], [32], [33], [34].

Spatial coupling (e.g., [35], [36]) is a well-known tool in the framework of low-density parity-check (LDPC) codes, a class of error correcting codes usually described by means of their Tanner graph, a sparse bipartite graph having a one-to-one correspondence with the code parity-check matrix. In coding theory, spatial coupling (SC) consists of starting from a set of unconnected Tanner graphs and of connecting them, according to certain rules, to obtain a graph chain ("spatially-coupled graph") with very remarkable error correction properties under iterative belief-propagation decoding. Such properties mainly rely on effect of *terminations*, i.e., of the edges of the chain, in which the encoded bits tend to be more protected from errors. When belief-propagation decoding is run on the whole spatially-coupled graph, a "correction wave" is triggered that propagates from terminations to the center of the chain. Due to the connections between CRA and codes on sparse graphs, there have been attempts to exploit the SC effect within CRA. In [37], a scheme based on a spatially-coupled super-frame was introduced and its asymptotic properties over a collision

Manuscript received 14 April 2022; revised 24 August 2022; accepted 3 October 2022. Date of publication 11 October 2022; date of current version 18 November 2022. An earlier version of this paper was presented in part at the IEEE Global Communications Conference, Madrid, Spain, December 2021 [DOI: 10.1109/GCWkshps52748.2021.9682156]. The associate editor coordinating the review of this article and approving it for publication was F. Verde. *(Corresponding author: Marco Chiani.)*

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3213279.

Digital Object Identifier 10.1109/TCOMM.2022.3213279

channel were analyzed, revealing an asymptotic performance over the collision channel very close to the theoretical limits developed in [20]. In [38] a frame asynchronous coded random access scheme with SC was described in which users activate on a slot-by-slot basis, again with performance close to the limit established in [20].

In this paper, we propose medium access control (MAC) layer protocols to address the above-mentioned challenges of next-generation MMA. The proposed schemes belong to the class of synchronous CRA, meaning that they are slotted and framed with slot- and frame-synchronous users, and aim at exploiting the SC effect. To cope with latency requirements of next-generation MMA, instead of resorting on super-frames we aim at triggering a SC effect within the single frame, by exploiting both the dimensions offered by orthogonal pilots and a massive number of antennas assumed available at the base station (BS). Furthermore, as opposed to previous works on spatially-coupled CRA schemes, where the analysis was conducted on the simple collision channel model and assuming ideal interference cancellation, we adopt a realistic wireless channel model and a realistic signal processing at the BS. We show that, when physical (PHY) layer processing is also considered, interference cancellation imperfectness may jeopardize propagation of the SC effect through the frame. Then, assuming the availability of a feedback channel, we introduce ACK-based coded random access as a technique to make spatial coupling effective and, contextually, to substantially enhance the energy efficiency of the system. The key contributions of the paper can be summarized as follows:

- we propose novel coded random access MAC protocols which make use of spatial coupling;
- we investigate scalability and efficiency improvements given by a feedback channel;
- we analytically derive performance bounds for the new schemes, as well as for those previously proposed.

This paper is organized as follows. Section II introduces preliminary concepts and models together with some background material. Section III describes the proposed MMA protocol, combining intra-frame SC and ACK messages, and develops analytical lower bounds to packet loss rate performance and average latency. Numerical results are shown in Section IV. Finally, conclusions are drawn in Section V. A subset of the results presented in this paper appeared in the conference version [33]. With respect to [33], the proposed schemes are here addressed in a more thorough way, providing all of the details. Moreover, a packet loss rate error floor estimation via analytical lower bound, an analysis of the average latency, as well as several numerical results have been added.

Notation: Throughout the paper, capital and lowercase bold letters denote matrices and vectors, respectively. Symbols $(\cdot)^T$ and $(\cdot)^H$ are used to indicate transposition and conjugate transposition, respectively, while $\|\cdot\|$ denotes Euclidean norm and $|\cdot|$ cardinality. Moreover, $\mathbb{E}\{\cdot\}$ denotes expectation, $\mathbb{V}\{\cdot\}$ variance, and $\mathbb{P}\{\cdot\}$ the probability of an event.

II. System Model

This section describes a baseline uncoordinated access protocol based on a random selection of orthogonal pilots performed by the active users. This protocol exploits the coded slotted ALOHA (CSA) principle in the framework of massive MIMO. Then, the physical channel model and the BS processing are described.

A. Coded Random Access With Massive MIMO and Random Pilot Selection

We consider a cell-based, slotted and framed MMA scenario with a very large number K of transmitters and one receiver, as shown in Fig. 1. The K users, hereafter referred to as "devices" or "users" interchangeably, are not all simultaneously active, since they are assumed to wake up unpredictably to transmit one data packet. The number of users that are simultaneously active and contending for the channel over the same frame is denoted by K_{a} and is not a priori known to the receiver. All slots have the same time duration T_{s}, the number of slots in a frame is denoted by N, the frame duration is $T_{\mathrm{F}} = NT_{\mathrm{s}}$, and all active users are frame- and slot-synchronous. The start of each frame is signaled by a BS beacon, which is useful for both synchronization and power control purposes. Perfect power control is assumed in this paper.

The BS is equipped with a massive number M of antenna elements, a key feature to enable multi-packet reception at the receiver (meaning that multiple packets can be successfully decoded in a single slot) provided reliable packet detection and channel estimation can be performed. The use of orthogonal preambles (or "pilot sequences", or simply "pilots") to be prepended to the active users' payloads represents a simple and popular approach to packet detection and channel estimation, but poses concerns in grant-free MMA applications. In fact: (i) As opposed to conventional multiple access, in MMA the user population size K is typically much larger than the number of available pilots, hence pre-assigning a pilot to each user is not feasible; (ii) Active users can coordinate neither with the BS nor with each other for pilot selection. A simple solution consists of a random pilot selection, that is of defining a set of $N_{\mathrm{P}} \ll K$ orthogonal pilot sequences and of letting each active user pick one pilot randomly from this set, without any coordination with the other active users, in every slot in which the user performs a transmission. Whenever a specific pilot is picked by one and only one active user in a slot, the user channel can be estimated very reliably in that slot. Otherwise, a pilot collision occurs; when this happens, the payload transmitted by any of the active users employing the same pilot in the same slot cannot usually be successfully decoded. These "unresolved collisions" can however be possibly resolved performing interference cancellation across slots according to the CRA principle.

Coded random access is a class of powerful grant-free MMA protocols. A general instance is represented by CSA, in which an information packet generated by an active user is split into a number of packet fragments that are encoded by the user via a packet-level erasure correcting code, to generate encoded fragments [20]. Encoded fragments are further

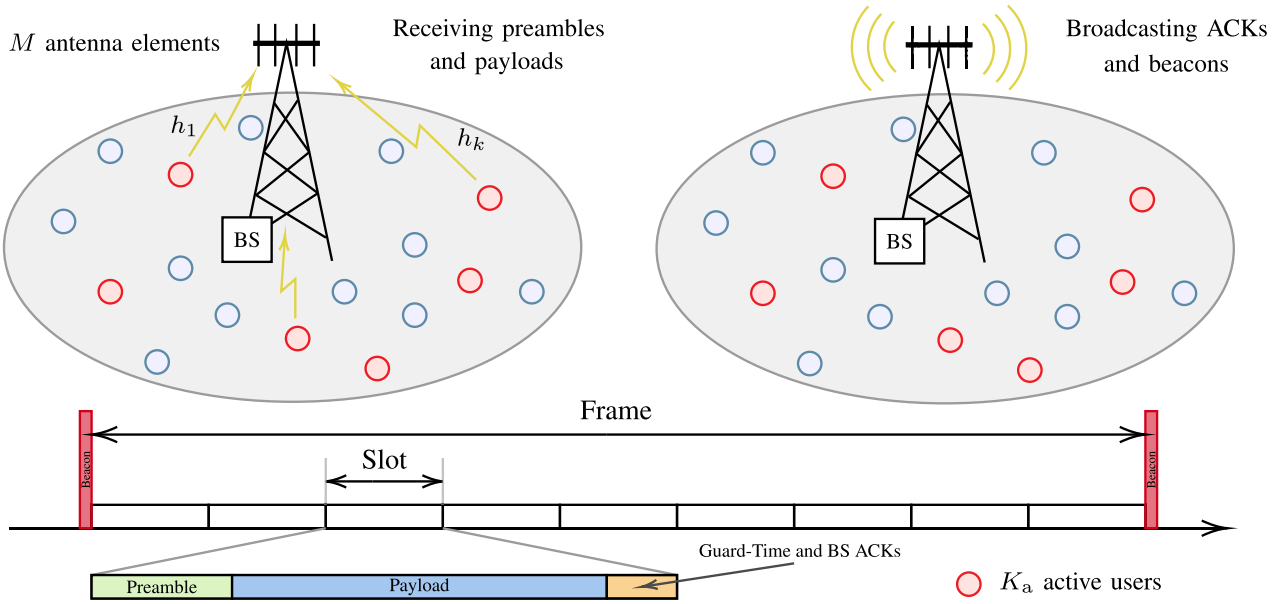

Fig. 1. Pictorial representation of the considered scenario. There are K_a simultaneously active users, out of K users (K very large), contending for grant-free uplink to a base station with M antennas. The time is framed and slotted, the users act in an uncoordinated fashion and may interfere with each other. The base station can send simple feedback messages such as beacons and acknowledgments.

individually encoded at PHY layer and transmitted over the frame in randomly-chosen sub-slots. User packets are recovered at the BS via a combination of successive interference cancellation (SIC) and packet-level erasure decoding. In this paper we use instances of CSA based on packet repetition, i.e., in which the above-mentioned packet erasure code is represented by a packet repetition code. Combined with random pilot selection, the baseline access protocol on device side from wake up to transmission may be summarized as follows. The active user chooses r different slots (i.e., r integers between 1 and N), uniformly at random without replacement, where r is the user repetition rate. For each such slot, the user picks uniformly at random one pilot sequence of length N_P symbols, out of the N_P available ones, and concatenates it with a payload of length N_D symbols, consisting of the information bits encoded by a PHY layer channel encoder and then mapped onto the complex symbols of the employed constellation. Next, the device waits for the start of the next frame, signaled by the BS beacon, and it finally transmits the r replicas in the r pre-selected slots. Note that, out of the r packet replicas transmitted by the active device, the payload remains the same, while the randomly-picked pilot may change from slot to slot. The data unit structure at PHY layer is illustrated in Fig. 1, where the presence of a guard time at the end of the slot to receive feedback messages from the BS, if foreseen, is also shown.

