IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 October 2023, accepted 10 November 2023, date of publication 17 November 2023,
date of current version 22 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3334256

RESEARCH ARTICLE

Multi-User IR-HARQ Latency and Resource Optimization for 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 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 under Grant 2022.08005.PTDC.

ABSTRACT The ultra-reliable low-latency communications (URLLC) tight latency requirements paired with transmission of small payload packets motivates the development of techniques that reduce or eliminate the need for dynamic scheduling. This justifies the study of grant free (GF) leveraged techniques in order to reduce both the latency and control signaling overhead. Previous works considered preallocating resources not only for the first transmission, but also for all possible IR-HARQ transmissions, effectively reducing the scheduling latency and control signaling overhead. However, this has several drawbacks, as it translates into wasted resources. To address these issues, we propose a group-based preallocation method combined with IR-HARQ. Initially, a pool of preallocated resources is assigned to a group of users, which then cooperatively use IR-HARQ feedback signals to distribute, on the fly, the resources amongst them without collisions. The proposed method has two phases: a preallocation phase that takes place once at the group formation stage and a transmission phase which happens at each uplink transmission. The transmission parameters for all possible transmission scenarios are selected at the preallocation stage, with the goal of reducing the latency under reliability and energy constraints. The transmission parameters are obtained through a constrained latency optimization procedure, which considers the stochastic nature of the underlying process. We prove that, asymptotically, the proposed scheme is able to reduce the latency, at least, down to the average latency of any single user (SU) HARQ. The numerical results show that the latency and resources wastage is significantly reduced comparatively to single user IR-HARQ with preallocated resources.

INDEX TERMS URLLC, low-latency, grant-free, multi-user, control-networks, multi-user diversity.

I. INTRODUCTION

Due to the envisioned high heterogeneity of 5G use cases, three broad types of services were defined, enhanced mobile broadband (eMMB), massive machine type communications (mMTC) and ultra-reliable low-latency communications (URLLC) [1]. The 5G URLLC aims to service mission critical applications, such as augmented reality, vehicular-to-vehicular communications and industrial automation [1]. Even within URLLC applications, we may find some service heterogeneity, as we can find both periodic and aperiodic traffic with varying latency and reliability requirements [2], [3], [4]. Nevertheless, all the applications require an exchange of small packets with simultaneous low-latency and high reliability requirements, making them difficult to comply [5]. The tight latency requirement combined with the fact that URLLC traffic uses small sized packets, forces the system to work on the finite block length (FBL) regime. This in turn implies a gap between the Shannon capacity and the effective sustainable transmission rate of the channel. In fact, assuming the Shannon capacity for short traffic, can greatly overestimate the effective capacity of the channel [6].

In [7] the authors showed that the gap between the FBL and the Shannon capacity is reduced when employing a feedback channel. The maximal reduction of this gap is achieved through the utilization of unequal

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

transmission/retransmission sizes [8], motivating the study of hybrid automatic repeat request (HARQ) optimization techniques [9], [10], [11], [12]. In [9] and [10] the authors focused on optimizing the average throughput of an incremental redundancy HARQ (IR-HARQ) scheme. In [9], the optimization was attained through the computation of the optimal sizes of the two transmission rounds of a HARQ scheme, while in [10] both the sizes and the transmission power of each of the M IR-HARQ transmission rounds were computed. Following this, the energy-latency trade-off of a IR-HARQ operating in FBL regime is studied in [11]. The results for different combinations of feedback latency and delay budget showed us that the optimal number of transmission rounds depends on the feedback latency. In fact, if the feedback latency is too high, it can become the main source of latency and using IR-HARQ might not be beneficial. Nevertheless, it was concluded that for reasonable feedback latency, an optimized IR-HARQ scheme is able to comply with lower delay budgets than a one-shot scheme using the same average energy. Therefore, despite its feedback delay overhead, the IR-HARQ scheme may have an important role on URLLC.

Another important source of latency on the uplink is the dynamic scheduling procedure that precedes every data transmission. Semi-persistent scheduling (SPS) was introduced in Long Term Evolution (LTE) to efficiently cope with this issue for voice over IP (VoIP) traffic [13]. In VoIP the packets are small and new transmissions occur periodically. Therefore, dynamic scheduling would lead to a high control signaling overhead [14]. SPS removes the need for scheduling requests, reducing the latency and the control signaling overhead. As several URLLC applications exchange small sized and periodic traffic, grant free (GF) access techniques, like SPS, can be important to realize such applications, as in VoIP. The GF removes the need for grant acquisition by preallocating the transmission resources to either a single user equipment (UE) or a group of UEs. Preallocating resources for all the possible HARQ rounds of a single UE translates into resource wastage [15], even in periodic traffic, as the retransmissions are not always necessary. On other hand, preallocating the resources to a group of UEs keeps the GF transmission benefits while reducing the resource wastage [15].

The works [15], [16], [17], [18], [19], [20], [21] consider group-based resource preallocation methods. Despite their distinct approaches, all of these schemes can substantially minimize resource wastage in comparison to single UE preallocation methods. Most of these works are described using high-level models that do not have an explicit unit of time, being the main focus of these works the reliability and resource efficiency. To the best of our knowledge, no group-based preallocation schemes, optimized for latency reduction, have been proposed so far.

In summary, this work builds upon two established points in the existing literature, 1) in single UE scenarios IR-HARQ reduces the latency in relation to the one-shot method improving energy efficiency but reducing resource usage efficiency; 2) group-based methods improve resource usage efficiency. In the latter, we may still divide in dynamic methods, which may turn ineffective due to the high signaling requirements, and blind methods that reduce signalling but introduce collisions, thus decreasing reliability. In this work the group-based preallocation method is combined with IR-HARQ with the aim of developing a multi-user IR-HARQ scheme. The goal is to improve resource usage efficiency, eliminate dynamic signaling overhead and collisions, and reduce latency, by exploiting multi-user diversity. These goals are achieved by optimizing both the sizes and transmission power of each of the M IR-HARQ transmission rounds under latency, reliability and energy constraints. This is a challenging problem due to the stochastic nature of the underlying process. Namely, the number of active users over the IR-HARQ rounds is a stochastic process which controls the partition of available resources among the UEs.

A. RELATED WORK

Assigning a configured grant to a group of UEs, rather than just a single UE, has emerged as a promising approach to efficiently meet the requirements of URLLC. The works [15], [16], [17], [18], [19], [20], [21] consider group-based resource preallocation methods. Shared [16], [17] or both dedicated (1st transmission) and shared resources (retransmissions) [15], [18], [19], [20], [21] can be preallocated. Shared resources can be accessed randomly leading to collisions, whereas dedicated resources suffer no collisions [15], [16], [17], [18], [19], [20], [21]. Collisions could be avoided by using base station (BS) HARQ signaling [15], solved through successive interference cancellation (SIC) [17], [18], [19], [20], [21], or the probability of collision minimized by optimally selecting the number of retransmissions [16] or the number of shared resources and UEs [21]. Despite their distinct approaches, all of these schemes can substantially minimize resource wastage in comparison to single UE preallocation methods.

Another promising approach being explored in the literature, is the joint-scheduling of eMBB and URLLC traffic [22], [23], [24]. Notably, a study in [22] revealed orthogonal slicing of eMBB and URLLC traffic to lead to an increase in packet drops, primarily due to insufficiently allocated resources. The problem of joint-scheduling of eMMB and URLLC traffic, through puncturing is studied in [23] and [24]. In [23] an optimization problem that maximizes the minimum expected achieved rate of eMMB UEs, while fulfilling the URLLC requirements, is formulated. Likewise, in [24] a deep reinforcement learning approach is taken in order to perform joint-scheduling of eMMB and URLLC traffic, through puncturing.

