IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 February 2024, accepted 26 February 2024, date of publication 1 March 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3371994

RESEARCH ARTICLE

Pipelined Multi-User IR-HARQ Scheme for Improved Latency Performance in URLLC

RAFAEL SANTOS, DANIEL CASTANHEIRA, ADÃO SILVA, AND ATÍLIO GAMEIRO
Instituto de Telecomunicações (IT), University of Aveiro, 3810-193 Aveiro, Portugal
Departamento de Electrónica, Telecomunicações e Informática (DETI), University of Aveiro, 3810-193 Aveiro, Portugal

Corresponding author: Rafael Santos (rafaelsantoscbt10@av.it.pt)

This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) through the Doctoral Program under Grant 2020/06241/BD, and in part by the REVOLUTION Project 2022.08005.PTDC.

ABSTRACT The demand for ultra-reliable low-latency communications (URLLC) has led to the adoption of grant-free (GF) access techniques by the 5G NR, with the goal of reducing uplink access time. When GF access is employed, the base station (BS) preallocates multiple transmission opportunities (TOs) that can be utilized by the user equipment (UE) as needed. However, this approach results in inefficient resource utilization as unused TOs are wasted. To overcome this inefficiency, the 5G NR allows the assignment of configured grants (CG) to a group of UEs instead of a single one. This development has led research into group-based CG (GCG) schemes, whose reliance on shared resources can result in collisions. The collisions can be prevented by the use of stop-and-wait IR-HARQ schemes. Nevertheless, the delay caused by feedback latency is also undesirable as it severely affects latency performance. This work proposes two new IR-HARQ GCG schemes to efficiently handle feedback latency. The first one is able to eliminate the feedback latency overhead and is proven to simultaneously achieve the latency of a one-shot transmission and the energy efficiency of IR-HARQ, even in the presence of non-instantaneous feedback signaling. The second one features both a feedback latency protection mechanism, similar to the first scheme, and a mechanism specifically designed to further reduce latency. The performance of the proposed schemes is compared with scenarios where each UE uses either an individual one-shot or an IR-HARQ scheme. These comparisons encompass scenarios with either power or energy constraints. The results have shown that the second scheme always outperforms the IR-HARQ scheme and that it is able to outperform the one-shot scheme on a wide interval of feedback latency values, achieving a lower latency both for power and energy constrained cases.

INDEX TERMS URLLC, low-latency, grant-free, multi-user, control-networks, multi-user diversity, real-time wireless communications.

I. INTRODUCTION

Ultra-Reliable Low-Latency Communications (URLLC) for 5G and beyond was established to support applications with stringent requirements in terms of reliability and low-latency. However, there is a high heterogeneity of requirements, even among use-cases falling within the URLLC umbrella [1]. Indeed, the URLLC traffic can be periodic or sporadic [2], with latency and reliability requirements ranging from 0.25ms to 1ms and $1 - 10^{-2}$ to $1 - 10^{-9}$, respectively [1]. The concept of grant-free (GF) access is considered a promising building block for URLLC due to its ability to reduce uplink access time [3], [4]. The configured grant (CG) is a new radio (NR) feature that allows periodic allocation of resources to an UE effectively enabling GF access [5]. The CG can be set with several transmission opportunities (TOs), where extra redundancy can be transmitted on each TO. This enables the pairing of CG with IR-HARQ [5]. This pairing can be an important enabler of URLLC due to two facts. First, the URLLC traffic will be mainly comprised of small-sized packets, requiring their system to operate within the finite blocklength limit (FBL). This operational constraint

The associate editor coordinating the review of this manuscript and approving it for publication was Xujie Li.

results in a notable performance gap relative to Shannon capacity [6]. Second, it was shown that this performance gap to capacity can be greatly mitigated by using a feedback channel [7]. This second point, prompted the development of several IR-HARQ optimization methods [8], [9], [10], [11]. The drawback of preallocating TOs to an IR-HARQ scheme, is that in the cases where the UE does not need all the TOs, these resources are effectively wasted, draining the system finite resources. To mitigate the amount of wasted resources, the NR enables a CG to be defined to a group of UEs instead of only one.

Several group CG (GCG) schemes can be found in the literature [12], [13], [14], [15], [16], [17], [18], [19], [20], and [21]. These GCG schemes have been optimized to achieve various benefits, including increased average throughput [12], improved resource efficiency [17] and enhanced latency performance [21]. All these works rely in some kind of retransmission scheme. To mitigate the latency overhead of waiting for feedback before initiating the next transmission, some works employ blind retransmission schemes on shared resources, which results in collisions. Conversely, others wait for the feedback, possibly avoiding collisions but suffering from feedback delay overhead. Hence, one has to choose between waiting for the feedback signal or allow collisions, which lowers the reliability.

In summary, sources of delay, such as the feedback latency and the uplink access time are undesirable [22], especially on URLLC. The access time can be eliminated through CG. The CG resource inefficiency is mitigated through group based CG paired with a feedback scheme, like IR-HARQ. To the best of the authors knowledge, no approach to mitigate the effect of the feedback latency has been proposed so far. In this work, we propose a GCG scheme that relies on a stop-and-wait IR-HARQ and is able to eliminate the feedback latency overhead, by pipelining the transmission of multiple users. We also consider the combination of this scheme with MU-HARQ, improving both the energy efficiency and the latency performance.

A. RELATED WORK

Configuring a grant to a group of UEs instead of a single UE, has emerged as promising building block for URLLC solutions. The works [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] propose different GCG schemes. Some rely only on shared resources [12], [13], [14], [15] while others consider both shared and dedicated resources [16], [17], [18], [19], [20], [21]. Approaches relying only on shared resources enable random access of sporadic traffic, leading to faster access times, but result in collisions which degrades the scheme latency/reliability performance. In [12] a slotted Aloha scheme paired with retransmissions is employed to enable random access to the pool of shared resources. The work studies the impact of the group size, the number of maximum retransmissions and the variance in physical location of the group members. In [14], the number of parallel transmissions of each UE are optimized such that the reliability requirements are fulfilled. However, as transmission resources are selected at random, increasing the number of parallel transmissions overloads the system, increasing the probability of collisions. This scheme was shown to outperform the K-Repetition approach, even though collisions are interpreted as erasures. Instead of considering collisions as erasures, in [15] a successive interference cancellation (SIC) technique is employed to remove the interference caused by the correctly decoded UEs. This effectively improves the system performance in overloaded scenarios. Approaches that use both dedicated and shared resources offer a balance between collision free transmissions (dedicated resources) and resource efficiency (shared resources) [16], [17], [18], [19], [20], [21]. In all these schemes, there is a first phase where all the transmissions are performed through user dedicated resources, and a second phase where transmissions are done in shared resources, where collisions can occur. The proposed schemes differ on how they handle possible collisions that may occur on the second phase. In the second phase, collisions can be handled through HARQ signalling [16], [21], successive interference cancellation (SIC) [17], [18], [19], [20], or have their probability minimized by optimally selecting the number of UEs [20]. In [21], a collision free GCG schemes optimized for latency performance was proposed. It was proven that this scheme, named multi-user HARQ (MU-HARQ), is able to achieve a latency as low as the average latency of any incremental redundancy hybrid automatic repeat request (IR-HARQ) scheme with improved resource efficiency. However, the MU-HARQ latency performance degrades equally fast, as the feedback latency increases.

B. CONTRIBUTIONS

In this paper, two different GCG schemes are proposed to eliminate the impact of feedback latency on IR-HARQ and improve the latency performance. The main contributions of this work are the following:

- A GCG scheme named pipeline IR-HARQ (PL-HARQ), where the UEs transmit serially through the entire bandwidth. This follows the same logic as the CPUs instruction pipelines [23]. This procedure removes feedback latency overhead, as the base station (BS) can decode the data and transmits the feedback to some group members, while receiving the uplink data of others.
- A second GCG scheme named pipelined MU-HARQ (PL-MU-HARQ), a fusion between PL-HARQ and MU-HARQ. This scheme balances MU-HARQ latency reduction and the PL-HARQ feedback latency protection.
- Proof of key PL-HARQ properties, namely its ability to eliminate overhead caused by feedback latency and the presence of lower feedback latency overhead compared to MU-HARQ.
- Derivation of the PL-MU-HARQ asymptotic performance.

The proposed schemes performance is analysed in the presence of either a transmission power constraint or an energy budget constraint, for different scenarios. The results show that the PL-MU-HARQ scheme is able to balance the benefits of MU-HARQ and PL-HARQ schemes and that it greatly outperforms both the IR-HARQ and one-shot latency performance on a wide range of feedback latency values.

C. PAPER OUTLINE

The document is organized as follows, in Section II the system model is described. In Section III, three schemes from the literature are described, necessary for benchmarking. Then, in Section IV two new schemes, PL-HARQ and PL-MU-HARQ, are described. In Section V, the description of two optimization problems and their corresponding algorithms, is carried out. In Section VI the numerical results are presented and discussed, while in Section VII the final conclusions are drawn.

