Received 11 September 2023; revised 31 October 2023; accepted 3 November 2023. Date of publication 7 November 2023; date of current version 29 November 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3330632

Semi-Grant-Free Orthogonal Multiple Access With Partial-Information for Short Packet Transmissions

ALBERTO RECH [1,2] (Member, IEEE), STEFANO TOMASIN [1,3] (Senior Member, IEEE), LORENZO VANGELISTA [1] (Senior Member, IEEE), AND CRISTINA COSTA [4] (Senior Member, IEEE)

[1]Department of Information Engineering, University of Padova, 35131 Padua, Italy
[2]Smart Networks and Services, Fondazione Bruno Kessler, 38100 Trento, Italy
[3]Department of Mathematics, University of Padova, 35131 Padua, Italy
[4]S2N National Lab, Consorzio Nazionale Interuniversitario per le Telecomunicazioni, 16121 Genoa, Italy

CORRESPONDING AUTHOR: A. RECH (e-mail: alberto.rech.2@phd.unipd.it)

This work was supported in part by the European Union through the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (Program "RESTART") under Grant PE0000001.
Part of this paper has been presented at the International Conference on Ubiquitous and Future Networks (ICUFN) [1] [DOI: 10.1109/ICUFN57995.2023.10200276].

ABSTRACT Traditional multiple access schemes, as well as more recent preamble-based schemes, cannot achieve the extremely low latency, complexity, and collision probability required by the next-generation Internet-of-Things (IoT) networks to operate. To address such issues and further reduce the latency and packet loss, we introduce a novel semi-grant-free multiple access protocol for short packet transmission, the partial-information multiple access (PIMA) scheme. PIMA transmissions are organized in frames, and in the partial information acquisition (PIA) sub-frame of each frame, the base station (BS) estimates the number of active devices, i.e., the devices having packets waiting for transmission in their queue. Based on this estimate, the BS chooses both the total number of slots to be allocated in the data transmission (DT) sub-frame and the respective user-to-slot assignment. Although collisions may still occur due to multiple users assigned to the same slot, they are drastically reduced with respect to slotted ALOHA-based schemes, while achieving lower latency than both time-division multiple-access (TDMA) and preamble-based protocols, due to the extremely reduced overhead of the PIA sub-frame. We assess the performance of PIMA under various activation statistics, proving the robustness of the proposed solution to the traffic intensity, also with traffic bursts.

INDEX TERMS Sixth-generation (6G), Internet-of-Things (IoT), Massive random access (MRA), Machine-type communications (MTC), multiple access, partial-information.

I. INTRODUCTION

MACHINE-TYPE communications (MTC) has gained increasing relevance in fifth-generation (5G) and beyond-5G (B5G) cellular networks. The introduction of the newest use cases in verticals such as industry, agritech, and smart buildings shows the progressive shift of the focus of mobile communications from humans to machines. To support this scenario, several new technological solutions should be adopted. In particular, the design of specific multiple access schemes can be used to meet the stringent requirements of both flavors of MTC envisioned by the emerging Internet-of-Things (IoT) applications and services: the massive MTC (mMTC) and the ultra-reliable low-latency communications (URLLC). They both show several features and challenges: URLLC use cases, for example, target a maximum latency of 1 ms and reliability of 99.99999% (e.g., mission-critical applications), while mMTC scenarios require supporting devices with density up to 1 million devices per km^2 (e.g., for industrial IoT).

In this paper, we focus on the medium access control of uplink transmissions in an MTC scenario, wherein users transmit to a common base station (BS).

A. RELATED WORK

In the latest research, multiple access procedures based on resource requests and grants have been discussed for MTC.

However, the sporadic nature of transmissions by a large number of users that characterizes the above use cases makes these schemes inefficient and pushes for the adoption of a grant-free (GF) or semi-GF solution. In this section, we provide an overview of the most relevant grant-based (GB), GF, and semi-GF multiple access schemes. Both GB and GF solutions can provide orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) for the packet transmission. In the case of NOMA [2], multiple users transmit simultaneously, and specific techniques are adopted at the receiver to mitigate the effects of collisions.

Grant-Based Multiple Access. In recent years, GB schemes have been proven to be suitable in IoT networks with moderate numbers of users [3], [4], [5]. A widely adopted solution is *multichannel ALOHA*, wherein active users choose a preamble from a common pool (codebook) of orthogonal preambles. Then, the BS can effectively differentiate between multiple users simultaneously accessing the network and schedule the users for data transmission in a second dedicated sub-frame, (typically of fixed length). However, such schemes require different preamble lengths based on the number of active users to perform optimally. Indeed, while long preambles support more active users, their entailed communication overhead grows rapidly. Multichannel ALOHA has also been coupled with NOMA [6] for data transmission, showing substantial performance improvement at the cost of higher decoding complexity. Moreover, the use of non-orthogonal preambles has been discussed to reduce overhead in the context of compressive random access (CRA) [7], [8], [9]; with CRA the BS resorts to compressed sensing (CS) [10] to identify active users. However, multiple measurements are typically needed to accurately estimate the preambles, therefore requiring multiple-antenna receivers or multiple transmission steps.

An extreme case of GB multiple access is the *fast-uplink grant* approach [11], wherein the BS schedules the user transmissions without any resource request. The exploitation of traffic statistics knowledge has been considered in the design of fast-uplink grant protocols, involving multiarmed bandits-based methods [12] and machine learning tools for traffic prediction [13].

Grant-Free Multiple Access. GF approaches typically make better use of the resources by letting users transmit immediately, without waiting for explicit approval from BS. Slotted ALOHA (SALOHA) is the simplest and most widely adopted GF OMA protocol: users transmit at the beginning of the first slot available after packet generation and, when collisions occur, collided packets are retransmitted with random delays. Collisions are further reduced by framed slotted-ALOHA (FSA), which divides time into frames (each split into slots) wherein users transmit at random: this solution is widely adopted in radio-frequency identification (RFID) systems [14], [15]. Other OMA schemes have been proposed for correlated user activity, which is typically due to a correlated underlying traffic generation [16]. Also, retransmissions (with the consequent accumulation of packets in user queues) introduce a correlation of transmissions among users: this further increases collisions, while it can also be exploited to coordinate multiple access. GF correlation-based schedulers have recently gained attention as a possible breakthrough for multiple access in MTC. Such schemes typically rely on the knowledge of traffic generation statistics [17], or learn the traffic correlation by tracking successes and collisions [18], or reinforcement learning techniques [19], [20].

NOMA-based GF outperform conventional OMA solutions in many scenarios. An effort has been made to unify and standardize the various NOMA schemes, whose performance gains over their orthogonal counterparts have also been investigated [21], [22]. However, NOMA requires advanced pairing and power allocation techniques, as well as powerful channel coding and interference cancellation mechanisms that only partially mitigate collision effects [23]. Under these conditions, the BS may become prohibitively complex to serve a large number of users. One simple NOMA solution is NOMA-ALOHA [24], [25], wherein users transmit simultaneously in a SALOHA fashion, using different power levels, and collision effects are mitigated with successive interference cancellation (SIC). Furthermore, unsourced random access (URA) has recently emerged to manage a massive number of devices [26]: at any time, a fraction of devices transmit simultaneously using the same channel codebook, i.e., encoding their messages into complex codewords that are known at both users and BS. The receiver decodes the arriving messages without knowing the identity of the transmitters. Although this approach is very effective in managing a large number of users, good performance can be achieved only for tiny payloads and with highly complex massive multiple input-multiple output (MIMO) receivers [27], [28], [29]. A downside of most of the works on URA is also that they assume prior knowledge of the number of active users, and only recently an information theoretical analysis of URA with random user activity has been proposed [30].

Semi-Grant-Free Multiple Access. Semi-GF multiple access combines both GB and GF approaches. As for the GB solutions, users compete for resources in a preliminary short sub-frame, then, they may receive a grant to transmit data without contention. Typically, users neither reveal their identity nor the BS deterministically schedules the transmission due to the limited acquired information. This hybrid approach aims at combining the efficiency of GB scheduling with the reduced overhead of GB access. Despite the increasing interest in semi-GF multiple access, the majority of the literature focuses on semi-GF NOMA solutions. For example, [31] and [32] proposed to opportunistically admit GF transmissions of a subset of users on the same resources reserved for GB data transmissions, exploiting the receive power diversity and SIC for decoding the messages. However, despite the research hype on NOMA, there are several motivations supporting the use of OMA. First, the separation of user signals in OMA eliminates multi-user

interference, significantly simplifying both transmission scheduling and receiver operations. NOMA indeed requires more sophisticated signal processing techniques to mitigate interference, which may be complex and expensive to deploy. Moreover, OMA is a well-established and relatively simple technique and many wireless networks already have the infrastructure and protocols in place for OMA. In such a context, a transition to NOMA would require significant changes and investments in network architecture and standardization. Therefore OMA remains a valuable choice to ensure compatibility with existing IoT networks and devices. A semi-GF OMA scheme was presented in [33], wherein the BS broadcast the colliding preambles, preventing the data transmissions of the users choosing such preambles.

B. PARTIAL-INFORMATION MULTIPLE ACCESS

In [1], we introduced a novel approach to multiple access, wherein the BS first acquires *partial information* on the state of the user and then schedules the transmissions. In particular, the BS estimates the *number of users* with packets to transmit (active users) without knowing their identities. Based only on this information, it then allocates users to slots. The allocation is non-exclusive, i.e., a slot is typically given to multiple users, whose transmissions may collide. Indeed, without knowledge of the *identity* of active users, collisions could only be avoided using time-division multiple-access (TDMA), which is extremely inefficient for sporadic traffic.