Concerning the choice of the repetition rate r, it may be deterministic and constant for all active users or randomly picked by each active user, any time it activates and without any coordination with the other users, by sampling an integer random variable with probability-generating function $\Lambda(x) = \sum_{r=2}^{r_{\max}} \Lambda_r x^r$. We refer to the case with a constant repetition rate as contention resolution diversity slotted ALOHA (CRDSA) [18], the first scheme proposing to combine constant packet repetitions (i.e., diversity ALOHA [39]) with SIC, and to the case with *irregular* repetition rates as irregular repetition slotted ALOHA (IRSA) [19], the first scheme where repetition rate irregularity was exploited.

Example 1: Fig. 2a provides a pictorial representation of a frame as an $N_P \times N$ grid in which each row corresponds to a pilot, each column to a slot, and each cell to a *resource*, i.e., a pilot-and-slot pair. In the specific example shown in Fig. 2a, we have $N_P = 4$ and $N = 9$; moreover, there are $K_a = 5$ active users, corresponding to the circles, all exploiting repetition rate $r = 3$. Each of them transmits r packet replicas into r resources, that must differ in the slot index, while they may or may not differ in the pilot index. Note that in the example, only two packet replicas do not experience a pilot collision, one in slot 2 and one in slot 7.

Next, we introduce the BS approach aimed at recovering the active users' messages. The simple description here provided is based on ideal PHY layer assumptions and only aims at conveying the general strategy, while the wireless channel model and the signal processing aspects at PHY layer are discussed in the remainder of the section. In the idealized setting: (i) The receiver can discriminate with no error between a resource where no packet replica has arrived (empty resource), a resource where a single packet replica has arrived (collision-free resource), and a resource where multiple packet replicas have arrived (collision resource); (ii) No information can be extracted from a collision resource, but any packet replica in a collision-free resource is decoded with zero error; (iii) When a packet replica is successfully decoded, information about the number and the positions (slot and pilot) of its twins can be extracted from it and the contribution of interference of the twins in the corresponding slots can be perfectly canceled by

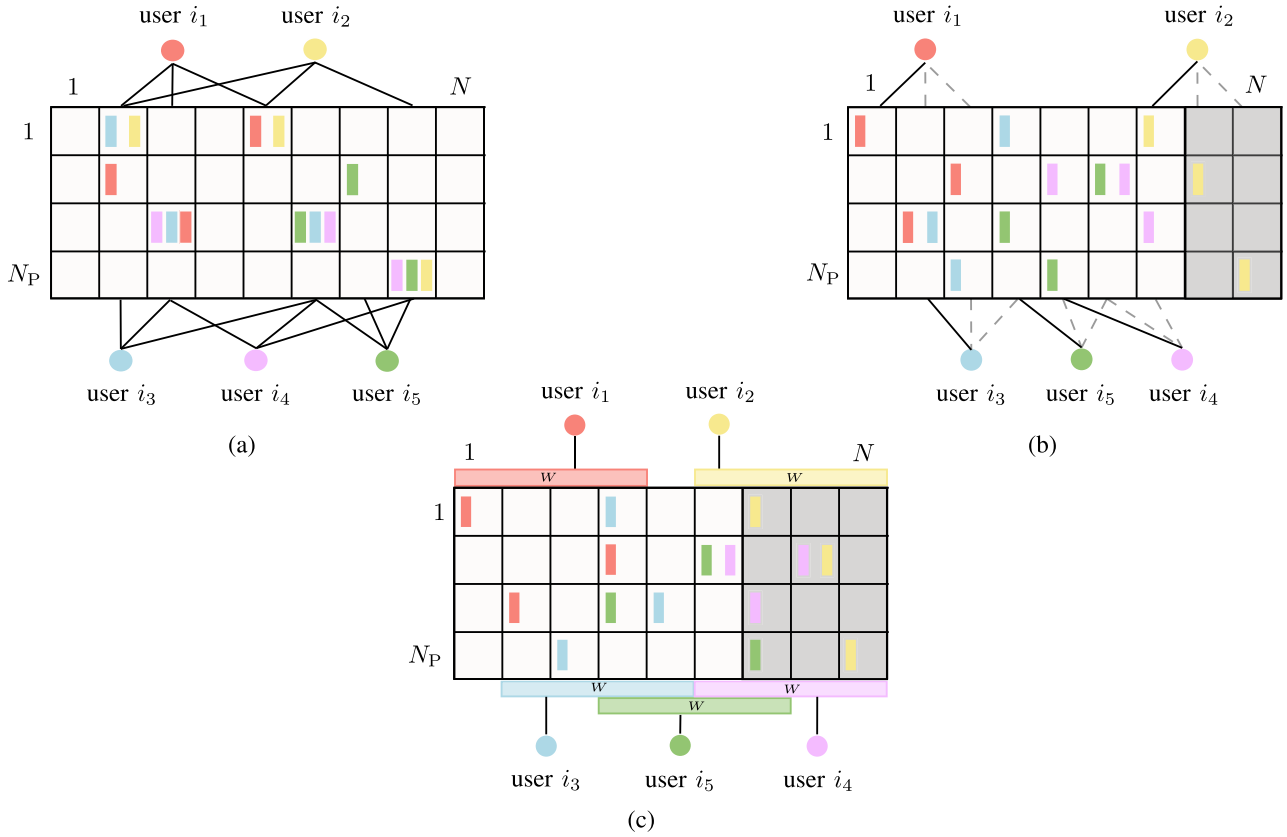

Fig. 2. Framed protocols with N slots, N_P orthogonal pilot sequences, and CSA with uniform repetition rate $r = 3$. (a) Standard CSA protocol. (b) CSA protocol with intra-frame SC. (c) CSA protocol with randomized intra-frame SC and window size $W = 4$.

subtraction. The BS performs an iterative message recovery strategy based on SIC, as follows. At every iteration, each packet replica in a collision-free resource is decoded, the slot index and the pilot index of its twins are extracted, and the interference generated by the twins is subtracted, possibly leading to new collision-free resources. The procedure is iterated until the frame is empty (decoding success) or only empty resources and collision resource are present (decoding failure).

Example 2: With reference again to Fig. 2a, the messages of active users i_1 and i_5 are decoded at the first iteration in slots 2 and 7, respectively. Interference subtraction allows cleaning the packet replica of user i_2 in slot 5; hence, the message of user i_2 is decoded at the second iteration. A further stage of interference subtraction allows decoding messages of users i_3 and i_4 in the third iteration.

B. Physical Channel Model

We consider a block Rayleigh fading channel model with additive white Gaussian noise (AWGN). The channel coherence time is denoted by T_c and is assumed equal to the slot duration T_s, which implies statistical independence of the channel coefficients of the same user across different slots. We do not consider large-scale propagation effects owing to the assumption of perfect power control.

Consistently with the above-described access protocol, each user active in a slot transmits a packet replica composed of one of the N_P orthogonal pilot sequences, of length N_P symbols, concatenated with a payload of length N_D symbols. Recalling that the number of BS antennas is denoted by M, the signal received in a slot by the BS may be expressed as $[\boldsymbol{P}, \boldsymbol{Y}] \in \mathbb{C}^{M\times(N_\mathrm{P}+N_\mathrm{D})}$ where

$$\boldsymbol{P} = \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{s}(k) + \boldsymbol{Z}_p \tag{1}$$

and

$$\boldsymbol{Y} = \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{x}(k) + \boldsymbol{Z}. \tag{2}$$

In (1) and (2), $\mathcal{A}$ is the set of users transmitting a packet in the considered slot, while $\boldsymbol{h}_k = (h_{k,1}, \ldots, h_{k,M})^T \in \mathbb{C}^{M\times 1}$ is the vector of channel coefficients of user $k \in \mathcal{A}$. The elements of $\boldsymbol{h}_k$ are modeled as zero-mean, circularly symmetric, complex Gaussian independent and identically distributed (i.i.d.) random variables, i.e., $h_{k,i} \sim \mathcal{CN}(0, \sigma_\mathrm{h}^2)$ for all $k \in \mathcal{A}$ and $i \in \{1, \ldots, M\}$. Moreover, $\boldsymbol{s}(k) \in \mathbb{C}^{1\times N_\mathrm{P}}$ in (1) and $\boldsymbol{x}(k) \in \mathbb{C}^{1\times N_\mathrm{D}}$ in (2) are the pilot sequence picked by user k in the current slot and the user payload, respectively. Finally, $\boldsymbol{Z}_p \in \mathbb{C}^{M\times N_\mathrm{P}}$ in (1) and $\boldsymbol{Z} \in \mathbb{C}^{M\times N_\mathrm{D}}$ in (2) are matrices whose elements are Gaussian noise samples. The elements of both $\boldsymbol{Z}_p$ and $\boldsymbol{Z}$ are i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_\mathrm{n}^2)$. Without loss of generality, we assume $\sigma_\mathrm{h}^2 = 1$ and unit energy symbols for both pilot and payload. Due to the

block fading channel assumption with coherence time $T_{\rm c} = T_{\rm s}$, the vector $\boldsymbol{h}_k$ is constant over a slot.

