A Flexible Scheme for Critical mMTC

A Flexible Scheme for Critical mMTC

Alessandro Mirri†, Diego Forlivesi†, Riccardo Schiavone×, Roberto Garello×, Marco Chiani†, Enrico Paolini†
† CNIT/WiLab, DEI, University of Bologna, Italy
Email: {alessandro.mirri7, diego.forlivesi2, marco.chiani, e.paolini}@unibo.it
× DET, Politecnico di Torino, Italy
{riccardo.schiavone, roberto.garello}@polito.it

***Abstract*—In this paper, a grant-free access scheme for massive machine-type communications (mMTC), along with a multiuser detection (MUD) aided processing for packet recovery, are investigated. The proposed grant-free access and receiver processing are flexible, in that they can be applied both in a slotted but unframed configuration and in a slotted and framed one, to achieve different tradeoff points between scalability, reliability, and latency. This looks particularly interesting for emerging "critical" mMTC applications, where reliability and latency requirements are more tightening than current 5G ones. An in-depth analysis of the considered scheme, performed through both numerical simulations and analytical tools, highlights a potential to support mMTC with diverse requirements. For example, reliable grant-free transmissions can be supported (e.g., packet-loss rates in the order of $10^{-4}/10^{-5}$) with an application-specific latency-scalability tradeoff.**

I. Introduction

The rise of the Internet of Things (IoT) has prompted an increasing attention on machine-type communications (MTC), meant as automated communication among devices, exchanging data over the network without any human intervention [1], [2]. More recently, as a consequence of the increasing pervasiveness of IoT in several application domains (e.g., smart factories, smart grid and metering, autonomous driving, public health), the expression massive machine-type communications (mMTC) has been coined to refer to situations where the density of connected devices becomes extremely large [3]–[5].

Machine-type devices in mMTC services have low duty cycle and generate data in a bursty, intermittent, and unpredictable way, where activity periods are used for transmission of one message, in the form of a short packet, to another device or to a remote server through the network. The main performance metrics in 5G mMTC services are device battery lifetime and scalability, the latter being the capability of the system to support increasing IoT node densities; requirements on reliability and latency are instead relaxed. A somewhat dual 5G service is ultra-reliable and low-latency communication (URLLC) [6], [7]. Also in URLLC machine-type devices transmit short packets, however, reliability and latency requirements are very severe while the device density is much smaller compared with mMTC.

The mMTC-URLLC duality affects not only the physical (PHY) layer, but several other functions of the communication stack, including the medium access control (MAC) layer. While URLLC traffic is subject to scheduling to avoid packet collisions that heavily deteriorate reliability and latency, grant-free access is attracting interest for mMTC, owing to its lightness on the device side and benefits on the device battery lifetime. The growing popularity of grant-free access is witnessed, for example by 5G-NR early data transmission (EDT) mode [8].

Recently, new emerging IoT use cases are however calling into question the rigid mMTC-URLLC separation. Such use cases, a prominent example of which is industrial IoT (besides, e.g., vehicle-to-infrastructure communications, and smart city) require convergence between the two services, with different trade-off points between scalability, latency, and reliability, depending on the specific scenario. With reference to convergent services, *critical mMTC* refers to mMTC services where scalability remains the main performance metric, but with relatively tightening (although non-URLLC) reliability and latency requirements [9].

Although there are attempts to cope with convergent IoT services, including 5G-NR Reduced Capacity (RedCap) [10], flexible and concurrent support to mMTC and URLLC remains an open problem. For this reason, in this paper we investigate a multi-user detection (MUD) scheme applicable in critical mMTC scenarios. We firstly propose a grant-free random access protocol that can flexibly switch between a slot-based and a frame-based configuration, depending on the latency requirements. Furthermore, we present a low-complexity MUD-aided decoder capable of de-multiplexing data streams of interfering active users. Analytical lower bounds have been derived and numerical simulations have been conducted in order to evaluate the performance of the investigated scheme.

The paper is organized as follows. The system model is introduced in Section II. The proposed MUD-aided processing and decoding is presented in Section III. The achieved performance is illustrated in Section IV via both analytical methods and Monte Carlo simulation. Finally, conclusions are drawn in Section V.

II. System Model

We consider a slotted and (possibly) framed massive multiple access (MMA) uplink system populated by a large number of machine-type devices, referred to as "users", that wake up sporadically and unpredictably to transmit short data packets in a grant-free manner to a common receiver (Rx). The Rx (base station or access point) may be equipped with multiple antennas, in the number of M, depending on the application. Out of the overall population, only a subset of the users,

whose number is a priori unknown at the receiver, starts the contention procedure at the beginning of each slot/frame. The system makes use of orthogonal subcarriers, in the number of K, according to an orthogonal frequency-division multiplexing (OFDM) transmission scheme [11]. The system configuration is flexible and, dependently on the scenario, switches from a *slot-based* access protocol (i.e., slotted ALOHA [12], [13]), to a *frame-based* one (i.e., coded slotted ALOHA [14]). Users are frame- and slot-synchronous owing to the presence of a beacon broadcast by the Rx at the beginning of the transmission, which may also be used for power control.