The resulting scheme is denoted as partial-information multiple access (PIMA). In detail, PIMA organizes time in frames of *variable length*, each split into two sub-frames. The first is the partial information acquisition (PIA) subframe, where all active users simultaneously send a signal to the BS. The BS performs a maximum a posteriori probability (MAP) estimation of the number of active users. Based on this knowledge, the BS then assigns one slot to each user in the system for transmission in the data transmission (DT) subframe. A major drawback of the scheme presented in [1] is that the PIA sub-frame introduced a significant overhead, as users must transmit a long random sequence to let the BS estimate the total received power, which provided information on the number of active users. Moreover, an in-depth analysis considering the differentiated scenarios of next-generation cellular networks of the scheme was needed to understand more clearly the behavior of the solution.

The PIMA protocol falls into the category of semi-GF multiple access solution. Indeed, with respect to other two-step multiple access schemes consisting of preamble and data transmission stages, PIMA acquires only partial information on the activation statistics in the PIA sub-frame, avoiding revealing the users' identities. Likewise, PIMA cannot be considered a fast uplink grant approach, due to its partial information acquisition at the BS. The main advantage of PIMA with respect to the existing semi-GF solutions is its extremely low overhead (few symbols) required to acquire the partial information and communicate the scheduling, which yields an extremely low latency and complexity.

In this paper, we partially re-design PIMA and provide an in-depth analysis of its performance in various scenarios. In particular, the main contributions of this paper are the following.

1) We focus on the short packet transmission scenario, with packets consisting of a few complex symbols each and we improve the user enumeration procedure. By assuming channel state information (CSI) is perfectly acquired at each user with a downlink beacon, users can precode their transmission so that they coherently combine at the BS, similarly to a pulse-amplitude modulation (PAM) signal. Therefore the BS can reliably count the active users with a significantly shorter PIA sub-frame, reducing the overhead.
2) We consider three different use cases represented by three different packet generation statistics, including a bursty traffic generation for massive random access (MRA), wherein the number of users in the system is arbitrarily large while the number of active users remains finite.
3) We prove that, under independent identically distributed (i.i.d.) activation, the allocation of each user to a single slot maximizes the resulting frame efficiency.
4) We prove that the maximum frame efficiency with MRA traffic is obtained by allocating a number of slots equal to the number of active users.
5) We significantly extend the numerical evaluation part, comparing PIMA also with state-of-the-art preamble-based approaches.

The remainder of the paper is organized as follows. In Sections II and III, we first introduce the system model and the packet generation processes, then we describe the frame structure and the PIMA protocol. Section IV provides details of the user enumeration task for partial information acquisition. Then, the optimal scheduler for i.i.d. activations is derived in Section V. In Section VII we discuss the numerical results and compare PIMA with some state-of-the-art OMA and NOMA schedulers. schedulers. Finally, Section VIII draws some conclusions.

Notation. Scalars are indicated in italic letters; vectors and matrices are indicated in boldface lowercase and uppercase letters, respectively. Sets are denoted by calligraphic uppercase letters and $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denotes the ceiling and floor operators, respectively. $\mathbb{P}(\cdot)$ denotes the probability operator, $\mathbb{E}[\,\cdot\,]$ denotes the statistical expectation, $\log(\cdot)$ indicates the natural logarithm function, and a mod b denotes the modulo operation, which returns the remainder of a/b. Lastly, $\binom{a}{b}$ denotes the binomial coefficient.

II. SYSTEM MODEL

We consider the uplink of a multiple access scenario with N single-antenna users transmitting to a common

single-antenna BS. We assume that the value of N is known at the BS.

Time Organization. Time is split into *frames*, each comprising an integer and variable number of *slots* and an additional short time interval, whose purpose is described in the following. Each slot has a fixed duration of T_s complex symbols, while each frame comprises a different number of slots. Perfect time synchronization at the BS is assumed, thus each user can transmit signals with specific times of arrival at the BS.

Each user may transmit at most one packet per frame, each of the duration of one slot. In the following, t denotes the frame index. We consider a multiple access protocol, where the same slot is in general assigned to multiple users for transmission.

Channel. Due to the scheduling strategy, collisions between packets may occur, being each slot assigned to multiple users. In particular, we assume that a collision occurs when two or more users transmit in the same slot and that such an event prevents the decoding of all collided packets by the BS. Analogously, we assume that the BS always correctly decodes the received packets in slots without collision and that the channel does not introduce other sources of communication errors. Successful transmissions are acknowledged by the BS at the beginning of the following frame, and upon collision, collided users retransmit their packets in the following frame.

A. PACKET GENERATION AND BUFFERING

Packets generated in frame t by user n are stored in its buffer and transmitted according to a first in, first out (FIFO) policy. In the following, we consider both the case of finite and infinite buffer capacity. In the case of finite-length queues, to ensure data freshness, whenever a new packet is generated while the buffer is at full capacity, the oldest packet is dropped. Let $B_n(t)$ be the number of packets in the queue of user n at the beginning of frame t. If $B_n(t) > 0$, the buffer of user n is non-empty, and n is said to be *active*. Instead, if $B_n(t) = 0$, its queue is empty and user n is considered *inactive*. The total number of active users at the beginning of frame t is $\nu(t)$.

B. ACTIVATION STATISTICS

The PIMA scheme has been designed to exploit the sporadic nature of transmissions in the presence of a large number of users. Users can stay dormant for long periods and transmissions are usually triggered by an event or preemptively scheduled. When this happens, packet transmissions are activated with different patterns. To catch the diverse nature of the existing use cases, we analyze the performance of PIMA in three different scenarios, depending on the users' activation statistics.

I.i.d. Activations. User activation times can be described with independent and identically distributed random variables when a) users have a queue for one packet only, and b) all colliding packets are dropped at the receiver after the first transmission. Such a scenario is typical of monitoring systems requiring frequent updates and data freshness [34] and will be discussed in detail in Section VI.

Correlated Activations. In this scenario, we assume that queues have infinite lengths and retransmissions of previously collided packets are allowed. In particular, we assume that collided packets at frame t are deterministically queued for retransmission in the following frames. Users' activations are, therefore, generally correlated due to the presence of queues and retransmissions. In this scenario, the stability condition of all queues is assumed.

For both the i.i.d. and the correlated activations cases, we assume that at each user the traffic generation, also denoted as *packet arrival process*, follows a Poisson distribution with parameter λ. The well-known properties of Poisson processes provide a total normalized arrival rate of $\Lambda_T = N\lambda/T_s$, and when $\Lambda_T = 1$, on average one packet is generated over a slot duration (T_s).

Bursty Activations. In this scenario, we assume a bursty traffic model, wherein a finite subset of users activates at the same time, with each user generating a single packet to transmit. Retransmissions of colliding packets are also allowed in this case, while the buffer capacity is unitary for all users. In particular, we assume that at each burst the number of generated packets follows a Poisson distribution with average Λ_B.

III. PARTIAL INFORMATION-MULTIPLE ACCESS

In this section, we provide a detailed description of the proposed PIMA protocol. Each frame is divided into two *sub-frames*, namely the *PIA sub-frame* and the *DT sub-frame*. The PIA sub-frame has a fixed length L_1 and it is used by the BS to estimate the number of currently active users. Based on this information, the BS decides the duration $L_2(t)$ (in slots) of the DT sub-frame and assigns each user to one slot, for uplink data transmission.

An example of the frame structure and packet transmissions of PIMA is reported in Fig. 1.

A. PARTIAL INFORMATION ACQUISITION SUB-FRAME

The beginning of frame t, and thus the start of the PIA sub-frame, is triggered by the reservation beacon (RB), which is transmitted in broadcast by the BS to all users. RBs contain also the acknowledgments of correctly received packets in the previous frame. RBs provide to each user an accurate CSI of its channel to the BS.