C. Physical Layer BS Processing

We now review a low-complexity signal processing strategy aimed at implementing, over the block fading wireless channel model, the message recovery algorithm that was previously introduced in an idealized setting. This approach, which exploits massive MIMO, is the one proposed in [13]. Here, the massive multiple antennas number plays two key roles: i) channel hardening, which guarantees stable SNRs despite of fading; ii) statistical orthogonality of propagation vectors (also called favorable propagation or multipath multiplexing), which enables multi-packet reception [40].

As mentioned above, the signal received in each slot of the frame may be expressed as $[\boldsymbol{P}, \boldsymbol{Y}]$. Soft samples received in all slots are stored by the BS. The algorithm performed by the BS can be split into two phases, of which the first phase is executed sequentially in all slots and the second phase starts at the end of the first one.

Phase 1: In each slot, the BS attempts channel estimation for all possible pilot sequences $\boldsymbol{s}_j \in \mathbb{C}^{1\times N_{\rm P}}$, $j = 1, \ldots, N_{\rm P}$, by computing $\boldsymbol{\phi}_j \in \mathbb{C}^{M\times 1}$ as

$$\boldsymbol{\phi}_j = \frac{\boldsymbol{P}\,\boldsymbol{s}_j^H}{\|\boldsymbol{s}_j\|^2} = \sum_{k\in\mathcal{A}^j} \boldsymbol{h}_k + \boldsymbol{z}_j \tag{3}$$

where $\mathcal{A}^j$ is the set of active devices employing pilot j in the current slot and $\boldsymbol{z}_j \in \mathbb{C}^{M\times 1}$ is a noise vector with i.i.d. entries characterized by distribution $\mathcal{CN}(0, \sigma_{\rm n}^2/N_{\rm P})$. Note that in absence of noise, when pilot j is picked by a single user in the slot, $\boldsymbol{\phi}_j$ equals the vector of channel coefficients for that user. Then, the BS computes the quantities $\boldsymbol{f}_j \in \mathbb{C}^{1\times N_{\rm D}}$ and $g_j \in \mathbb{R}^+$ as

$$\begin{aligned}\boldsymbol{f}_j &= \boldsymbol{\phi}_j^H\,\boldsymbol{Y} \\ &= \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \sum_{k\in\mathcal{A}^j}\sum_{m\in\mathcal{A}\setminus\{k\}} \boldsymbol{h}_k^H\,\boldsymbol{h}_m \boldsymbol{x}(m) + \tilde{\boldsymbol{z}}_j\end{aligned} \tag{4}$$

and

$$g_j = \|\boldsymbol{\phi}_j\|^2 = \sum_{k\in\mathcal{A}^j}\left(\|\boldsymbol{h}_k\|^2 + \sum_{m\in\mathcal{A}^j\setminus\{k\}} \boldsymbol{h}_k^H\,\boldsymbol{h}_m\right) + \tilde{n}_j \tag{5}$$

where $\tilde{\boldsymbol{z}}_j \in \mathbb{C}^{1\times N_{\rm D}}$ and $\tilde{n}_j$ are noise terms; this is conventional maximal ratio combining. Under the hypothesis that all cross-terms (i.e., terms involving a product $\boldsymbol{h}_k^H\,\boldsymbol{h}_m$ with $k \neq m$) can be neglected, (4) and (5) can be approximated as

$$\boldsymbol{f}_j \approx \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \tilde{\boldsymbol{z}}_j \tag{6}$$

$$g_j \approx \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 + \tilde{n}_j. \tag{7}$$

Therefore, under the same hypothesis, when a single user ℓ employs pilot j in the current slot, $\boldsymbol{f}_j/g_j$ can be taken as an estimate, $\hat{\boldsymbol{x}}(\ell)$, of the user's payload. Symbol demapping and channel decoding is then performed on the estimated payload $\boldsymbol{f}_j/g_j$: If channel decoding returns a valid codeword and a CRC test is passed, then a message decoding success is declared and the message is stored in a buffer.

Phase 2: All messages buffered during Phase 1 are used in the second phase of the algorithm, the SIC phase, triggered at the end of the frame. In Phase 2, the BS processes all buffered packets; for each such packet, the BS subtracts its contribution of interference from the samples of the slot where the packet has been detected and the contribution of interference of its replicas from the samples of the corresponding slots. Consider one packet in the buffer, let ℓ be the index of the corresponding user, $\boldsymbol{x}(\ell)$ be its payload, and $\|\boldsymbol{h}_\ell\|$ be the user channel in the slot where the user's packet replica has been successfully decoded. Next, consider a second slot where a packet replica from the same user has arrived, using pilot sequence $i \in \{1, \ldots, N_{\rm P}\}$. In this second slot, the BS updates $\boldsymbol{f}_i$ and g_i as

$$\boldsymbol{f}_i \leftarrow \boldsymbol{f}_i - \|\hat{\boldsymbol{h}}_\ell\|^2\,\boldsymbol{x}(\ell) \tag{8}$$

$$g_i \leftarrow g_i - \|\hat{\boldsymbol{h}}_\ell\|^2 \tag{9}$$

where $\hat{\boldsymbol{h}}_\ell$ is an estimate of the channel and then it attempts again demapping and decoding on $\boldsymbol{f}_i/g_i$. The processed packet is removed from the buffer, while any newly successfully decoded one is added to it. The process continues until the buffer is empty.

Discussion: Approximations (6) and (7) are justified in [13] and [33] and analyzed in [34] with the use of a massive MIMO BS. In particular, they heavily rely on the statistical orthogonality of the propagation vectors. To provide more insights on these approximations, here we note that we have $\mathbb{E}\{\|\boldsymbol{h}_k\|^2\} = M$ for any length-M vector of channel coefficients and $\mathbb{E}\{\boldsymbol{h}_k^H\,\boldsymbol{h}_m\} = 0$, $\mathbb{V}\{\boldsymbol{h}_k^H\,\boldsymbol{h}_m\} = M$, for any pair $(\boldsymbol{h}_k, \boldsymbol{h}_m)$, $m \neq k$. Thus, since each cross-term $\boldsymbol{h}_k^H\,\boldsymbol{h}_m$ has a null mean and a standard deviation equal to $\sqrt{M}$, while the expectation of each $\|\boldsymbol{h}_k\|^2$ term is M, it is reasonable to neglect cross-terms in (4) and (5) for large M but only if their number is not too large. Considering that the number of cross-terms in (4) is $|\mathcal{A}^j|(|\mathcal{A}| - 1)$, under independence a condition for validity of the approximations may be formulated as $[\,|\mathcal{A}^j|(|\mathcal{A}| - 1)M]^{1/2} \ll |\mathcal{A}^j|M$; for $|\mathcal{A}^j| = 1$, we obtain the simple criterion $|\mathcal{A}| \ll M + 1 \approx M$. When the number of arrivals in a slot $|\mathcal{A}|$ is not small compared to the number of antennas M, a replica may not be decodable even if it is the only one using a specific pilot; this was actually observed in simulations.

Also the updating rules (8) and (9) are justified in [13] as a consequence of large M, and in particular of the temporal "stability" of channel powers, i.e., $\|\boldsymbol{h}_k\|^2$ remaining stable across slots. We note however that over a block fading channel model (also assumed in [13]), the channel vectors of the same user in different slots are statistically independent, which poses some concern about (8) and (9), where the channel power estimated in one slot is used in another slot. For this reason, in this paper we use the modified updating rules

$$\boldsymbol{f}_i \leftarrow \boldsymbol{f}_i - \mathbb{E}\{\|\boldsymbol{h}_\ell\|^2\}\,\boldsymbol{x}(\ell) = \boldsymbol{f}_i - M\boldsymbol{x}(\ell) \tag{10}$$

$$g_i \leftarrow g_i - \mathbb{E}\{\|\boldsymbol{h}_\ell\|^2\} = g_i - M \tag{11}$$

which yield a slightly better performance according to simulations while keeping low complexity.

As an additional observation, the updating rules (8) and (9) (or (10) and (11)) require knowledge of the number of replicas, in the slots where replicas have been transmitted, and of the corresponding pilot indexes. This can be implemented in several ways, for example using a dedicated field in the packet or, as suggested by several authors, by letting the number of replicas, their positions, and their pilot indexes be functions of the random message.

III. Proposed MMA Protocols

In this section we present the proposed new protocols. More specifically, we describe coded random access with intra-frame SC, a "diluted" version of it, and another version, improving both transmit energy efficiency and performance, relying on simple feedback messages broadcast by the BS. We also provide a lower bound on the packet loss probability that turns very useful to estimate the error floor in the performance curve (packet loss probability versus the number of simultaneously active users per frame) of these protocols. The combination of intra-frame SC and ACKs improves considerably the performance and the energy efficiency, as illustrated in Section IV.

A. CSA With ACK Messages

We firstly introduce feedback-based CSA which exploits simple ACK messages broadcast by the BS at the end of each slot to improve scalability and efficiency. Such messages may, for example, simply carry the indexes of the pilots associated with packet replicas that have been successfully decoded in the current slot. To work properly, this ACK message format requires that a packet replica is never correctly decoded when the corresponding pilot has been chosen by multiple users, which is always verified under power control assumptions. Pilot-based ACK messages are an efficient feedback channel implementation compared, for example, to ID-based ACKs [41].[1] Availability of such acknowledgments is not an issue, even in the current 5G systems. As discussed in the following and as confirmed by numerical results, this very simple mechanism can substantially improve both system performance and its energy efficiency.

When a device performs transmission of a packet replica without running into a pilot collision and the replica is successfully decoded, the user becomes aware of successful transmission from the ACK message broadcast by the BS. The device can now immediately stop transmissions of the remaining replicas, which yields a two-fold benefit: (1) The device consumes less transmit energy; (2) The device does not generate unnecessary interference in subsequent slots.

