Partial-Information Multiple Access Protocol for Orthogonal Transmissions

Alberto Rech*‡, Stefano Tomasin*†, Lorenzo Vangelista*, and Cristina Costa§
*Department of Information Engineering, University of Padova, Italy.
†Department of Mathematics, University of Padova, Italy.
‡Smart Networks and Services, Fondazione Bruno Kessler, Trento, Italy.
§S2N National Lab, CNIT, Genoa, Italy.
alberto.rech.2@phd.unipd.it, stefano.tomasin@unipd.it,
lorenzo.vangelista@unipd.it, cristina.costa@cnit.it

***Abstract*—With the stringent requirements introduced by the new sixth-generation (6G) internet-of-things (IoT) use cases, traditional approaches to multiple access control have started to show their limitations. A new wave of grant-free (GF) approaches have been therefore proposed as a viable alternative. However, a definitive solution is still to be accomplished. In our work, we propose a new semi-GF coordinated random access (RA) protocol, denoted as partial-information multiple access (PIMA), to reduce packet loss and latency, particularly in the presence of sporadic activations. We consider a machine-type communications (MTC) scenario, wherein devices need to transmit data packets in the uplink to a base station (BS). When using PIMA, the BS can acquire partial information on the instantaneous traffic conditions and, using compute-over-the-air techniques, estimate the number of devices with packets waiting for transmission in their queue. Based on this knowledge, the BS assigns to each device a single slot for transmission. However, since each slot may still be assigned to multiple users, collisions may occur. Both the total number of allocated slots and the user assignments are optimized, based on the estimated number of active users, to reduce collisions and improve the efficiency of the multiple access scheme. To prove the validity of our solution, we compare PIMA to time-division multiple-access (TDMA) and slotted ALOHA (SALOHA) schemes, the ideal solutions for orthogonal multiple access (OMA) in the time domain in the case of low and high traffic conditions, respectively. We show that PIMA is able not only to adapt to different traffic conditions and to provide fewer packet drops regardless of the intensity of packet generations, but also able to merge the advantages of both TDMA and SALOHA schemes, thus providing performance improvements in terms of packet loss probability and latency.**

***Index Terms*—Machine-type communications (MTC), Orthogonal multiple access (OMA), Partial-information, Internet-of-things (IoT).**

I. Introduction

Two are the main categories of applications scenarios for machine-type communications (MTC) foreseen to be enabled by fifth-generation (5G) networks: ultra-reliable low-latency communications (URLLC) and massive MTC (mMTC). Both of them show several distinct features and challenges, that stem from the necessity to address the needs of the emerging internet-of-things (IoT) applications and services. The performance requirements of each category are quite challenging, e.g., URLLC use cases target a maximum latency of 1 ms and reliability of 99.99999% (e.g. mission-critical applications), while mMTC scenarios require supporting devices with a density up to 1 million devices per square km (e.g. Industrial IoT). For such demanding targets, that will eventually be more strict in sixth-generation (6G) networks, several technical advances should be adopted at all layers. In particular, designing appropriate multiple access techniques and protocols is particularly relevant due to their impact on the limited resources and capabilities of the transmitting devices. Before the advent of 5G, adopted multiple access protocols were based on resource requests and grants [1] thus incurring on signaling overhead. Grant-free (GF) approaches address this issue by allowing users to transmit the data immediately, without the need to wait for the grant approval of the base station (BS). GF approaches include several different techniques, which can be classified into uncoordinated and coordinated random access (RA) [2].

Uncoordinated RA: these approaches can deal effectively with collisions while requiring limited communication overhead. Users transmit at random time instants, and specific techniques are adopted at the receiver to mitigate the effects of collisions. Among the uncoordinated RA solutions, non-orthogonal multiple access (NOMA) [3] has been widely advocated as the most promising and as an alternative to grant-based orthogonal multiple access (OMA). However, NOMA requires advanced pairing and power allocation techniques, as well as powerful channel coding and interference cancellation mechanisms that only partially mitigate the collision effects. Under these conditions, the BS may become prohibitively complex to serve a large number of users. In recent years, unsourced RA has been proposed as an effective solution to manage a massive number of devices [4]. In this paradigm, at any time, a fraction of devices transmit simultaneously, making use of the same channel codebook. The receiver decodes arriving messages without knowing the identities of the transmitters. Although this approach is very effective in managing many users, good performance can be achieved only for very small payloads and with high-complexity massive multiple input-multiple output (MIMO) receivers [5], [6].