In the PIA sub-frame the BS obtains the estimate $\hat{\nu}(t)$ of the number of active users $\nu(t)$, as described in detail in Section IV. The BS, knowing $\hat{\nu}(t)$, schedules the transmissions for the next sub-frame. Let $\boldsymbol{q}(t) = [q_1(t), \ldots, q_N(t)]$ be the *slot selection vector*, collecting the slot indices assigned to each user; then, the length of the DT sub-frame $L_2(t)$ can be derived from $\boldsymbol{q}(t)$ as

$$L_2(t) = \max_n q_n(t). \tag{1}$$

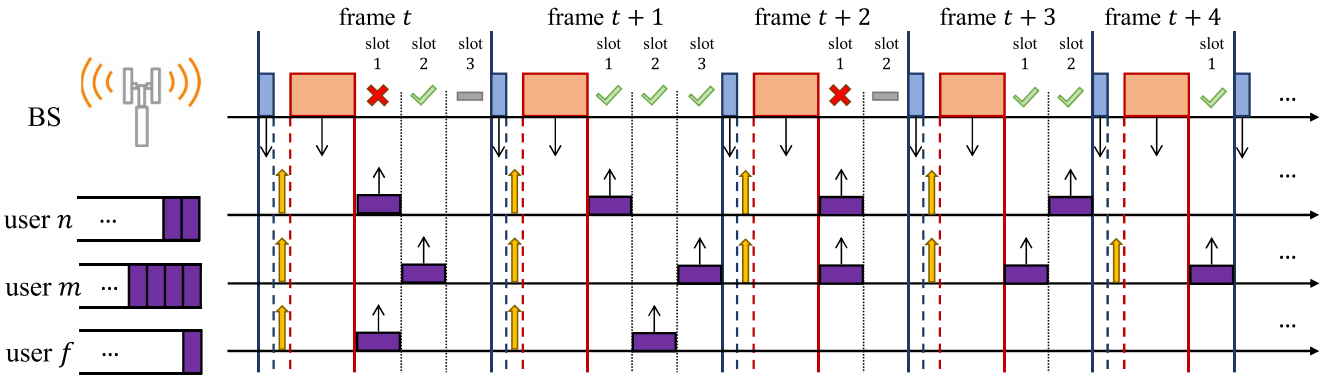

FIGURE 1. Example of the PIMA protocol and its frame structure. In this example, the user *n*, *m*, and *f* are active at the beginning of frame *t*, each with a different number of packets to transmit (purple rectangles). The RB and the SB transmitted by the BS in the downlink are, respectively, represented by the blue and orange rectangles, while the yellow arrow represents the AI signal for the user enumeration. For drawing simplicity, in this example, no packets are generated after the beginning of frame *t*.

To end the PIA sub-frame and trigger the beginning of the following DT sub-frame, the BS transmits the scheduling beacon (SB), which contains the slot selection vector $\boldsymbol{q}(t)$, encoded as described in Section III-C. The PIA sub-frame duration L_1 accounts for all the symbols transmissions in both uplink and downlink.

Note that, to maintain synchronization, inactive users could a) wake up and wait for the next downlink RB when generating a packet, or b) always wake up when RBs and SBs are transmitted (this can be achieved by collecting timing information in the beacons).

B. DATA TRANSMISSION SUB-FRAME

In the DT sub-frame, users transmit their packets, according to the scheduling set by the BS in the SBs. Note that packets generated by user n during the DT sub-frame are delayed and transmitted in the following frame, to reduce collisions, since the DT frame length is derived only based on the number of users active in the PIA sub-frame. Indeed, data transmissions of users who do not request resources in the PIA subframe would introduce uncertainty in the optimization of $\boldsymbol{q}_n(t)$, limiting the ability of PIMA to adapt to instantaneous traffic conditions.

Since all the derivations in the following are related to each frame separately, to simplify notation, we drop the frame index t from all the variables.

C. PIMA BEACONS OVERHEAD

We consider two options for the coding of $\boldsymbol{q}$. The first option provides that all the users in the system receive the explicit indication of the allocated slot in the DT sub-frame. Consider a codebook of N codewords (each of length $\log_2 N$ bits) representing all possible sorting of user indices, where if a user is in position i_n, its assigned slot is $k = i_n \mod L_2 + 1$. With this codebook, the SB introduces an overhead of $R_{\text{SB}} = (N+1)\log_2 N$ bits, since an additional codeword is needed to indicate the length of the DT sub-frame, L_2. Note that for a large number of users, this encoding strategy can significantly increase the length of SB, deteriorating the performance of PIMA.

Therefore, we consider a second strategy, wherein all users know a list of random sequences (each of $N\log_2 N$ bits) of length J, indicating the order of user service. In this case, the BS transmit in the SB only the index corresponding to the scheduling sequence and the codeword to indicate L_2, providing an overhead of $\log_2 J + \log_2 N$ bits. In the following, we adopt this second strategy to reduce the SB overhead.

IV. ESTIMATION OF THE NUMBER OF ACTIVE USERS

To obtain an estimate of the number of active users at the BS, each active user transmits an activation information (AI) signal immediately after receiving the RB. Note that we neglect here the propagation time between BS and the user, which can be easily accounted for by considering a transition (silent) time between the RB and AI transmissions. The set of users transmitting the AI signals during the PIA sub-frame is

$$\mathcal{N}_{\text{a}} = \{n{:}B_n > 0\}, \tag{2}$$

with $|\mathcal{N}_{\text{a}}| = \nu$. We stress that the BS does not know the identity of the active users, since the AI signals do not contain such information, to make them shorter.

We consider that each user transmits a single complex symbol γ_n in the PIA sub-frame. Assuming perfect synchronization, the received signal, at the BS, is the superposition of all the symbols transmitted by the users, i.e.,

$$y = \sum_{n\in\mathcal{N}_{\text{a}}} h_n\gamma_n + w, \tag{3}$$

where h_n is the channel coefficient between user n and the BS and w is the additive white Gaussian noise (AWGN) term with zero mean and variance σ_w^2. Assuming that perfect CSI is obtained by users through the RB downlink transmission, each user n perfectly inverts the channel, setting $\gamma_n = 1/h_n$, therefore the BS receives

$$y = \nu + w. \tag{4}$$

In low-signal-to-noise ratio (SNR) scenarios, the AI signal may last several symbols, all with the same structure, and the BS will first average the received samples (obtaining (4) but with less noise) and then proceed with the estimate of the number of active users.

The MAP estimate of the number of active users is then

$$\hat{\nu} = \arg\max_{b} \; p_{y|\nu}(y|b)\, p_{\nu}(b), \tag{5}$$

where, for the AWGN channel, the probability density function (PDF) of the received signal conditioned to the number of active users is

$$p_{y|\nu}(y|b) = \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{|y-b|^2}{\sigma_w^2}}. \tag{6}$$

This criterion establishes decision regions on the received signal. Define δ_b as the distance from b (the value obtained without noise) and the region associated with the decision $b+1$. Then, when y falls in the region $[b-\delta_b, b+1-\delta_{b+1}]$, the decision on the number of active devices is b. Since the distance between b and $b+1$ is 1, we have that the distance from b (the value obtained without noise) and the region associated with the decision $b+1$ is $1-\delta_{b+1}$. According to the MAP criterion (5), the optimal regions satisfy the following equation

$$p_{b|\nu}(b+\delta_b|b)p_{\nu}(b) = p_{b+1|\nu}(b+1-(1-\delta_b)|b+1)p_{\nu}(b+1), \tag{7}$$

since $\delta_{b+1} = 1-\delta_b$. Replacing (6) into (7), after some algebraic steps we have

$$\delta_b = \frac{1}{2} + \frac{\sigma_w^2}{2} \ln \frac{p_{\nu}(b)}{p_{\nu}(b+1)}. \tag{8}$$

The error probability is the probability of falling out of the correct decision region, thus, conditioned on the fact that $\nu = b$ users are active, we have

$$P_e(b) = \mathrm{Q}\left(\frac{\delta_b}{\sigma_w/\sqrt{2}}\right) + \mathrm{Q}\left(\frac{1-\delta_{b+1}}{\sigma_w/\sqrt{2}}\right), \tag{9}$$

where $\mathrm{Q}(\cdot)$ is the tail distribution function of the standard normal distribution.

I.i.d. Activations. In the case of i.i.d. activations, the activation process coincides with the packet generation process. Assuming that each user generates packets according to a temporal Poisson process with parameter $\lambda_n = \lambda, \forall n$, we obtain

$$p_{\nu}(b) = \binom{N}{b} b\left(1-e^{-\lambda T_{\mathrm{a}}}\right)(N-b)e^{-\lambda T_{\mathrm{a}}}, \tag{10}$$

where T_{a} is the time interval considered for the packet generations and $\binom{N}{b}$ counts for all the possibilities of having exactly b active users. Note that for $N \to \infty$ we have $p_{\nu}(b) \approx p_{\nu}(b+1)$, thus $\delta_b \to \frac{1}{2}$ and all regions have the same size. In this asymptotic scenario, we also have

$$P_e(b) \to 2\mathrm{Q}\left(\frac{1}{\sqrt{2}\sigma_w}\right). \tag{11}$$

A. PIA WITH IMPERFECT CSI

In case of imperfect CSI, the estimated channel at user n is

$$\hat{h}_n = h_n + \epsilon_n, \tag{12}$$

where ϵ_n accounts for the estimation error. User n transmit $\gamma_n = 1/\hat{h}_n$, and the received samples becomes, from (3),

$$y = \nu - \sum_{n\in\mathcal{N}_{\mathrm{a}}} \frac{\epsilon_n}{h_n+\epsilon_n} + w = \nu + w'. \tag{13}$$

Assuming $\epsilon_n \sim \mathcal{CN}(0,\sigma_\epsilon)$ for all n, the terms $\frac{\epsilon_n}{h_n+\epsilon_n}$ follow a complex Gaussian ratio distribution, which has been widely studied in both cases of correlated and independent numerator and denominator (see [35] and [36]).

When either the SNRs is high (by the Gaussian ratio statistics) or n is large (by the central limit theorem), the summation in (13) can be approximated as a Gaussian r.v., thus w' is also Gaussian. In these scenarios, the imperfect CSI case is modeled as additional noise, and the analysis of the previous subsection still holds. For other scenarios, numerical methods (including Monte Carlo simulations) should be used to evaluate the error probability in the counting process.

V. FRAME EFFICIENCY-BASED SCHEDULING

In this section, we propose a time-resource scheduling conditioned on the number of active users estimated in the PIA sub-frame.