Remarkably, aborting the transmission of the not yet sent replicas can be equivalently interpreted as an *ideal cancellation* of the interference these replicas would generate in the slots where they would be transmitted in absence of ACK messages. Although this provides no performance advantages over simple surrogate channels, such as the collision channel, where interference cancellation is always assumed as ideal, when modeling the system including a realistic wireless channel model, noise, and accurate physical layer processing, these ideal cancellations can boost the system performance. Clearly, when the BS broadcasts an ACK message with a list of pilot indexes for which successful decoding occurred in that slot, the BS de-activates interference cancellation of all corresponding packet replicas in future slots. Through the paper we assume a perfect feedback mechanism, i.e., no user receives false ACKs and users are always able to receive their ACK messages. In real systems we can have two types of feedback errors: i) a user cannot decode the ACK, flagging the error (detected error); ii) a user receives a false ACK accepting it as correct (undetected error). The perfect feedback assumption is justified by the fact that the feedback channel is noise-limited (as opposed to the uplink one that is impaired by interference), so that the detected error probability can be lowered by channel coding and with a sufficient BS transmit power. Moreover, the problem of false ACKs can be effectively tackled using error detection codes such as cyclic redundancy check (CRC) ones, making the undetected error probability even smaller than the detected one.

B. CSA With Intra-Frame Spatial Coupling

In conventional CSA with repetition codes and uniform repetition rate (i.e., CRDSA), each active user chooses r different slot indexes in the set $\{1, \ldots, N\}$, independently and without replacement. In contrast, in the proposed scheme, an active user only picks one slot index randomly in the set $\{1, \ldots, N-(r-1)\}$. Denoting by n the drawn index, the user then transmits its r packet replicas in slots n, $n+1$, $\ldots$, $n+r-1$. In each such slot, the packet payload is the same, while the pilot is still chosen randomly from a set of N_{P} orthogonal pilots. This is exemplified in Fig. 2b for a repetition rate $r = 3$.

Importantly, compared with conventional repetition-based CSA with repetition factor r, the introduced access strategy yields a lower degree of randomization in the choice of the slots by each user, which might jeopardize the iterative interference cancellation process. In fact, if multiple users pick the same index n, they necessarily transmit all replicas in the same r slots, increasing the probability that all transmissions from the same user experience a pilot collision. The proposed access protocol, however, also potentially brings substantial performance advantages in terms of packet loss probability due to its capability to trigger an effect similar to the SC one in the framework of LDPC coding. This effect exploits slots having a lower error probability to help message decoding in neighboring slots through SIC. In our scheme, this happens since the physical load in the first and in the last $r-1$ slots of the frame (the physical load in a slot being defined as the number of replicas arriving in it) is on average lower than the physical load in the other slots, increasing the probability of successful decoding in those slots. More specifically, given that there are K_{a} active devices, the average physical load in

[1] In a scenario with imperfect power control, using the proposed pilot-based ACK messages may lead to a performance degradation. If the power unbalance is not severe, the deterioration is negligible. Otherwise, ID-based ACKs should be considered.

slot n, denoted by $G_{\mathrm{phy}}(n|K_{\mathrm{a}})$, is given by

$$G_{\mathrm{phy}}(n|K_{\mathrm{a}}) = c_r(n)\,\frac{K_{\mathrm{a}}}{N-r+1} \tag{12}$$

where $c_r(n)$ is a discrete trapezoidal function

$$c_r(n) = \min(n,\, N-r+1) - \max(1,\, n-r+1) + 1. \tag{13}$$

In fact, the contribution to the physical traffic load in the first slot is $K_{\mathrm{a}}/(N-r+1)$ because only active users choosing that slot may transmit in it. Next, observing that only users selecting the first or the second slot may contribute to the load in the second slot, an so on for the successive slots, it is easy to derive the trapezoidal behavior of the average physical traffic load according to (13). Owing to the reduced load on the frame *terminations*, i.e., on the $r-1$ initial and the $r-1$ final slots of the frame, termination slots are characterized by a lower pilot collision probability; as a consequence, they exhibit a higher probability of successful packet decoding. Thus, applying the iterative PHY layer BS processing described in Section II-C, the reduced load on the frame terminations is potentially able to trigger an "interference cancellation wave" propagating from the edges of the frame towards the center of it, where previously decoded packets foster interference cancellation in adjacent slots in which new packets can be successfully decoded. The trigger and propagation of such a "wave" heavily relies on the additional dimension offered by the N_{P} orthogonal pilots and by the multi-packet reception capability enabled by the large number of BS antennas.

As we will show in Section IV, this mechanism can provide very remarkable enhancements to the system performance in terms of packet loss rate (PLR) versus the number of simultaneously active users K_{a} when ACK-based CSA is enabled. In fact, ACK-based CSA synergizes very well with the intra-frame SC strategies introduced in Section III-B and Section III-C due to "ideal" interference cancellation which favors propagation of the forward interference cancellation wave. Performance in terms of PLR can achieve values that are close to URLLC ones, provided a sufficient number of pilots is available and interference is effectively cancelled out across slots.

C. Randomized Intra-Frame Spatial Coupling

The lower degree of randomization of the CSA protocol with intra-frame SC, presented in Section III-B, can be mitigated introducing a window of W slots for each transmitting user as depicted in Fig. 2c. In this diluted variant of the scheme, called CSA with *randomized* intra-frame SC, each active device initially chooses one *offset slot* index n_{off} at random in the set $\{1, 2, \ldots, N-W+1\}$. Then, it randomly picks the r slots in which to perform transmissions of its r packet replicas in the set $\{n_{\mathrm{off}}, \ldots, n_{\mathrm{off}}+W-1\}$, uniformly and without replacement. Note that the user might not choose the slot n_{off} to transmit a replica.

The window size W can range from $W=r$ to $W=N$, where $W=r$ corresponds to the scheme of Section III-B and $W=N$ is equivalent to conventional CRDSA. Given that the number of active users is K_{a}, the average physical load in slot n, $G_{\mathrm{phy}}(n|K_{\mathrm{a}})$, may in this case be expressed as

$$G_{\mathrm{phy}}(n|K_{\mathrm{a}}) = c_W(n)\,\frac{r}{W}\,\frac{K_{\mathrm{a}}}{N-W+1}. \tag{14}$$

where

$$c_W(n) = \min(n,\, N-W+1) - \max(1,\, n-W+1) + 1. \tag{15}$$

The expression (14) is justified in a way that is analogous to (12). The r/W factor accounts for the probability that a user, which may transmit a replica in a slot, actually performs a replica transmission in it. As expected, (14) provides (12) for $W=r$ and is constant for all n when $W=N$. Despite W can assume values in the range $[r, N]$, in order to keep the advantages achieved using SC, W has to be chosen close to r value as it will be shown in Section IV. In the next section we will analyze the error floors of the discussed MAC access protocols, proving that randomized intra-frame SC is able to lower the error floor introduced by standard intra-frame SC.

D. Error Floor Analysis

While at high load (large K_{a}) the protocols may fail due to several causes such as imperfect SIC, inaccurate channel estimation, and unresolvable collisions, in the low load regime the few error events are caused essentially by the unresolvable interference between two users choosing exactly the same resources (slot and pilot pairs). In this section, we derive a lower bound on the packet loss probability that turns very tight for small K_{a}, i.e., in the error floor region. We focus our derivation on the case of randomized intra-frame SC protocol with uniform repetition rate r, which includes the other protocols for particular choices of W.

In the considered scheme, the r replicas are placed in different slots and, for each such slot, one pilot is chosen randomly out of N_{P} available ones. Hence, the r replicas correspond to a resource r-tuple. An error event which cannot be resolved, even in the idealized setting introduced at the end of Section II-A, occurs when at least two users choose the same resource r-tuple.

We start by counting the total number of admissible resource r-tuples (i.e., resource r-tuples a device may pick with nonzero probability) in case of window size W. This quantity, denoted by $\mathcal{A}_W$ in the following, is equal to the total number of admissible slot r-tuples multiplied by N_{P}^r. For any offset slot n there are $\binom{W}{r}$ admissible slot r-tuples in the corresponding window. Some of them, however, are included also in windows starting at different offset slots. (Example: take $r=3$ and $W=4$; the slots triplet $(2,3,4)$ is included in the windows starting at $n=1$ and $n=2$.) To correctly enumerate the admissible slot r-tuples, we slide the window of size W from the first offset slot $(n=1)$ to the last one $(n=N-W+1)$, counting each time only the r-tuples that cannot be included in the subsequent window positions. For all windows starting at some offset slot $n < N-W+1$, these r-tuples are the ones that include slot n; their number is $\binom{W-1}{r-1}$. Only for the

last window position, starting at offset slot $n = N-W+1$, we need to count all $\binom{W}{r}$ r-tuples. Hence, we have

$$\mathcal{A}_W = \left[\binom{W-1}{r-1}(N-W) + \binom{W}{r}\right] N_{\mathrm{P}}^r \tag{16}$$

which also yields

$$\mathcal{A}_r = (N-r+1)\, N_{\mathrm{P}}^r \tag{17}$$

and

$$\mathcal{A}_N = \binom{N}{r} N_{\mathrm{P}}^r \tag{18}$$

in the two special cases of SC protocol without randomization ($W = r$) and conventional CRDSA ($W = N$), respectively.