A. Slot-based Configuration

In the slot-based configuration, U_{S} users (with U_{S} random and unknown to the Rx) contend in a slot for transmission of one information message $\mathsf{W} \in \{1, \ldots, 2^{\kappa}\}$ each, where κ is the number of message bits. All devices employ the same channel code and modulation. Each active user processes its message with the channel encoder of rate R_{c} and maps its encoded bits to a quadrature phase-shift keying (QPSK) constellation. Then, it randomly chooses f subcarriers (or "resources") out of the K available ones and, for each of them, randomly picks up a pilot sequence from a pool of N_{P} orthogonal pilots[1]; this choice defines a resource-pilot pair. Finally, for each of the f selected resources, it transmits a packet composed of the concatenation of pilot symbols (possibly different from resource to resource) and payload symbols (identical in all the f subcarriers). The frequency-repetition rate f is the same for all users. At the Rx, the decoding procedure occurs in a slot-by-slot fashion and makes use of an intra-slot successive interference cancellation (SIC) procedure to recover multiple packets per slot.

B. Frame-based Configuration

In the frame-based configuration, U_{F} users (with U_{F} random and unknown to the Rx) are active at the beginning of a frame, composed of N_{S} slots, and contend for transmission of an information packet each, as follows. The transmission procedure is analogous to the one described for the slot-based configuration until mapping of bits to QPSK symbols. Then, each user randomly chooses r slots out of the N_{S} available ones and, for each such slot, it randomly selects f resources. Again, for each of them it picks a pilot sequence and appends it to the data payload prior to transmission. Overall, each user transmits a total of $r \cdot f$ packet replicas, exploiting both frequency diversity (with repetition degree f) and time diversity (with repetition degree r). This is pictorially represented in Fig. 1. In principle, the repetition degree r could be the same for all the users transmitting in the frame [15] or it may follow a certain probability distribution from a set of possible ones [16]. The Rx performs both an intra-slot SIC (analogous to the one executed in the slot-based scheme) and an inter-slot SIC procedure, exploiting the time diversity offered by packet replicas across slots.

[1]In each resource, the generic user may pick up any of the available pilots. The set of pilots is known at the Rx.

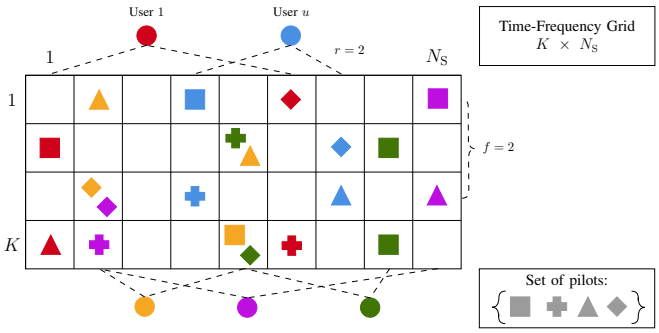

Fig. 1. Pictorial representation of the frame-based transmission scheme. In the example, each user picks $r = 2$ slots and $f = 2$ subcarriers for packet transmission. As for each user the payload sequence is the same in all the selected resources, we only display the drawn pilot sequences.

C. Channel Model

We consider a Rayleigh block fading channel model where the channel coherence time equals the slot time. Thus, all the packet symbols from the same user experience the same channel gain over the same resource. For a positive integer K, let $[K] = \{1, \ldots, K\}$. Let also $\mathcal{A}_k$ be the set of users active in resource $k \in [K]$ of a generic slot. The received signal in subcarrier k is $[\boldsymbol{P}_k, \boldsymbol{Y}_k] \in \mathbb{C}^{M \times (N_{\mathrm{P}} + N_{\mathrm{D}})}$ where

$$\boldsymbol{P}_k = \sum_{j \in \mathcal{A}_k} \boldsymbol{h}_k^j \boldsymbol{s}_k^j + \boldsymbol{Z}_k \tag{1}$$

and

$$\boldsymbol{Y}_k = \sum_{j \in \mathcal{A}_k} \boldsymbol{h}_k^j \boldsymbol{x}^j + \boldsymbol{W}_k. \tag{2}$$