First, we introduce a performance metric that takes into account both the packet latency and the collision probability, whose optimization aims at finding the right balance between the two. Let $l \in \{1,\dots,L_2\}$ be the slot index within the frame (in the DT sub-frame). We define the success indicator function in slot l as $c_l = 1$ if a successful transmission occurs in slot l and $c_l = 0$ otherwise. Note that the latter case considers both the collision and non-transmission cases. Then, the *conditional frame efficiency* is defined as the ratio between the number of successes in the frame and the length of the DT sub-frame, i.e.,

$$\eta(\nu) = \frac{1}{L_2}\sum_{l=1}^{L_2} \mathbb{E}[c_l|\nu]. \tag{14}$$

The adaptive maximization of this metric provides the proper balance between the DT sub-frame length and the successful transmission probability.

At each frame, immediately after the end of the PIA sub-frame, the BS solves the following optimization problem:

$$\max_{\boldsymbol{q}} \; \eta(\nu), \tag{15a}$$

$$\text{s.t.} \quad q_n \in \{1,\dots,L_2\}. \tag{15b}$$

The optimization problem (15) is one of mixed integer non-linear programming (MINLP), and its solution quickly becomes infeasible with long queues or many users. For these reasons, in the following, we focus on the analysis of the i.i.d. activation scenario, designing the parameters of the PIMA scheduler based on its basic assumptions. Note that,

while the i.i.d. activation scenario could substantially differ from the correlated activations one at high traffic, it still represents a good approximation in low traffic conditions, wherein retransmissions due to collisions occur sporadically. Moreover, the analysis under i.i.d. activations is useful when no information is available on the transmission correlation at the BS or when obtaining such information is too expensive.

VI. SCHEDULING WITH I.I.D. ACTIVATIONS

We now design PIMA for the i.i.d. activation scenario, under the assumption of perfect user counting in the PIA sub-frame. First, we observe that, since activations of users are i.i.d., we only have to determine how many users are assigned to each slot, as any specific assignment satisfying this constraint will yield the same collision probabilities, thus the same expected frame efficiency. Note that in case of decoding failure of multiple packets, the user scheduling should be randomized to avoid the repetition of the same collisions.

To minimize the number of users assigned to the same slot, given a length L_2, we assign to slot l the following number of users

$$u_l = \begin{cases} \left\lceil \frac{N}{L_2} \right\rceil & \text{if} \quad l \leq N \bmod L_2, \\ \left\lfloor \frac{N}{L_2} \right\rfloor & \text{if} \quad l > N \bmod L_2, \end{cases} \tag{16}$$

where we possibly schedule one more user in the first $\lfloor \frac{N}{L_2} \rfloor$ slots to minimize the transmission delay.

With this scheduling policy, the slot success random variable c_l can be rewritten as a function of u_l, as it only depends on the number of users scheduled in slot l. Thus, the optimization problem (15) is reduced to the optimization of the DT sub-frame length, L_2, and from (14), we have

$$L_2^* = \arg\max_{L_2} \frac{1}{L_2} \sum_{l=1}^{L_2} \mathbb{E}[c_l|\nu, u_l], \tag{17a}$$

$$\text{s.t. } L_2 \in \mathbb{N}/\{0\}. \tag{17b}$$

Now, given ν, the probability that user n is the one and only active user assigned to slot l is derived by considering all cases of active users, where user n is active and all other users assigned to slot l are, instead, inactive. Consider the matrix $\lceil N/L_2 \rceil \times L_2$, having entry (r, c) equal to one when in slot c only user r (among those assigned to that slot) is active. The number of favorable cases wherein we have a single "1" in column l is given by all the possibilities to place the "1" of the remaining $\nu - 1$ users in any of the $N - u_l$ entries of the matrix excluding column l. Therefore, the favorable case is given by all the combinations of $\nu - 1$ objects taken from a set of $N - u_l$ objects. Instead, the total number of cases is given by all the combinations of ν active users chosen from the N scheduled users. The probability of having a successful transmission in slot l is therefore

$$\mathbb{E}[c_l|\nu, u_l] = u_l \frac{\binom{N-u_l}{\nu-1}}{\binom{N}{\nu}}, \tag{18}$$

where factor u_l counts the users assigned to slot l. Note that the numerator of (18) counts the number of combinations giving exactly one active user assigned to slot l, while the denominator counts the total number of possible combinations of active users. The probability of collision in slot l is therefore $1 - \mathbb{E}[c_l|\nu, u_l]$.

Problem (17) is a MINLP problem, and its solution strictly depends on the number of users in the system. If the number of active users is comparable with N, (17) is not solvable by continuous relaxation of L_2, as the rounding functions are not differentiable. However, it is possible to find the optimal frame length L_2^* with complexity $O(\log N)$, using a binary search algorithm, or alternatively using a discrete gradient ascent algorithm. In any case, L_2^* depends only on ν, thus can be computed offline and then stored in a table. Instead, if $N \to \infty$ and $\nu << N$ is finite, the following result holds:

Theorem 1: For $N \to \infty$, under i.i.d. activations, given a finite number of active users ν, the DT sub-frame length L_2 maximizing frame efficiency is exactly ν.

Proof: Since $N \to \infty$ and $1 < \nu << N$, $u_l = \frac{N}{L_2}$ for all $l = 1, \ldots, L_2$. From (18) we have

$$\begin{aligned} \eta(\nu) &= \mathbb{E}\left[c_l \middle| \nu, \frac{N}{L_2}\right] = \frac{N}{L_2} \binom{N - \frac{N}{L_2}}{\nu - 1} \Big/ \binom{N}{\nu} \\ &= \frac{\nu}{L_2} \frac{\left(N - \frac{N}{L_2}\right)!(N-\nu)!}{\left(N - \frac{N}{L_2} - \nu - 1\right)!(N-1)!}, \end{aligned} \tag{19}$$

then, using the Stirling factorial approximation $\alpha! = \sqrt{2\pi\alpha}(\frac{\alpha}{e})^\alpha$, whose validity is verified with good accuracy even for small values of α, with some algebraic steps we obtain

$$\eta(\nu) = \frac{\nu}{L_2} \overbrace{\sqrt{\frac{N\left(1 - \frac{1}{L_2}\right)(N-\nu)}{\left[N\left(1 - \frac{1}{L_2}\right) - \nu + 1\right](N-1)}}}^{\mu(N,\nu,L_2)} \underbrace{\times \frac{\left[N\left(1 - \frac{1}{L_2}\right)\right]^{N\left(1-\frac{1}{L_2}\right)} (N-\nu)^{N-\nu}}{\left[N\left(1 - \frac{1}{L_2}\right) - \nu + 1\right]^{N\left(1-\frac{1}{L_2}\right)-\nu+1} (N-1)^{N-1}}}_{\xi(N,\nu,L_2)}. \tag{20}$$

Taking the limit for $N \to \infty$ we have

$$\lim_{N\to\infty} \mu(N, \nu, L_2) = 1, \tag{21a}$$

$$\lim_{N\to\infty} \xi(N, \nu, L_2) = \left(1 - \frac{1}{L_2}\right)^{\nu-1}. \tag{21b}$$

Therefore the maximum frame efficiency only depends on ν and L_2, and it is given by

$$\eta(\nu) = \frac{\nu}{L_2}\left(1 - \frac{1}{L_2}\right)^{\nu-1}. \tag{22}$$

Its stationary points are derived from the first order derivative with respect to L_2 as

$$\frac{d}{dL_2}\left[\frac{\nu}{L_2}\left(1-\frac{1}{L_2}\right)^{\nu-1}\right]=0 \Leftrightarrow L_2=\nu \vee L_2=1, \quad (23)$$

thus, while $L_2=1$ is trivially the global minimum for $\nu>1$, the frame efficiency is maximized for $L_2=\nu$. ■

A. ON THE OPTIMALITY OF THE SINGLE SLOT SCHEDULING

Throughout the paper, we have assumed that each user n is assigned to a single slot q_n in the frame. However, we may wonder if this scheduling policy is optimal or if it is preferable to assign multiple slots to each user. Focusing on the case of i.i.d. activations, we have the following result.

Theorem 2: Under i.i.d. activations, the assignment of a single slot to each user in the DT sub-frame is optimal, i.e., it maximizes the expected frame efficiency.

Proof: Since the collision probability is the same for all users and depends only on the number of other users transmitting in the same slot, by allocating more slots to each user, we increase the collision probability. Hence, single-slot scheduling is optimal in this case. ■

VII. NUMERICAL RESULTS

In this section, we compare our PIMA protocol with the state-of-the-art OMA schedulers in three different scenarios: a) i.i.d. activation scenario, with $N=50$ users, b) correlated activation scenario, also with $N=50$ users, and c) bursty activation scenario, with a large N. Following the assumption of *short packet transmission*, the number of symbols in each packet (slot duration) is $T_s=10$ unit symbol duration (usd). [1] Furthermore, an SNR of 10 dB is assumed at the BS. Note that while the analysis presented in Section VI assumes a perfect estimation of the number of active users, in the following we report the performance of PIMA with a perfect estimation of ν (denoted PIMA ideal) and with estimation affected by noise (simply denoted PIMA).

For performance comparison, we consider a) the standard TDMA, which provides fixed-duration frames of N slots, with one user assigned per slot deterministically, b) a stabilized version of the SALOHA protocol, c) the NOMA-ALOHA protocol, and d) the CRA-2 protocol of [9] with preambles of length $M_p=N/2$ and N.