Let P_{u} be the probability that at least two active users, out of the K_{a} ones, choose the same resource r-tuple, under the assumption of a uniform probability distribution on the $\mathcal{A}_W$ resource r-tuples. It is immediate to see that at least a collision occurs with probability

$$P_{\mathrm{u}} = 1 - \prod_{i=0}^{K_{\mathrm{a}}-1} \frac{\mathcal{A}_W - i}{\mathcal{A}_W} = 1 - \frac{(\mathcal{A}_W - K_{\mathrm{a}} + 1)_{K_{\mathrm{a}}}}{A_W^{K_{\mathrm{a}}}} \tag{19}$$

where the compact expression (19) uses the Pochhammer symbol. Next, let P_{c} be the probability that at least two active users choose the same resource r-tuple, under the actual probability distribution on the $\mathcal{A}_W$ resource r-tuples, induced by the protocol. The probability P_{c} is lower bounded by P_{u}, i.e., we have

$$P_{\mathrm{c}} \geq P_{\mathrm{u}}. \tag{20}$$

The lower bound (20) follows immediately from the "birthday problem" where the worst case is given by a uniform probability distribution of the birth dates, as proved in [42]. Notably, (20) holds with equality in the two special cases $W = r$ (SC without randomization) and $W = N$ (conventional CRDSA).

The event that at least two users choose the same resource r-tuple is not the sole error event in the protocol. Therefore, the frame error probability P_{e}, defined as the probability that not all messages transmitted over a frame are correctly received, is lower bounded by P_{c}, hence by P_{u} through (20). Moreover, the packet loss probability given that there are K_{a} users simultaneously active over a frame can be lower bounded as

$$P_{\mathrm{L}} = \sum_{k=0}^{K_{\mathrm{a}}} \mathbb{P}\{k \text{ collisions}\}\, \frac{k}{K_{\mathrm{a}}} \geq P_{\mathrm{c}}\, \frac{2}{K_{\mathrm{a}}} \geq P_{\mathrm{u}}\, \frac{2}{K_{\mathrm{a}}}. \tag{21}$$

It is possible to prove that the derived lower bound on the packet loss probability is monotonically decreasing with W. This is shown in Appendix V. However, increasing W could attenuate the benefits of the SC for large number of users as will be shown in the Section IV by numerical analysis. Furthermore, in Section IV it will be illustrated that the bound (21) is tight even when we consider the actual physical layer processing over the wireless channel, provided that the number of active users K_{a} is not too large.

Discussion: The error floor analysis provides useful system design guidelines. For example, using this tool it is possible to: i) discard schemes having an error floor that is above the target PLR without running extensive Monte Carlo simulations; ii) tune the randomized spatial coupling window W to set the floor below a specific target; iii) predict extremely low error floors in conventional CRA where simulations would take very long running times.

E. Latency Analysis

The proposed schemes have a minimum latency equal to one time slot, T_{s}, for the users which wake up just at the beginning of a frame and transmit successfully on the first slot. On the other hand, they have a maximum latency equal to two time frame, $2\,T_{\mathrm{F}}$, for the users which wake up right after the beginning of a frame (so they have to wait for the next one) and moreover are resolved only in the SIC phase.

We can also derive a lower bound for the average latency of the different MAC protocols. Similarly to Section III-D, we derive this bound for the randomized intra-frame SC. The other protocols can be seen as particular cases. The results derived in this section can be very useful, for example, in the design of the upper protocol layers, where the PHY and MAC layers are often regarded as "black boxes" and modeled through simple formulas. In that case, the derived lower bound can be used as a best case or as an approximation of the average delay.

In the randomized intra-frame SC, each user chooses the offset slot n_{off} uniformly at random in the set $\{1, 2, \ldots, N-W+1\}$. Thus, the probability distribution of the offset slot n_{off} is

$$P_{\mathrm{off}}(n_{\mathrm{off}}) = \begin{cases} \dfrac{1}{N-W+1} & 0 < n_{\mathrm{off}} \leq N-W+1 \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Then, each user picks r transmission slots in the set $\{n_{\mathrm{off}}, n_{\mathrm{off}}+1, \ldots, n_{\mathrm{off}}+W-1\}$. At this point, the probability that the first transmission occurs in slot n, given the window size W, can be derived as

$$P(n; W) = \sum_{n_{\mathrm{off}}=1}^{N} P(n \,|\, n_{\mathrm{off}})\, P_{\mathrm{off}}(n_{\mathrm{off}}), \quad n = 1, 2, \ldots, N \tag{23}$$

where $P(n \,|\, n_{\mathrm{off}})$ is the probability that the first transmission is in slot n given an offset slot n_{off}. This probability can be obtained by a bins and balls problem formulation. Specifically, let us consider this problem: given W bins and r balls which have to be placed inside the bins without repetitions, find the probability that the first occupied bin is the n-th. Defining the event $\mathcal{E}_1$ as "one ball in bin n" and $\mathcal{E}_2$ as "the other $r-1$ balls are located in bins after the n-th", we can derive the above mentioned probability as

$$\begin{aligned} p(n; r, W) &= \mathbb{P}\{\mathcal{E}_1, \mathcal{E}_2\} = \mathbb{P}\{\mathcal{E}_2 | \mathcal{E}_1\}\, \mathbb{P}\{\mathcal{E}_1\} \\ &= \frac{r}{W} \prod_{i=0}^{r-2} \frac{W-n-i}{W-1-i}. \end{aligned} \tag{24}$$

Considering also that $P(n\,|\,n_{\rm off})$ is zero outside the interval $n_{\rm off} \le n \le n_{\rm off} + W - r$, we have

$$P(n\,|\,n_{\rm off}) = \begin{cases} p(n - n_{\rm off} + 1;\, r, W) & \text{if } n_{\rm off} \le n \le n_{\rm off} + W - r \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

Then, through the change of index $w = n - n_{\rm off}$ in (23), we can write

$$\begin{aligned} P(n; W) &= \sum_{w=n-1}^{n-N} P(n\,|\,n-w)\, P_{\rm off}(n-w) \\ &= \frac{1}{N - W + 1} \sum_{w=w_0}^{w_1} p(w+1;\, r, W) \end{aligned} \tag{26}$$

where $w_0 = \max(0,\, n - N + W - 1)$ and $w_1 = \min(n - 1,\, W - r)$ are introduced to keep only the non-zero terms in (22) and (25). In particular, we can study the baseline MAC protocol without SC by rewriting (26) with $W = N$, obtaining

$$P(n; N) = p(n;\, r, N) = \frac{r}{N} \prod_{i=0}^{r-2} \frac{N - n - i}{N - 1 - i} \tag{27}$$

while, for intra-frame SC without randomization ($W = r$), we have

$$P(n; r) = \begin{cases} \dfrac{1}{N - r + 1} & 0 < n \le N - r + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

The next step consists in using the probability derived in (26) to lower bound the average latency of the MAC protocols. Hereafter, we denote with $\Omega_{\rm avg}$ ($\Omega'_{\rm avg}$) the average delay between the user wake up (start of the frame) and the actual decoding instant. Due to the assumption of random arrivals it is

$$\Omega_{\rm avg}(W) = \frac{T_{\rm F}}{2} + \Omega'_{\rm avg}(W). \tag{29}$$

To study $\Omega'_{\rm avg}$, considering each user successfully decoded at its first transmission we write

$$\begin{aligned} \Omega'_{\rm avg}(W) &\ge T_{\rm s} \sum_{n=1}^{N} n\, P(n; W) \\ &= T_{\rm s}\, \frac{1}{N - W + 1} \sum_{n=1}^{N} n \sum_{w=w_0}^{w_1} p(w+1;\, r, W). \end{aligned} \tag{30}$$

From (27) and (28) we can then compute the average latency lower bounds for these particular cases as

$$\Omega'_{\rm avg}(N) \ge T_{\rm s} \sum_{n=1}^{N} n\, p(n;\, r, N), \quad \Omega'_{\rm avg}(r) \ge \frac{N - r + 2}{2}\, T_{\rm s}. \tag{31}$$

Finally, we get the bound on the overall average latency $\Omega_{\rm avg}(W)$ by substituting (30) in (29).

IV. Performance Analysis

A. Simulation Setup

In order to make the number of slots per frame compliant with a latency requirement Ω, we adopt the following procedure. We start from the information message size and the physical layer channel code rate, yielding the number of coded bits per message. Next, given the constellation we compute the number of data symbols per message, $N_{\rm D}$, and the total number of transmitted symbols $N_{\rm P} + N_{\rm D}$, where $N_{\rm P}$ is the pilot length. Then the time slot $T_{\rm s}$ shall fulfill

$$\frac{N_{\rm P} + N_{\rm D} + N_{\rm a}}{B_{\rm s}} \le T_{\rm s} < T_{\rm c} \tag{32}$$

where $B_{\rm s}$ is the symbol rate, $T_{\rm c}$ is the channel coherence time, and $N_{\rm a}$ is a parameter that accounts for extra delays such as ACK time and processing time. In (32), the lower bound captures the fact that the slot duration can in principle be larger than the packet transmission time, e.g., to accommodate guard times. Moreover, the upper bound is due the fact that, in case the slot time exceeds the coherence time, the channel estimates based on the pilot sequence cannot be applied to the whole payload. Finally, when a valid slot time $T_{\rm s}$ is chosen, the number of slots needed to be compliant with the latency requirement is given by

$$N = \left\lfloor \frac{T_{\rm F}}{T_{\rm s}} \right\rfloor = \left\lfloor \frac{\Omega}{2T_{\rm s}} \right\rfloor \tag{33}$$

where $T_{\rm F}$ is the frame duration. In (33) we use $\Omega = 2\,T_{\rm F}$ because, when a user wakes up, it needs to wait the beginning of the next frame to start transmissions. Hence in the worst case, a device waking up right after the beginning of a frame and picking a slot at the end of the next frame experiences a delay $\Omega = 2\,T_{\rm F}$.