B. CONTRIBUTIONS

In this paper we propose a multi-user IR-HARQ scheme. The proposed method considers both time and multi-user

diversity in order to achieve a trade-off between latency and energy dimensions, in contrast to single-user IR-HARQ, which only takes time diversity into account. The scheme operates on the uplink and assumes periodic URLLC traffic, as it is expected in several URLLC applications, specially in industrial networks [3], [4], [19], [25]. In summary, the main contributions of this work include:

- Proposal of an multi-user IR-HARQ scheme, where the UEs use the IR-HARQ feedback signals to distribute the available resources. The proposed scheme merges the feedback channel of all the cooperating UEs and multicasts the feedback information of each UE to the entire cooperating group. As such, collisions are eliminated, and feedback overhead is identical to that of a single user IR-HARQ scheme.
- The proposed multi-user IR-HARQ scheme achieves a latency that is at least as low as the average latency of any single-user IR-HARQ scheme.
- Proof that the proposed group-based IR-HARQ scheme achieves a latency lower than the one achieved by the single-user IR-HARQ scheme (even lower than the average latency) for a common energy efficiency target.
- Formulation of a stochastic programming problem to optimize the size and power of each multi-user IR-HARQ round. The main challenge i the stochasticity of the underlining process together with the interplay among different problem dimensions (latency, energy, and reliability).
- Proposal of a projected gradient method to solve the considered optimization problem. For the case of two IR-HARQ rounds, an optimal method is devised which exploits the problem convexity.

The results show that the group based scheme reduces the latency and minimizes the wasted resources, even for a small group size when compared to the single-user counterpart for the same average consumed energy in the finite blocklength regime.

II. PRELIMINARIES

The goal of this section is to introduce the problem of SU-HARQ solutions and the motivation to develop proposed MU-HARQ scheme.

As previously stated, the IR-HARQ is a promising URLLC building block. The basic IR-HARQ operation involves a transmitter and a receiver communicating over a wireless channel. The IR-HARQ operation can be summarised as follows:

1) Transmitter encodes the data in a codeword.
2) The codeword is sent over the air to the receiver.
3) The receiver decodes the codeword and checks for errors.
4) If errors are found, the receivers sends a NACK to the transmitter, otherwise an ACK is sent.
5) The transmitter receives the feedback from the receiver. If either an ACK is received or the maximum number of transmissions were already performed performed, the process terminates. Otherwise, the transmitter encodes the original data, generating extra redundancy which is sent to the receiver.
6) The receiver jointly decodes the original codeword with the extra redundancy received so far, and checks for errors.
7) Go to step 4.

In such a scheme, due to the possibility of early termination when the transmitter receives an ACK, some of the preallocated resources may go unused, leading to inefficiency.[1] This issue becomes more significant as the number of parallel SU-HARQ schemes increases. If however the set of resources is jointly allocated to several URLLC users, the vacant resources can be reused according to the needs of the remaining traffic. This is illustrated in Fig. 1 which shows that the combined bandwidth of individual SU-HARQ schemes (dotted blue line) is considerably higher compared to the bandwidth of each individual SU-HARQ scheme (dashed blue line). Both remain constant since the resources are statically pre-allocated. Furthermore, the green line reveals a decreasing trend in the number of active SU-HARQ schemes after each transmission round.While it is unpredictable which SU-HARQ scheme will fail during the first transmission, we can estimate the probability of failure for a certain number of schemes. This behavior is depicted by the decreasing green line in Fig. 1, indicating a reduction in the number of active User Equipment (UE) after each transmission, resulting in resource wastage. This expected reduction in active UEs can be considered a form of multi-user diversity and has been effectively explored in prior studies on group-based preallocation techniques [15], [16], [17], [18], [19], [20], [21], in order to mitigate resource wastage. In the case of SU-HARQ, if the bandwidth allocated is kept fixed, then an optimization can be achieved by defining the duration of the different retransmissions and respective power. When resources in the frequency domain increase, additional degrees of freedom can be used to optimize either the energy per bit or latency. To illustrate these ideas, let us assume we have resources in a time-frequency grid of dimension w_T. In the context of uncoordinated SU-HARQ, aiming to meet reliability and latency requirements, the average resource utilization is denoted as a fraction α of w_T. Then a coordinated allocation scheme, if fully efficient, could either accommodate $\frac{1}{\alpha}$ more users or reduce the latency by a factor close to α. Although such reasoning is based on averages and, due to the randomness in the availability of the extra resources, we will not reach exactly this value, it allows us to anticipate that with the proper coordination schemes relevant improvements may be achieved.

III. SYSTEM MODEL

This work considers a SISO uplink AWGN channel, where a group of G UEs, with similar channel statistics, is formed.

[1]The resources may be used by other non-URLLC traffic at the cost of a more complex scheduling but is wasted for other URLLC users.

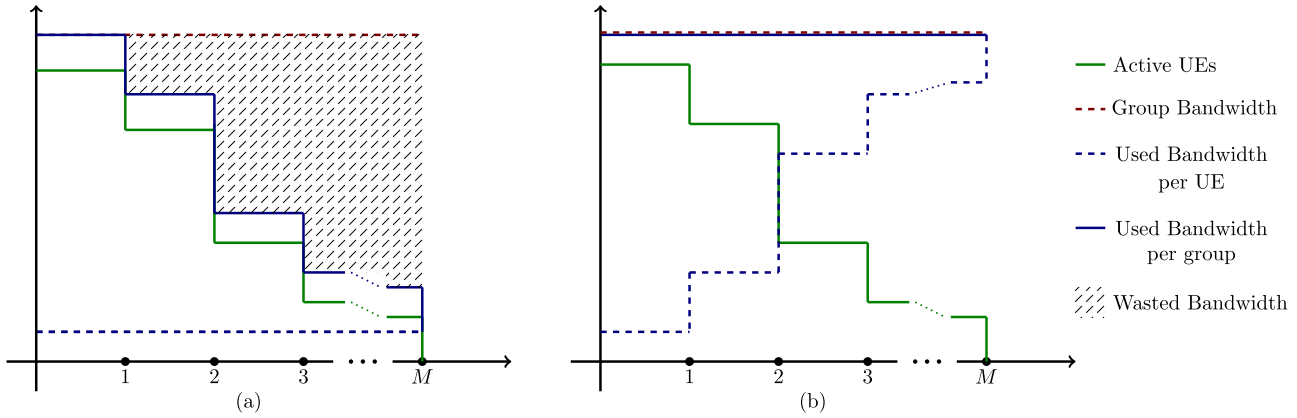


FIGURE 1. The time evolution of the number of active UEs, used bandwidth per UE and wasted resources when using several SU-HARQ in parallel (a) vs what we want to achieve (b).

TABLE 1. Summary of notations.

Notation	Definition
MU-HARQ	The proposed multi-user IR-HARQ.
SU-HARQ	Regular single user IR-HARQ.
MSU-HARQ	Multiple SU-HARQ operating in parallel.
$X \sim B(n, \epsilon)$	Binomial RV with n trials and probability of success ϵ.
$P_{bin}(n, x, \epsilon)$	Probability of x successes out of n when their probability is 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).
G	Group size.
$\mathcal{U}$	IDs of every group member.
$U^{(g)}$	Group member with ID g.
$\mathcal{X}^{(m)}$	R.V. modeling the number of 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.
$x^{[m]}$	Realization of $\mathcal{X}$ up to the mth round, such that $x^{[M]} = x$.
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 transmission of realization x.
$t_{x^{(m)}}$	Time duration of the mth transmission for realization $x^{(m)}$.
$\mathcal{N}_{x^{[m]}}$	Channel uses of the first m transmissions for realization $x^{[m]}$.
$\mathcal{P}_{x^{[m]}}$	Transmission power of the first m transmissions of x.
$\mathcal{T}_{x^{[m]}}$	Transmission duration of the first m transmissions of x.
$\mathcal{E}_{x^{[m]}}$	Probability of error of the first m transmissions when $\mathcal{X} = x$.
$\Theta_{x^{[m]}}$	MU-HARQ parameters; $\Theta_{x^{[m]}} = (\mathcal{N}_{x^{[m]}}, \mathcal{P}_{x^{[m]}})$
Δ_Θ	Average latency
E_x	Average energy expended when $\mathcal{X} = x$.
Γ_x	Wasted resources when $\mathcal{X} = x$.
$\preceq, \nabla$	Vector wise comparison and vector gradient, respectively.
$\|\|.\|\|$	Vector norm.
$a \| b$	Divergence between the distributions a and b.
*	Used to to denote a optimal solution of a given problem.

Each UE transmits periodically B new bits of information to a BS, while complying with the URLLC QoS constraints (latency, reliability) and an energy budget. Transmissions are performed over a preallocated bandwidth, which is then split among UEs. In the following subsection, the proposed MU-HARQ scheme is described in more detail. The notation is developed throughout the text, but for readers' convenience, we have summarized it in Table 1.