In (1) and (2), $\boldsymbol{h}_k^j = (h_{k,1}^j, \ldots, h_{k,M}^j) \in \mathbb{C}^{M \times 1}$ is the vector of channel coefficients in resource k of user $j \in \mathcal{A}_k$, whose elements are independent and identically distributed (i.i.d.) random variables with complex Gaussian distribution $\mathcal{CN}(0, \sigma_{\mathrm{h}}^2)$. Assuming perfect power control, without loss of generality we take $\sigma_{\mathrm{h}}^2 = 1$. $\boldsymbol{s}_k^j \in \mathbb{C}^{1 \times N_{\mathrm{P}}}$ is the pilot transmitted by user $j \in \mathcal{A}_k$ in resource k, while $\boldsymbol{x}^j \in \mathbb{C}^{1 \times N_{\mathrm{D}}}$ is user j data payload. Finally, $\boldsymbol{Z}_k \in \mathbb{C}^{M \times N_{\mathrm{P}}}$ and $\boldsymbol{W}_k \in \mathbb{C}^{M \times N_{\mathrm{D}}}$ are matrices of noise samples distributed as $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2)$. Recall that r is the number of replicas across slots, f is the number of employed resources per slot, M is the number of receiver antennas, and R_{c} is the code rate; letting Q be the constellation order, E_{b} be the energy per message bit and N_0 be the one-sided noise power spectral density, we can write

$$\frac{E_{\mathrm{b}}}{N_0} = \frac{r f M}{R_c \left(\log_2 Q\right) \sigma_{\mathrm{n}}^2}. \tag{3}$$

We remark that in the slot-based access scheme (Section II-A) we have $r = 1$.

III. MUD-Aided Decoder for Critical mMTC

Regardless of the system configuration, slotted or framed, the Rx processing is divided into two sequential steps, referred to as initialization (or "init") phase and SIC phase. Some parts of the decoding algorithm have been derived from [17],

Algorithm 1: Slot-by-Slot Processing (Init Phase)

input : $[\boldsymbol{P}_k, \boldsymbol{Y}_k] \in \mathbb{C}^{M\times(N_\mathrm{P}+N_\mathrm{D})}$, for all $k \in [K]$

$\mathcal{L} \leftarrow \emptyset$;
flagSlotProcessing $= 1$;
while flagSlotProcessing $== 1$ **do**
  flagSlotProcessing $= 0$;
  forall $k \in [K]$ **do**
    $\mathcal{S}_k \leftarrow \emptyset$; $\delta_k = 0$;
    forall $i \in [N_\mathrm{P}]$ **do**
      $\Lambda(k,i) \leftarrow$ `pilot_detection`$(\boldsymbol{s}_i)$;
      if $\Lambda(k,i) > \eta$ **then**
        $\mathcal{S}_k = \mathcal{S}_k \cup \{\boldsymbol{s}_i\}$;
    $\delta_k = |\mathcal{S}_k|$;
    if $\delta_k \leqslant T$ **then**
      forall $\ell \in [\mathcal{S}_k]$ **do**
        $\boldsymbol{\phi}_k^\ell \leftarrow$ `channel_estim`$(\boldsymbol{P}_k, \boldsymbol{s}_\ell)$;
      $[\hat{\boldsymbol{x}}_k^{u_1}, \ldots, \hat{\boldsymbol{x}}_k^{u_{\delta_k}}] \leftarrow$ `MUD_detection`$(\boldsymbol{Y}_k, [\boldsymbol{\phi}_k^{(\boldsymbol{s}_1)}, \ldots, \boldsymbol{\phi}_k^{(\boldsymbol{s}_{\delta_k})}])$;
      forall $\hat{\boldsymbol{x}} \in [\hat{\boldsymbol{x}}_k^{u_1}, \ldots, \hat{\boldsymbol{x}}_k^{u_{\delta_k}}]$ **do**
        $[\hat{\mathsf{W}}, \mathrm{ACK}] \leftarrow$ `BCH_decoder`$(\hat{\boldsymbol{x}})$;
        if ACK **then**
          if $\hat{\mathsf{W}} \notin \mathcal{L}$ **then**
            $w = |\mathcal{L}| + 1$;
            $\mathcal{L}_w = \hat{\mathsf{W}}$;
            $\mathcal{F}_{w,k} = 0$; $\forall k = 1, \ldots, K$
          else
            $w =$ `find_w`$(\mathcal{L}, \hat{\mathsf{W}})$;
          $\mathcal{F}_{w,k} = 1$;
          flagSlotProcessing $= 1$;
          $\boldsymbol{P}_k = \boldsymbol{P}_k - \boldsymbol{\phi}_k^{(\boldsymbol{s})} \boldsymbol{s}$;
          $\boldsymbol{Y}_k = \boldsymbol{Y}_k - \boldsymbol{\phi}_k^{(\boldsymbol{s})} \hat{\boldsymbol{x}}$;