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

$$N_{\rm ACK} = N_{\rm P,ACK} + \frac{(N_{\rm P} + N_{\rm CRC})}{\mathcal{R}_{\rm a} \log_2(\mathsf{M})} \tag{34}$$

where $N_{\rm P,ACK}$ is the ACK preamble length, $N_{\rm CRC}$ is the number of CRC bits protecting the ACK message, and $\mathcal{R}_{\rm a}$ is the code rate. For example, considering $N_{\rm P,ACK} = 4$, $N_{\rm CRC} = 16$, $\mathsf{M} = 256$, and $\mathcal{R}_{\rm a} = 2/3$, we end up with $N_{\rm ACK} = 19$ when $N_{\rm P} = 64$ and $N_{\rm ACK} = 31$ when $N_{\rm P} = 128$. A simple way to include ACK time is imposing $N_{\rm a} = N_{\rm ACK}$ in equation (33). In order not to stick to a particular ACK implementation, we consider $N_{\rm a} = 0$ in the following section. Nevertheless, when schemes using ACKs will be compared to scheme without ACKs some considerations will be drawn using (34).

B. Numerical Results

We present numerical results for the schemes described in Section III as well as the baseline scheme (conventional repetition-based CSA) reviewed in Section II. The results are obtained by Monte Carlo simulation for different choices of the pilot sequence length N_{P}, repetition rate r, and window size W. The metrics we consider are the packet loss rate (PLR) P_{L}, the average number of transmitted replicas per active user (related to the average transmission energy per active user), and the maximum supported number of active users per frame K_{a}^{*} at a target PLR P_{L}^{*}. All the results are computed with the same maximum latency $\Omega = 50$ ms, in accordance with the requirement.

In particular, we consider a system where users transmit messages of length 421 bits, encoded with a 10-error-correcting Bose–Chaudhuri–Hocquenghem (BCH) code of length 511 bits. We assume that messages contain a CRC field that allows validating them after decoding. This check is necessary to avoid error propagation in the SIC procedure. Each codeword is padded with a null bit and then modulated using a quadrature phase-shift keying (QPSK) constellation with Gray mapping, yielding $N_{\mathrm{D}} = 256$ symbols. Simulations have been carried out with $B_{\mathrm{s}} = 1$ Msps, BS with $M = 256$ antennas, and $\sigma_{\mathrm{n}}^2 = 0.1$. Since the system failure is mainly caused by interference, numerical results obtained throughout this section remain practically the same for all $\sigma_{\mathrm{n}}^2 \leq 1$. Then, the number of slots in a frame is computed according to (33) for given N_{P}. The orthogonal pilot sequences are constructed using Hadamard matrices. Regarding the arrival model, for each simulated point we have chosen to keep constant the number of active users K_{a} per frame, without loss of generality. In fact, it is always possible to retrieve the PLR of a particular arrival model from the PLR conditioned to the number of active users. For example, for a Poisson model with parameter λ, we can write

$$\mathbb{P}\{\text{error}\,|\,\lambda\} = \sum_{K_{\mathrm{a}}} \mathbb{P}\{\text{error}\,|\,K_{\mathrm{a}}\}\,\mathbb{P}\{K_{\mathrm{a}}\,|\,\lambda\} \tag{35}$$

where $\mathbb{P}\{K_{\mathrm{a}}\,|\,\lambda\} = e^{-\lambda}\lambda^{K_{\mathrm{a}}}/(K_{\mathrm{a}})!$ and $\mathbb{P}\{\text{error}\,|\,K_{\mathrm{a}}\}$ is the PLR we evaluate in this paper. This allows us not to stick to a particular model of packet arrivals.

Fig. 3 compares different MAC protocols in terms of PLR versus the number of active users per frame K_{a}. The considered protocols are framed ALOHA, diversity ALOHA, conventional CRDSA and the CSA with intra-frame SC introduced in Section III-B, with and without feedback, for different choices of the number of available pilots N_{P} and repetition rate r. From the plot we can observe, as expected, that framed ALOHA is not well-suited for access schemes having reliability constraints. In fact, transmitting a single packet within the frame, the PLR is limited by the probability that two users pick the same pair slot-pilot. An improvement is achieved by diversity ALOHA where the user transmits r replicas in the frame. The difference between this access scheme and CRDSA r, is the absence of a SIC phase. Diversity ALOHA is able to achieve a better performance compared to framed ALOHA due to the fact that the scheme is now

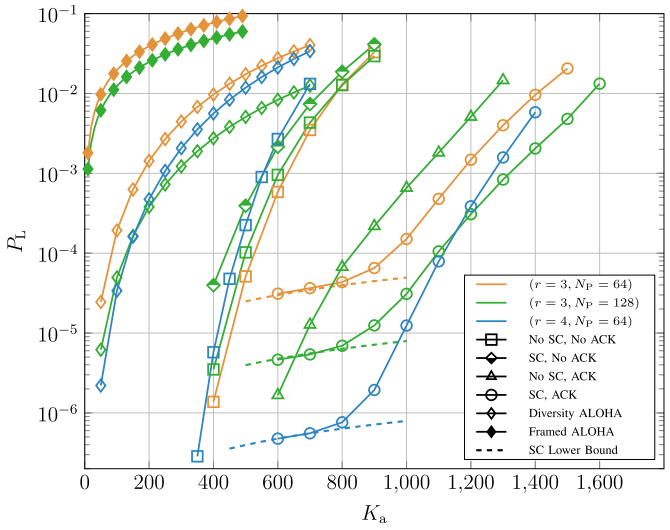

Fig. 3. Packet loss rate comparison between schemes characterized by different combinations of acknowledgments and intra-frame SC activation, for CSA with $r \in \{3, 4\}$ and $(N_{\mathrm{P}}, N) \in \{(64, 78), (128, 62)\}$. Framed ALOHA uses $r = 1$ despite of the curve color. Error floors derived in Section III-D are reported in dashed line.

limited by the probability that all r replicas are collided. However, a high quality of service is achievable only for a limited number of active users per frame. In order to improve scalability having constraints in latency and reliability, well-designed SIC algorithms are required.

Let us focus at first on the $N_{\mathrm{P}} = 128$ case. As we can see, use of SC without ACK messages tends to worsen performance with respect to conventional CRDSA. A closer inspection reveals that this effect is associated with failures in SIC physical layer processing due to an increased number of active devices choosing the same r slots, even with different pilots, which makes the cross terms in (4) and (5) not negligible. This observation is supported by the fact that, when ACK messages from the BS are enabled, the proposed SC scheme exhibits the most pronounced performance boost. In fact, intra-frame SC gives rise to a lower number of resource collisions in the first slots, which stops a higher number of replica transmissions in subsequent slots and reduces interference in them. For example, at $P_{\mathrm{L}} = 10^{-3}$, the baseline scheme supports $K_{\mathrm{a}} = 600$ active users per frame, which are pushed to more than $K_{\mathrm{a}} = 1300$ active users per frame combining of intra-frame SC and ACK messages. Accounting for the realistic ACK delay assumptions that have been discussed in Section IV-A, leading to $N = 60$ when $N_{\mathrm{ACK}} = 31$ ($N_{\mathrm{P}} = 128$), the performance of the intra-frame SC and ACK scheme degrades from $K_{\mathrm{a}} = 1300$ to approximately $K_{\mathrm{a}} = 1250$. Nevertheless, the improvement with respect to the baseline remains remarkable. Notably, the intra-frame SC protocol with ACK messages performs better in the $N_{\mathrm{P}} = 128$ case than is the $N_{\mathrm{P}} = 64$ one, despite the fact that the number of slots N decreases according to (33). In fact, with $N_{\mathrm{P}} = 128$, resource collisions in the first slots are less likely, making the ACK-based procedure more effective when the SC scheduling is adopted. On the contrary, increasing N_{P} in the baseline scheme worsens the system performance, as the cross-term

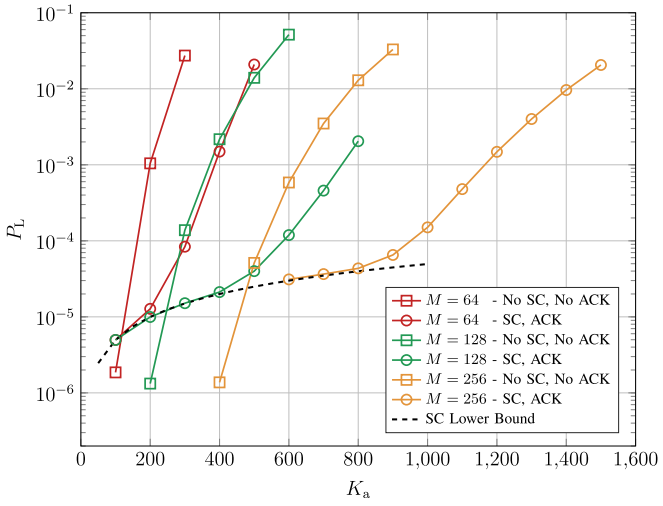

Fig. 4. Packet loss rate comparison between schemes characterized by different combinations of acknowledgments and intra-frame SC activation for some typical antenna size M, using CSA with $r = 3$, $N_P = 64$, $N = 78$.

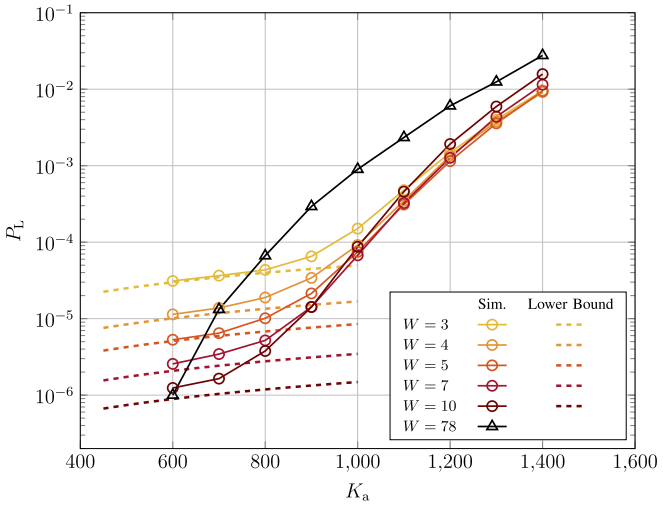

Fig. 5. Packet loss rate comparison between schemes adopting randomized intra-frame SC and ACK messages, for CSA with $r = 3$, $N_P = 64$, $N = 78$, and $W \in \{3, 4, 5, 7, 10, 78\}$. The baseline scheme is represented by the case $W = N = 78$, while the standard intra-frame SC by the case $W = r =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

interference increases due to the reduction in the number of slots and the non-ideality of the SIC processing.