Coordinated RA: these solutions typically divide time into slots, each with the duration of one packet. Slotted ALOHA (SALOHA) is the simplest and most widely adopted coordinated RA protocol: users transmit at the beginning of

the first slot available after packet generation. When collisions occur, a random delay is added before re-transmitting the collided packets. Typically, the random delay has the same statistics for all users, and the coordination is limited to slot synchronization. One of its variants, the framed slotted-ALOHA (FSA) protocol, has been widely adopted in radio-frequency identification (RFID) systems [7], [8]. In FSA, time is divided into frames, and each of these is split into slots. Each user is allowed to transmit in only one slot per frame. Whenever an uplink packet is generated, the user postpones its transmission until the next frame, then it selects a specific slot uniformly at random. Unlike the standard SALOHA, this protocol reduces collisions. Coordinated RA solutions are particularly useful when user activations are highly correlated, for example as a result of correlated underlying traffic generation [9]. Also, re-transmissions (with the consequent accumulation of packets in user queues) may yield correlated transmissions among different users. If on the one hand, such correlation further increases the chances of collisions; on the other hand, it can be exploited to indirectly coordinate RA. In the literature, 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 [10], [11], or learn the traffic correlation by exploiting the capabilities of machine learning tools [12], [13]. Lastly, an extreme case of coordinated RA is *fast uplink grant* [14], wherein the BS schedules one slot for each user, without any resource request. Note that in this case, the access randomness is removed, while coordination remains. Slots are usually shared by multiple users, thus collisions may still occur in case of simultaneous transmissions.

In this paper, we introduce a new semi-GF coordinated RA protocol, named partial-information multiple access (PIMA). In PIMA, time is organized into frames of *variable length*, each divided into two sub-frames. The first is the partial information acquisition (PIA) subframe, where active users (having packets to transmit) send a signal to the BS. Using a compute-over-the-air approach [15]–[17], the BS measures the received power and estimates the number of active users. Based on this knowledge, the BS then assigns one slot to each user for the transmission in the data transmission (DT) sub-frame. We stress that with respect to other two-step RA-access schemes consisting of preamble and data transmission stages [1], PIMA acquires only partial information on the activation statistics in the PIA sub-frame, avoiding to reveal the users' identities.

The rest of the paper is organized as follows. In Section II, we first introduce the system model and the packet generation processes. Then, in Section III, we describe the frame structure and the PIMA protocol. Section IV provides the procedure used to perform the partial information estimation, while Section V presents the scheduling optimization problem conditioned on this information. In Section VI we discuss the numerical results and compare PIMA with conventional time-division multiple-access (TDMA) and SALOHA schedulers. Finally, in Section VII we draw some conclusions.

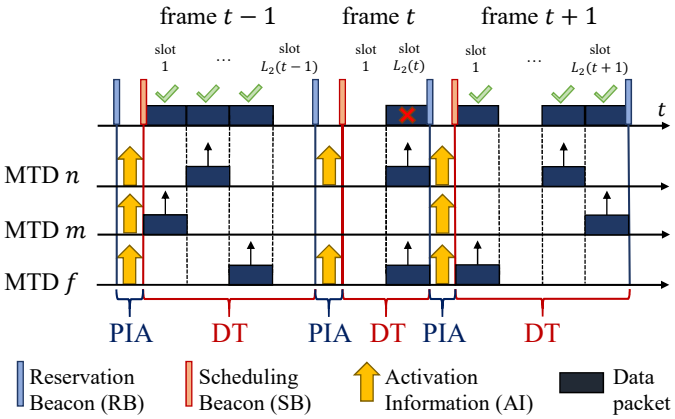

Figure 1. Example of the PIMA protocol and its frame structure.