Algorithm 2: Slot-by-Slot Processing (SIC Phase)

input : $\mathcal{L} = \{\hat{\mathsf{W}}_1, \ldots, \hat{\mathsf{W}}_{|\mathcal{L}|}\}$, $\mathcal{F}$

forall $w \in [\mathcal{L}]$ **do**
  $\mathsf{R} \leftarrow$ `msg2resource`$(\mathcal{L}_w)$;
  forall $k \in \mathsf{R}$ **do**
    if $\mathcal{F}_{w,k} = 0$ **then**
      $\hat{\boldsymbol{h}}_k = \boldsymbol{Y}_k \hat{\boldsymbol{x}}^H / ||\hat{\boldsymbol{x}}^H||^2$;
      $\boldsymbol{P}_k = \boldsymbol{P}_k - \hat{\boldsymbol{h}}_k \boldsymbol{s}$;
      $\boldsymbol{Y}_k = \boldsymbol{Y}_k - \hat{\boldsymbol{h}}_k \hat{\boldsymbol{x}}$;

where a low-complexity decoder for an SCMA-based grant-free access scheme has been elaborated.

A. Init Phase

During the init phase, the Rx systematically processes the signal samples received in each resource of the current slot, employing a MUD technique together with channel decoding to retrieve packets transmitted by active users. The procedure executed in each slot, summarized in Algorithm 1, is iterated until no more packets can be decoded from the slot.

1) Pilot Detection: In each resource $k \in [K]$, the pilot detection procedure (`pilot_detection` in Algorithm 1) allows detecting the pilot sequences transmitted by the users in the resource. Since pilot detection must precede channel estimation, non-coherent detection shall be performed. A possible solution consists of adopting a generalized likelihood ratio test (GLRT) approach. Concretely, this consists of performing, for each pilot $i \in [N_\mathrm{P}]$ the test

$$\Lambda(k,i) = \left\| \frac{\boldsymbol{P}_k \boldsymbol{s}_i^\mathrm{H}}{\|\boldsymbol{s}_i\|^2} \right\|^2 \underset{\mathcal{D}_0}{\overset{\mathcal{D}_1}{\gtrless}} \eta \tag{4}$$

where $\mathcal{D}_0$ and $\mathcal{D}_1$ are the decisions for the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding to noise only and noise plus pilot samples, respectively, $\boldsymbol{P}_k$ is the pilot signal matrix received in the k-th resource, $\boldsymbol{s}_i$ is the i-th pilot and η is the test threshold. We remark that the GLRT threshold η can be fixed by analysis, to achieve a given tradeoff between false alarm and misdetection probabilities.

The pilot detection step is of fundamental importance for the overall procedure, since a correct estimation of the number of users supposed to be active in a resource heavily affects the MUD algorithm effectiveness. At the end of the pilot detection phase the Rx possesses a list $\mathcal{S}_k$ of the estimated transmitted pilots for each resource $k \in [K]$. Denoting by δ_k the cardinality of this list, the Rx estimates the number of active users in the resource as δ_k. Consequently, if a collision occurs (i.e., two users adopt the same pilot in the same resource), the number of active users is underestimated.

2) Channel Estimation: Following the expression in (4), the Rx performs a maximum likelihood (ML) channel estimation (`channel_estim` in Algorithm 1) for each pilot $\ell \in \mathcal{S}_k$, as

$$\boldsymbol{\phi}_k^\ell = \frac{\boldsymbol{P}_k \, \boldsymbol{s}_\ell^\mathrm{H}}{\|\boldsymbol{s}_\ell\|^2} = \sum_{j \in \mathcal{A}_k^\ell} \boldsymbol{h}_k^j + \boldsymbol{z}_k \,. \tag{5}$$

In (5), $\boldsymbol{\phi}_k^\ell = (\phi_{k,1}^\ell, \ldots, \phi_{k,M}^\ell) \in \mathbb{C}^{M\times 1}$ is the vector of estimated channel coefficients and $\mathcal{A}_k^\ell$ is the subset of active users employing pilot ℓ in resource k. When there is only one such user ($|\mathcal{A}_k^\ell| = 1$), $\boldsymbol{\phi}_k^\ell$ is an ML estimate of the user's channel gain in that resource.