A. MU-HARQ PROCEDURE

The MU-HARQ scheme comprises two distinct phases: an activation phase and an operating mode. In the activation phase, the base station (BS) allocates a bandwidth of w_T to a group of G users and assigns them identification (ID) numbers, such that $\mathcal{U} = \{U^{(1)}, \cdots, U^{(G)}\}$ is the group, and $U^{(1)}$ is the UE with ID 1. In the subsequent operating phase, time is divided into periods of size t_T. Each t_T is subdivided into M transmission rounds. During each period, the G users utilize ACK/NACK indicators distributed via a multicast channel by the BS, to distributively reach a consensus on how to divide the resources amongst themselves. The activation phase occurs only once, making its long-term signalling overhead negligible.

In the first transmission round, each of the G users transmit their messages of duration $t^{(1)}$ through orthogonal bandwidth $w^{(1)} = w_T/G$ with power $p^{(1)}$. In subsequent rounds, users who have not achieved success reconfigure their bandwidth, power, and time duration to transmit incremental redundancy, which at the BS is combined with the data received from previous rounds. The amount of bandwidth allocated depends on the number of users that are still active. The BS transmits all ACK/NACK indicators for active users through a multicast channel, enabling each user to compute their available bandwidth. Indeed, if at the end of round $m-1$, there are still $x^{(m)}$ unsuccessful users, the available bandwidth per users becomes $w^{(m)} = w_T/x^{(m)}$. Consequently, the current active UE with lowest ID uses the first transmission band $[f_0, f_0 + w_T/x^{(m)}]$ while the one with lowest kth ID uses frequency band $[f_0 + (k-1)w_T/x^{(m)}, f_0 + kw_T/x^{(m)}]$. After computing the bandwidth to be used, each user refers to a pre-compute table where the power and duration of the next $(m+1)$th round are defined for each possible scenario $(x^{(m)})$. This process repeats until the M transmission rounds

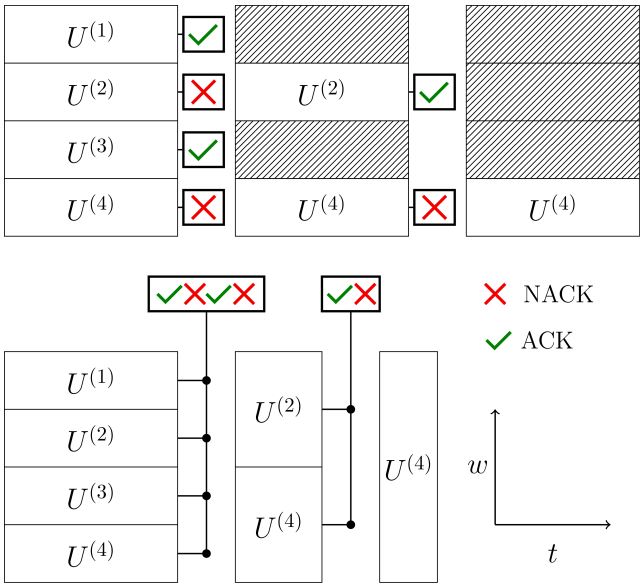


FIGURE 2. The same realization of MSU-HARQ (top) and MU-HARQ (bottom) for $G = 4$ and $M = 3$. All UEs transmit the same amount of channel uses in both cases, but with different transmission times. Dashed lines represent the MSU-HARQ not-used (wasted) resources.

are exhausted, and restarts on the subsequent operating phase.

An possible outcome of the previously described process is exemplified for $G = 4$ in Fig. 2. All the 4 UEs perform the first transmission in parallel through dedicated resources. After the first transmission, $U^{(1)}$ and $U^{(3)}$ receive an ACK and will not need a second transmission, while $U^{(2)}$ and $U^{(4)}$ receive a NACK and are set to perform a second transmission. Note that the BS multicasts the feedback signals of all UEs to the entire group, meaning that $U^{(2)}$ and $U^{(4)}$ know that they are the only active UEs for the second transmission. Knowing this, they are able to divide the entire bandwidth w_T into two, where $U^{(2)}$ gets the first half of the bandwidth as it has the lowest ID of the two. On the third transmission, $U^{(4)}$ is the only active UE, and therefore knows that it can transmit through the entire bandwidth w_T. Regarding the necessary signalling, it can be seen in Fig. 2 that the signalling transmitted by the BS is identical in both the MU-HARQ and MSU-HARQ schemes. The key difference lies in the way the feedback signals are delivered: whereas in the MSU-HARQ they are unicasted to each respective UE, in the MU-HARQ they are multicasted to the entire group.

In this description, the decision-making process occurs at the user premises. However, an equivalent implementation can be achieved with all computations performed at the BS. In such case, the bandwidth, power and round duration information would be sent through the multicast feedback channel.

In the following subsections, we will present the MU-HARQ scheme model and its generalized notation.

B. MU-HARQ SCHEME

In the proposed MU-HARQ scheme the available resources are shared among G UEs. A bandwidth w_T over a time span

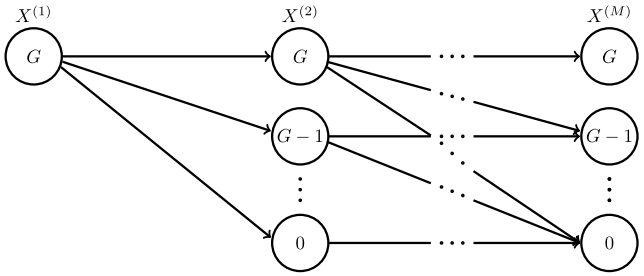


FIGURE 3. Stochastic process representing the number of active UEs over the M rounds of a MU-HARQ scheme.

t_T is preallocated by the BS to a group of G statistically identical UEs. Each UE employs an IR-HARQ scheme with a maximum number of M transmissions, with the goal of meeting the target block error rate ϵ_T under the time constraint t_T. All active UEs transmit in parallel through a dedicated frequency band. The IR-HARQ splits t_T into M parts, named rounds. However, not all UEs transmit in the mth round, only those whose codeword was not successfully decoded by the BS in all previous rounds do so. The number of active UEs at the mth round is defined by the R.V. $X^{(m)}$ with realizations $x^{(m)} \in \{0, 1, \ldots, G\}$. To avoid resource wastage and improved latency performance, the full bandwidth w_T is split among active UEs, being assigned a bandwidth $W^{(m)} = \frac{w_T}{X^{(m)}}$ to each active UE in the mth round. The history of the number of active UEs forms the stochastic process (SP),

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

represented in Fig. 3, whose state space (all possible SP values) is

$$\mathcal{S} = \{x^{(1)}, \cdots, x^{(M)} : 0 \leq x^{(i)} \leq x^{(j)} \leq G,\ j \leq i \leq M \in \mathbb{N}\} \tag{2}$$

meaning that each SP realization $x \in \mathcal{S}$. We further define $x^{[m]} = \{x^{(1)}, \cdots, x^{(m)}\} \subseteq x \in \mathcal{S}$ to be used as an indexing set. For each $x \in \mathcal{S}$, there is a corresponding $\mathcal{W}_{x^{[m]}}$ realization $w_{x^{[m]}} = \{w_{x^{(1)}}, \cdots, w_{x^{(m)}}\} = \{\frac{w_T}{x^{(1)}}, \cdots, \frac{w_T}{x^{(m)}}\}$ being the bandwidth normalized such that $\frac{w_T}{G} = 1$.