Notation: Scalars are denoted by italic letters, vectors, and matrices by boldface lowercase and uppercase letters, respectively. Sets are denoted by calligraphic uppercase letters and $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. $\mathbb{P}(\cdot)$ denotes the probability operator and $\mathbb{E}[\cdot]$ denotes the statistical expectation.

II. System Model

We consider an uplink multiple access scenario with K users transmitting in the uplink to a common BS. We assume that the value of K is known at the BS, while this knowledge is not needed by any user.

Time is divided into *frames*, each comprising an integer number of *slots* and an additional short time interval, whose purpose is described in the following. Each slot has a fixed duration, 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. Users transmit packets, each with the duration of a slot, and each user can transmit at most one packet per frame. In the following analysis, τ denotes the generic time instant, t the frame index, and $\tau_0(t)$ the start time of frame t. In the considered setup, the same slot is in general assigned to multiple users for transmission.

Channel: Due to the scheduling of the same slot to multiple users, collisions between packets may occur. We assume that when two or more users transmit in the same slot, a collision occurs, preventing the decoding of all collided packets by the BS. Successful transmissions are acknowledged by the BS at the beginning of the following frame, and in the event of collisions, users retransmit their packets in the following frame. We assume that the BS always correctly decodes the received packets in slots without collision, thus the channel does not introduce other sources of communication errors.

Packet Generation and Buffering: Packets generated in frame t by user k are stored in its buffer and transmitted in the next slots, according to a first in-first out (FIFO) policy. We assume that all users have a limited buffer capacity $\bar{B}$. To ensure data freshness, whenever a new packet is generated while the buffer is at full capacity, the oldest packet is dropped. Let

$B_k(\tau)$ be the length of user k queue at time τ. We also define the *activation vector* as $\boldsymbol{B}(\tau) = [B_1(\tau), B_2(\tau), \ldots, B_k(\tau)]$. If $B_k(\tau) > 0$, the buffer of user k is non-empty at time τ, and k is said to be *active*, instead, if $B_k(\tau) = 0$, its queue is empty and user k is considered *inactive*. The total number of active users at the beginning of frame t is $\nu(t)$. Finally, we define the *activation probability* of user k in frame t as

$$\phi_n(t) = \mathbb{P}(B_k(\tau_0(t)) > 0). \tag{1}$$

III. Partial-Information Multiple Access Protocol

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 is used to estimate (at the BS) the number of currently active users. Based on this information, the BS decides the duration (in slots) of the DT sub-frame and assigns each user to one slot, for possible uplink data transmission. Such scheduling information is transmitted in unicast to each of the users at the end of the PIA sub-frame. An example of PIMA is shown in Fig.1.

A. Partial Information Acquisition Sub-Frame

The beginning of frame t, and thus of the PIA sub-frame, is triggered by the reservation beacon (RB), which is transmitted in broadcast by the BS to all users. RBs are transmitted to mark the start of the PIA sub-frame and contain the acknowledgments of the transmissions that occurred during the previous frame. Moreover, the RB allows each user to estimate the large-scale fading coefficient (LSFC) coefficient of the channel to the BS, denoted as g_k. In the PIA sub-frame the BS obtains the estimate $\hat{\nu}(t)$ of the number of active users $\nu(t)$. While this problem has already been discussed in the literature and solved with deep neural networks [18], here we propose a novel low-complexity estimate inspired by computing over-the-air [15]. For a duration L_1, the users transmit signals to let the BS estimate the number of active users $\nu(t)$. More details on this procedure are provided in Section IV.

At the end of the PIA sub-frame, the BS, knowing $\hat{\nu}(t)$, schedules the transmissions for the next sub-frame. Let $\boldsymbol{q}(t) = [q_1(t), \ldots, q_K(t)]$ be the *slot selection vector*, collecting the slot indices assigned to each user; then, the length $L_2(t)$ of the DT sub-frame can be derived from $\boldsymbol{q}(t)$ as

$$L_2(t) = \max_k q_k(t). \tag{2}$$

To end the PIA sub-frame and trigger the beginning of the following DT sub-frame, the BS transmits the scheduling beacons (SBs), which contains the slot selection vector $\boldsymbol{q}(t)$ [1].