II. SYSTEM MODEL

This work considers a SISO uplink AWGN channel, where a group of G UEs, with similar traffic characteristics and channel statistics, is formed. In this work, and similarly to [17] and [20], it is assumed that the group is formed using an appropriate classification procedure for group formation [24], [25]. The group of G UEs is denoted as $\mathcal{U} = \{U^{(1)}, \cdots, U^{(G)}\}$, where $U^{(i)}$, $i \in [1, G]$ is the UE with group ID i. Each UE has to periodically transmit B new bits of information to a BS, as expected in sensor-to-controller communications within wireless control networks [2] and various other URLLC applications [18], [26], [27]. The goal of every group member is to transmit the B information bits while meeting URLLC QoS, i.e., complying with a target probability of error ϵ_T and delay budget t_T. A configured grant is set to the entire group, such that the BS periodically preallocates a bandwidth w_T and allows a maximum of M TOs. At each TO, extra redundancy is transmitted to the BS with a predetermined transmission power, i.e., at the mth TO a transmission of size $n^{(m)}$ channel uses is carried out with transmission power $p^{(m)}$. The total allocated bandwidth is normalized to group size G such that $\frac{w_T}{G} = 1$. The number of channel uses of a given transmission is proportional to its bandwidth $w^{(m)}$ and time duration $t^{(m)}$ [28]. The time duration of mth TO transmission, can now be defined as

$$t^{(m)} = \frac{n^{(m)}}{w^{(m)}}, \tag{1}$$

where $t^{(m)} = n^{(m)}$ if all the G UEs perform the one-shot transmission through their dedicated bandwidth $\frac{w_T}{G} = 1$, i.e., in parallel. When $M = 1$ the UEs operate with a one-shot scheme. When $M > 1$, the transmission scheme relies on ACK/NACK mechanism. This ACK/NACK mechanism operates as follows, at each uplink transmission the BS receives extra redundancy, jointly decodes it with all the redundancy received so far and checks if the data was correctly decoded (through CRC). If the data was correctly decoded, the BS transmits an ACK to the UE, otherwise it transmits a NACK. Upon reception of the feedback signal, the UE transmits on the next TO only if a NACK was received and has not yet used all the M TOs. The probability of error at the mth TO can be obtained as [10]

$$\epsilon^{(m)} = Q\left(\frac{\sum_{i=1}^{m} n^{(i)} \log(1+p^{(i)}) - B\log(2)}{\sqrt{\sum_{i=1}^{m} \frac{n^{(i)} p^{(i)} (2+p^{(i)})}{(1+p^{(i)})^2}}}\right), \tag{2}$$

which is an approximation of the original PPV bound [6] for AWGN with unit power variance. In this work, the feedback latency, t_f, is defined as the elapsed time (in channel uses) between the reception of the uplink signal by the BS and the reception of the feedback signal by the UE. The notation is developed throughout the text, but for readers' convenience, we have summarized it in Table 1.

TABLE 1. Summary of notations.

Notation	Definition
OS	One-shot scheme (no feedback).
SU-HARQ	Single user IR-HARQ.
MU-HARQ	The multi-user IR-HARQ proposed in [21].
PL-HARQ	The proposed pipelined IR-HARQ.
PL-MU-HARQ	The proposed pipelined MU-HARQ.
$P_{bin}(n, x, \epsilon)$	Probability of x successes out of n with probability ϵ.
B	Number of information bits per user.
M	Maximum number of transmission rounds.
t_T	Delay budget (maximum tolerable latency).
w_T	Group allocated bandwidth.
ϵ_T	Target probability of error.
E_T	Energy budget (average energy spent).
t_{fo}	Feedback latency overhead.
$\mathcal{F}$	Total latency performance (feedback + transmission).
G	Group size.
K	Number of PL-MU-HARQ subgroups.
C	Number of elements in each PL-MU-HARQ subgroup.
$\mathcal{U}$	IDs of every group member.
$\mathcal{U}^{(u)}$	IDs of members within the uth subgroup.
$U^{(g)}$	Group member with ID g.
$\mathcal{X}^{(m)}$	R.V. modeling active UEs in the mth round.
$\mathcal{X}$	SP of the number of active UEs at each round.
$\mathcal{S}$	State space of $\mathcal{X}$.
$W^{(m)}$	R.V. modeling the available bandwidth at mth round.
W	SP of available bandwidth at all rounds.
$x, x^{(m)}$	Realization of $\mathcal{X}$ and $\mathcal{X}^{(m)}$, respectively.
$t^{(m)}$	Time duration of the mth round for realization $x^{(m)}$.
$x^{[m]}$	Realization of $\mathcal{X}$ up to the mth round, where $x^{[M]} = x$.
$w_x, w_{x^{(m)}}$	Realization of W and $W^{(m)}$ given x, respectively.
$n_{x^{[m]}}$	Channel uses of the mth transmission for realization x.
$p_{x^{[m]}}$	Power of the mth transmission for realization x.
$\epsilon_{x^{[m]}}$	Error probability of the mth round of realization x.
$\mathcal{T}$	Transmission duration of every transmission round.
$\mathcal{T}^{[m]}$	Transmission duration of the first m rounds.
$\mathcal{N}_{x^{[m]}}$	Channel uses of the first m rounds for realization $x^{[m]}$.
$\mathcal{P}_{x^{[m]}}$	Transmission power of the first m rounds of x.
$\mathcal{E}_{x^{[m]}}$	Error probability for the initial m rounds with $\mathcal{X} = x$.
$\Theta_{x^{[m]}}$	MU-HARQ parameters; $\Theta_{x^{[m]}} = (\mathcal{N}_{x^{[m]}}, \mathcal{P}_{x^{[m]}})$
E_x	Average energy expended when $\mathcal{X} = x$.
$\star$	Used to to denote a optimal solution of a given problem.

III. LITERATURE SCHEMES: FROM ONE-SHOT TO MULTI-USER IR-HARQ

In this section, three schemes from the literature are described. The first one, named one-shot (OS), is characterized by the absence of feedback signaling. Conversely, the second and third schemes, identified as single-user IR-HARQ and MU-HARQ, respectively, are reliant on feedback signaling. In the latter case (MU-HARQ), resources are effectively managed from a group perspective, leading to a substantial reduction in latency.

A. ONE-SHOT SCHEME

In a one-shot scheme the group configured grant is set such that the UEs perform a one-shot transmission in parallel, through a dedicated bandwidth $\frac{w_T}{G} = 1$, as shown in Fig. 1(a). Considering a transmission power P, then the transmission has to be of size N, being N the lowest value such that

$$\epsilon_T \geq Q\left(\frac{N\log(1+P) - B\log(2)}{\sqrt{\frac{NP(2+P)}{(1+P)^2}}}\right), \tag{3}$$

is verified. The G UEs transmit in parallel through dedicated resources, meaning that the transmission duration is $t_{OS} = N\frac{G}{w_T} = N$ and since it is not influenced by t_f, the overall latency denoted $\mathcal{F}_{OS}$, is

$$\mathcal{F}_{OS} = t_{OS}. \tag{4}$$

The average energy expended by each UE E_{OS} is equal to

$$E_{OS} = NP. \tag{5}$$

B. INCREMENTAL REDUNDANCY HARQ

In a IR-HARQ scheme, a group configured grant is set such that each UE applies an IR-HARQ scheme with M TOs through a dedicated bandwidth $\frac{w_T}{G} = 1$, see Fig. 1(b). This is equivalent to a scenario where each UE uses an IR-HARQ with an individual CG. This scheme is denoted by single-user IR-HARQ (SU-HARQ) to better convey its absence of cooperation. Whenever a UE receives an ACK before the Mth TO, the remaining dedicated preallocated resources are wasted. The SU-HARQ transmissions are parameterized by $\Theta = (\mathcal{N}, \mathcal{P})$, being $\mathcal{N} = \{n^{(1)}, n^{(2)}, \cdots, n^{(M)}\}$ and $\mathcal{P} = \{p^{(1)}, p^{(2)}, \cdots, p^{(M)}\}$, where $n^{(m)}$ and $p^{(m)}$ define the mth TO size and power, respectively. Let $\mathcal{T} = \{t^{(1)}, t^{(2)}, \cdots, t^{(M)}\}$ define the time duration of all the TOs, then $\mathcal{T} = \mathcal{N}$ since the transmission bandwidth is constant and equal to $\frac{w_T}{G} = 1$ (1).