Since all group members are assumed to be statistically identical, the optimal MU-HARQ procedure is considered identical for all users. The MU-HARQ scheme is considered a parameterizable method, being the parameterizable parameters denoted by $\Theta \in \Lambda$, and Λ the feasible parameters set. The parameters are defined in the preallocation phase, before any transmission. The parameterization includes a set of parameters for each of the M rounds, where for a given round, both the size (number of channel uses) and power of the sub-codeword to be transmitted are parameterizable. The MU-HARQ parameters Θ influence the underlying stochastic process $\mathcal{X}$ probability distribution. To reflect this aspect, the probability distribution of the stochastic process $\mathcal{X}$ is assumed to belong to the parametric family $P_\Theta = \{P(\bullet; \Theta) | \Theta \in \Lambda\}$. The probability distribution can be

expressed with more detail in the form

$$P(\mathcal{X}=x;\Theta) = \prod_{m'=1}^{M-1} P(X^{(m'+1)}=x^{(m'+1)}|X^{(m')}=x^{(m')};\Theta_{x^{[m']}}) \quad (3)$$

where $\Theta_x = \{\theta_{x^{[1]}}, \theta_{x^{[2]}}, \cdots, \theta_{x^{[M]}}\}$ denotes the selected MU-HARQ parameters on the corresponding SP realization $x \in \mathcal{S}$, $\theta_{x^{[m]}}$ represents the selected mth round MU-HARQ parameter and $\Theta_{x^{[m]}} = \{\theta_{x^{[1]}}, \theta_{x^{[2]}}, \cdots, \theta_{x^{[m]}}\}$ the parameters of the first m transmissions. Notice that, there is a set of parameters for each different realization of $\mathcal{X}$, as different available bandwidths imply different optimal transmission parameters. The expression (3) follows from the chain rule for probability, where the transition kernel $P\left(X^{(m')}=x^{(m)}|X^{(m'-1)}=x^{(m'-1)};\Theta_{x^{[m'-1]}}\right)$ models the impact of the MU-HARQ scheme on the number of active users in the next round, given the MU-HARQ parameters selected in previous round. Let $x, x' \in \mathcal{S}$ be two different stochastic process realizations overlapping up to the mth round, $\left(x^{[m]}=x'^{[m]}\right)$, then due to the causal nature of the process, the parameterization of both realizations up to the m transmission is equal $\left(\Theta_{x^{[m]}}=\Theta_{x'^{[m]}}\right)$. Therefore, one has to compute $\binom{G+m-2}{G-1}$ parameters at the mth round, resulting in a total $\sum_{m=1}^{M}\binom{G+m-2}{G-1}$ parameters. In the following sections $\Theta_{\mathcal{S}^{(m)}} = \{\theta_{x^{[m]}} : \forall x \in \mathcal{S}\}$ denotes all the $\binom{G+m-2}{G-1}$ parameters that can be used on the mth transmission. Considering all rounds, the total number of parameters grows asymptotically as $\mathcal{O}(M^G)$ when the number of users is fixed, and as $\mathcal{O}(G^{M-1})$ when the number of rounds is fixed. Note that when $X^{(m)}=0$ (data of all UEs successfully decoded) no transmission takes place and no additional parameter is required. When this happens, the remaining preallocated resources are unused, effectively wasting these resources.

C. MU-HARQ METRICS

The MU-HARQ parameters must be selected with the aim of maximizing a given figure of merit under system specific constraints, e.g. energy, reliability and latency. For MU-HARQ the average energy E_Θ, average probability of error ϵ_Θ, latency t_Θ, average latency Δ_Θ and average wasted resources ψ_Θ are given by

$$E_\Theta = \mathbb{E}_{\mathcal{X}\sim P_\Theta}[E_\mathcal{X}] = \sum_{x\in\mathcal{S}} P(\mathcal{X}=x)E_x \quad (4)$$

$$\epsilon_\Theta = \mathbb{E}_{\mathcal{X}\sim P_\Theta}[\epsilon_\mathcal{X}] = \sum_{x\in\mathcal{S}} P(\mathcal{X}=x)\epsilon_x \quad (5)$$

$$\Delta_\Theta = \mathbb{E}_{\mathcal{X}\sim P_\Theta}[\Gamma_\mathcal{X}] = \sum_{x\in\mathcal{S}} P(\mathcal{X}=x)\Gamma_x \quad (6)$$

$$\psi_\Theta = \mathbb{E}_{\mathcal{X}\sim P_\Theta}[\psi_\mathcal{X}] = \sum_{x\in\mathcal{S}} P(\mathcal{X}=x)\psi_x \quad (7)$$

$$t_\Theta = \max_{\mathcal{X}\sim P_\Theta}[\Gamma_\mathcal{X}] = \max_{x\in\mathcal{S}} \Gamma_x \quad (8)$$

where E_x, ϵ_x, Γ_x and ψ_x denotes the energy, probability of error, latency and wasted resources of some realization $x \in \mathcal{S}$.

Let a parameter be divided in two parts $\theta = (n, p)$, where n and p are the sub-codeword size and power allocation, respectively, allowing the definition of $\mathcal{N}_{x^{[m]}} = \{n_{x^{[1]}}, \cdots, n_{x^{[m]}}\}$ and $\mathcal{P}_{x^{[m]}} = \{p_{x^{[1]}}, \cdots, p_{x^{[m]}}\}$ which are analogous to $\Theta_{x^{[m]}}$. The energy, reliability, latency and average wasted resources for a SP realization $x \in S$ are given by

$$E_x = \frac{1}{G}\sum_{m=1}^{M} x^{(m)} n_{x^{[m]}} p_{x^{[m]}} \quad (9)$$

$$\epsilon_x = \epsilon(\Theta_x) \quad (10)$$

$$\Gamma_x = \sum_{m=1}^{M} t_{x^{[m]}} \quad (11)$$

$$\psi_x = t_\Theta w_T - \sum_{m=1}^{M} x^{(m)} n_{x^{[m]}}. \quad (12)$$

For the probability of error (10), we will use the following expression from [11],

$$\epsilon_{x^{[m]}}(\Theta_{x^{[m]}}) = Q\left(\frac{\sum_{i=1}^{m} n_{x^{[i]}}\log(1+p_{x^{[i]}}) - B\log(2)}{\sqrt{\sum_{i=1}^{m}\frac{n_{x^{[i]}}p_{x^{[i]}}\left(2+p_{x^{[i]}}\right)}{\left(1+p_{x^{[i]}}\right)^2}}}\right). \quad (13)$$

This expression is an approximation of the original PPV bound [6] for AWGN channels with unit power variance. It represents the achievable error rate with optimal channel coding. In the previous expressions the relation between the sub-codeword size and time-bandwidth product is

$$n_{x^{[m]}} = t_{x^{[m]}} w_{x^{(m)}} = t_{x^{[m]}}\frac{w_T}{x^{(m)}}. \quad (14)$$

From (13)(14) we verify that $\Theta_{x^{[m]}}$ implicitly defines $\mathcal{E}_{x^{[m]}} = \{\epsilon_{x^{[1]}}, \cdots, \epsilon_{x^{[m]}}\}$ and $\mathcal{T}_{x^{[m]}} = \{t_{x^{[1]}}, \cdots, t_{x^{[m]}}\}$, i.e., the error probability and transmission duration of each round. Having defined ϵ_x the transition kernel may be parameterized as follows

$$P\left(X^{(m)}=x^{(m)}|X^{(m-1)}=x^{(m-1)};\Theta_{x^{[m-1]}}\right) = P_{bin}\left(x^{(m)}, x^{(m-1)}, \frac{\epsilon_{x^{[m-1]}}}{\epsilon_{x^{[m-2]}}}\right), \quad x \in \mathcal{S}. \quad (15)$$

In section V a method is proposed to optimize the MU-HARQ parameters. First the corresponding problem is formulated and then a projected gradient based method is proposed to find the best MU-HARQ parametrization.

D. PARTICULAR MODEL INSTANCES

The general MU-HARQ was introduced in the previous section. In the following sub-sections we present and discuss two particular model examples to allow a better understanding of the model and establish a link to a method from the literature [11].

1) SINGLE-USER IR-HARQ

The SP state space for the mth round of a single user scenario ($G = 1$) is $\mathcal{S} = \{(x^{(1)} = 1, \cdots, x^{(m)} = 1, x^{(m+1)} = 0, \cdots, x^{(M)} = 0) : m \leq M\}$. When $m < M$ then no parameter is required for the remainder $M - m$ rounds as the transmission procedure already ended (BS successfully decoded user data). As M rounds are considered and $G = 1$, then only M parameters are required, one new parameter per round. Let $\theta_{x'[m]} = (n_{x'[m]}, p_{x'[m]})$ denote such a parameter, then from (4), (5) and (8) it follows that

$$E_\Theta = n_{x'[1]} p_{x'[1]} + \sum_{m=2}^{M} \epsilon_{x'[m-1]} n_{x'[m]} p_{x'[m]}, \tag{16}$$

$$\epsilon_\Theta = \epsilon\left(\Theta_{x'}\right), \tag{17}$$

$$t_\Theta = \sum_{m=1}^{M} t_{x'[m]} = \frac{w_T}{G} \sum_{m=1}^{M} n_{x'[m]} = \sum_{m=1}^{M} n_{x'[m]} \tag{18}$$

where $x' = \{x^{(1)} = 1, \ldots, x^{(M)} = 1\}$. The proposed model extends the one presented in [11] by adding the notion of transmission bandwidth, which is necessary to draw comparisons between multi-user and single-user scenarios. Indeed, looking at (16)(17)(18) we see that the model described in [11] is a particular case of our model as the expressions become equivalent when applied to a single user IR-HARQ with a constant transmission bandwidth equal to $\frac{W_T}{G} = 1$.