[1]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 the 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.

If a packet is generated by the user k during the DT sub-frame, the packet is delayed and transmitted in the following frame. This feature is needed to ensure low collision probability, as the DT frame length is derived only based on the number of users active in the PIA sub-frame.

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 of duration L_1 immediately after receiving the RB. The transmit power is such that the signal from each user is received with the same (unitary) power at the BS. Note that we are neglecting here the propagation time between the BS and the user, which can be easily accommodated by considering a transition (silent) time between the RB and AI transmissions.

In particular, given a total system bandwidth W, we assume that each user transmits $M_1 = WL_1$ complex Gaussian symbols in the PIA sub-frame with zero mean and power $1/g_k$. The BS then measures the total received power and estimates the number of active users. The set of users transmitting the AI signals during the PIA sub-frame is

$$\mathcal{K}_{\mathrm{a}}(t) = \{k : B_k(\tau_0(t)) > 0\}, \tag{3}$$

with $|\mathcal{K}_{\mathrm{a}}(t)| = \nu(t)$. The BS does not know the identity of the active users, since the AI signals do not contain such information, to make them shorter.

Letting the received samples at frame t be $\tilde{\gamma}_\ell(t)$, $\ell = 1, \ldots, M_1$, the estimated total power is

$$\hat{P}(t) = \frac{1}{M_1}\sum_{\ell=1}^{M_1}\tilde{\gamma}_\ell(t) = \frac{1}{M_1}\sum_{\ell=1}^{M_1}\left| w_\ell(t) + \sum_{k\in\mathcal{K}_{\mathrm{a}}(t)}\gamma_{k,\ell}(t)\right|^2, \tag{4}$$

where $\gamma_{k,\ell}(t)$ is the signal received from user k and $w_\ell(t)$ is the additive white Gaussian noise (AWGN) term with zero mean and variance σ_w^2.

Let us indicate the probability density function (PDF) of the received power given that $\nu(t) = b$ users are active as $p_{\hat{P}|\nu(t)}(a|b)$, and the probability that $\nu(t) = b$ users are active as $p_{\nu(t)}(b)$. The maximum a posteriori probability (MAP) estimate of the number of active users is then

$$\hat{\nu}(t) = \underset{b}{\mathrm{argmax}}\ p_{\hat{P}|\nu}(\hat{P}(t)|b)\, p_\nu(b). \tag{5}$$

The value of $\hat{\nu}(t)$ is obtained from (5) by splitting the set of real numbers into K properly designed intervals $\mathcal{I}(b) = [\epsilon_{b-1}, \epsilon_b]$, $b = 1, \ldots, K$, and finding the region where $\hat{P}(t)$ is falling. Note that the first and last intervals are special, as for the first we have $\mathcal{I}(0) = [0, \epsilon_0]$, while for the last we have $\mathcal{I}(K) = [\epsilon_k, \infty]$. Decision regions may have different lengths, thus providing different estimations according to b.

Optimal decision regions $\mathcal{I}(b)$, $b = 1, \ldots, K$ are intervals with boundaries at the intersections between the adjacent

Gaussian curves; in particular, the boundary ϵ_b between interval $\mathcal{I}(b)$ and $\mathcal{I}(b+1)$ must solve

$$\frac{p_\nu(b)}{\sigma_P(b)\sqrt{2\pi}} e^{-\frac{(\epsilon_b-\bar{P}(b))^2}{\sigma_P^2(b)}} = \frac{p_\nu(b+1)}{\sigma_P(b+1)\sqrt{2\pi}} e^{-\frac{(\epsilon_b-\bar{P}(b+1))^2}{\sigma_P^2(b+1)}}. \tag{6}$$

Equation (6) is equivalent to the quadratic equation $A\epsilon_b^2 + B\epsilon_b + C = 0$, with

$$\begin{aligned} A &= \frac{1}{\sigma_P^2(b+1)} - \frac{1}{\sigma_P^2(b)}, \\ B &= \frac{2\bar{P}(b)}{\sigma_P^2(b)} - \frac{2\bar{P}(b+1)}{\sigma_P^2(b+1)}, \\ C &= \frac{\bar{P}(b+1)^2}{\sigma_P^2(b+1)} - \frac{\bar{P}(b)^2}{\sigma_P^2(b)} - \log\left(\frac{p_\nu(b+1)\sigma_P(b)}{p_\nu(b)\sigma_P(b+1)}\right). \end{aligned} \tag{7}$$

Among the solutions of the quadratic equation, we must choose that falling between $\bar{P}(b)$ and $\bar{P}(b+1)$.