Stabilized Slotted ALOHA: For the SALOHA protocol, we consider Rivest's stabilized SALOHA [37, Ch. 4], [38], where all users generating packets at slot l are backlogged with equal probability. The backlog probability is computed for each user through a pseudo-Bayesian algorithm based on an estimate of the number of backlogged nodes $G(l)$ as

$$\alpha(l)=\min\left(1,\frac{1}{G(l)}\right), \quad (24)$$

where

$$G(l)=\begin{cases} G(l-1)+N\theta+(e-2)^{-1} & \text{if } c_l=0,\\ \max(N\theta, G(l-1)+N\theta-1) & \text{if } c_l=1,\end{cases} \quad (25)$$

is the estimated number of backlogged users (with $G(0)=0$) and $\theta=1-e^{-\lambda}$ is the packet generation of probability at slot l.

NOMA-ALOHA: For the NOMA-ALOHA [24], we assume that active users transmit with high (H) or low (L) power chosen with equal probability. According to [24] a relatively large number of users, i.e., up to a few dozen, can transmit with L without interfering with the users transmitting with level H. In such an implementation, a collision occurs when two or more users transmit with the same power level. Moreover, if two or more active users choose H it is impossible to decode the packets transmitted with both H and L, as the interference cancellation fails. Note that this is an ideal assumption, as we do not consider the noise term but only the interference, assuming high SNR. The same backlogging mechanism (25) of SALOHA is adopted.

Modified CRA-2 Protocol: The CRA-2 protocol was proposed in [9]. Similarly to PIMA, each frame includes two sub-frames, and in the first sub-frame active users are identified with a temporary identifier, rather than just counted. To this end, each active user uniformly randomly chooses and transmits a sequence of complex symbols of length M_p (*preamble*), from a preamble pool known to both users and BS. The BS receives all preambles simultaneously and detects them. Preambles are used as temporary identifiers for the active users, and the BS schedules the data transmission in the second sub-frame by allocating one slot per each detected preamble. Here, we assume that preambles are orthogonal; therefore, the number of preambles equals their length M_p. When $M_p=N$, each user is uniquely assigned to a preamble, while for $M_p=N/2$, the preamble choice is random. In the latter case, if two or more users transmit the same preamble (and the BS detect it), they are assigned the same slot and collide. For both schemes, we consider the probability of misdetection in the preamble in the first subframe $P_{md}=0.1$. Note that in [9] preambles are assumed to be non-orthogonal, and a CS-based algorithm is applied. However, preamble detection in the presence of noise is not considered and a fixed misdetection probability is assumed. The impact on system performance of different CS algorithms has been discussed for the case of multiple measurements (e.g., multiple antennas in BS) in [39]. Although on one hand CS-based detection allows increasing the number of preambles and reducing the probability of collision, on the other hand, it also produces a high probability of misdetection when BS is equipped with a single antenna (single measurement) and a large number of users are active [40].

Comparing PIMA with CRA-2, we have two main differences: a) the first sub-frame is shorter for PIMA than for CRA-2, and b) the second sub-frame has the same length

1. The usd is the inverse of the bandwidth if the Nyquist sampling rate is used. All the lengths of the sequences, slots, and beacons are given in usd.

TABLE 1. Overhead comparison of the considered protocols.

Protocol	Uplink Overhead	Downlink Overhead
PIMA	1 usd	2 usd
Standard PIMA [1]	$\propto N^2$ usd	2 usd
CRA-2	M_{p} usd	$\hat{\nu}$ usd

for both approaches. Thus, on the one hand, overall the PIMA frame is shorter, reducing the average number of packets generated in each frame, and this reduces the average number of packets accumulated in the buffers before transmission in the next frame. On the other hand, in PIMA users may collide in the second sub-frame due to a non-exclusive allocation, which increases (with respect to CRA-2) the number of packets to be re-transmitted in the next frame (thus increasing the average number of users in buffers). Lastly, a wrong counting of users in the PIA sub-frame or a wrong identification in the first CRA-2 sub-frame increases the collision probability of both schemes. The numerical results presented in this section will compare the performance of both schemes, taking into account all these effects.

Overhead Comparison: Since the packet acknowledgments are neglected for all schemes, both TDMA and SALOHA do not entail any overhead. For both PIMA and the preamble-based solutions, we assume that a 64-QAM modulation is used to modulate the SB, and the list of the scheduling sequence is set to $J = 64$. Under the aforementioned assumptions, the total overhead induced by the PIA sub-frame is constant and equal to $L_1 = 3$ usd (one for AI and two for SB, as the overhead of the RB is neglected). Instead, for preamble-based approaches, the overhead is given by the preamble of M_{p} usd and the BS feedback, which is typically longer than the SB. In particular, assuming that ν preambles are detected by the BS, the identifiers of these preambles are fed back in broadcast, in the order of slot allocation (for the subsequent data transmission), providing a total overhead of $M_{\mathrm{p}} + \hat{\nu}$ usd. The overhead of RB is neglected and does not play any role in the performance comparison, as its length is comparable for all the considered schedulers. The overhead of the compared schemes (in usd) is summarized in Table 1.

Performance metrics: For the i.i.d. and correlated activation case, performance is assessed in terms of both *average frame efficiency* $\bar{\eta}$ and *average latency* $\bar{D}$. The former metric is the average of the conditional frame efficiency for $\nu > 0$, i.e., $\bar{\eta} = \mathbb{E}[\eta(\nu)|\nu > 0]$. In the latter metric, the average is computed among all successfully delivered packets. Moreover, in the case of i.i.d. activations, we also consider the packet dropping probability P_{drop}, counting the packets dropped due to both collisions and replacements in the unit-length buffers when generations occur. This probability is 0 in the correlated activation case for all schemes, due to the possible retransmission and the infinite-length queues.

Finally, for bursty activations, the performance of the system is evaluated in terms of the *burst transmission time* D_{B}, i.e., the time needed to transmit all the packets generated in a traffic burst, and its average $\bar{D}_{\mathrm{B}} = \mathbb{E}[D_{\mathrm{B}}]$, computed over many bursts.

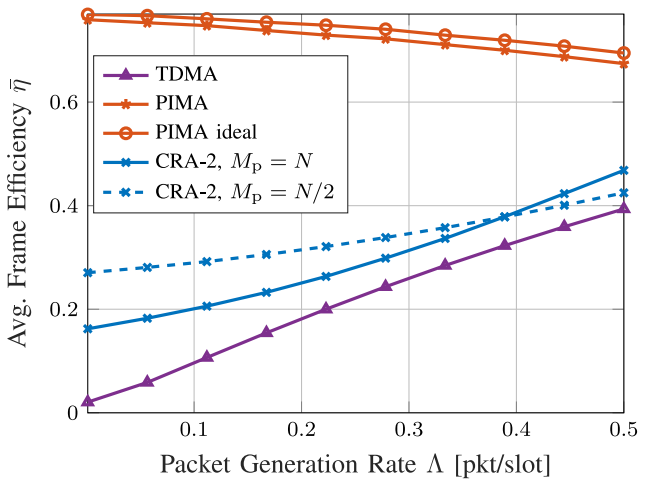

FIGURE 2. Average frame efficiency versus the total packet generation rate for $N = 50$ and i.i.d. activations.

A. I.I.D. ACTIVATIONS

We first report and discuss the results obtained for i.i.d. user activity and $N = 50$ users. In this activation scenario, retransmissions are not allowed. Therefore, SALOHA does not include backlogging, and each user attempts the transmission immediately upon the packet generation, i.e., $\alpha(l) = 1, \forall l$.

Firstly, a comparison of the average frame efficiency achieved by each of the schemes is shown in Fig. 2, as a function of the total packet generation rate Λ. While this metric cannot be defined for SALOHA and NOMA-ALOHA as they do not divide time into frames, we observe that TDMA, adopting the constant frame length, provides a very low frame efficiency. Moreover, since only the frames with $\nu > 0$ are considered in the average, PIMA achieves its highest frame efficiency at extremely low traffic due to its low overhead, and the efficiency slightly decreases at higher traffic due to collisions. As the traffic intensity increases, PIMA always achieves the highest frame efficiency, outperforming both the TDMA and the preamble-based schemes, whose overhead severely affects performance. Fig. 3, instead, illustrates the relationship between SNR and average frame efficiency of PIMA, showing a monotonically increasing trend, in all the considered traffic conditions. Notably, for SNR values exceeding 10 dB, the curves approach the PIMA ideal upper bound.

Figs. 4 and 5 show the effect of the packet generation rate on the average latency and packet dropping probability, respectively. In this context, all packets generated during a frame transmission wait, on average, $N/2$ slots in low traffic conditions, therefore TDMA shows the highest latency. Still, latency decreases as the traffic increases, since the

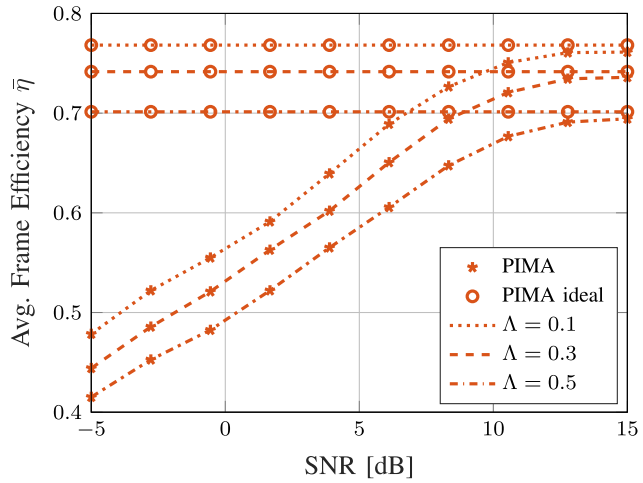

FIGURE 3. Average frame efficiency versus the SNR for $N = 50$ and $\Lambda = \{0.1, 0.3, 0.5\}$, under i.i.d. activations.