3) MUD-Aided Detection: The Rx attempts the detection of superimposed users through an MUD procedure (`MUD_detection` in Algorithm 1) that is iteratively performed for each received payload symbol $d \in [N_\mathrm{D}]$ and that consists of an ML decision performed over a super-constellation in the complex plane; this super-constellation is built by the Rx for each antenna $m \in [M]$ based on the previously estimated channel gains. More in detail, each super-constellation point is obtained as a linear combination

of δ_k QPSK symbols, one for each user supposed active in that resource, rotated and scaled by the corresponding channel coefficient estimate. Hence, the super-constellation order is Q^{δ_k}, where Q is the single-user constellation order (e.g., $Q = 4$ for QPSK). Next, the Rx computes and stores the distances between the d-th received payload sample in resource k and each super-constellation point. This *likelihood* metric is obtained for each receiving antenna and the final decision is made merging the metrics of all antennas. The above-described processing could become computationally onerous as the super-constellation order grows exponentially with δ_k. Hence, we set a parameter T that determines the maximum number of users that can potentially be detected in each resource by the MUD procedure. Such parameter can be tuned to trade-off performance and complexity.

Explicitly, let $\hat{\mathcal{A}}_k = \{u_1, \ldots, u_{\delta_k}\}$ be the set of detected users superimposed in resource k. If the estimated number of active users in the resource, δ_k, fulfills $\delta_k > T$, no detection is triggered. Otherwise, if $\delta_k \leqslant T$, the Rx estimates the channel coefficients vectors and builds a super-constellation with Q^{δ_k} points for each antenna $m \in [M]$. Here, all possible Q^{δ_k} sums of payload symbols, each multiplied by the corresponding estimated channel gain, are generated. Based on the computed super-constellation, the receiver executes an ML decision for each symbol $n_{\mathrm{d}} \in [N_{\mathrm{D}}]$, as

$$(\hat{x}_{k,n_{\mathrm{d}}}^{u_1}, \ldots, \hat{x}_{k,n_{\mathrm{d}}}^{u_{\delta_k}}) = \underset{\tilde{\boldsymbol{x}}^{u_1},\ldots,\tilde{\boldsymbol{x}}^{u_{\delta_k}}}{\operatorname{argmin}} \sum_{m=1}^{M} \left| y_{k,n_{\mathrm{d}}}^{m} - \sum_{i=1}^{\delta_k} \phi_{k,m}^{u_i} \tilde{x}^{u_i} \right|^2, \quad (6)$$

where $\tilde{\boldsymbol{x}} \in \mathbb{C}^{1\times Q}$ is the vector of QPSK symbols, $y_{k,n_{\mathrm{d}}}^{m}$ is the received data symbol at antenna m and $\phi_{k,m}^{u_i}$ is the channel coefficient estimate on the m-th antenna for the i-th user.

4) BCH Decoding: After collecting all payload samples, for each user $u_i \in \hat{\mathcal{A}}_k$ the estimated sequence $\hat{\boldsymbol{x}}_k^{u_i} \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ undergoes decoding using a $[n, \kappa, d]$ BCH decoder (`BCH_decoder` in Algorithm 1). If the number of channel errors is fewer than the error-correction capability, denoted as $t = \lfloor (d-1)/2 \rfloor$, the codeword is successfully recovered and the decoder returns an acknowledge message (ACK). In this case both the estimated user payload $\hat{\mathsf{W}}$, and its corresponding decoding resource k are stored to allow further processing during the successive interference cancellation phase. Finally, the contribution of the decoded user is subtracted from the received signal, both in its pilot and payload components (lines 27 and 28 of Algorithm 1). In order to do this, the Rx employs the channel estimate performed with the corresponding pilot picked by the user in that resource (pilot-based channel estimation).

5) List Updating: The Rx employs two supporting lists, dubbed $\mathcal{L}$ and $\mathcal{F}$, to store the information that will be used during the SIC procedure. The list $\mathcal{L}$ contains the users payload sequences estimated during the init phase, while $\mathcal{F}$ stores the subcarrier indexes in which each payload message has been decoded. Thus, if a payload has been correctly estimated (i.e., the ACK is sent from the BCH decoder), the Rx scrolls through $\mathcal{L}$ checking if that payload has been previously detected in other resources. If the check is unsuccessful, the Rx adds a new row in $\mathcal{L}$ (index $w = |\mathcal{L}| + 1$) and appends $\hat{\mathsf{W}}$ to the list. Then, it also generates a new row in the list $\mathcal{F}$. Otherwise, it stores in the index w the position of the estimated payload $\hat{\mathsf{W}}$ in the list $\mathcal{L}$ (`find_w` in Algorithm 1). In the end, the Rx stores in $\mathcal{F}$ the resource index k in which the estimated payload has been lastly decoded ($\mathcal{F}_{w,k} = 1$).