The SU-HARQ probability of error ϵ_{SU}, average expended energy E_{SU} and transmission time t_{SU} can be obtained as

$$E_{SU} = n^{(1)}p^{(1)} + \sum_{m=2}^{M} \epsilon^{(m-1)} n^{(m)} p^{(m)} \tag{6}$$

$$\epsilon_{SU} = \epsilon^{(M)} \tag{7}$$

$$t_{SU} = \sum_{m=1}^{M} t^{(m)}, \tag{8}$$

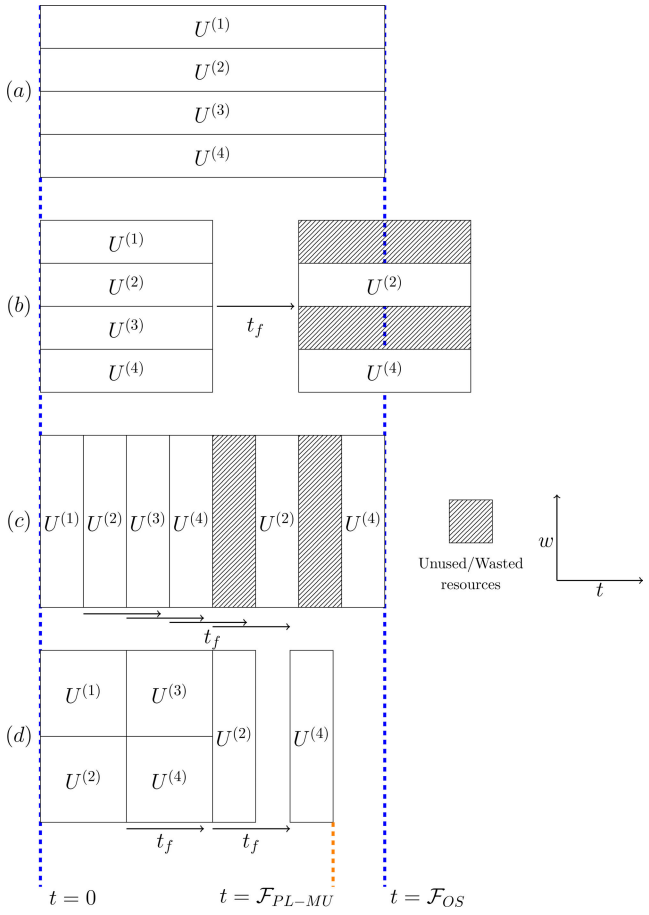


FIGURE 1. Latency performance of OS (a), SU-HARQ (b), PL-HARQ (c) and PL-MU-HARQ (d) with $G = 4$. In the feedback scenarios (b),(c),(d) both $U^{(1)}$ and $U^{(3)}$ do not need the second TO (received ACK). All schemes have to transmit the same number of symbols due to the equal power constraint. The value of t_f is equal in (b), (c), and (d). However the effective latency overhead in (c) is null and the one-shot latency is met, while in (d), despite a small delay overhead, the one-shot latency is beaten due to the increased transmission bandwidth on the second TO.

where $\epsilon^{(m)}$ can be obtained through (2). The SU-HARQ total latency $\mathcal{F}_{SU}$ can now be defined as

$$\mathcal{F}_{SU} = t_{SU} + (M-1)t_f. \tag{9}$$

C. MULTI-USER INCREMENTAL REDUNDANCY HARQ

The MU-HARQ is a cooperative group-based IR-HARQ scheme, which leverages on the common IR-HARQ feedback signals, in order to distributively reach a consensus on how to distribute the preallocated group resources amongst the UEs. In this scheme, the feedback signals of the entire group are multicasted to the group itself. This means that, after each TO, each group member knows which other members are still active, i.e. need next TO, or not. Knowing this, the currently active UEs are able to divide the entire bandwidth w_T amongst themselves, possibly increasing the transmission bandwidth and avoiding collisions. The number of active UEs at the mth TO is modeled by the R.V. $X^{(m)}$ being

$x^{(m)} \in \{0, 1, \cdots, G\}$ its realization. The bandwidth used at the mth TO is also a R.V

$$W^{(m)} = \frac{w_T}{X^{(m)}}. \tag{10}$$

To model the MU-HARQ one has to consider the stochastic process (SP)

$$\mathcal{X} = \{X^{(1)}, \cdots, X^{(M)} : M \in \mathbb{N}\}, \tag{11}$$

whose state space $\mathcal{S}$ is

$$\mathcal{S} = \{x^{(1)}, \cdots, x^{(M)} : 0 \leq x^{(i)} \leq x^{(j)} \leq G,\ j \leq i \leq M \in \mathbb{N}\} \tag{12}$$

meaning that each possible SP realization $x \in \mathcal{S}$. The notation $x^{[m]} = \{x^{(1)}, \cdots, x^{(m)}\} \subseteq x \in \mathcal{S}$ is introduced to be used as an indexing set, being $x = x^{[M]}$. The MU-HARQ transmissions are parameterized by $\Theta \in \Lambda$, being Λ the feasible parameter set. Each TO is parameterized by its transmission size and power. The MU-HARQ needs a set of M parameters for all possible SP realization $x \in \mathcal{S}$. Let $\mathcal{N}_x = \{n_{x^{[1]}}, \cdots, n_{x^{[M]}}\}$ and $\mathcal{P}_x = \{p_{x^{[1]}}, \cdots, p_{x^{[M]}}\}$ be the transmission size and power used on realization $x \in \mathcal{S}$, where $n_{x^{[m]}}$ and $p_{x^{[m]}}$ represents the transmission size and power of the mth TO of the corresponding SP realization x. The pair $(\mathcal{N}, \mathcal{P})$ has the same information as Θ, however Θ is used in order to simplify the notation. In this work, the time duration of each TO is independent of the current realization x. The duration of each MU-HARQ transmission is defined as $\mathcal{T} = \{t^{(1)}, t^{(2)}, \cdots, t^{(M)}\}, \forall x \in \mathcal{S}$. This means that each transmission has a size equal to

$$n_{x^{[m]}} = \frac{t^{(m)} w_T}{x^{(m)}}, \tag{13}$$

as the available bandwidth depends on $x^{(m)}$ (10). This means that $\mathcal{N}$ is completely defined by $\mathcal{T}$ and $\mathcal{S}$ (13), making $\mathcal{T}$ and $\mathcal{P}$ the only undetermined parameters. In a scenario where the transmission power is predefined and constant across TOs, only $\mathcal{T}$ is left undetermined, which further reduces the number of optimizable MU-HARQ parameters to M.

The MU-HARQ probability of error ϵ_{MU}, average expended energy E_{MU} and transmission time t_{MU} can be obtained as

$$E_{MU} = \sum_{x \in \mathcal{S}} P(\mathcal{X} = x) E_x \tag{14}$$

$$\epsilon_{MU} = \sum_{x \in \mathcal{S}} P(\mathcal{X} = x) \epsilon_x \tag{15}$$

$$t_{MU} = \sum_{m=1}^{M} t^{(m)}. \tag{16}$$

where the energy E_x and probability of error ϵ_x, for a SP realization $x \in S$ are given by

$$E_x = \frac{1}{G} \sum_{m=1}^{M} t^{(m)} p_{x^{[m]}} \tag{17}$$

$$\epsilon_x = Q\left(\frac{\sum_{m=1}^{M} \frac{w_T}{x^{(m)}} t^{(m)} \log(1 + p_{x^{[m]}}) - B \log(2)}{\sqrt{\sum_{m=1}^{M} \frac{\frac{w_T}{x^{(m)}} t^{(m)} p_{x^{[m]}} \left(2 + p_{x^{[m]}}\right)}{\left(1 + p_{x^{[m]}}\right)^2}}} \right). \tag{18}$$

Having defined ϵ_x, the SP transition kernel can be parameterized as

$$P\left(X^{(m)} = x^{(m)} | X^{(m-1)} = x^{(m-1)}\right) = P_{bin}\left(x^{(m)}, x^{(m-1)}, \frac{\epsilon_{x^{[m-1]}}}{\epsilon_{x^{[m-2]}}}\right), x \in \mathcal{S}, \tag{19}$$

where $P_{bin}(n, x, \epsilon)$ denotes the probability of having x successes out of n trials, being the probability of success at each trial ϵ. The MU-HARQ total latency $\mathcal{F}_{MU}$ can now be defined as

$$\mathcal{F}_{MU} = t_{MU} + (M - 1) t_f. \tag{20}$$

IV. PROPOSED SCHEMES

The IR-HARQ schemes, particularly MU-HARQ, show considerable latency improvement compared to OS when feedback delays are low to moderate. However, this efficiency degrades as feedback delays become excessively high. Motivated by these results, two new schemes, PL-HARQ and PL-MU-HARQ, are introduced. These schemes aim to combine the insensitivity of OS to feedback delays with the high performance of feedback schemes, particularly for low delays.