Estimation Error Probability: We define the average error probability

$$\bar{p}_\mathrm{e} = \mathbb{E}[\nu \neq \hat{\nu}] = \sum_b p_\mathrm{e}(b) p_\nu(b), \tag{8}$$

where $p_\mathrm{e}(b) = \mathbb{P}[\hat{\nu} \neq b | \nu = b]$.

Assuming $\gamma_{k,\ell}(t)$ and $w_\ell(t)$ *independent identically distributed (i.i.d.)* $\forall t, k, \ell$, as the square modulus of the sum of complex Gaussian random variables is exponentially distributed, $\hat{P}(t)$ follows an Erlang distribution with shape M_1 and rate $1/(\nu(t)+\sigma_w^2)$. For large values of M_1, $\hat{P}(t)$ can be well approximated as a Gaussian variable with mean $\bar{P}(\nu(t)) = \nu(t) + \sigma_w^2$ and variance

$$\sigma_P^2(\nu(t)) = \frac{[\nu(t)+\sigma_w^2]^2}{M_1}. \tag{9}$$

Then, the conditional estimation error probability is

$$\begin{aligned} p_\mathrm{e}(b) &= \mathbb{P}[\hat{P}(t) \notin \mathcal{I}(b) | \nu(t) = b] \\ &= \mathrm{Q}\left(\sqrt{M_1}\frac{\epsilon_b - b + \sigma_w^2}{(b+\sigma_w^2)}\right) \\ &\quad + \mathrm{Q}\left(\sqrt{M_1}\frac{-(\epsilon_{b-1} - b + \sigma_w^2)}{(b+\sigma_w^2)}\right), \end{aligned} \tag{10}$$

where $\mathrm{Q}(\cdot)$ is the tail distribution function of the standard normal distribution. For $b = 0$ and $b = K$, the first and second terms in (10) are zero, respectively.

To minimize L_1, we should derive the minimum M_1 that guarantees the achievement of a target p_e. However, the computation of the optimal decision regions from (8) is in general quite complex, due to its dependency on $p_\nu(b)$, which in turn depends on the duration of the previous frame(s), as well as on the previous transmission outcomes, therefore being time-variant and strictly dependent on the traffic generation statistics. In the following, for simplicity, we assume that the BS only knows the number of active users and performs the time resource scheduling conditioned on this partial information.

From a practical perspective, since M_1 is a design parameter, it should be time-invariant and should not depend on the user activations statistics. Hence, by defining $\bar{P}(\nu(t)) = \nu(t) + \sigma_w^2$ and choosing $\epsilon_b = \bar{P}(b) + \frac{1}{2}$, $\forall b$ we approximate (10) as

$$p_\mathrm{e}(b) \approx 2\mathrm{Q}\left(\frac{\sqrt{M_1}}{2(b+\sigma_w^2)}\right). \tag{11}$$

As $\mathrm{Q}(\cdot)$ is a monotonically decreasing function, the maximum estimation error probability is achieved if all users are active in the frame t. Then, we derive M_1 in this worst-case scenario, which provides a target error probability $\tilde{p}_\mathrm{e} = p_\mathrm{e}(K)$, as

$$M_1 = \left[2(K+\sigma_w^2)\mathrm{Q}^{-1}\left(\frac{\tilde{p}_\mathrm{e}}{2}\right)\right]^2. \tag{12}$$