2) MU-HARQ WITH TWO ROUNDS

Let us now consider a scenario with two rounds and multiple users. For this case, the SP state space is $\mathcal{S} = \{(x^{(1)} = G, x^{(2)} = y) : 0 \leq y \leq G\}$. From (4), (5) and (8) follows that

$$E_\Theta = n_{x[1]} p_{x[1]} + \frac{1}{G} \sum_{x \in \mathcal{S}} x^{[2]} n_{x[2]} p_{x[2]} P_{bin}(x^{(2)}; G, \epsilon_{x[1]}), \tag{19}$$

$$\epsilon_\Theta = \sum_{x \in \mathcal{S}} \epsilon_{x[2]} P_{bin}(x^{(2)}; G, \epsilon_{x[1]}), \tag{20}$$

$$t_\Theta = \frac{1}{w_T} \max_{x \in \mathcal{S}} \left(G n_{x[1]} + x^{(2)} n_{x[2]} \right). \tag{21}$$

This scenario corresponds to the simplest model for the MU-HARQ case ($M = 2$) and is used in the following section, where it is shown that the proposed method asymptotically achieves the SU-HARQ average latency.

IV. ASYMPTOTIC PERFORMANCE

In this section we prove that the proposed MU-HARQ scheme can, asymptotically, reduce the latency down to the average latency of any SU-HARQ scheme. We first prove it for a two transmission round scheme i.e. $M = 2$, and then show that it can be directly generalized for a generic number of transmissions.

Consider a SU-HARQ scheme with $M = 2$ rounds transmitting through a bandwidth $\mathcal{W}^{SU}_{x'} = \{w_{x'^{(1)}} = 1, w_{x'^{(2)}} = 1\}$ with parameterization $\Theta^{SU}_{x'}$, defining the probability of error $\mathcal{E}^{SU}_{x'}$ and transmission duration $\mathcal{T}^{SU}_{x'} = \mathcal{N}^{SU}_{x'}$ due to (13) and (14), respectively. We further assume that the SU-HARQ scheme satisfies the target QoS, i.e., $\epsilon^{SU}_\Theta = \epsilon_T$ and $t_T = t^{SU}_\Theta$. We can apply the SU-HARQ parameterization to a MU-HARQ scheme, meaning that $\Theta^{MU}_x = \Theta^{SU}_{x'}$ resulting in $\epsilon^{MU}_x = \epsilon^{SU}_{x'}$ and $t^{MU}_x = \{t^{SU}_{x'[1]}, t^{SU}_{x'[2]}/w^{MU}_{x[2]}\}$ (13) (14). Therefore, the MU-HARQ transmission duration depends on the available bandwidth. In an asymptotic regime of a infinite number of group members, the available bandwidth at the mth transmission round is described by the following theorem,

Theorem 1: Let $\epsilon^{(m)}$ be the error probability of the mth transmission round, identical for all $x \in \mathcal{S}$ SP realizations, and $W^{(m)}$ the R.V. describing the available bandwidth for the mth MU-HARQ round. Then, in the asymptotic regime of an infinite group size G, $W^{(m)}$ is equal to $\frac{w_{x[1]}}{\epsilon^{(m-1)}}$ with probability 1, i.e., for any $\zeta \in \mathbb{R}^+$.

$$\lim_{G \to \infty} P\left(|W^{(m)} - \frac{w_{x^{(1)}}}{\epsilon^{(m-1)}}| < \zeta \right) = 1 \quad \forall m \in [1, M], \tag{22}$$

Proof: Appendix.B. ∎

From Theorem 1 we know that for an asymptotically infinite group size,

$$w^{MU}_{x^{(m)}} = \frac{w^{SU}_{x'^{(m)}}}{\epsilon^{SU}_{x'[m-1]}} = \frac{w^{MU}_{x^{(m)}}}{\epsilon^{MU}_{x[m-1]}} = \frac{w_T}{G \epsilon^{MU}_{x[m-1]}} = \frac{1}{\epsilon^{MU}_{x[m-1]}}, \tag{23}$$

where $\frac{w_T}{G} = 1$. This means that in the asymptotic regime, the MU-HARQ with parametrization $\Theta^{MU}_x = \Theta^{SU}_{x'}$ has a transmission latency defined by $t^{MU}_\Theta = t^{SU}_{x'[1]} + \epsilon^{SU}_{x'[1]} t^{SU}_{x'[2]}$ (14), which is equal to the average latency of the original SU-HARQ scheme. Likewise, for a MU-HARQ with M transmission rounds, the obtained latency is

$$t^{MU}_\Theta = t^{SU}_{x[1]} + \sum_{m=2}^{M} \epsilon^{SU}_{x[m-1]} t^{SU}_{x[m]}, \tag{24}$$

which is equal to the average SU-HARQ latency with M transmissions. The latency reduction mechanism is exemplified in Fig. 2. In the second transmission, both $U^{(2)}$ and $U^{(4)}$ have twice the bandwidth compared to the MSU-HARQ scheme, as illustrated in Fig. 2. Therefore, in MU-HARQ, $U^{(2)}$ and $U^{(4)}$ can transmit the same number of symbols in half the time it takes in MSU-HARQ.

The general result (24) is achieved by simply using the SU-HARQ parameters. However, in the optimization framework we optimize all these parameters for each $x \in \mathcal{S}$, in order to achieve the SU-HARQ average latency with a realist group size.

V. LATENCY MINIMIZATION PROBLEM AND OPTIMIZATION METHODS

In this section we describe a latency minimization method that considers a URLLC target reliability paired with an energy budget constraint. The direct latency optimization problem is complex and to circumvent it, we show that the optimal solution of the latency optimization problem can

be obtained by optimally solving an energy minimization problem, whose solution is less complex to compute. This is followed by the description of two energy optimization algorithms. The first algorithm, presented in subsection V-C2, is proven to be optimal for any group size G. The second algorithm, presented in subsection V-C2, is a low-complexity sub-optimal solution that is better suited for a variable number of transmissions ($M > 2$). The optimal algorithm importance is two fold, it offers a optimal low-complexity solution for two transmission HARQ and at the same time sets the performance baseline of the sub-optimal algorithm.

A. LATENCY OPTIMIZATION PROBLEM FORMULATION

The latency minimization problem can be mathematically formulated as follows,

Problem 1:

$$T^{\star}_{\Theta}(E_T) = \min_{\Theta=(\mathcal{N},\mathcal{E})} \quad t_{\Theta} \tag{25}$$

$$s.t. \quad E_{\Theta} \leq E_T \tag{26}$$

$$\epsilon_{\Theta} \leq \epsilon_T \tag{27}$$

$$\Theta \in \Lambda \tag{28}$$

The inequality (26) defines the average energy budget constraint, (27) is the reliability constraint and (28) defines the parameters constraint, where

$$\Lambda = \{((n_{x[1]}, \epsilon_{x[1]}), \cdots (n_{x[i]}, \epsilon_{x[i]}), \cdots, (n_{x[M]}, \epsilon_{x[M]})) : 1 \leq n_{x[i]},\ 0 < \epsilon_{x[i+1]} < \epsilon_{x[i]} < 0.5,\ i \in]1, M[,\ \forall x \in \mathcal{S}\}. \tag{29}$$

Following the approach in [12], (25) is optimized over (n, ϵ), meaning that from now on $\theta = (n, \epsilon)$. Since it is not possible to explicitly define a function $p_{x[m]}\left(\mathcal{N}_{x[m]}, \mathcal{E}_{x[m]}\right), x \in \mathcal{S}$ [12], one has to obtain the resulting value of $p_{x[m]}$ with some iterative algorithm. We follow the approaches in [11] and [12] and use the successive bisection algorithm to compute $p_{x[m]}$ given $\Theta_{x[m]} = \left(\mathcal{N}_{x[m]}, \mathcal{E}_{x[m]}\right)$.