A. PL-HARQ

As described in previous sections, both in the IR-HARQ and the MU-HARQ, all the active UEs transmit at the same time through orthogonal frequency bands. This means that the overall system operates in two states. One state where all the UEs are transmitting, the channel bandwidth is fully occupied, and the BS is listening. Another where the BS is processing all the received signals, while both the channel and the UEs are idle and waiting for the BS feedback. In the PL-HARQ the tasks are partitioned to avoid idle states of the UEs, the channel and the BS. The procedure is reminiscent of the CPUs' instruction pipelines [23]. To make it clear, in the PL-HARQ each UE transmits successively (one at a time) through the entire bandwidth, as shown in Fig. 1(c). The transmission ordering is decided by the user ID, i.e., $U^{(1)}$ transmits first, followed by $U^{(2)}$ and so on. Conceptually, when $U^{(4)}$ is performing its transmission, the BS could be decoding $U^{(3)}$'s data and transmitting $U^{(2)}$'s feedback while $U^{(1)}$ could, depending on the feedback, be preparing the next transmission. This removes the idle time and reduces the user load in the channel and the BS. For CPUs pipelining enables

lower clock periods, while in communications it enables lower round-trip latency. Contrary to the MU-HARQ, the bandwidth used on each TO is constant and equal to $w_T = G$. This means that each UE is able to perform a transmission of size $n^{(1)}$ with a time duration $t^{(1)} = \frac{n^{(1)}}{w_T}$. However, since they transmit one at a time, $U^{(G)}$ (the last to transmit), ends its first transmission at time instant $Gt^{(1)} = n^{(1)}$, which is the time needed by the one-shot scheme, to perform a transmission of size $n^{(1)}$. Hence, there is no gain or loss in doing this sequential transmission reordering. The difference lies on the fact that, when $U^{(G)}$ ends the first transmission, the BS had received the first transmission of $U^{(1)}$ $(G-1)t^{(1)}$ units of time ago. Hence, if $t_f < (G-1)t^{(1)}$, then $U^{(1)}$ can start the second transmission, if necessary, right after $U^{(G)}$ first transmission. If this happens for all group members, then from the group perspective, the channel is never idle and $U^{(G)}$ (the last to transmit) has the same latency performance as it would have on a one-shot scheme with equal transmission power P. Hence, the PL-HARQ is a stop-and-wait scheme, that is able to exhibit the one-shot latency performance even for $t_f > 0$. This motivated the formulation of the following theorem

Theorem 1: Let $t^{(m)}$ be the time duration of the mth transmission round of a PL-HARQ solution where G is the UE group size, and t_f the elapsed time between the reception of the transmission round by the BS and the reception of the feedback signal by the UE. Then, if the condition

$$t^{(m)} > \frac{t_f}{G-1} \; \forall m \in [1, M], \tag{21}$$

is verified, the proposed scheme has the same latency performance as an single-user IR-HARQ scheme operating through a bandwidth $w_T/G = 1$ with instantaneous feedback.

Proof: Appendix. A. □

Even if the condition (21) is not met, PL-HARQ can provide benefits as it is always able to mitigate the feedback latency overhead. The feedback latency overhead at the mth TO is quantified in the following theorem

Theorem 2: Let $t^{(m)}$ and $t^{(m+1)}$ be the time duration of the mth and $m+1$th transmission rounds of a PL-HARQ solution with group size G. Let t_f be the elapsed time between the reception of the transmission round by the BS and the reception of the feedback signal by the UE. In such case, the feedback latency overhead at the mth transmission round is equal to

$$t_{fo}^{(m)} = \max\left[\left(t_f - \min\left(t^{(m)}, t^{(m+1)}\right)(G-1)\right), 0\right]. \tag{22}$$

which is lower than t_f as long as $\min\left(t^{(m)}, t^{(m+1)}\right) > 0$.

Proof: Appendix. B. □

The PL-HARQ total latency $\mathcal{F}$ can now be defined as

$$\mathcal{F}_{PL} = \sum_{m=1}^{M} Gt^{(m)} + \sum_{i=1}^{M-1} t_{fo}^{(i)}. \tag{23}$$

Since $t_{fo}^{(i)} < t_f \;\; \forall i \in [1, M-1]$, the PL-HARQ has a lower feedback latency overhead than SU-HARQ (9) and MU-HARQ (20).

B. PL-MU-HARQ

The PL-MU-HARQ is a combination between MU-HARQ and PL-HARQ. The motivation to pair these schemes arises from the fact that MU-HARQ is able to obtain a better latency performance than a OS scheme for $t_f = 0$, even when using the same transmission power, but its latency performance is rapidly degraded by t_f (20). The PL-HARQ can never achieve a latency lower than the one-shot scheme when using the same transmission power, however, contrary to MU-HARQ its latency performance is resilient against t_f. In order to pair the two schemes, the initial group of size G is further divided into K equally sized sub-groups $\mathcal{U} = \{\mathcal{U}^{(1)}, \ldots, \mathcal{U}^{(K)}\}$ each of size $C = \frac{G}{K}$, being C an integer. Each sub-group will transmit one a time through the whole bandwidth and use the MU-HARQ scheme among sub-group members. One should note that with such UE configuration, the PL-HARQ scheme appears as an inter sub-group scheme, while the MU-HARQ is set as a intra sub-group scheme Fig. 1(d). This can be observed in Fig. 1, where the 4 UEs are divided in two subgroups, one formed by $\{U^{(1)}, U^{(2)}\}$ and the other formed by $\{U^{(3)}, U^{(4)}\}$. This establishes K as a new design variable, which balances the PL-MU-HARQ scheme between the PL-HARQ and MU-HARQ effects. Indeed, when $K = 1$, the PL-MU-HARQ is equivalent to a MU-HARQ with G UEs while if $K = G$ it is equivalent to a PL-HARQ of group size G. Therefore, the PL-MU-HARQ scheme is a broader framework that incorporates both the PL-HARQ and MU-HARQ schemes, as possible configurations. The performance metrics are equal to a MU-HARQ scheme with a group size C such that

$$E_x = \frac{1}{C}\sum_{m=1}^{M} t^{(m)} p_{x[m]} \tag{24}$$

$$\epsilon_x = Q\left(\frac{\sum_{m=1}^{M} \frac{w_T}{x^{(m)}} t^{(m)} \log(1+p_{x[m]}) - B\log(2)}{\sqrt{\sum_{m=1}^{M} \frac{\frac{w_T}{x^{(m)}} t^{(m)} p_{x[m]}\left(2+p_{x[m]}\right)}{\left(1+p_{x[m]}\right)^2}}}\right) \tag{25}$$

being $x^{(m)} \in [0, 1 \ldots, C]$. The values E_Θ, ϵ_Θ and t_Θ can then be obtained as in (14)(15)(16), respectively. The total delay $\mathcal{F}$ defined by the combined influence of $t_\mathcal{T}$ and t_f, can be computed through

$$\mathcal{F}_{PL-MU} = Kt_{PL-MU} + \sum_{m=1}^{M-1} t_{fo}^{(m)}, \tag{26}$$

$$t_{fo}^{(m)} = \max\left[\left(t_f - \min\left(t^{(m)}, t^{(m+1)}\right)(K-1)\right), 0\right], \tag{27}$$

which is equal to a PL-HARQ scheme with K users (23).

The PL-MU-HARQ total delay in the asymptotic regime of an infinite group size $\mathcal{F}_{PL-MU}$ is described by the following theorem.

Theorem 3: Let $n^{(m)}$ and $n^{(m+1)}$ be the transmission size and $\epsilon^{(m)}$ and $\epsilon^{(m+1)}$ the error probability of the mth and $m+1$th transmission rounds of a PL-MU-HARQ solution. Let t_f be the elapsed time between the reception of the transmission round by the BS and the reception of the feedback signal by the UE. In the asymptotic group size regime, where $C \to \infty$ and $K \to \infty$, the PL-MU-HARQ total latency becomes

$$\mathcal{F}^{\infty}_{PL-MU} = \sum_{m=1}^{M} n^{(m)}\epsilon^{(m-1)} + t^{(m)}_{fo} \tag{28}$$

$$t_{fo} = \max\left[t_f - \min\left(n^{(m)}\epsilon^{(m-1)}, n^{(m+1)}\epsilon^{(m)}\right), 0\right] \tag{29}$$

Proof: Appendix. C. □

Furthermore, from [21, Theorem 1] one knows that in the asymptotic regime, there is only one possible SP realization, which is equal to $x = \{K, \frac{K}{\epsilon^{(1)}}, \cdots, \frac{K}{\epsilon^{(M-1)}}\}$. Therefore, in the asymptotic regime, the PL-MU-HARQ has $2M$ parameters, the transmission size and power of each TO. The PL-MU-HARQ performance bound is obtainable through the computation of the $2M$ parameters that minimize (28).

V. LATENCY OPTIMIZATION

In this section, two latency optimization algorithms are proposed, both designed for the PL-MU-HARQ scheme described in the previous section. The first considers a predefined and constant transmission power P for all TOs, while the second considers an average energy constraint and allows the optimization of the transmission power of each TO.