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. Let $l \in \{1, \ldots, L_2(t)\}$ be the slot index within frame t (in the DT sub-frame). We define the success indicator function at slot l as $c_l = 1$, if a successful transmission occurs at slot l and $c_l = 0$ otherwise. Then, the *conditional frame efficiency* is defined as the ratio between the number of successes in frame t and the length of the DT sub-frame, i.e.,

$$\eta(t) = \frac{1}{L_2(t)} \sum_{l=1}^{L_2(t)} \mathbb{E}[c_l | \nu(t)]. \tag{13}$$

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

In frame t, immediately after the end of the PIA, the BS solves the following optimization problem:

$$\max_{\boldsymbol{q}(t)} \eta(t), \tag{14a}$$

$$\text{s.t.} \quad q_k(t) \in \{1, \ldots, L_2(t)\}. \tag{14b}$$

Note that, without making any assumption on the activation probability distribution, the problem is not solvable in closed form; therefore we focus on the case of i.i.d. activation probabilities are equally distributed, without considering the correlations caused by the retransmission attempts and the packets generated during the DT 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 and thus the same expected frame efficiency.

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

$$u_l(t) = \begin{cases} \left\lceil \frac{K}{L_2(t)} \right\rceil & \text{if} \quad l \leq K \bmod L_2(t), \\ \left\lfloor \frac{K}{L_2(t)} \right\rfloor & \text{if} \quad l > K \bmod L_2(t), \end{cases} \tag{15}$$

where we may schedule one more user in the first $\left\lfloor \frac{K}{L_2(t)} \right\rfloor$ slots to minimize the transmission delay.

Note that the slot success random variable c_l can be rewritten as a function of $u_l(t)$, as it only depends on the number of users scheduled in slot l. Now, the optimization problem (14) is reduced to the optimization of the second sub-frame length, $L_2(t)$, i.e., from (13), we have

$$L_2^*(t) = \underset{L_2(t)}{\operatorname{argmax}} \frac{1}{L_2(t)} \sum_{l=1}^{L_2(t)} \mathbb{E}[c_l|\nu(t), u_l(t)], \tag{16a}$$

$$\text{s.t. } L_2(t) \in \mathbb{K}/\{0\}. \tag{16b}$$

Now, given $\nu(t)$, the probability of user k being the one and only active user assigned to slot l is derived by considering all cases of active users, where user k is active and all other users assigned to slot l are, instead, inactive. The number of favorable cases is given by all the possibilities to put $\nu(t)-1$ objects on a chessboard with $K-u_l(t)$ places, i.e., all the combinations of $\nu(t)-1$ elements taken from $K-u_l(t)$. For each combination, there are $K-u_l(t)$ possibilities to put the active user in slot l. Therefore, the collision probability in slot l is $1-\mathbb{E}[c_l|\nu(t), u_l(t)]$, where

$$\mathbb{E}[c_l|\nu(t), u_l(t)] = \frac{u_l(t)\binom{K-u_l(t)}{\nu(t)-1}}{\binom{K}{\nu(t)}}, \tag{17}$$

is the probability of having a successful transmission in slot l. Note that, in (17), the numerator 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.

Note that (16) is a mixed integer non-linear programming (MINLP) problem, and is not solvable by continuous relaxation of $L_2(t)$, as the rounding functions are not differentiable. However, it is possible to find the optimal frame length $L_2^*(t)$ with complexity $O(\log K)$, using a binary search algorithm. In any case, $L_2^*(t)$ depends only on $\nu(t)$, thus can be computed offline and then stored in a table.

VI. Numerical Results

In this section, we present the numerical results, comparing our PIMA approaches with the TDMA and SALOHA schedulers.