The latency minimization problem is hard to solve due to the constraint (26) which is a non-convex and non-linear function of the parameterization. In the next subsection we establish the relationship between the optimal solutions of *Problem 1* with a MU-HARQ energy minimization problem. This allow us to obtain the optimal solution of *Problem 1* through a energy minimization problem, which we show to be simpler to optimize.

B. LATENCY MINIMIZATION THROUGH ENERGY OPTIMIZATION

Let $T^{\star}_{\Theta}(E_T)$ be the optimal latency (25) as a function of the energy budget E_T. Likewise, let $E^{\star}_{\Theta}(t_T)$ be the minimal achievable average energy as a function of the delay budget t_T. Hence, $E^{\star}_{\Theta}(t_T)$ can be obtained by solving the following optimization problem,

Problem 2:

$$E^{\star}_{\Theta}(t_T) = \min_{\Theta_{\mathcal{S}}} \quad E_{\Theta} \tag{30}$$

$$s.t. \quad t_{\Theta} \leq t_T \tag{31}$$

$$\epsilon_{\Theta} \leq \epsilon_T \tag{32}$$

$$\Theta \in \Lambda \tag{33}$$

The inequality (31) defines the latency constraint, which is the only one not defined in *Problem 1*. Several works [11], [12] studied FBL energy optimization problems similar to *Problem 2*. From [11, Lemma 2], we know that the optimal solution of *Problem 2*, satisfies the constraint (32) with equality, allowing us to change it to

$$\epsilon_{\Theta} = \epsilon_T, \tag{34}$$

which simplifies *Problem 2*. Likewise, we know that in both frameworks [11], [12], $E^{\star}_{\Theta}(t_T)$ is proven to be a monotonic decreasing function. The proofs are not applicable for MU-HARQ, however the monotonicity property is numerically verifiable for a MU-HARQ scheme, as shown for $M = 2$ in *Section* VI. This motivates the formulation of the following conjecture,

Conjecture 2: Let $E^{\star}_{\Theta}(t_T)$ represent the optimal energy value of Problem 2 as a function of the delay budget t_T given the number of information bits B, target probability of error ϵ_T, number of transmission rounds M and group size G. Then, $E^{\star}_{\Theta}(t_T)$ is a strict monotonic decreasing function.

From *Conjecture 2* follows that $E^{\star}_{\Theta}(t_T)$ is an injective function meaning that the optimal solution of *Problem 2* satisfies (31) with equality. We set the duration of the Mth round for every $x \in \mathcal{S}$ as a dependent variable

$$\mathcal{T}_x = \left\{t_{x[1]}, \cdots, t_{x[M-1]}, t_T - y : y = \sum_{m=1}^{M-1} t_{x[m]}\right\},\ \forall x \in \mathcal{S}, \tag{35}$$

removing the latency constraint that is incorporated into the objective function, further simplifying the problem. We define $\mathcal{K}$ as the set of parameters that comply with all the simplified *Problem 2* constraints. For these reasons *Problem 2* is comparatively less complex to solve than *Problem 1*. The relationship between $E^{\star}_{\Theta}(t_T)$ and $T^{\star}_{\Theta}(E_T)$ is established in the following theorem,

Theorem 3: Let $T^{\star}_{\Theta}(E_T)$ be the optimal solution of Problem 1 as a function of the target energy budget E_T and, $E^{\star}_{\theta}(t_T)$ the optimal solution of Problem 2 as a function of the delay budget t_T. Then, if Conjecture 2 is true, the following relationship is verifiable $E^{\star}_{\Theta}(t_T) = {T^{\star}_{\Theta}}^{-1}(t_T)$, i.e., $E_T = E^{\star}_{\Theta}(T^{\star}_{\Theta}(E_T))$.

Proof: Appendix. C ■

Considering these results, we know that

1) From the monotonic property of $E^{\star}_{\Theta}(t_T)$ follows that the function $f_{\Theta}(t_T) = E^{\star}_{\Theta}(t_T) - E_T$ has, at most, a single zero.
2) From *Theorem 3*, follows that the zero of $f(t_T)$ is equal to $T^{\star}_{\Theta}(E_T)$.

Considering both these points, one can obtain the optimal solution of *Problem 1*, i.e. $T^{\star}_{\Theta}(E_T)$, by obtaining the zero of $f_{\Theta}(t_T)$. For this, we apply Brent's method to $f_{\Theta}(t_T)$ due

to its balance between reliability and convergence speed. Whenever Brent's method requires a query to $f_\Theta(t_T)$ (and consequently $E^\star_\Theta(t_T)$), we solve *Problem 2* using the corresponding t_T as the latency constraint. This approach pushes all the complexity of solving *Problem 1* into *Problem 2*.

C. ENERGY OPTIMIZATION FRAMEWORK

In this section two algorithms to solve *Problem 2* are proposed. The first algorithm is optimal for a two-transmission ($M = 2$) MU-HARQ system with arbitrary group size. The second algorithm is suboptimal but has significantly lower complexity, making it well-suited for systems with an arbitrary number of transmissions ($M > 2$).

1) OPTIMAL ALGORITHM

Considering a two transmission MU-HARQ, one can formulate the following theorem,

Theorem 4: For a two transmissions MU-HARQ scheme ($M = 2$) *with probability of error and number of channel uses in the first round* ϵ'_1 *and* n'_1*, and an arbitrary group size* ($G \geq 1$)*, the problem (30) is convex over the set of probabilities of error of the second transmission* $\mathcal{E}_{\mathcal{S}^{(2)}}$ *if* $\max(\mathcal{E}_{\mathcal{S}^{(2)}}) < 0.5$.

Proof: Appendix.D. ∎

From *Theorem 4*, one can employ a convex optimization algorithm and obtain the optimal $\mathcal{E}_{\mathcal{S}^{(2)}}$ given a pair $\Theta_{\mathcal{S}^{(1)}} = (n_{\mathcal{S}^{(1)}}, \epsilon_{\mathcal{S}^{(1)}})$. Therefore, this optimization algorithm should be computed for each possible $\Theta_{\mathcal{S}^{(1)}}$. Being both upper and lower bounded (14)(31)(33), one has to introduce an adequate discretization step for $\mathcal{E}_{\mathcal{S}^{(1)}}$ denoted $\Delta\mathcal{E}$. This approach is viable for two reasons. First, the possible number of pairs $\Theta_{\mathcal{S}^{(1)}}$ does not scale with the group size G. Second, as $\mathcal{N}_{\mathcal{S}^{(2)}}$ is defined through the relationship (14)(35), $\mathcal{E}_{\mathcal{S}^{(2)}}$ are the only remaining unknowns and can be obtained through convex optimization - *Theorem 4*. Hence, we iterate through all possible combinations of $(n_{\mathcal{S}^{(1)}}, \epsilon_{\mathcal{S}^{(1)}})$ and for each pair we solve *Problem 2* by employing the log barrier method [26] on the inequality constraints and the Newton's method to obtain the KKT satisfying solution [26].

To ease the search over $\mathcal{E}_{\mathcal{S}^{(1)}}$, we do the steps on $C_{\mathcal{S}^{(1)}}$ domain, $C_{\mathcal{S}^{(1)}} = Q^{-1}(\mathcal{E}_{\mathcal{S}^{(1)}})$, as it allows a simple linear step.

2) SUB-OPTIMAL ALGORITHM

In order to eliminate the exhaustive numerical search, we perform joint optimization over all design variables $(\mathcal{N}_{\mathcal{S}[M-1]}, \mathcal{E}_{\mathcal{S}})$ using the projected gradient descent method. Due to the nature of objective function, *Problem 2* has a intrinsic barrier on the inequality constraints Λ. Thus, if the gradient step is small enough, the inequality constraints are preserved and only the equality constraint (34) remains problematic. Nevertheless, after each gradient step, we have to perform a projection step in order to push the current solution back onto the feasible set. These procedures are described in *Algorithm 1*, where we see that on the main loop, the gradient step (line 1) is always followed by the gradient projection (line 2). Since the equality constraint only depends on $\mathcal{E}_{\mathcal{S}^{(M)}}$, the projection step onto the equality constraint is only performed on $\mathcal{E}_{\mathcal{S}^{(M)}}$ (line 2.4)(line 2.5). However, after the projection step some elements of $\mathcal{E}_{\mathcal{S}^{(M)}}$ might violate their inequality constraints (line 2.6). When this happens we undo the projection onto the equality constraint, and set the violating elements on the edge off their corresponding violated constraint (line 2.7). At this point, if the resulting vector lies on the feasible set the projection operator returns, otherwise it repeats the previously described process without considering the elements in $\overline{\mathcal{A}}$, as they already are on the edge of the constraint. This process is repeated until the resulting solution is on the feasible set $\mathcal{K}$.