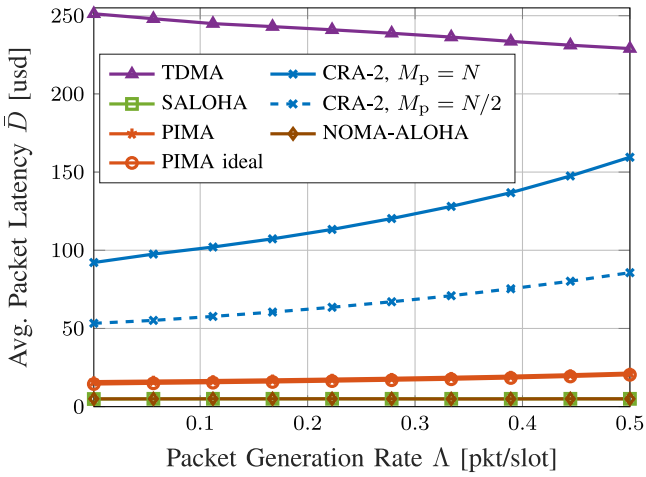

FIGURE 4. Average latency versus the total packet generation rate for $N = 50$ and i.i.d. activations.

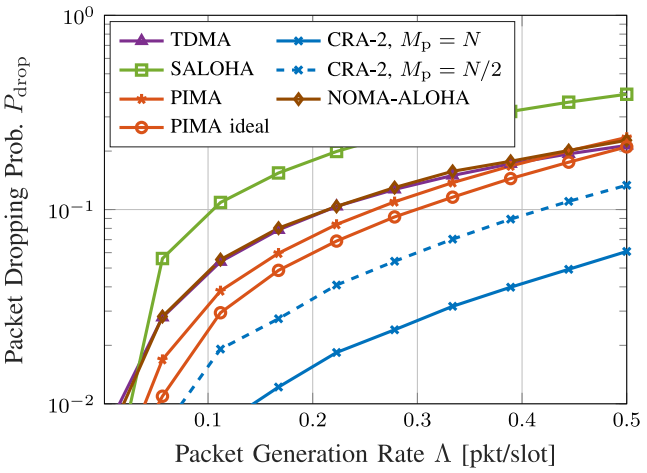

FIGURE 5. Average packet dropping probability versus the total packet generation rate for $N = 50$ and i.i.d. activations.

buffering delay is reduced by the new packets replacing the older ones in the queue. However, the dropping probability increases up to over 0.1 in high-traffic conditions. Instead, SALOHA and NOMA-ALOHA attain the lowest latency in this scenario, transmitting all packets immediately after their generation. Indeed, these schemes provide a lower bound on the latency, as all colliding packets are discarded and do not contribute to its evaluation. However, dropped packets increase the dropping probability, which approaches 1 for SALOHA at large Λ and is comparable to both TDMA and NOMA-ALOHA. The CRA-2 scheduler with N preambles, instead, is collision-free, and it drops a reduced number of buffered packets due to its shorter DT sub-frame with respect to TDMA. Indeed, CRA-2 achieves the lowest P_{drop} among the considered approaches: this improvement comes at the cost of higher latency, due to the longer time needed for the first sub-frame. Lastly, while the already mentioned schemes drop packets due either to collision (SALOHA) or new packet generations (TDMA and CRA-2 with $M_p = N$), both PIMA and CRA-2 with $M_p = N/2$ merge the advantages of the aforementioned solutions. On one hand, PIMA provides a higher collision probability than the preamble collision probability of CRA-2. On the other hand, collisions are compensated by a reduced dropping probability of newly generated packets, thanks to a shorter first sub-frame. Thus, while CRA-2 achieves a considerably lower dropping probability than PIMA, its latency is higher than that of PIMA, which guarantees close-to-minimum latency.

B. CORRELATED ACTIVATIONS

We now assess the performance of the correlated user activation scenario, assuming infinite queue lengths and an infinite number of possible (re)transmission attempts of collided packets. We consider a total number of $N = 50$ users and performance results are shown as a function of the traffic intensity $0.001 \leq \Lambda \leq 0.35$ pkt/slot.

First, Fig. 6 shows the average frame efficiency $\bar{\eta}$, as a function of the total packet generation rate. The performance is almost the same as in the i.i.d. activation scenario. This is mostly due to the fact that, in stability conditions, few retransmissions are performed multiple times. While all the observations on Fig. 2 still hold, we also note a very slight degradation of PIMA, as it is designed for i.i.d. activations and does not take into account previous collisions.

Second, Fig. 7 shows the average packet latency as a function of the packet generation rate for $N = 50$ users. Still, the latency of TDMA is much higher than that of other schemes, due to its frame length (N, the maximum among all schemes). We also observe that, at low traffic, the PIMA scheduler achieves extremely low latency, comparable to the minimum achieved by SALOHA. In higher traffic conditions, instead, the SALOHA backlogging mechanism prevents users from transmitting their buffered packets immediately, thus increasing the average latency. This effect is mitigated in PIMA, whose latency is reduced, due to its better ability

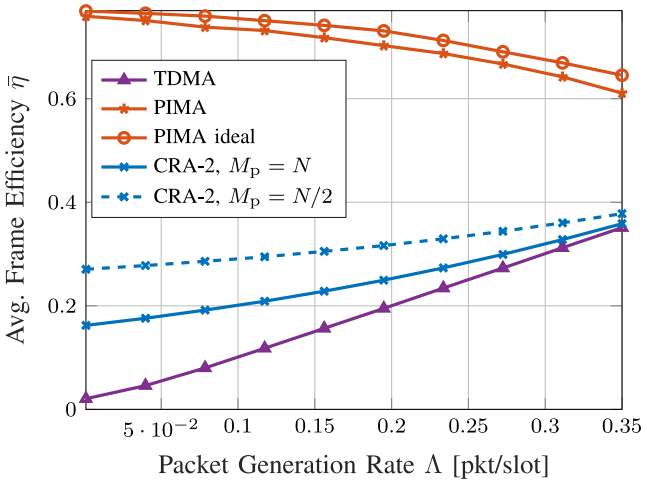

FIGURE 6. Average frame efficiency versus the total packet generation rate for $N = 50$ and correlated activations.

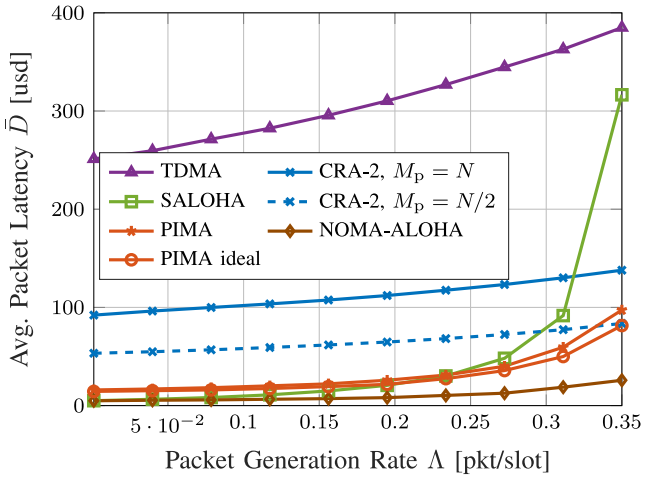

FIGURE 7. Average packet latency versus the total packet generation rate for $N = 50$ and correlated activations.

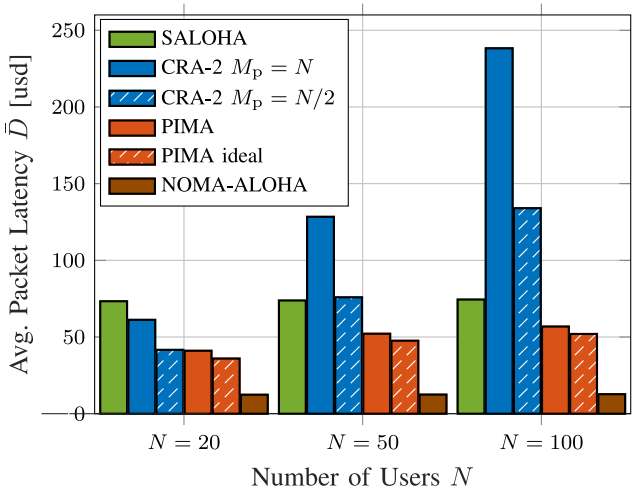

FIGURE 8. Average packet latency versus the number of users N, for $\Lambda = 0.3$ pkt/slot and correlated activations.

to adapt to instantaneous traffic load. For the preamble-based approach, instead, we observe that overhead plays a crucial role in overall latency, and shorter preambles yield better performance, while PIMA is still outperforming CRA-2. Finally, NOMA-ALOHA is the most effective protocol in terms of latency. However, we stress that with respect to the other compared schemes, its implementation requires high SNR at the BS to perform SIC, which is here assumed to be ideal.