A. POWER CONSTRAINED TRANSMISSIONS

The latency minimization problem with constrained transmission power, can be formulated as follows

Problem 1:

$$\min_{\mathcal{T}} \quad \mathcal{F} \tag{30}$$

$$\text{s.t.} \quad \epsilon_\Theta \le \epsilon_T \tag{31}$$

$$\mathcal{T} \in \mathbb{R}^M_+ \tag{32}$$

The inequality (31) defines the reliability constraint and (32) defines the constraint on the round duration. Since the transmission power is predefined and constant across all TOs, the goal is to find $\mathcal{T}$ that minimizes $\mathcal{F}$, given P,B, ϵ_T and t_f. This means that this scheme has a total of M parameters. A PL-MU-HARQ with M parameters to define, allows *Problem 1* to be solved through a simple numerical search on $\mathcal{T} = \{t^{(1)}, \cdots, t^{(M)}\}$. Indeed, given $\mathcal{T}$ and t_f, one can readily obtain $\mathcal{F}$. Likewise, given $\mathcal{T}^{[M-1]}$, t_f and $\mathcal{F}$, one can obtain $t^{(M)}$. The optimization procedure follows the latter approach, given the system t_f an initial target $\mathcal{F}$ is set. Then a numerical search is performed on $\mathcal{T}^{[M-1]}$, being $t^{(M)}$ obtained for each $\mathcal{T}^{[M-1]}$. If all the combinations of $\mathcal{T}^{[M-1]}$ that produce a positive $t^{(M)}$ are exhausted and no solution satisfying (31)(32) was found, then the current $\mathcal{F}$ cannot be satisfied. In this case, $\mathcal{F}$ is increased and a new numerical search over $\mathcal{T}^{[M-1]}$ is carried out until a $\mathcal{F}$ with a valid solution is found. This way, the first solution to be found corresponds to the minimum achievable $\mathcal{F}$, making it optimal in terms of latency.

B. ENERGY CONSTRAINED TRANSMISSIONS

The latency minimization problem with constrained energy budget, can be mathematically formulated as follows

Problem 2:

$$\min_{\Theta=(\mathcal{T},\mathcal{E})} \quad \mathcal{F} \tag{33}$$

$$\text{s.t.} \quad E_\Theta \le E_T \tag{34}$$

$$\epsilon_\Theta \le \epsilon_T \tag{35}$$

$$\Theta \in \Lambda. \tag{36}$$

The inequality (34) defines the average energy budget constraint, (35) is the reliability constraint and (36) defines the parameters constraint, being

$$\Lambda = \{((n_{x[1]}, \epsilon_{x[1]}), \cdots (n_{x[i]}, \epsilon_{x[i]}), \cdots, (n_{x[M]}, \epsilon_{x[M]})) : 1 \le n_{x[i]},\ 0 < \epsilon_{x[i+1]} < \epsilon_{x[i]} < 0.5,\ i \in]1, M[,\ \forall x \in \mathcal{S}\}. \tag{37}$$

Following the approaches in [11], [21], (33) is optimized through (t, ϵ), meaning that from now on $\Theta = (\mathcal{T}, \mathcal{E})$. The function $p_{x[m]} = (\mathcal{N}_{x[m]}, \mathcal{E}_{x[m]}), x \in \mathcal{S}$ does not have an explicit form, nevertheless $p_{x[m]}$ can be obtained through an iterative algorithm. In this paper, the bissection algorithm is used to obtain $p_{x[m]}$ [10], [11].

As shown in [21], *Problem. 2* can be solved through an energy minimization problem paired with a root finding method. The energy minimization problem can be formulated as follows,

Problem 3:

$$\min_{\Theta} \quad E_\Theta \tag{38}$$

$$\text{s.t.} \quad \mathcal{F} \le t_T \tag{39}$$

$$\epsilon_\Theta \le \epsilon_T \tag{40}$$

$$\Theta \in \Lambda \tag{41}$$

Specifically, *Problem 2* can be solved by applying a root finding method to the function $f(t_T) = E^\star(t_T) - E_T$ where each query of $f(t_T)$ is performed by solving *Problem 3*. Hence, the root finding method is important as queries are computationally heavy. For this reason, Brent's method is used to find the root of $f(t_T)$ due to its balance between reliability and convergence speed. This approach pushes all the complexity of solving *Problem 2* into *Problem 3*, which is simpler to solve. To solve *Problem 3* the projected gradient descent method described in [21] applied with the necessary changes to the gradient functions.

VI. RESULTS

The numerical results, presented in this section, are divided into two subsections. The first subsection deals with a constant transmission power scenario. The second corresponds to the energy constrained case. The proposed schemes are compared with the OS and SU-HARQ schemes for that two cases. The power-constrained case is of interest to evaluate the latency performance improvements brought by the proposed schemes with respect to the OS scheme, since, due to its stop-and-wait nature, an SU-HARQ scheme can never outperform an OS scheme in terms of latency. The energy constrained case is important for the uplink, as too high energy expenditure leads to shorter battery life, hindering the portability of mobile devices.

A. POWER CONSTRAINED TRANSMISSIONS

The results for the power constrained case are presented in this subsection and consider a constant transmission power $P = 1$. The objective is to understand the impact of the feedback latency t_f on the latency performance of PL-MU-HARQ considering varying K, G, M, B, and ϵ_T. To accomplish this, the performance of PL-MU-HARQ configurations is assessed as a function of t_f. Then, the latency and energy performance of the PL-MU-HARQ scheme is compared with that of OS. Finally, the latency performance of the best PL-MU-HARQ configuration is compared with the OS scheme, by setting the transmission bandwidth accordingly to 5G numerology.

In Fig. 2 the $\mathcal{F}$s of both the OS scheme and three out of the six possible PL-MU-HARQ configurations of $G = 32$, are presented. The $\mathcal{F}$ of the OS is obviously independent of t_f. The SU-HARQ latency performance exhibits a linear degradation with t_f, but is never able to outperform the OS. As for PL-MU-HARQ, one is able to verify that all the displayed configurations exhibit different $\mathcal{F}$ to t_f evolution. To understand this results, consider first the two extreme configurations $(K = 1, C = 32)$ and $(K = 32, C = 1)$. When $G = 32$, the former corresponds to the MU-HARQ and the latter to the PL-HARQ. The MU-HARQ configuration $(K = 1, C = 32)$ exhibits the best $\mathcal{F}$ performance for $t_f = 0$, but it increases linearly with t_f (20). On the other hand, the PL-HARQ configuration $(K = 32, C = 1)$, shows a constant performance, identical to the one of the OS, up to a t_f threshold and then degrades linearly with the increase of the delay. It is then intuitive to expect that with intermediate K and C values (such that $KC = G$), one obtains a curve exhibits a mixture of the two extremes behaviors, as shown with the curve $(K = 4, C = 8)$. So with the correct choice of K and C, one can get the best of these two types of behavior given a range of t_f. This justifies the results in Fig. 3, that shows that the optimal K increases with t_f for the two combinations of B and ϵ_T. These results indicate that the ideal PL-MU-HARQ configuration is contingent upon the value of t_f. In situations where t_f is exceptionally low, a MU-HARQ behavior is more favorable. Conversely, when t_f is significantly high, opting for a PL-HARQ strategy yields

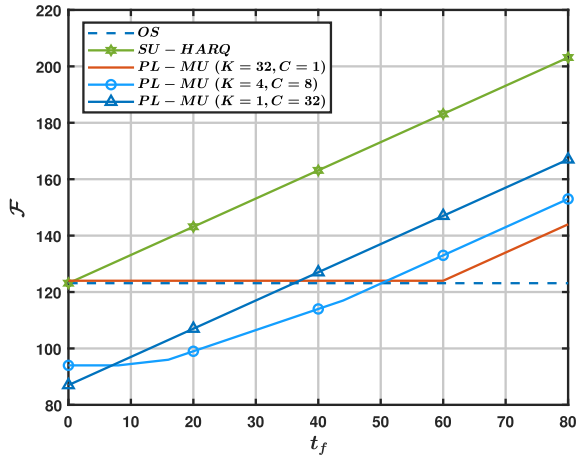


FIGURE 2. Comparing $\mathcal{F}$ of a one-shot scheme with all possible PL-MU-HARQ configurations of size $G = 32$ and SU-HARQ, considering $\{B = 64, \epsilon_T = 10^{-5}\}$.

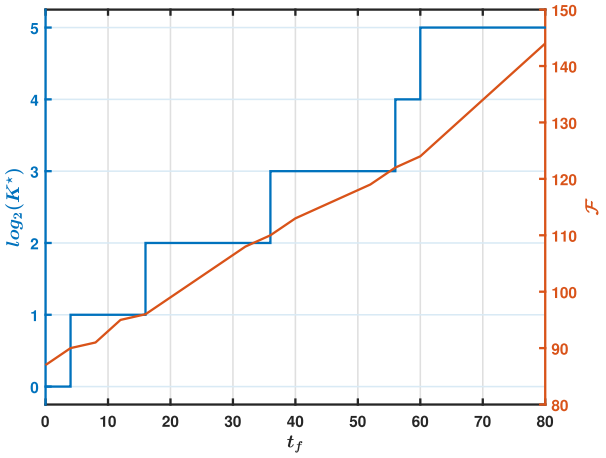


FIGURE 3. Optimal K for $G = 32$ and $M = 2$ with transmission parameters $\{B = 64, \epsilon_T = 10^{-5}\}$. The blue line represents the best PL-MU-HARQ configuration $(K^\star, C^\star)$. The orange line showcases the obtained latency performance.

better performance. For moderate values of t_f, a behavior in between these two extremes is preferable. By selecting the optimal PL-MU configuration for each t_f, one can obtain the latency curve displayed in orange.