In Fig. 3 we also report the PLR bound (21), applied to the SC schemes. As we can see, the bound is tight in the error floor region, which makes it useful for design purposes. Moreover, despite the fact that (21) has been derived without considering the wireless channel effects, it is remarkable that it well-fits the behavior of the schemes also under realistic physical layer processing. The different behavior of the PLR in the waterfall and floor regions highlights the different trade-offs achieved by the two schemes (with and without SC). However, for well-designed system parameters, the error floor of SC schemes is below the PLR targets in MMA scenarios (e.g., $P_L = 10^{-4}$, considered as a tightening requirement in MMA applications).

We then analyze the effect of the repetition rate, comparing schemes with $r = 3$ and $r = 4$. In this case we assume $N_P = 64$ and, as a consequence, $N = 78$. The behavior of the schemes for $r = 4$ is similar to the previously described one. Increasing the number of replicas has a negative effect on the performance of CRDSA, as expected from an asymptotic analysis recently carried out in [43]. However, when active users' transmissions are scheduled according to intra-frame SC, increasing r brings benefits in terms of PLR. Under the setting considered in the figure, for target PLR values between 10^{-6} and 10^{-4}, the best performance is achieved with $r = 4$, intra-frame SC, and ACK messages enabled.

In Fig. 4 we show simulations for typical array sizes $M \in \{64, 128, 256\}$ in both baseline and intra-frame SC with ACK assumptions. Other parameters such as $r = 3$, $N_P = 64$, and $N = 78$. As it can be observed from the plot, the improvement provided by intra-frame SC with ACK, with respect to the baseline, is constant varying M. In fact, fixing a target $P_L^* = 10^{-3}$ we observe that the number of supported user K_a^* is almost double in each configuration. In general, we observe how increasing M improves the performance of our CRA schemes. This is motivated by the signal processing adopted in this paper heavily relying on channel hardening.

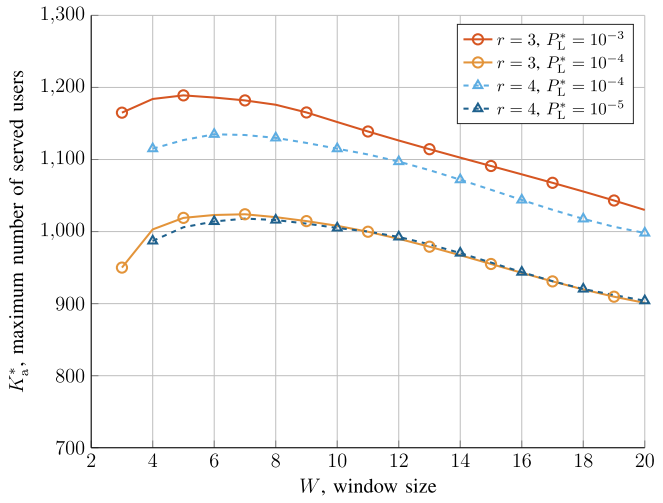

Fig. 6. Window size W analysis at $P_L^* \in \{10^{-3}, 10^{-4}, 10^{-5}\}$ for schemes adopting randomized intra-frame SC and ACK messages, using $N = 78$, $N_P = 64$, $M = 256$, $N_D = 256$, $r \in \{3, 4\}$. In the y-axis is reported the number of served user per frame given the target PLR.

Next, in Fig. 5 we focus on the randomized SC protocol of Section III-C, assuming ACK messages enabled, number of pilots $N_P = 64$, and repetition rate $r = 3$. Again, the total number of slots in the frame is $N = 78$. As expected from the discussion about the dependence of the lower bound (21) on W, the schemes using a larger window size W exhibit a lower error floor. In this regard, we observe that (21) remains tight also in the case where $r < W < N$ despite the fact that the equality in (20) does not hold. As another important observation, for small $W > r$ window size, the benefits in error floor region come at no loss in terms of waterfall performance, where a small improvement is even observed. Then, depending on the target PLR P_L^*, we can note that there exists an optimal value of W. This is shown explicitly in Fig. 6, where we plot the maximum number of served users

TABLE I

COMPARISON BETWEEN BASELINE, BASELINE WITH ACKS, PROPOSED INTRA-FRAME SC WITH ACKS, AND RANDOMIZED INTRA-FRAME SC WITH ACKS. TOP TABLE: FIXED PACKET LOSS RATE. BOTTOM TABLE: FIXED NUMBER OF SERVED USER PER FRAME

System Parameters	
Maximum latency, Ω [ms]	50
Orthogonal pilots, N_{P}	64
Slots per frame, N	78
Number of antennas, M	256
Information bits, k	421
Payload symbols, N_{D}	256
Packet repetition rate, r	3
Frame time, T_{F} [ms]	25
Slot time, T_{s} [μs]	320
SNR [dB]	0
CRC bits	32
Additional Symbols, N_{a}	0
QPSK modulation	-

Scenario 1, $\mathbf{P_L^* = 10^{-4}}$	No SC, No ACK	No SC, ACK	SC, ACK	SC, ACK, $W = 7$
Served users K_{a}	525	830	950	1025
Avg. Tx packets	3	1.36	1.25	1.4
Avg. Latency, Ω_{avg} [ms]	21.10	21.45	25.05	24.87
Lower Bound on Ω_{avg} [ms]	18.82	18.82	24.82	24.5
Sum Rate, γ [info bit/ch. use]	8.18	12.93	14.80	15.67

Scenario 2, $\mathbf{K_a^* = 800}$	No SC, No ACK	No SC, ACK	SC, ACK	SC, ACK, $W = 7$
PLR, P_{L}	$1.30 \cdot 10^{-2}$	$6.95 \cdot 10^{-5}$	$4.30 \cdot 10^{-5}$	$5.16 \cdot 10^{-6}$
Error Floor	$4.01 \cdot 10^{-8}$	$4.01 \cdot 10^{-8}$	$3.98 \cdot 10^{-5}$	$2.77 \cdot 10^{-6}$
Avg. Tx packets	3	1.34	1.23	1.23
Avg. Latency, Ω_{avg} [ms]	22.30	21.27	24.97	24.73
Lower Bound on Ω_{avg} [ms]	18.82	18.82	24.82	24.5

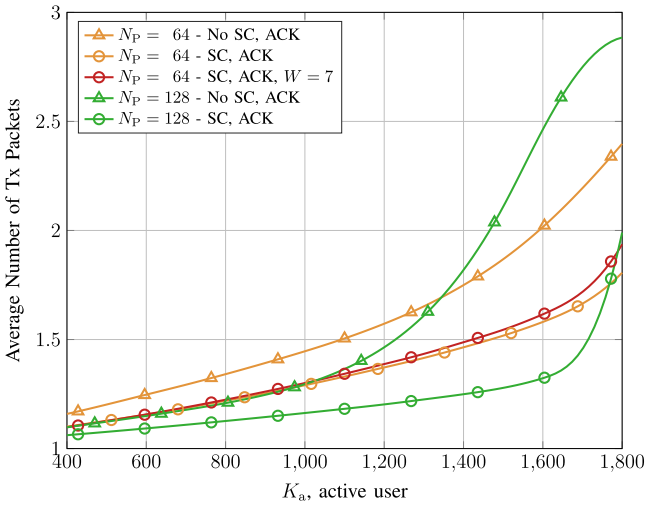

Fig. 7. Average number of transmitted packets by the users. The comparison is carried out between schemes characterized by different combinations of acknowledgments and intra-frame SC, for CSA with $r = 3$, $(N_{\mathrm{P}}, N) \in \{(64, 78), (128, 62)\}$. Randomized SC is reported using $W = 7$ in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

versus the window size W for given P_{L}^*, rate r, and number of available pilots N_{P}.

Another fundamental metric in MMA protocols is the energy efficiency. In Fig. 7 we plot the average number of transmitted packet replicas per active user (proportional to the average transmit energy per active user) versus the number of active users per frame, for the different access schemes. Clearly, this value equals r for all schemes not exploiting ACK messages from the BS. For example, with reference to Fig. 3 and Fig. 7, considering the intra-frame SC protocol with ACK messages enabled, $r = 3$, and $N_{\mathrm{P}} = 64$, each user transmits on the average less than 1.4 replicas per frame at $P_{\mathrm{L}}^* = 10^{-3}$. We see from the figure that exploiting ACK messages provides substantial savings, with an average number of transmitted packet replicas below 1.5, over a large range of K_{a}. Concerning randomized SC schemes, for small window sizes W (which are the values of interest, as pointed out in Fig. 6), the average number of transmitted packet replicas does not increase significantly. It can be verified that, when W is significantly larger than r, the energy efficiency worsens compared to the $W = r$ case.

Finally, we compare in Table I the different protocols in two scenarios, the first one characterized by a reliability constraint and the second one by a fixed number of served active users. In the first setting we consider $P_{\mathrm{L}}^* = 10^{-4}$ as reliability constraint and we evaluate the maximum number of served user per frame, the average number of transmitted packets per user, and the average latency for each access scheme. In the second scenario, we constrain the number of served user per frame to $K_{\mathrm{a}}^* = 800$ and we evaluate the performance (PLR, average number of transmitted packets, latency), again for the different schemes. In particular, in Table I we analyze four schemes: i) the baseline exploiting neither SC nor ACKs; ii) the baseline with ACK messages enabled; iii) the intra-frame SC scheme with ACK messages enabled; iv) the randomized intra-frame SC scheme with $W = 7$ and ACK messages enabled. From Table I it is possible to appreciate the improvements provided by the proposed schemes. As an example, for a target $P_{\mathrm{L}}^* = 10^{-4}$ the number of served users is almost doubled from the baseline ($K_{\mathrm{a}} = 525$) to the randomized SC scheme ($K_{\mathrm{a}} = 1025$). In scenario 2, where the number of served users is $K_{\mathrm{a}}^* = 800$, we can see that the PLR decreases from approximately 10^{-2} of the baseline, to $5 \cdot 10^{-6}$ offered by randomized SC.