The average latency comparison for different numbers of users in the system (N) is shown in Fig. 8, for $\Lambda = 3$ pkt/slot. The performance of SALOHA and NOMA-ALOHA remains constant for all the considered values of N, as the performance of such protocols is only dependent on the traffic intensity. Differently, the latency provided by the CRA schemes strictly depends on the number of users in the system, as preamble lengths need to be tuned accordingly. Indeed, the larger the number of users in the system, the longer the preamble is required to ensure user identification, with a direct impact on the average packet delay. The proposed PIMA protocol achieves a low latency for all the considered values of N. Still, performance significantly changes for different numbers of users. This is mainly due to the increased number of collisions occurring for large N. Indeed, the optimal DT sub-frame duration $L_2(t)$ solving (17) is proportional to the number of active users in the system $\nu(t)$. However, with the allocation of (16), once $L_2(t)$ is fixed, more users are assigned to the same slots for large N, increasing both the collision probability and (as a consequence) the average packet latency.

Lastly, in the correlated activation scenario, we have considered infinite-length queues, thus we may wonder if such queues are stable for the considered traffic intensity. However, since collision statistics in this scenario are quite difficult to compute due to the correlated nature of activations, we leave an in-depth analysis of the stability for future study.

C. BURSTY ACTIVATIONS

We now investigate the performance in the bursty activations scenario. We first assess the performance obtained for a single burst of intensity Λ_B, and then discuss some constraints on the burst interarrival time. As discussed in Section II-B, here we assume that a random number of active users ν, out of an arbitrarily large number of users in the system N, generate a single short packet to be transmitted in the following time slots. As we consider an arbitrary large N, the length of the DT sub-frame in PIMA is here derived according to Theorem 1.

The number of packet generations in a burst follows a Poisson distribution, with average $\Lambda_B \in [10, 10000]$ pkt/burst. For comparison purposes, we consider CRA-2 with fixed preamble lengths $M_p = 1000$ and 10000. We also consider an ideal solution, wherein the length of the preamble is adapted to the average number of generated packets, that is, $M_p = \Lambda_B$. Note that this ideal solution

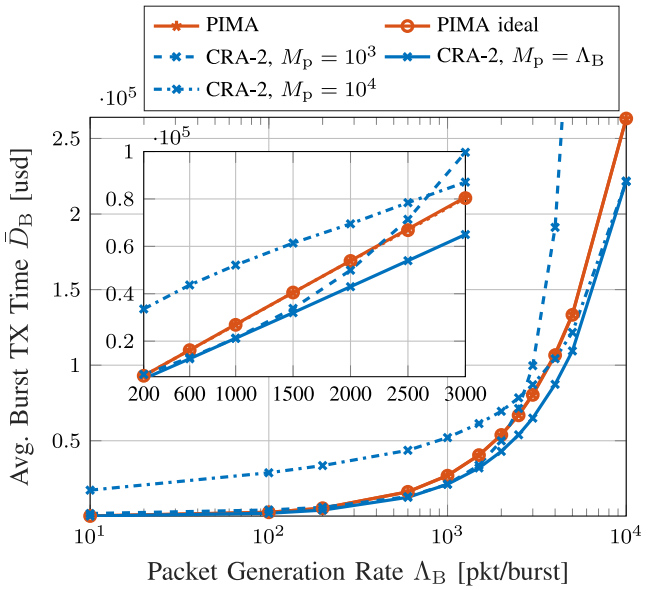

FIGURE 9. Average burst transmission time versus the packet generations rate (in log scale) in a bursty scenario. The zoomed plot reports results for 600 pkt/burst $\leq \Lambda_B \leq$ 3000 pkt/burst, in linear scale.

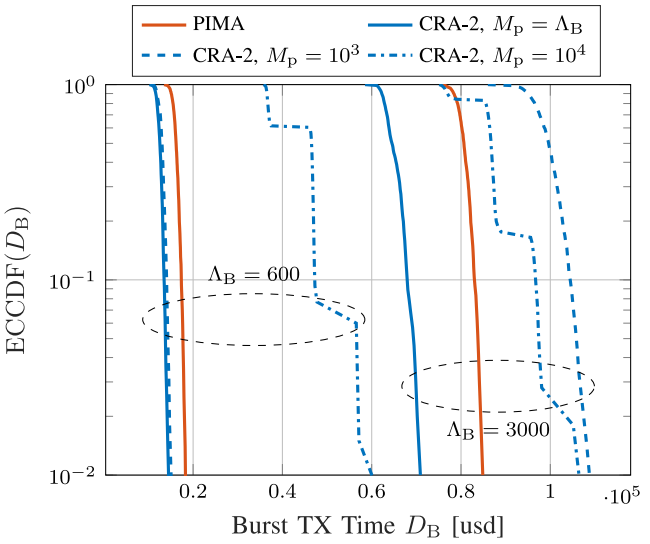

FIGURE 10. ECCDF of the burst transmission time.

is hardly implementable, as it requires different preamble pools based on the traffic generation rate. Moreover, we do not consider SALOHA and NOMA-ALOHA, as all packets simultaneously generated would collide with extremely high probability (1 for SALOHA). We also exclude TDMA, as a large N results in very long frame lengths. Note that a longer list of random sequences (that is, a longer overhead) for the scheduling vector encoding (Section III-C) is necessary as the number of users in the system increases. In the following, we still consider a PIA sub-frame overhead of $L_1 = 3$ symbols, which can be easily accommodated by adopting a higher modulation order at the BS for SB transmission.

Fig. 9 shows the average burst transmission time as a function of the average number of packet generations. First, we observe a constant gap between PIMA and the ideal preamble-based solution with $M_p = \Lambda_B$, while fixed-length preambles achieve better performance when the number of active users is comparable to the preamble length. In particular, the CRA-2 solution with $M_p = 1000$ preambles achieves a lower burst transmission time only for 400 pkt/burst $\leq \Lambda_B \leq$ 2000 pkt/burst, while at least $\Lambda_B = 4000$ pkt/burst is needed when $M_p = 1000$. For a low average number of packet generations, PIMA achieves the best performance due to its reduced overhead. For faster packet generations, both fixed-length preamble approaches suffer from preamble collisions, which implies a much higher packet transmission time due to retransmissions. Therefore, while PIMA adapts to all traffic conditions, preamble-based approaches should adopt different preamble pools depending on the traffic intensity to achieve a good performance.

Finally, the ECCDF of the burst transmission time is shown by Fig. 10 for $\Lambda_B = 600$ and 3000 pkt/burst. The results confirm the comments on Fig. 9, with the ideal case $M_p = \Lambda_B$ always attaining the lowest transmission time and the fixed-preamble approaches outperforming PIMA only for a preamble length comparable to the number of packet generations.

The figure gives also an indication of the minimum burst interarrival time τ_B that minimizes the probability of overlap between two traffic bursts. For example, from Fig. 10, the probability of overlap of PIMA is 10^{-2} if $\tau_B = D_B \approx 19 \cdot 10^3$ usd for $\Lambda_B = 600$ pkt/burst, and if $\tau_B \approx 85 \cdot 10^3$ usd for $\Lambda_B = 3000$ pkt/burst. In this comparison, PIMA is shown to be effective in minimizing burst transmission time, being very close to the CRA-2 solutions with $M_p = 1000$ and $M_p = \Lambda_B$ in the low traffic scenario, while outperforming both fixed-length preamble solutions for $\Lambda_B = 3000$ pkt/burst.

D. ON THE ESTIMATION OF ν

Note that, all Figs. 2-9 show a negligible performance gap between PIMA (where estimation of ν is subject to errors) and PIMA where users are perfectly counted in the PIA sub-frame. Therefore, the optimal scheduling analysis provided in Section VI nicely approximates the case of imperfect estimation for the considered SNR. In Fig. 9 we have not included the performance of PIMA ideal to improve the readability, while, based on the results shown in Fig. 9, performances are almost identical to those of PIMA.

VIII. CONCLUSION

In this paper, we discussed how the fulfillment of the requirements of mMTC and URLLC use cases in 5G networks in terms of latency and efficiency are impacted by different design choices. To overcome the current limitations of present schemes in these conditions, we proposed the PIMA protocol, a semi-GF multiple access scheme for short packet transmission, based on the knowledge of the number of users that have packets to transmit. We derived the optimal scheduling in the case of i.i.d. activations and assessed its

performance with different users' activation statistics. The numerical results obtained in such scenarios show that PIMA achieves extremely low latency with respect to state-of-the-art OMA multiple access solutions due to its low overhead and adapts to different activation conditions by exploiting the partial knowledge of the instantaneous traffic load.

REFERENCES

[1] A. Rech, S. Tomasin, L. Vangelista, and C. Costa, "Partial-information multiple access protocol for orthogonal transmissions," in *Proc. Int. Conf. Ubiquitous Future Netw. (ICUFN)*, 2023, pp. 271–276, *arXiv:2304.12057*.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. VTC Spring*, 2013, pp. 1–5.

[3] M. Centenaro, L. Vangelista, S. Saur, A. Weber, and V. Braun, "Comparison of collision-free and contention-based radio access protocols for the Internet of Things," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3832–3846, Sep. 2017.

[4] S. Riolo, D. Panno, and L. Miuccio, "Modeling and analysis of tagged preamble transmissions in random access procedure for mMTC scenarios," *IEEE Trans. Commun.*, vol. 20, no. 7, pp. 4296–4312, Jul. 2021.