Algorithm 1 Sub-Optimal Algorithm

1 **Input:** G, ϵ_T, t_T, B, M
2 **Output:** Θ
3
4 *# Energy Minimization* : Main Loop
5 **while** *Stop Condition* == *0* **do**
6 ** Gradient step.**
7 **1** $\Theta' \leftarrow \Theta - \alpha \nabla E_\Theta$
8 ** Projection step.**
9 **2** $\Theta \leftarrow P(\Theta')$
10
11 *# Projection Step* : $P(\Theta')$
12 **2.1** $\overline{\mathcal{A}} \leftarrow \varnothing$
13 **2.2** $\mathcal{E}_{\mathcal{S}^{(M)}} \leftarrow \mathcal{E}'_{\mathcal{S}^{(M)}}$
14 **while** $\mathcal{E}_{\mathcal{S}^{(M)}} \notin \mathcal{K}$ **do**
15 ** Active elements indices : only these are used when projecting onto the equality constraint.**
16 **2.3** $\mathcal{A} \leftarrow \mathcal{S}^{(\mathcal{M})} \setminus \overline{\mathcal{A}}$
17 ** Compute β using $\mathcal{A}$.**
18 **2.4** $\beta \leftarrow \frac{\epsilon_{\Theta'} - \epsilon_T}{\| \nabla_{\mathcal{E}_{\mathcal{S}^{(M)}}} \epsilon_{\Theta'} \|^2}$
19 ** Project active elements onto equality constraint.**
20 **2.5** $\mathcal{E}''_{\mathcal{A}} \leftarrow \mathcal{E}'_{\mathcal{A}} - \beta \nabla_{\mathcal{E}'_{\mathcal{A}}} \epsilon_{\Theta'}$
21 */*Update Index of inequality violating elements. */*
22 **2.6** $\overline{A} \leftarrow \overline{A} \cup \{\mathcal{A}' : \mathcal{E}''_{\mathcal{A}'} \notin \Lambda\}$
23 */*Elements are set in the edge of the violated*
24 *constraint */*
25 **2.7** $\mathcal{E}'_{\overline{\mathcal{A}}} \leftarrow \arg\min_{\mathcal{E} \in \Lambda} ||\mathcal{E} - \mathcal{E}''_{\overline{\mathcal{A}}}||$
26 ** Set the value to the projection step output.**
27 **2.8** $\mathcal{E}_{\mathcal{S}^{(M)}} \leftarrow \mathcal{E}'_{\mathcal{S}^{(M)}}$

D. COMPLEXITY ANALYSIS

In this section, we evaluate the complexity of both the optimal and sub-optimal algorithms.

1) OPTIMAL ALGORITHM

The optimal algorithm relies on Brent's method. On each Brent's method iteration, two main sequential operations are performed:

1) Numerical search on the $\zeta = 2\sum_{m=1}^{M-1}\binom{G+m-2}{G-1}$ parameters of the first $M-1$ transmission rounds.
2) Newton method to obtain the $\zeta^{(M)} = \binom{G+M-2}{G-1}$ error probabilities of the last round.

The numerical search needs a quantization step on the continuous variables. The number of different parameter values one has to sweep during the numerical search is inversely proportional to the quantization step. Then, the numerical search has a complexity $\mathcal{O}\left((\frac{1}{\Delta})^{\zeta}\right)$, where Δ is the applied quantization step size. The Newton method is used to obtain the values of $\zeta^{(M)}$ parameters, which translates into a complexity of $\mathcal{O}\left((\zeta^{(M)})^{3.5}\right)$. Hence, the overall complexity of the optimal algorithm is $\mathcal{O}\left((\frac{1}{\Delta})^{\zeta}(\zeta^{(M)})^{3.5}k_{Brent}\right)$, where k_{Brent} represents the number of iterations required for Brent's method to converge. Analysing ζ, one is able to verify that $\zeta = 2$ for $M = 2$, independently of the group size G, and that it exhibits a fast increase when the inequalities $M > 2$ and $G > 1$, are satisfied. This rapid increase in complexity for $M > 2$, motivated the development of the sub-optimal algorithm.

2) SUB-OPTIMAL ALGORITHM

The sub-optimal algorithm relies on the Brent's method as well. On each iteration of Brent's method, two main sequential operations are performed:

1) Gradient descent on all $\zeta + \zeta^{(M)}$ parameters.
2) Projection step on all $\zeta^{(M)}$.

The gradient descent step has the cost of computing $\zeta + \zeta^{(M)}$ derivatives and updating the values of $\zeta + \zeta^{(M)}$ variables. Hence, the complexity of the gradient step is $\mathcal{O}\left(C(\zeta + \zeta^{(M)})\right)$, where C is the cost of computing each derivative and updating the values of the parameters. The projection step is a simple linear projection, being its complexity $\mathcal{O}\left(C\zeta^{(M)}\right)$. Therefore, the overall complexity of the sub-optimal algorithm is $\mathcal{O}\left(C(\zeta + 2\zeta^{(M)})k_{PGD}k_{Brent}\right)$ where k_{PGD} represents the number of iterations required for the projected gradient descent algorithm to converge.

From previous complexity analysis, one may verify that the complexity of the sub-optimal algorithm scales linearly with the number of parameters, whereas the optimal algorithm demonstrates exponential complexity scaling.

As a final remark, we would like to emphasize that, since the algorithm outputs the parameters for every $\mathcal{X}$ realization, it does not have to be executed before every transmission, provided that these parameters are saved in a look-up table (LUT). Indeed, the algorithm has to be executed only when there are changes in either the QoS, the constraints or the quality of the channel. In any other scenario, the UEs just have to access the LUT to determine the optimal transmission parameters, given the current group state.

VI. RESULTS

In this section we present the numerical results. The main objective of this work is to assess the latency performance and resource efficiency gains of using the proposed MU-HARQ scheme for G UEs instead of the MSU-HARQ scheme. In the MSU-HARQ setup, G optimal SU-HARQ schemes operate concurrently on dedicated bandwidths $\frac{w_T}{G} = 1$. To achieve this assessment, we use two schemes from the existing literature [10], [11] as benchmark. We selected these schemes because they have been proven to offer optimal performance in terms of latency [11] and average latency [10]. By comparing our proposed MU-HARQ scheme to the optimal latency performance of MSU-HARQ, we can quantify the achievable latency gains. Furthermore, by comparing our results to the optimal average-latency performance, we can assess how closely our system approaches the asymptotic limit discussed in Section IV.

TABLE 2. Simulation parameters.

Parameter	Values
M	$\{2, 3, 4\}$
G	$\{3, 10, 50\}$
$\{B, \epsilon_T\}$	$\{256, 10^{-5}\}, \{64, 10^{-9}\}$
E_T/B	$[0.85, 1.3]$

Due to the optimal latency performance and non-cooperative nature of the benchmark scheme, the obtainable gains can only be attributed to the inherent MU-HARQ cooperation. Indeed, in order to attain a fair comparison, the minimal obtainable latency of both schemes are compared for the same w_T, E_T and ϵ_T. As the feedback latency increases, the latency performance of both schemes degrades equally, resulting in a constant performance difference between the two for varying feedback latency values. Therefore, all the comparisons are drawn considering zero feedback latency. We focus on two different transmission parameters $\{B = 256, \epsilon_T = 10^{-5}\}$ as these are used in [10] and [11] and $\{B = 64, \epsilon_T = 10^{-9}\}$, which is a more extreme scenario, aligned with industrial networks [1]. The simulation parameters are laid out throughout the text, but we have summarized them in Table 2 for readers' convenience.