B. SIC Phase

During this step, the information extracted from the decoded payloads is employed to subtract the residual interference contribution with the chance to decode new packets, previously not found due to unsolvable events such as pilot collisions. This process is outlined in Algorithm 2. Specifically, for each payload of the overall ones stored in the list $\mathcal{L}$, the Rx gets the information about the resources in which the user has transmitted[2] (`msg2resource` in Algorithm 2). Then, it checks if any of the employed resources is not in the list $\mathcal{F}$, i.e., the packet has not been previously decoded in that resource ($\mathcal{F}_{w,k} = 0$). If that is the case, the channel estimation is performed through the decoded payload and the interference contribution (both in pilot and payload received signal) is successively removed [18]. Since that the payload sequences are not orthogonal, the payload aided based (PAB)-SIC removes imperfectly the interference, leaving a residual term over the received signal. As one possible solution, the availability of multiple antenna elements at the Rx (with the consequent "favorable propagation" effect [19], [20]) allows the decoding of multiple packets in the same slot. The above-mentioned SIC technique can be easily extended to the frame-based scenario, including for each transmitted user payload the information about the chosen slots. Again, the Rx can remove the interference contribution of undecoded active users exploiting the PAB algorithm.

IV. Performance Evaluation

A. Analysis in the High Signal-to-Noise Ratio Regime

In this section we address the packet loss rate (PLR) performance of the slot-based scheme described in Section II, in the high E_{b}/N_0 regime. The expression is derived by neglecting noise and assuming that no "capture effect" can hold. This means that when multiple users collide in a resource-pilot pair, they can never be decoded exploiting unbalance in their channel gains.

Let firstly define $\mathcal{E}_{\mathrm{S}}$ as the error event indicating that a generic user (out of the total U_{S} ones active in the slot) experiences at least one pilot collision in all of its f employed resources. Assuming independence between the subcarriers, we focus on the error event per resource, letting $\mathcal{E}_k$ be the outcome of an unsolvable pilot collision (at least one) experienced by a user in one of its employed resources. Then, we let $\mathbb{P}\{\mathcal{E}_k|j\}$ be the probability of user's pilot collision

[2]For example, a simple way to do this consists of make the selected resources a function of the payload message, that is random.

in the resource given that j interfering users are active in that resource, and $\mathbb{P}\{j|U_{\mathrm{S}}\}$ be the probability that exactly j users choose that resource for the packet transmission, given that U_{S} users are overall active in the slot. Under the above assumptions, we can lower bound P_{L} as

$$\begin{aligned} P_{\mathrm{L}} &\geq \mathbb{P}\{\mathcal{E}_{\mathrm{S}}|U_{\mathrm{S}}\} \\ &= \left(\mathbb{P}\{\mathcal{E}_k|U_{\mathrm{S}}\}\right)^f \\ &= \left(\sum_{j=0}^{U_{\mathrm{S}}-1} \mathbb{P}\{\mathcal{E}_k|j\}\ \mathbb{P}\{j|U_{\mathrm{S}}\}\right)^f \\ &= \left\{\sum_{j=0}^{U_{\mathrm{S}}-1}\left[1-\left(\frac{N_{\mathrm{P}}-1}{N_{\mathrm{P}}}\right)^j\right]\right. \\ &\quad \left.\times \binom{U_{\mathrm{S}}-1}{j}\left(\frac{\binom{K-1}{f-1}}{\binom{K}{f}}\right)^j\left(1-\frac{\binom{K-1}{f-1}}{\binom{K}{f}}\right)^{U_{\mathrm{S}}-1-j}\right\}^f \end{aligned} \quad (7)$$

where N_{P} is again the size of the pool of orthogonal pilots and K is the number of orthogonal subcarriers.

B. Simulation Setup

We provide Monte Carlo simulation results for the proposed grant-free access schemes, where each transmitted message has a size of $\kappa = 421$ bits. Each active user encodes its message via an $(n = 511, \kappa = 421, t = 10)$ binary BCH code and appends a null padding bit to the BCH codeword. Next, the encoded bits are mapped to a QPSK constellation, resulting into $N_{\mathrm{D}} = 256$ payload symbols per user per resource. The number of orthogonal subcarriers is set to $K = 4$. The frequency-domain repetition rate is assumed constant for all transmitting users and equal to $f = 2$, both in the slot-based and the frame-based systems. With reference to the frame-based protocol, each frame has a duration of $N_{\mathrm{S}} = 50$ slots, and the time-domain repetition degree r is assumed to be the same for all active devices. Simulations are conducted for two distinct sets of orthogonal pilots: $N_{\mathrm{P}} = 64$ and $N_{\mathrm{P}} = 128$. Each orthogonal pilot sequence in the set has a length of N_{P} symbols. In line with common practice in short packet communications, we use the PLR as reliability performance indicator, rather than the bit error rate.