In addition, we can define the achievable sum rate γ, in information bits per channel use, at a given target P_{L}^* as

$$\gamma = (1 - P_{\mathrm{L}}^*)\, K_{\mathrm{a}}^* \, \frac{N_{\mathrm{D}} \log_2(\mathsf{M})\, \mathcal{R} - N_{\mathrm{extra}}}{N\,(N_{\mathrm{P}} + N_{\mathrm{D}})} \tag{36}$$

where N_{extra} accounts for payload bits which are not used for information data as CRC and zero padding bits, $\mathcal{R}$ is the code rate, and M is the modulation order. In the table we report sum rates using $N_{\mathrm{extra}} = 33$, $\mathcal{R} = 421/511$, and $\mathsf{M} = 4$. Besides scheme comparisons, the sum rate in information bits per second, $\gamma_b = \gamma B_{\mathrm{s}}$, can be useful for backhaul network design. In Table I we also show the error floor on the PLR given by (21) at a given K_{a}^*, and the average latency derived in Section III-E. About the latter, we observe from the simulation that the adoption of SC slightly increases

the average latency, which however is close to the lower bound, and well-approximated by T_{F}. Overall, considering the fact that a maximum latency Ω is guaranteed for all schemes under examination, the adoption of SC substantially improves the performance in terms of number of served users, PLR, and energy.

V. Conclusion

Grant-free massive multiple access, where a large number of uncoordinated users interfere which each other, is possible with guaranteed quality when massive MIMO is available. In this paper we have shown new schemes able to ameliorate the scalability of massive MIMO random access protocols, when constraints in latency and packet loss rate are given. For example, choosing a latency of 50 ms and a target packet loss rate $P_{\mathrm{L}} = 10^{-3}$, the proposed scheme adopting intra-frame spatial coupling and ACKs is able to double the number of served users with respect to the baseline. Furthermore, we have shown that these techniques also considerably reduce the energy spent by each user.

Appendix
Error Floor Lower Bound Monotonicity

In this appendix we prove that the lower bound (21) on the packet loss probability is monotonically decreasing with W. To this aim, let us first prove that $\mathcal{A}_W$ in (16) is monotonically increasing with W. Considering the difference between $\mathcal{A}_{W+1}$ and $\mathcal{A}_W$ for $1 < r \leq W < N$ and normalizing with respect to N_{P}^r, we obtain

$$\begin{aligned}\frac{\mathcal{A}_{W+1} - \mathcal{A}_W}{N_{\mathrm{P}}^r} &= \left[\binom{W}{r-1} - \binom{W-1}{r-1}\right](N-W) \\ &= \binom{W-1}{r-2}(N-W) > 0. \end{aligned} \tag{37}$$

This proves that $\mathcal{A}_W$ is monotonically increasing with W. To show that P_{u} in (19) is monotonically decreasing with W we observe that

$$\begin{aligned} &P_{\mathrm{u}}(W+1) - P_{\mathrm{u}}(W) \\ &= \left[1 - \prod_{i=0}^{K_{\mathrm{a}}-1}\left(1 - \frac{i}{\mathcal{A}_{W+1}}\right)\right] - \left[1 - \prod_{i=0}^{K_{\mathrm{a}}-1}\left(1 - \frac{i}{\mathcal{A}_{W}}\right)\right] \\ &= \prod_{i=0}^{K_{\mathrm{a}}-1}\left(1 - \frac{i}{\mathcal{A}_{W}}\right) - \prod_{i=0}^{K_{\mathrm{a}}-1}\left(1 - \frac{i}{\mathcal{A}_{W+1}}\right) \\ &= \prod_{i=0}^{K_{\mathrm{a}}-1} a_i - \prod_{i=0}^{K_{\mathrm{a}}-1} b_i. \end{aligned} \tag{38}$$

From (38), $P_{\mathrm{u}}(W+1) - P_{\mathrm{u}}(W)$ can be expressed as the difference of two quantities, each given by the product of K_{a} numbers denoted as $a_1, a_2, \ldots, a_{K_{\mathrm{a}}}$ and $b_1, b_2, \ldots, b_{K_{\mathrm{a}}}$, respectively. Since for all i we have $a_i < b_i$ (because $\mathcal{A}_{W+1} > \mathcal{A}_W$) we can conclude that $P_{\mathrm{u}}(W+1) < P_{\mathrm{u}}(W)$.

Acknowledgment

This work has been carried out in the framework of the CNIT National Laboratory WiLaboratory and the WiLaboratory-Huawei Joint Innovation Center. The authors wish to thank Alberto Faedi for his work on software simulator implementation.

References

[1] S. Henry, A. Alsohaily, and E. S. Sousa, "5G is real: Evaluating the compliance of the 3GPP 5G new radio system with the ITU IMT-2020 requirements," *IEEE Access*, vol. 8, pp. 42828–42840, 2020.

[2] C. Bockelmann et al., "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[3] T. Jacobsen et al., "System level analysis of uplink grant-free transmission for URLLC," in *Proc. IEEE Global Commun. Conf. Workshops*. Singapore, Dec. 2017, pp. 1–6.

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

[5] C. Kalalas and J. Alonso-Zarate, "Massive connectivity in 5G and beyond: Technical enablers for the energy and automotive verticals," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.

[6] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796–131813, 2020.

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

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

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

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

[11] 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.

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

[13] J. H. Sorensen, 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.

[14] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-free massive random access with a massive MIMO receiver," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 23–30.

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

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

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

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

[19] 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.

[20] 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.

[21] E. Paolini, C. Stefanovic, 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.

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

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

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

[25] N. H. Mahmood, H. Alves, O. A. Lopez, M. Shehab, D. P. M. Osorio, and M. Latva-Aho, "Six key features of machine type communication in 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.
[26] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
[27] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[28] L. Sanguinetti, A. A. D'Amico, M. Morelli, and M. Debbah, "Random access in massive MIMO by exploiting timing offsets and excess antennas," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6081–6095, Dec. 2018.
[29] 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.
[30] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.
[31] V. Shyianov, F. Bellili, A. Mezghani, and E. Hossain, "Massive unsourced random access based on uncoupled compressive sensing: Another blessing of massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 820–834, Mar. 2021.
[32] U. K. Ganesan, E. Bjornson, and E. G. Larsson, "An algorithm for grant-free random access in cell-free massive MIMO," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.
[33] L. Valentini, A. Faedi, M. Chiani, and E. Paolini, "Coded random access for 6G: Intra-frame spatial coupling with ACKs," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2021, pp. 1–6.
[34] 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.*, May 2022, pp. 1–6.
[35] 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.
[36] D. G. M. Mitchell, M. Lentmaier, and D. J. Costello, "Spatially coupled LDPC codes constructed from protographs," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4866–4889, Sep. 2015.
[37] G. Liva, E. Paolini, M. Lentmaier, and M. Chiani, "Spatially-coupled random access on graphs," in *Proc. IEEE Int. Symp. Inf. Theory*. Cambridge, MA, USA, Jul. 2012, pp. 478–482.
[38] E. Sandgren, A. G. I. Amat, and F. Brännström, "On frame asynchronous coded slotted ALOHA: Asymptotic, finite length, and delay analysis," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 691–704, Feb. 2017.
[39] G. Choudhury and S. Rappaport, "Diversity ALOHA-A random access scheme for satellite communications," *IEEE Trans. Commun.*, vol. COM-31, no. 3, pp. 450–457, Mar. 1983.
[40] 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, Nov. 2017.
[41] 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.
[42] D. M. Bloom, "A birthday problem," *Amer. Math. Monthly*, vol. 80, no. 10, pp. 1141–1142, 1973.
[43] 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.* Rio di Janeiro, Brazil, Dec. 2022, pp. 1–6.

Lorenzo Valentini (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 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in electronics, telecommunications and information technologies engineering. From 2019 to 2020, he has been with the Interdepartmental Centre for Industrial ICT Research, University of Bologna, working on Internet of Things. His research interests include communication theory, wireless sensor networks, massive multiple access protocols, and quantum information.

Marco Chiani (Fellow, IEEE) received the Dr.Ing. degree (summa cum laude) in electronic engineering and the Ph.D. degree in electronic and computer engineering from the University of Bologna, Italy, in 1989 and 1993, respectively. Since 2003, he has been a Frequent Visitor at the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, where he currently holds a Research Affiliate appointment. He is also a Full Professor of telecommunications at the University of Bologna. His research interests include information theory, wireless systems, statistical signal processing, and quantum information. He received the 2011 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems, the 2012 IEEE Communications Society Fred W. Ellersick Prize, and the 2012 IEEE Communications Society Stephen O. Rice Prize in the Field of Communications Theory. He served as an Elected Chair (2002–2004) for the Radio Communications Committee of the IEEE Communication Society and as an Editor (2000–2007) of Wireless Communication for the IEEE TRANSACTIONS ON COMMUNICATIONS.

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. While working towards the Ph.D. degree, he was a Visiting Research Scholar with the Department of Electrical Engineering, University of Hawai'i at Manoa, 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 Chair of the ITSoc Italy Section Chapter and the Secretary of the IEEE ComSoc Radio Communications Committee. He was an Editor of IEEE COMMUNICATIONS LETTERS from 2012 to 2015 and the 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