In Fig. 4, the $\mathcal{F}$s and energy performance of both the optimal PL-MU-HARQ schemes $(K^\star, C^\star)$ with different Ms and the $\mathcal{F}$ of OS scheme, are presented. This enables one to assess the impact/benefit of using more or less TOs and compare the latency performance against the OS for varying values of t_f. In both transmission scenarios $\{B = 64, \epsilon_T = 10^{-9}\}$ and $\{B = 256, \epsilon_T = 10^{-5}\}$, there exists an initial interval of t_f where the latency performance is nearly identical across all values of M. However, beyond this interval, there is another range where the degradation (increase) in $\mathcal{F}$ occurs more rapidly in schemes with higher M. This fastest degradation on schemes with higher M is expected, as it means that there are $M - 1$ moments where the UEs have to wait for a feedback before proceeding to the next transmission. The

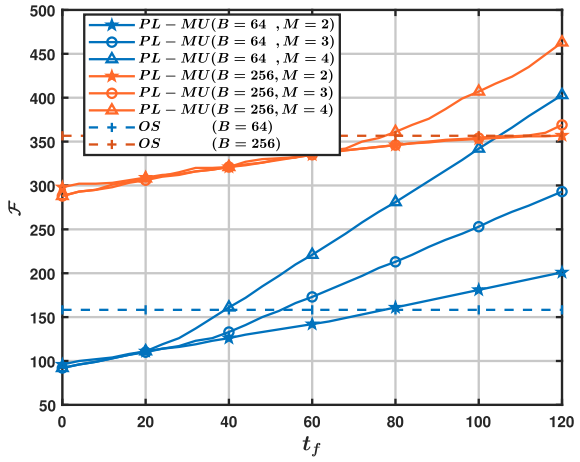


FIGURE 4. Comparing the impact of the number of transmission on the achievable $\mathcal{F}$.

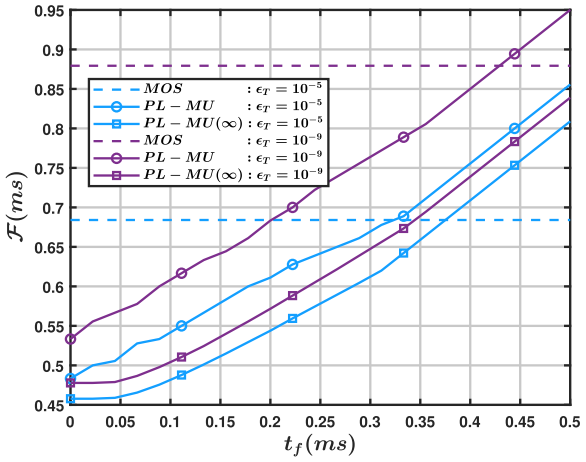


FIGURE 6. $\mathcal{F}$ considering $\frac{W_T}{G} = 180kHz$ and $B = 64$.

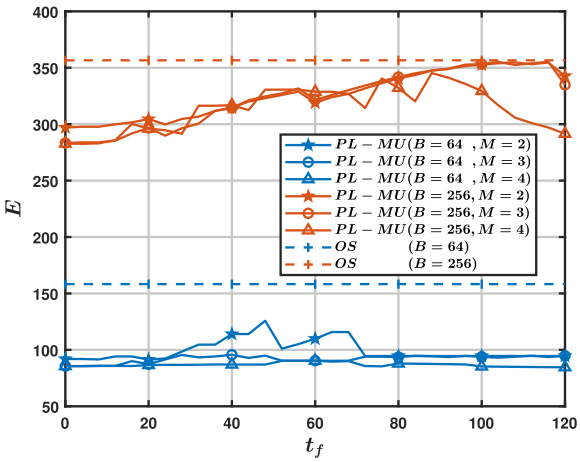


FIGURE 5. The corresponding energy expenditure E of the latency optimal schemes of different M.

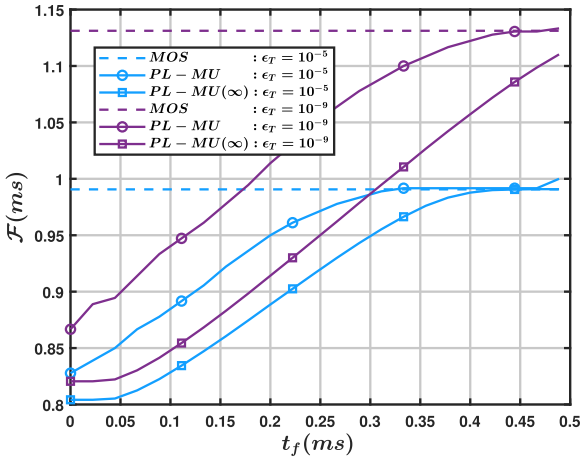


FIGURE 7. $\mathcal{F}$ considering $\frac{W_T}{G} = 360kHz$ and $B = 256$.

fact that schemes with higher M are not able to convincingly outperform schemes with lower M, makes $M = 2$ the most interesting parameterization due to its simplicity and performance. This conclusion is further reinforced by Fig. 5, where it is shown that the energy expenditure does not vary much with M. Nevertheless, for an AWGN there is no substantial gain in using more that 2 TOs, justifying the fact that the remainder of the results consider $M = 2$. Comparatively to the OS scheme, both Fig. 4 and Fig. 5 show that by using the PL-MU-HARQ instead of a OS scheme, one is able to obtain a simultaneous latency and energy reductions.

In Fig. 6 the $\mathcal{F}$s of a OS scheme, a PL-MU-HARQ with $G = 32$ and a PL-MU-HARQ with $G = \infty$ are plotted, considering $\frac{W_T}{G} = 12 \times 15kHz$. This corresponds to the LTE 12 sub-carriers separated by $15kHz$, which is also a valid 5G configuration. It can be seen that, even for considerable high t_f, PL-MU-HARQ can reduce the overall latency and help the system meet the URLLC latency requirement. The same analysis can be performed in Fig. 7 for $B = 256$ but considering a sub-carrier separation of $30kHz$, enabled by the 5G numerology. In this case, one can even find that OS surpasses the URLLC general 1ms target, and that it could be satisfied by the PL-MU-HARQ if t_f was not too high.

B. ENERGY CONSTRAINED TRANSMISSIONS

The results regarding the energy constrained case, are presented in this subsection. The objective is to quantify the latency gains of the PL-MU-HARQ comparatively to the one-shot latency when operating within an energy budget. this the subsection follows a similar structure to the power constrained case. First, the performance of PL-MU-HARQ configurations (K, C, M) is assessed as a function of t_f. Then, the latency performance of the PL-MU-HARQ scheme is compared with that of OS, for different energy budgets and t_fs. Finally, the latency performance of the best PL-MU-HARQ configuration is compared with the OS, by setting the transmission bandwidth accordingly to 5G numerology.

Fig. 8 illustrates the latency performance of both the SU-HARQ and three PL-MU-HARQ configurations of

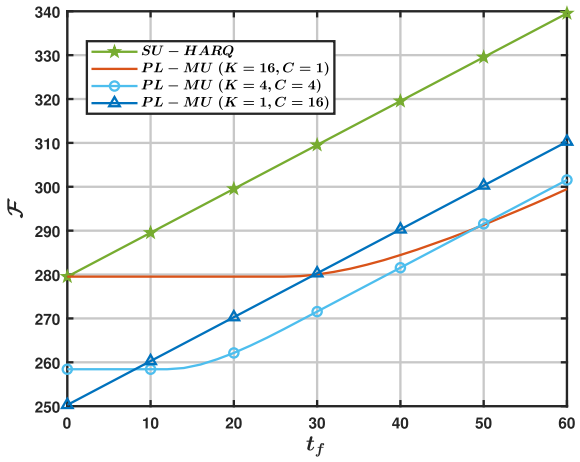


FIGURE 8. Comparing $\mathcal{F}$ of a one-shot scheme with three possible PL-MU-HARQ configurations of size $G = 16$, considering $\{B = 256, \epsilon_T = 10^{-5}\}$.

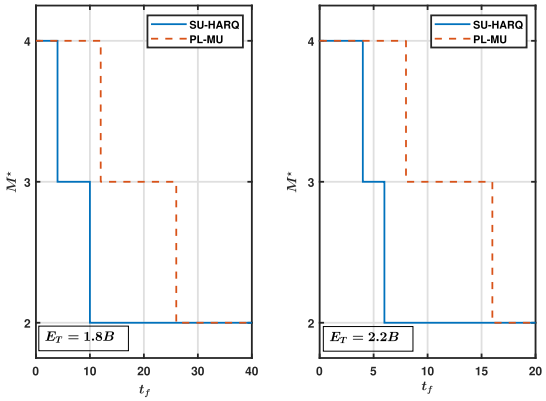


FIGURE 9. Optimal M for $G = 16$ and $\{B = 256, \epsilon_T = 10^{-5}\}$ and two different energy budgets.