We assume that the traffic generation, also denoted as *packet arrival process*, at each user follows a Poisson distribution with parameter λ. We consider $K = 20$ users. The well-known properties of the Poisson processes provide a total arrival rate of $\Lambda = K\lambda$. For a performance comparison, we consider a) the standard TDMA, which provides frames of fixed duration of K slots, with one user assigned per slot, deterministically, and b) the SALOHA protocol. In basic SALOHA, in every time interval l, users transmit their packets immediately upon generation unless they are *backlogged* after a collision, in which case they transmit with a backoff probability. Instead, we consider Rivest's stabilized SALOHA [19, Chapter 4], wherein all users generating packets in slot l are backlogged with the same backoff probability. The backoff probability is computed for each user through a pseudo-Bayesian algorithm

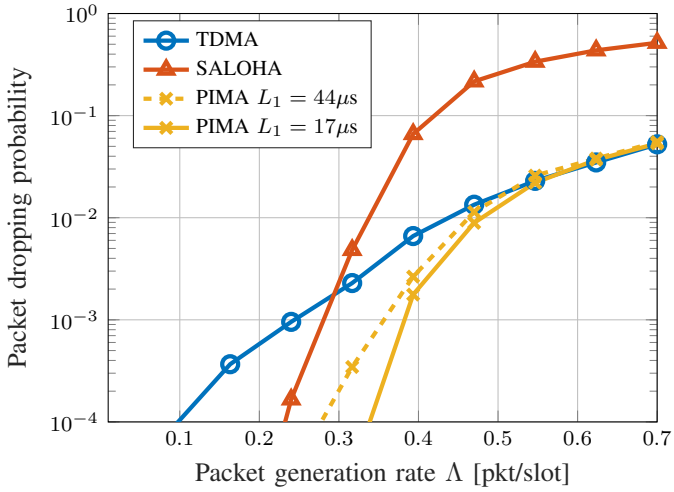

Figure 2. Average packet dropping probability versus the total packet generation rate, for $K = 20$ and $\bar{B} = 3$.

based on an estimate of the number of backlogged nodes $G(l)$ as

$$\alpha(l) = \min\left(1, \frac{1}{G(l)}\right), \tag{18}$$

where

$$G(l) = \begin{cases} G(l-1) + N\theta + (e-2)^{-1} & \text{if } c_{l-1} = 0, \\ \max(N\theta, G(l-1) + N\theta - 1) & \text{if } c_{l-1} = 1, \end{cases} \tag{19}$$

is the estimated number of users backlogged (with $G(0) = 0$) and θ is the probability packet generation in slot l. Note that, despite being conventional multiple access solutions, SALOHA and TDMA remain the ideal solutions for OMA in the time domain in case of low and high traffic conditions, respectively. In the following, the comparison is made in terms of packet loss probability and average packet latency.

A. PIA Parameters and Results

Although the analysis of Section V assumes a perfect estimation of the number of active users, results shown in this section are obtained using the estimated $\hat{\nu}(t)$, thus including the effects of an estimation error.

In the PIA sub-frame, M_1 has to be large enough to make the approximation (12) valid and ensure low estimation error probability (11). During the PIA sub-frame we assume that AI signals transmitted by the active users have unitary power, while the noise power is $\sigma_w^2 = -10$ dB. Furthermore, we adopt the third numerology of the new radio (NR) specification, which provides DT time slots of 0.125 ms [20], and assume a total system bandwidth $W = 100$ MHz. Moreover, we neglect the duration of downlink beacons. In the following, either $L_1 = 17$ μs or 44 μs, such that the target error probability (14) is either $\tilde{p}_e = 0.1$ or $\tilde{p}_e = 0.3$, respectively.

In Fig. 2, we show the empirical packet dropping probability as a function of the packet generation rate Λ, for $K = 20$ users with buffer length $\bar{B} = 3$. Such metric measures the probability of packet replacement in the buffers when fresher

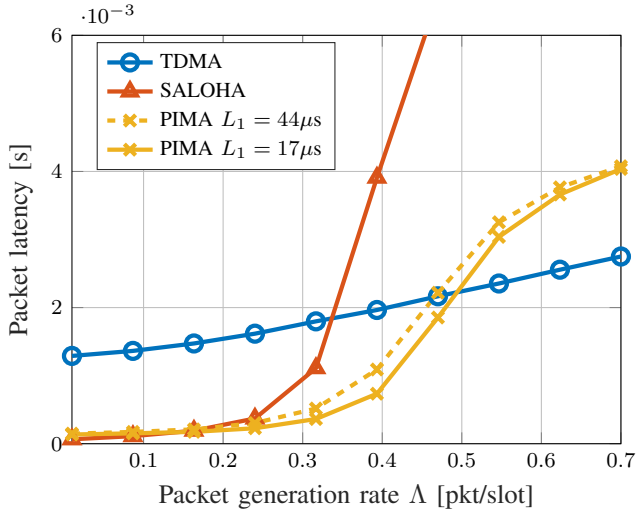

Figure 3. Average packet latency versus the total packet generation rate, for $K = 20$ and $\bar{B} = 3$.