[5] P. Popovski et al., "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[6] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.

[7] G. Wunder, P. Jung, and C. Wang, "Compressive random access for post-LTE systems," in *Proc. IEEE ICC WKSHPS*, 2014, pp. 539–544.

[8] H. Seo, J.-P. Hong, and W. Choi, "Low latency random access for sporadic MTC devices in Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5108–5118, Jun. 2019.

[9] J. Choi, "On throughput of compressive random access for one short message delivery in IoT," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3499–3508, Apr. 2020.

[10] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[11] S. Ali, N. Rajatheva, and W. Saad, "Fast uplink grant for machine type communications: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 97–103, Mar. 2019.

[12] S. Ali, A. Ferdowsi, W. Saad, and N. Rajatheva, "Sleeping multi-armed bandits for fast uplink grant allocation in machine type communications," in *Proc. IEEE GLOBECOM WKSHPS*, 2018, pp. 1–6.

[13] M. Shehab, A. K. Hagelskjar, A. E. Kalør, P. Popovski, and H. Alves, "Traffic prediction based fast uplink grant for massive IoT," in *Proc. IEEE PIMRC*, 2020, pp. 1–6.

[14] S.-R. Lee, S.-D. Joo, and C.-W. Lee, "An enhanced dynamic framed slotted ALOHA algorithm for RFID tag identification," in *Proc. Int. Conf. Mobile Ubiquitous Syst. Netw. Services*, 2005, pp. 166–172.

[15] J. Su, Z. Sheng, D. Hong, and G. Wen, "An effective frame breaking policy for dynamic framed slotted aloha in RFID," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 692–695, Apr. 2016.

[16] "Study on RAN improvements for machine-type communications," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 37.868, version v.0.8.1, Oct. 2014.

[17] A. E. Kalør, O. A. Hanna, and P. Popovski, "Random access schemes in wireless systems with correlated user activity," in *Proc. IEEE SPAWC*, 2018, pp. 1–5.

[18] F. Moretto, A. Brighente, and S. Tomasin, "Greedy maximum-throughput grant-free random access for correlated IoT traffic," in *Proc. VTC Fall*, 2021, pp. 1–5.

[19] A. Rech and S. Tomasin, "Coordinated random access for Industrial IoT with correlated traffic by reinforcement-learning," in *Proc. IEEE GLOBECOM WKSHPS*, 2021, pp. 1–6.

[20] A. Destounis, D. Tsilimantos, M. Debbah, and G. S. Paschos, "Learn2MAC: Online learning multiple access for URLLC applications," in *Proc. IEEE INFOCOM WKSHPS*, 2019, pp. 1–6.

[21] Y. Chen et al., "Toward the standardization of non-orthogonal multiple access for next generation wireless networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 19–27, Mar. 2018.

[22] X. Meng et al., "Advanced NOMA receivers from a unified variational inference perspective," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 934–948, Apr. 2021.

[23] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[24] Y. Jin and T.-J. Lee, "Throughput analysis of NOMA-ALOHA," *IEEE Trans. Mobile Comput.*, vol. 21, no. 4, pp. 1463–1475, Apr. 2022.

[25] J. Choi, "On throughput bounds of NOMA-ALOHA," *IEEE Commun. Lett.*, vol. 11, no. 1, pp. 165–168, Jan. 2022.

[26] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE ISIT*, 2017, pp. 2523–2527.

[27] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925–2951, May 2021.

[28] A. Decurninge, I. Land, and M. Guillaud, "Tensor-based modulation for unsourced massive random access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 552–556, Mar. 2021.

[29] A. Rech, A. Decurninge, and L. G. Ordóñez, "Unsourced random access with tensor-based and coherent modulations," 2023. *arXiv:2304.12058*.

[30] K.-H. Ngo, A. Lancho, G. Durisi, and A. Graell i Amat, "Unsourced multiple access with random user activity," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4537–4558, Jul. 2023.

[31] Z. Ding, R. Schober, and H. V. Poor, "A new QoS-guarantee strategy for NOMA assisted semi-grant-free transmission," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7489–7503, Nov. 2021.

[32] N. Jayanth, P. Chakraborty, M. Gupta, and S. Prakriya, "Performance of semi-grant free uplink with non-orthogonal multiple access," in *Proc. IEEE PIMRC*, 2020, pp. 1–6.

[33] J. Ding, M. Feng, M. Nemati, and J. Choi, "Performance analysis of massive MIMO assisted semi-grant-free random access," in *Proc. IEEE CCNC*, 2021, pp. 1–7.

[34] A. Munari, "On the value of retransmissions for age of information in random access networks without feedback," in *Proc. IEEE GLOBECOM*, 2022, pp. 4964–4970.

[35] Y. Li and Q. He, "On the ratio of two correlated complex gaussian random variables," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2172–2176, Dec. 2019.

[36] S. Wu, "Moments of complex Gaussian ratios," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 88–91, Jan. 2019.

[37] D. Bertsekas and R. Gallager, *Data Networks*. Nashua, NH, USA: Athena Sci., 2021.

[38] R. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. 33, no. 3, pp. 323–328, May 1987.

[39] J. Choi, "Stability and throughput of random access with CS-based MUD for MTC," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2607–2616, Mar. 2018.

[40] S. Semper, F. Römer, T. Hotz, and G. DelGaldo, "Sparsity order estimation from a single compressed observation vector," *IEEE Trans. Signal Process.*, vol. 66, no. 15, pp. 3958–3971, Aug. 2018.

ALBERTO RECH (Member, IEEE) received the master's degree in ICT for Internet and multimedia engineering from the University of Padova, Italy, in 2020, where he is currently pursuing the Ph.D. degree in information engineering with joint support from the Fondazione Bruno Kessler, Trento, Italy. From September 2019 to August 2020, he was a double degree student with the National Taiwan University, Taipei, Taiwan. From October 2022 to April 2023 and from October to December 2023, he served as a Research Intern with the Advanced Wireless Technology Lab, Huawei, Paris, France. His research interests include signal processing and protocol design for smart electromagnetic environment systems and random access schemes for machine-type communications. In 2023, he was awarded with the Best Workshop Paper Award at the IEEE Wireless Communications and Networking Conference, and the Best Paper Award at the International Conference on Ubiquitous and Future Networks.

STEFANO TOMASIN (Senior Member, IEEE) received the Ph.D. degree from the University of Padova, Padua, Italy, in 2003. During his studies he did internships with IBM Research, Switzerland, and Philips Research, The Netherlands. He joined the University of Padova, where he was an Assistant Professor from 2005 to 2015, an Associate Professor from 2016 to 2022, and has been a Full Professor since 2022. He was a Visiting Faculty with Qualcomm, San Diego, CA, USA, in 2004, Polytechnic University, Brooklyn, NY, USA, in 2007, and the Mathematical and Algorithmic Sciences Laboratory of Huawei, Paris, France, in 2015. His research interests include physical layer security, security of global navigation satellite systems, signal processing for wireless communications, synchronization, and scheduling of communication resources. He was an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGIES from 2011 to 2016 and the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2017 to 2020. He has been an Editor of the *EURASIP Journal of Wireless Communications and Networking* since 2011 and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY since 2020. He has been the Deputy Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY since January 2023. He has been a member of EURASIP since 2011.

LORENZO VANGELISTA (Senior Member, IEEE) received the Laurea and Ph.D. degrees in electrical and telecommunication engineering from the University of Padua, Padua, Italy, in 1992 and 1995, respectively. He subsequently joined the Transmission and Optical Technology Department, CSELT, Torino, Italy. From December 1996 to January 2002, he was with Telit Mobile Terminals, Trieste, Italy, and then, he was with Microcell A/S, Copenaghen, Denmark, until May 2003. In July 2006, he joined the Worldwide Organization of Infineon Technologies as a Program Manager. From October 2006 to October 2021, he was an Associate Professor of Telecommunication with the Department of Information Engineering, Padua University, where he is currently a Full Professor. His research interests include signal theory, multicarrier modulation techniques, cellular networks, and the Internet of Things connectivity, with a special focus on low power wide area networks.

CRISTINA COSTA (Senior Member, IEEE) received the Laurea degree in electronical engineering from the University of Genoa, Genoa, Italy, in 1996, the master's degree in telematics and multimedia applications from COREP, Turin, in 1997, and the Ph.D. degree in information and telecommunications technologies from the University of Trento, Italy, in 2005. She currently holds the position of Head of Research with the Smart and Secure Networks National Laboratory, CNIT, Genoa. She started her career with CSELT, Turin, Italy, in 1997, where she joined the Internet Services Department. In 2004, she joined CREATE-NET, where she was involved in several research projects both at a national and international level, gaining experience both as a researcher and as project coordinator, in the fields of multimedia communications, interfaces and interaction and wireless and mobile networks. From 2020 to 2022, she was with Fondazione Bruno Kessler, where she has been coordinating the Smart Network and Services Research Unit since February 2021. Her current research interests include 5G networks, edge computing, AI, and IoT. She served as a member in the organizing committees of various conferences, in particular of European Wireless, Intetain, and UCMedia. She is an Active Member of the IEEE Women In Engineering AG Italy section.

Open Access funding provided by 'Università degli Studi di Padova' within the CRUI CARE Agreement