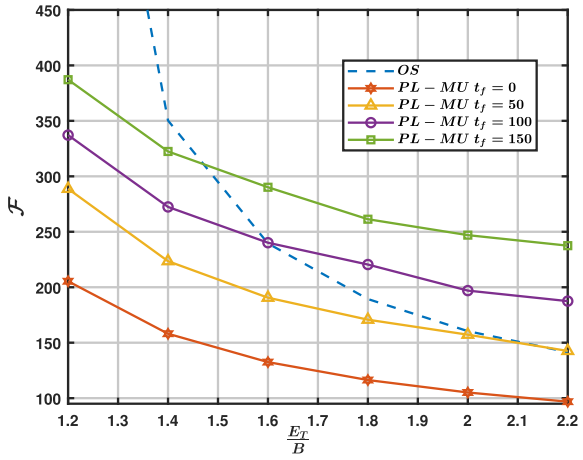


FIGURE 10. Comparing the $\mathcal{F}$ of the optimal PL-MU-HARQ scheme with the OS.

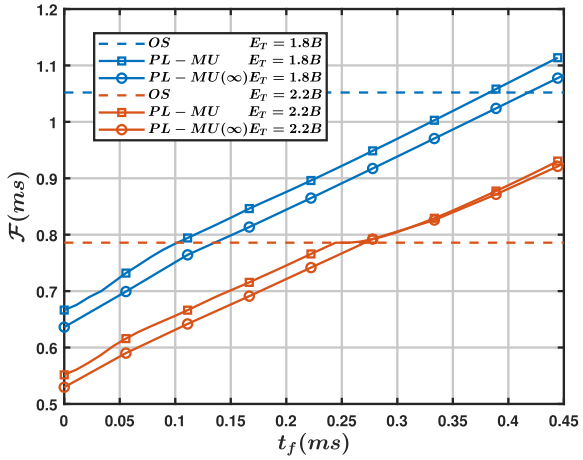


FIGURE 11. Unnormalized $\mathcal{F}$ considering $\{B = 256, \epsilon_T = 10^{-5}\}$ and $\frac{W_T}{G} = 12 \times 15kHz$.

size 16,[1] taking into account an energy budget $\frac{E_T}{B} = 1.2$. The behavior is very similar to the one observed on the constant transmission power case, since as t_f increases, the optimal K increases, making the optimal PL-MU-HARQ gradually transition from a MU-HARQ into a PL-HARQ behavior.

In Fig. 9 the optimal M ($M^\star$) of a SU-HARQ scheme and the best PL-MU-HARQ ($K^\star$, $C^\star$) parameterization is plotted. As the feedback latency overhead rises with both M and t_f, a predictable decrease in $M^\star$ is noted in both SU-HARQ and PL-MU-HARQ, as t_f increases. Moreover, it can be seen that the PL-MU-HARQ $M^\star$ is never lower than the SU-HARQ one, which is justified by *Theorem* 2 which states that the PL-MU-HARQ always has lower feedback latency overhead than the SU-HARQ. This result can be important for transmission schemes where extra TOs lead to increased diversity. One example is schemes that incorporate frequency hopping between TOs.

[1] A group of 16 is chosen as it already closely approximates the scheme's asymptotic performance, as will be evident in the results.

In Fig.10 the latency of the optimal PL-MU-HARQ ($K^\star$, $C^\star$, $M^\star$) scheme is compared with the OS latency for varying values of t_f and $\frac{E_T}{B}$. It is possible to verify that for lower energy budgets, the PL-MU-HARQ is able to outperform the OS scheme, even for high values of t_f. As the energy budget increases, so does the relative performance of the OS scheme. Nevertheless, if t_f is not too high (in this case $t_f < 50$), the OS is never able to outperform the PL-MU-HARQ scheme.

The unnormalized asymptotic latency of OS, PL-MU-HARQ with $G = 16$ and PL-MU-HARQ with $G = \infty$ is illustrated in Fig. 11. These results assume a bandwidth $\frac{W_T}{G} = 12 \times 15kHz$, which aligns with the common 5G/LTE numerology. These results highlight the advantages of PL-MU-HARQ. Indeed, they demonstrate that PL-MU-HARQ can outperform the OS scheme in practical feedback latency scenarios. More precisely, the $\frac{E_T}{B} = 1.8$ curves, show that the PL-MU-HARQ can meet the 1ms target latency even in situations where the OS scheme falls short. Additionally, it is noteworthy that PL-MU-HARQ with $G = 16$ operates

closely to the asymptotic performance of an infinite group size, indicating that the asymptotic performance is practically achievable. Since the original results are taken with a normalized unit of time, considering different values of sub-carrier separation changes the time unit, but not the proportional latency reduction between OS and PL-MU-HARQ shown in Fig. 11. Taken together, these findings demonstrate that PL-MU-HARQ exhibits superior latency performance, particularly in the low-energy regime. Using low energy budgets in the uplink is crucial, as it has direct impact on the UEs battery runtime. Another observation is that the latency gains of the PL-MU-HARQ become more pronounced when t_f decreases. This means that increasing the computational resources of the BS (such that t_f decreases) as a mean to reduce the uplink latency, is particularly relevant in the PL-MU-HARQ scheme.

VII. CONCLUSION

In this paper, two new GCG schemes were proposed. The first, named PL-HARQ is able to mitigate feedback latency overhead. Indeed, is is shown that the PL-HARQ is able to attain, at the same time, the latency performance of a OS scheme and the efficiency of a stop-and-wait feedback scheme. The second GCG scheme, PL-MU-HARQ, is a fusion between PL-HARQ and MU-HARQ. This pairing was motivated by the fact that their strengths and weaknesses complement each other. MU-HARQ is able to improve the one-shot latency, however it is heavily impacted by the feedback latency. PL-HARQ is able to mitigate feedback latency overhead, but has no latency minimization mechanism. The results demonstrated that PL-MU-HARQ successfully achieved a weighted balance between PL-HARQ and MU-HARQ behaviors as required, with the optimal behavior depending on the feedback latency value. The latency reduction decreases with the feedback latency, which is highly dependent on the BS processing time. Furthermore, the latency reduction does not evolve linearly with feedback delay as the PL-MU-HARQ is able to mitigate feedback latency overhead due to its PL-HARQ component. It was also shown that PL-MU-HARQ scheme with realistic group size is able to attain a latency performance very close to the asymptotic one.

In summary, the PL-MU-HARQ is able to outperform the OS scheme on a wide rage of feedback latency values, while relying only on regular IR-HARQ feedback signals. This makes the PL-MU-HARQ scheme a promising option for complying with the strict requirements of URLLC.

APPENDIX A
PROOF OF THEOREM 1

Consider a group of G UEs employing a PL-HARQ scheme with 2 TOs per UE. In this case, if all the following G conditions

$$(G-g)t^{(1)} + (g-1)t^{(2)} > t_f \quad \forall g \in [1, G], \tag{42}$$

are true, then all UEs receive the feedback signal before their turn to transmit, meaning that from the group perspective, the channel is never idle. The lowest left side value of (42) is either $(G-1)t^{(1)}$ or $(G-1)t^{(2)}$ meaning that if $(G-1)t^{(1)} > t_f$ and $(G-1)t^{(2)} > t_f$ are both true, then all the remaining $G-2$ conditions are also true. Therefore, if

$$\min\left(t^{(1)}, t^{(2)}\right)(G-1) > t_f \tag{43}$$

is true, then the conditions (42) are also true. Moreover, when (43) is verified, the Gth UE terminates its second transmission at time instant $G(t^{(1)} + t^{(2)}) = n^{(1)} + n^{(2)}$, which is equal to the OS latency. Therefore, if condition (43) is satisfied, the scheme eliminates the feedback latency overhead.

Condition (43) can be generalized for M transmissions as

$$t^{(m)} > \frac{t_f}{G-1}. \tag{44}$$

APPENDIX B
PROOF OF THEOREM 2

Consider a group of G UEs using a PL-HARQ scheme, where each UE has 2 TOs of duration $t^{(1)}$ and $t^{(2)}$, respectively. The gth UE performs the first transmission at time instant $t_i^{(1)}(g) = (g-1)t^{(1)}$ and ends it at time instant $t_e^{(1)}(g) = gt^{(1)}$. If there is no feedback delay overhead, the gth UE starts its second transmission at time instant $t_i^{(2)}(g) = Gt^{(1)} + (g-1)t^{(2)}$. The feedback signal of the first TO arrives at the UE at time instant $t_f^{(1)}(g) = gt^{(1)} + t_f$. Therefore, the feedback latency overhead of the gth UE can be obtained as

$$t_{fo}^{(1)}(g) = t_f^{(1)}(g) - t_i^{(2)}(g) = t_f - (G-g)t^{(1)} - (g-1)t^{(2)} \tag{45}$$

Looking at expression (45), one can infer that the UE with higher $t_{fo}^{(1)}(g)$ is either $g=1$ in the case where $(t^{(1)} < t^{(2)})$ or $g=G$ when $(t^{(1)} > t^{(2)})$. Hence, if after the first transmission of $U^{(G)}$ the entire group waits for

$$t_{fo}^{(1)} = \max\left[t_f - \min(t^{(1)}, t^{(2)})(G-1), 0\right] \tag{46}$$

before starting the second series of transmissions, then the whole group is able to receive their respective feedback signal before their transmission timing. The max operator is necessary to ensure that $t_{fo}^{(1)}$ is not negative and is in accordance to *Theorem* 1. A negative value of t_fo is not possible as even if the UEs receives the feedback signal before their transmission timing, they have to wait for their turn to transmit. Similarly,

$$t_{fo}^{(m)} = \max\left[t_f - \min(t^{(m)}, t^{(m+1)})(G-1), 0\right] \tag{47}$$

quantifies the amount of waiting (latency overhead) necessary on the mth TO. Furthermore, since $t^{(m)} > 0 \forall m \in [1, M]$, then $t_{fo}^{(m)} < t_f$ (47), meaning that the PL-HARQ feedback latency overhead is always lower than the MU-HARQ and SU-HARQ.