packets are generated at users. In low-traffic conditions, very few packets are dropped for both SALOHA and PIMA. A higher number of packets are dropped with TDMA, as the constantly adopted maximum frame length leads to time waste. In high-traffic conditions, instead, the dropping probability provided by SALOHA keeps growing, performing even worse than TDMA, as the backlogging system prevents most users from transmitting to avoid collisions. The advantages of both conventional schemes are merged in PIMA, which is able to adapt to traffic conditions to provide fewer packet drops regardless of the intensity of packet generation. In the worst-case scenario, PIMA matches TDMA, with a slight gap due to the overhead of PIA.

Finally, Fig. 3 shows the impact of the packet generation rate on the average latency. As expected, the TDMA and SALOHA protocols perform well at low and high packet generation rates, respectively. In fact, at low traffic intensity TDMA suffers from a long waiting due to deterministic slot allocation. SALOHA instead has a very low latency at low traffic, as packets are immediately transmitted, while more collisions occur as the probability of packet generation increases, increasing the latency. Again, PIMA merges the two advantages to achieve great results in all traffic conditions. However, a slight performance degradation is observed at high traffic, due to both the PIA overhead and the additional delay introduced when the packet is delayed to the next frame upon generation.

Finally, note that as L_1 decreases, packet loss and latency are reduced. However, the short PIA duration of the subframe could imply a lower M_1, leading to erroneous estimation of $\nu(t)$, and thus making the BS perform inefficient scheduling.

VII. Conclusions

For addressing the challenging requirements of the emerging 5G/6G IoT use cases, we have proposed a new semi-GF coordinated multiple access scheme, the PIMA protocol, based on the knowledge of the number of users that have packets to transmit. To this end, PIMA organizes time into frames, and each frame includes a preliminary phase (the PIA sub-frame), wherein active users transmit a compute-over-the-air signal that enables the BS to estimate the number of active users. Then, we analyzed the protocol, indicating the policies to be used to minimize packet losses and latency. From the analysis and the numerical results obtained in a generic MTC scenario, we conclude that PIMA with proper scheduling significantly reduces packet loss and latency with respect to TDMA and SALOHA, particularly when dealing with sporadic activations.

References

[1] 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.

[2] M. El-Tanab and W. Hamouda, "An overview of uplink access techniques in machine-type communications," *IEEE Network*, vol. 35, no. 3, pp. 246–251, Mar. 2021.

[3] 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. IEEE VTC Spring*, 2013, pp. 1–5.

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

[5] 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.

[6] 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, Oct. 2020.

[7] 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. on Mobile and Ubiquitous Systems: Networking and Services*, 2005, pp. 166–172.

[8] 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.

[9] 3GPP, "Study on RAN improvements for machine-type communications," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 37.868, 10 2014, version v.0.8.1.

[10] 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.

[11] 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.

[12] A. Rech and S. Tomasin, "Coordinated random access for industrial IoT with correlated traffic by reinforcement-learning," in *Proc. IEEE Globecom Workshops*, 2021, 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. 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.

[15] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3863–3877, Sep. 2013.

[16] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5488–5502, Aug. 2020.

[17] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive IoT," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 57–65, Apr. 2021.

[18] M. U. Khan, E. Paolini, and M. Chiani, "Enumeration and identification of active users for grant-free NOMA using deep neural networks," *IEEE Access*, vol. 10, pp. 125 616–125 625, Oct. 2022.

[19] D. Bertsekas and R. Gallager, *Data networks*. Athena Scientific, 2021.

[20] 3GPP, "5G-NR; physical channels and modulation," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.211, 7 2020, version v.16.2.0.