C. Numerical Results

1) Performance Analysis of the Slot-based Scheme: In Fig. 2, we evaluate the PLR performance of the proposed slot-based scheme varying the E_{b}/N_0. The simulated curves are obtained for different numbers of Rx antenna elements $M \in \{1, 2, 4, 8\}$, and assuming orthogonal pilots of size $N_{\mathrm{P}} \in \{64, 128\}$. The number of active users per slot is set to $U_{\mathrm{S}} = 6$, along with the complexity parameter T. We also report in dashed lines the analytical lower bound formulated in Section IV-A. Except for the single-antenna scenario, the simulated curves slightly outperform the analytically bounded ones in the high E_{b}/N_0 regime. This is due to the capture effect, which occurs when multiple users collide in the same resource-pilot pair. For example, if two users select the same

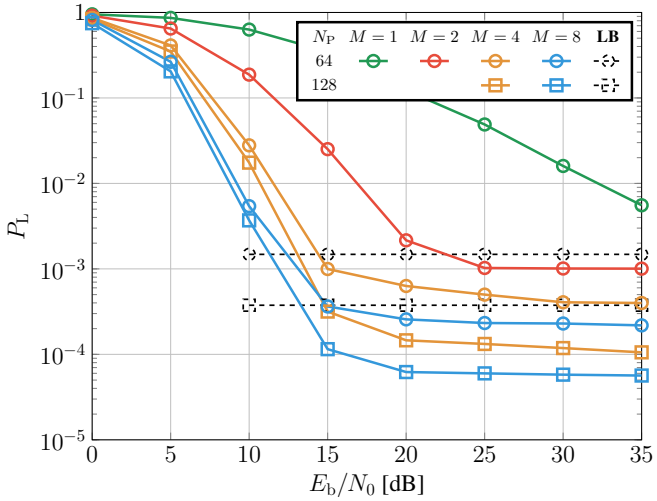

Fig. 2. Packet loss rate versus E_{b}/N_0 achieved by the slot-based scheme. The simulations are performed for different numbers of Rx antennas $M \in \{1, 2, 4, 8\}$ and pilot sequences $N_{\mathrm{P}} \in \{64, 128\}$. The complexity parameter is set to $T = U_{\mathrm{S}} = 6$. Dashed: Lower bound predictions computed by (7).

pilot in a certain resource, the MUD processing will not necessarily fail. In fact, if the channel gains experienced by the two users are appreciably unbalanced (i.e., one user is 'stronger' than the other), the stronger user can be correctly decoded. Then, after removing the interference contribution, the weaker user can also hopefully be decoded. Such an effect is amplified by the use of multiple antennas, as the reduction of the "off-pilot" users interference [18] allows for an additional boost in decoder performance. Notably, increasing M improves the system performance in the "waterfall" region (i.e., for medium to low values of PLR). For example, assuming a target PLR of $P_{\mathrm{L}}^* = 10^{-2}$, the scheme with $M = 8$ antennas outperforms the one with $M = 2$ antennas by within 9 dB. Furthermore, the use of a higher number of pilot sequences allows for better performance in the "error floor" region (i.e, for low values of PLR above the target).

2) Performance Analysis of the Frame-based Scheme: The numerical results obtained for the frame-based protocol are illustrated in Fig. 3. In this figure, the PLR is evaluated for different numbers of active users per frame U_{F}, which represents the system scalability parameter. The simulations are conducted by varying the time-domain user repetition degree $r \in \{1, 2, 3\}$ while the frequency repetition rate is fixed to $f = 2$ for all transmitters. The simulated number of Rx antennas is $M = 8$, the set of pilots has a dimension of $N_{\mathrm{P}} = 64$, and the E_{b}/N_0 is set to 25 dB. The complexity parameter is again set to $T = 6$. In the dash-dotted lines, we illustrate the system performance after the initialization phase only, before enabling both intra- and inter-slot SIC. The dashed curves depict the results for the initialization plus intra-slot SIC phases. The solid curves highlight the overall system performance when both intra-slot and inter-slot SIC techniques are performed. We remark that, in the scheme with $r = 1$, each active user chooses only one slot for the packet