APPENDIX C
PL-MU-HARQ ASYMPTOTIC PERFORMANCE

Consider the expression for $\mathcal{F}_{PL-MU}$

$$\mathcal{F}_{PL-MU} = \sum_{m=1}^{M} Kt^{(m)} + t_{fo}^{(m)} \tag{48}$$

where

$$t^{(m)} = \frac{n^{(m)}}{w^{(m)}} \tag{49}$$

$$t_{fo}^{(m)} = \max\left[t_f - \min\left(\frac{n^{(m)}}{w^{(m)}}, \frac{n^{(m+1)}}{w^{(m+1)}}\right)(K-1), 0\right] \tag{50}$$

Consider the asymptotic case where both $C \to \infty$ and $K \to \infty$, meaning that $G \to \infty$. From [21, Theorem 1], one knows that $\lim_{C\to\infty} w^{(m)} = \frac{w^{(1)}}{\epsilon^{(m-1)}}$. Since $w^{(1)} = \frac{G}{C} = K$, then

$$\lim_{C\to\infty} t^{(m)} = \frac{n^{(m)}\epsilon^{(m-1)}}{K} \tag{51}$$

$$\lim_{K\to\infty} t_{fo}^{(m)} = \max\left[t_f - \min\left(n^{(m)}\epsilon^{(m-1)}, n^{(m+1)}\epsilon^{(m)}\right), 0\right] \tag{52}$$

meaning that in the asymptotic regime of a infinite group size, $\mathcal{F}_{PL-MU}$ is equal to

$$\mathcal{F}_{PL-MU}^{\infty} = \sum_{m=1}^{M} n^{(m)}\epsilon^{(m-1)} + \max\left[t_f - \min\left(n^{(m)}\epsilon^{(m-1)}, n^{(m+1)}\epsilon^{(m)}\right), 0\right]. \tag{53}$$

REFERENCES

[1] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.

[2] B. Galloway and G. P. Hancke, "Introduction to industrial control networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 860–880, 2nd Quart., 2013.

[3] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Müller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[4] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System level analysis of uplink grant-free transmission for URLLC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[5] M. Gerami and B. Singh, "Configured grant for ultra-reliable and low-latency communications: Standardization and beyond," *IEEE Commun. Standards Mag.*, vol. 6, no. 4, pp. 40–47, Dec. 2022.

[6] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[7] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[8] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[9] A. Avranas, M. Kountouris, and P. Ciblat, "Throughput maximization and IR-HARQ optimization for URLLC traffic in 5G systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[10] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475–2485, Nov. 2018.

[11] S. Xu, T.-H. Chang, S.-C. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[12] D. Tyrovolas, A. Chrysologou, G. Chondrogiannis, S. Tegos, P.-V. Mekikis, P. Diamantoulakis, S. Ioannidis, C. Liaskos, N. Chatzidiamantis, and G. Karagiannidis, "Slotted Aloha with code combining for IoT networks," in *Proc. IEEE Int. Medit. Conf. Commun. Netw. (MeditCom)*, Sep. 2023, pp. 164–168.

[13] A. Dataesatu, K. Sanada, H. Hatano, K. Mori, and P. Boonsrimuang, "Adaptive K-repetition transmission employing site diversity reception for 5G NR uplink grant-free URLLC," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–5.

[14] Q. Song, C. She, and F.-C. Zheng, "Optimization of repetition scheme for URLLC with diverse reliability requirements," in *Proc. IEEE 95th Veh. Technol. Conf.*, Jun. 2022, pp. 1–5.

[15] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.

[16] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2017, pp. 1–5.

[17] R. Abreu, G. Berardinelli, T. Jacobsen, K. Pedersen, and P. Mogensen, "A blind retransmission scheme for ultra-reliable and low latency communications," in *Proc. IEEE Veh. Technol. Conf.*, Jun. 2018, pp. 1–5.

[18] Z. Zhou, R. Ratasuk, N. Mangalvedhe, and A. Ghosh, "Resource allocation for uplink grant-free ultra-reliable and low latency communications," in *Proc. IEEE 87th Veh. Technol. Conf.*, Jun. 2018, pp. 1–5.

[19] G. Sun, S. Paris, Y. Hu, and K. Pedersen, "Iterative resolution and optimal scheduling of blind retransmissions for multi-user URLLC," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2021, pp. 1–6.

[20] Q. He, Y. Zhu, P. Zheng, Y. Hu, and A. Schmeink, "Multi-device low-latency IoT networks with blind retransmissions in the finite blocklength regime," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12782–12795, Dec. 2021.

[21] R. Santos, D. Castanheira, A. Silva, and A. Gameiro, "Multi-user IR-HARQ latency and resource optimization for URLLC," *IEEE Access*, vol. 11, pp. 129994–130009, 2023.

[22] E. Castañeda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.

[23] D. A. Patterson and J. L. Hennessy, *Computer Organization and Design ARM Edition: The Hardware Software Interface*, 5th ed. San Francisco, CA, USA: Morgan Kaufmann, 2013.

[24] W. Chen, S. Zhao, R. Zhang, H. Chen, and L. Yang, "Generalized user grouping in NOMA: An overlapping perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2876–2887, May 2021.

[25] C. Wang, R. Zhang, and B. Jiao, "Hybrid user grouping with heterogeneous devices in NOMA-enabled IoT networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9564–9578, Dec. 2023.

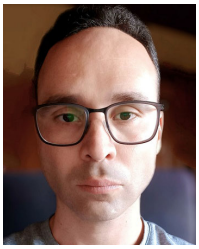
[26] W. Liu, G. Nair, Y. Li, D. Nesic, B. Vucetic, and H. V. Poor, "On the latency, rate, and reliability tradeoff in wireless networked control systems for IIoT," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 723–733, Jan. 2021.

[27] T.-K. Le, U. Salim, and F. Kaltenberger, "Enhancing URLLC uplink configured-grant transmissions," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–5.

[28] 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.

RAFAEL SANTOS received the M.Sc. degree in electrical engineering from the University of Porto, in 2018. He is currently pursuing the Ph.D. degree with the MAP-Tele Doctoral Program, University of Aveiro.

He has industry experience in digital circuit design, synthetic aperture radar, and signal processing for sensors. His research interests include signal processing for wireless communications, ultra-reliable low-latency communications, cooperative communications, real-time wireless communications, and short-blocklength channel coding.



DANIEL CASTANHEIRA received the Licenciatura (ISCED level 5) and Ph.D. degrees in electronics and telecommunications from the University of Aveiro, in 2007 and 2012, respectively. He is currently an Auxiliary Researcher with Instituto the Telecomunicações, Aveiro, Portugal. He has been involved in several national and European Projects, namely RETIOT, SWING2, PURE-5GNET, HETCOP, COPWIN, and PHOTON, within the FCT Portuguese National Scientific Foundation, and CODIV, FUTON, and QOSMOS with the FP7 ICT. In 2011, he was with Departamento de Eletrónica, Telecomunicações e Informática, Aveiro University, as an Assistant Professor. His research interests include signal processing techniques for digital communications and radar, with an emphasis for physical layer issues, including channel coding, precoding/equalization, and interference cancelation.

ADÃO SILVA received the M.Sc. and Ph.D. degrees in electronics and telecommunications from the University of Aveiro, in 2002 and 2007, respectively. He is currently an Associate Professor with DETI, University of Aveiro, and a Senior Researcher with Instituto de Telecomunicações. He has participated in several national and European projects. He has led several research projects in broadband wireless communications at the national level. He served as a member for TPC at several international conferences. Currently, he is an Associate Editor of IEEE Access and *IET Signal Processing*. He has published over 150 technical papers in international journals and conference proceedings. His research interests include multicarrier-based systems, cooperative networks, precoding, multiuser detection, and massive MIMO.

ATÍLIO GAMEIRO received the Licenciatura and Ph.D. degrees from the University of Aveiro, in 1985 and 1993, respectively. He is currently an Associate Professor with the Department of Electronics and Telecommunication Engineering, University of Aveiro, and a Researcher with Instituto de Telecomunicações, Pólo de Aveiro, where he is also the Head of the Group. His industrial experience includes a period of one year with the BT Laboratories and one year with NKT Elektronik. His research interests include signal processing techniques for digital communications and communication protocols, within this research line he has done work for optical and mobile communications, and either at the theoretical and experimental level. He has published over 200 technical papers in international journals and conferences. His current research interests include space-time-frequency algorithms for the broadband wireless systems and cross-layer design. He has been involved and has led IT or University of Aveiro participation on more than 20 national and European projects.