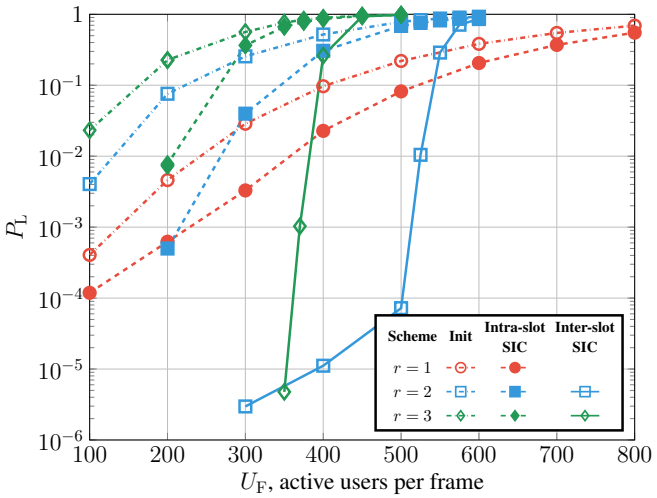

Fig. 3. Packet loss rate versus active users per frame U_{F} achieved by the frame-based protocol. The simulations are performed assuming $M = 8$ Rx antennas, $N_{\mathrm{P}} = 64$ pilot sequences, $N_{\mathrm{S}} = 50$ slots and $E_{\mathrm{b}}/N_0 = 25$ dB. The complexity parameter is set to $T = 6$. Dash-dotted: Init phase only. Dashed: Init plus intra-slot SIC. Solid: Init plus intra-and inter-slot SIC.

transmission, with no opportunity to retrieve the information message through the inter-slot SIC procedure. For a target PLR of $P_{\mathrm{L}}^{*} = 10^{-3}$, the protocol with $r = 2$ exhibits the best waterfall performance, demonstrating scalability gains as high as 35% and 50% over the schemes with $r = 3$ and $r = 1$, respectively. Remarkably, we emphasize the following points: i) the introduction of time-domain packet replicas allows for improved scalability with respect to the slot-based system, at the cost of increased latency; ii) the relationship between the physical system load and the complexity parameter T significantly impacts the performance. In fact, it is observed that the scheme with $r = 3$ performs worse than the scheme with $r = 2$ because, in many cases, the number of transmitted packets for each subcarrier does not meet the constraints for super-constellation processing due to the limitation imposed by T. Consequently, the SIC phase cannot be triggered, resulting in a degradation of the entire system's performance.

V. Conclusions

Grant-free protocols look like an appealing solution for next-generation IoT applications, especially in critical mMTC scenarios, where they are required to flexibly support different scalability and latency constraints, depending on the application. To this aim, we developed an hybrid MUD-aided scheme that, switching from a slot-based to a frame-based configuration, provides a very good latency-scalability tradeoff without degrading the reliability performance. The proposed scheme will be further investigated in future works, exploiting some useful analytical tools (e.g, density evolution) that will enable a more rigorous system performance validation. Furthermore, novel features, such as non-orthogonal pilot sequences and more powerful channel encoders, will be additionally exploited to further improve the performance.

Acknowledgements

Supported by the European Union under the Italian National Recovery and Resilience Plan of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - "RESTART").

References

[1] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: Potential, challenges, and solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, Mar. 2012.

[2] J. Sachs, P. Popovski, A. Höglund, D. Gozalvez-Serrano, and P. Fertl, "Machine-type communications," in *5G Mobile and Wireless Commun. Technology*. Cambridge University Press, 2016.

[3] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. J antti, and A. Ghosh, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.

[4] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[5] C. Bockelmann et al., "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28 969–28 992, 2018.

[6] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[7] P. Popovski, Č. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[8] A. Hoglund *et al.*, "3GPP release-16 preconfigured uplink resources for LTE-M and NB-IoT," *IEEE Commun. Standards Mag.*, vol. 4, no. 2, pp. 50–56, Jun. 2020.

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

[10] S. Veedu *et al.*, "Toward smaller and lower-cost 5G devices with longer battery life: An overview of 3GPP release 17 RedCap," *IEEE Commun. Standards Mag.*, vol. 6, no. 3, pp. 84–90, Sep. 2022.

[11] R. W. Chang, "Synthesis of band-limited orthogonal signals for multichannel data transmission," *Bell System Technical Journal*, vol. 45, no. 10, pp. 1775–1796, 1966.

[12] N. Abramson, "The ALOHA system: Another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*, 1970, pp. 281–285.

[13] L. G. Roberts, "ALOHA packet system with and without slots and capture," *SIGCOMM Comput. Commun. Rev.*, vol. 5, no. 2, p. 28–42, Apr. 1975.

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

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

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

[17] A. Mirri, D. Forlivesi, L. Valentini, M. Chiani, and E. Paolini, "An SCMA-based grant-free access scheme," in *Proc. 2024 IEEE Wireless Commun. and Networking Conf. Workshops*, Dubai, United Arab Emirates, Apr. 2024.

[18] L. Valentini, M. Chiani, and E. Paolini, "Interference cancellation algorithms for grant-free multiple access with massive MIMO," *IEEE Trans. Commun.*, vol. 71, no. 8, pp. 4665–4677, Aug. 2023.

[19] ——, "Massive grant-free access with massive MIMO and spatially coupled replicas," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7337–7350, Nov. 2022.

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