We denote a MSU-HARQ scheme optimized with procedures [10] and [11] by SU and $\mathbb{E}[SU]$, respectively. Let us define t_{MU} and t_{SU} as the lowest achievable latency by the MU-HARQ and SU, given an average energy budget and a reliability target. Likewise, let $\mathbb{E}[t_{SU}]$ be the lowest achievable average latency by $\mathbb{E}[SU]$, given the same conditions. With this second approach, we are able to evaluate the minimum group size required to achieve the asymptotic performance discussed in Section IV. It is now possible to quantify the latency reduction incurred by switching from MSU-HARQ to MU-HARQ as $1 - \frac{t_{MU}}{t_{SU}}$. Notice that, when comparing MU-HARQ and SU with $\mathbb{E}[SU]$, the y axis should be interpreted as the delay budget t_Θ for the first two, and as average delay Δ_Θ for the latter. We also compare the resource wastage of SU with our MU-HARQ scheme. Resource wastage happens when some of the statically allocated resources are not used. We use ψ_{SU} and ψ_{MU} to

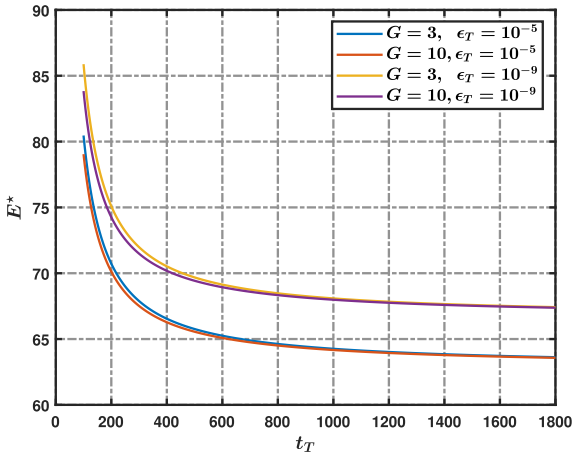


FIGURE 4. $E^\star(t_T)$ is strict monotonic decreasing for $B = 64$ and $M = 2$.

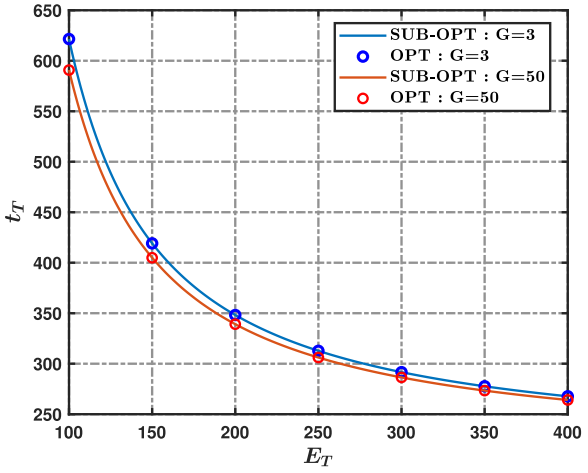


FIGURE 5. Optimal Algorithm vs Sub-Optimal Algorithm for $M = 2$, $B = 256$ and $\epsilon_T = 10^{-5}$.

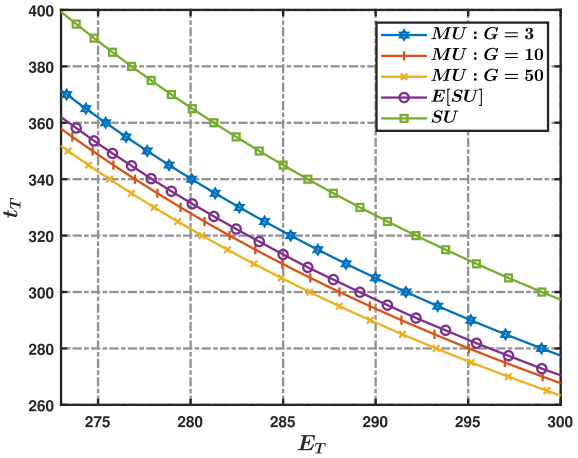


FIGURE 6. Comparing MU-HARQ solution $M = 2$ with SU and $\mathbb{E}[SU]$ for $M = 2$, B = 256, $\epsilon_T = 10^{-5}$.

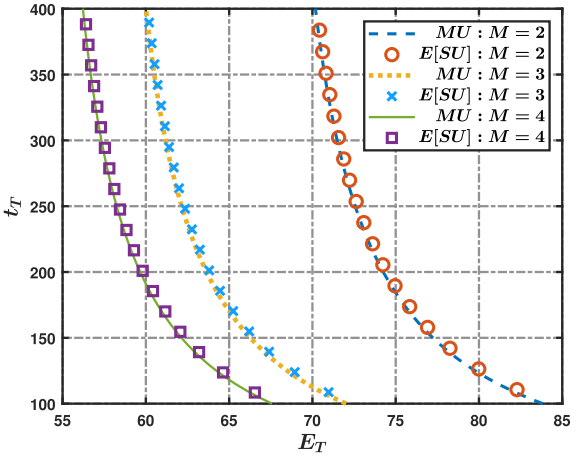


FIGURE 7. Comparing $\mathbb{E}[SU]$ with MU-HARQ for B = 64, $\epsilon_T = 10^{-9}$ and $G = 10$.

denote the average number of channel uses wasted (unused) by SU and MU-HARQ, respectively. On the MSU-HARQ scheme, resources are wasted when the UE achieves early success and does not need to do the M transmissions Fig. 2. On the MU-HARQ scheme, there are wasted resources when at some point, all the G UEs achieved early transmission success. The resource wastage reduction can be quantified through $1 - \frac{\psi_{MU}}{\psi_{SU}}$.

In Fig. 4 the function $E^\star(t_T)$ for $M = 2$ is presented for several different parameter combinations. We can see that the function is strictly decreasing. This indicates that *Conjecture 2* is true for $M = 2$, meaning *Problem 1* can be optimally solved for these parameters by applying the Brent's method paired with the optimal energy optimization algorithm, as previously described. Results for $M > 2$ were also obtained, whose results reinforced the belief that *Conjecture 2* is indeed true.

Furthermore, it can be observed that both the optimal and sub-optimal algorithms yield identical solutions for $M = 2$, $B = 256$, and $\epsilon_T = 10^{-5}$ (the parameterization used in [11]), which is evidenced by Fig. 5, thereby attesting to the sub-optimal algorithm excellent performance.

In Fig. 6 we compare the performance of the optimal MU-HARQ solution with SU and $\mathbb{E}[SU]$, for $M = 2$. It is possible to observe that the MU-HARQ with $G = 10$ already outperforms $\mathbb{E}[SU]$. This means that, the MU-HARQ with a group size of 10 is already able to meet the SU-HARQ optimal average latency. In fact, for a group size of $G = 50$ the MU-HARQ scheme widens the gap between the optimal average latency and optimal SU-HARQ average latency. This is justified by the result of *Section IV*, which proved that as G increases to infinity it is possible to reduce the latency to the average latency of any SU-HARQ scheme. The proof of section *Section IV* focuses on the asymptotic case and does not preclude that for finite G we may get a solution below this limit. Since $G = 10$ outperforms the proven asymptotically achievable results, we further investigate the performance of this group size. In Fig. 7, we observe that even for $B = 64$ and $\epsilon_T = 10^{-9}$, a group size of 10 is enough to reduce the latency down to the

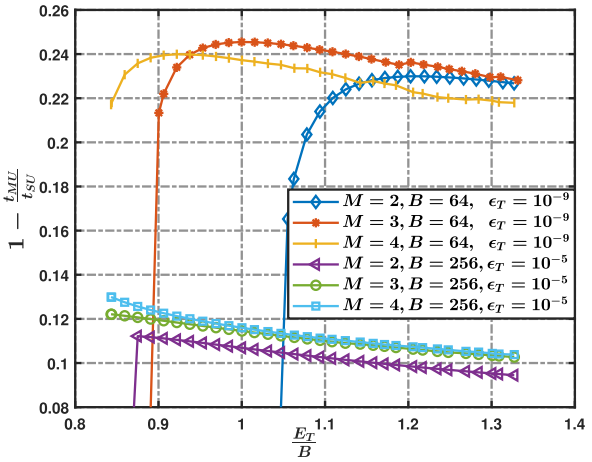


FIGURE 8. Latency reduction when using MU-HARQ instead of *SU* [11] for $G = 10$.

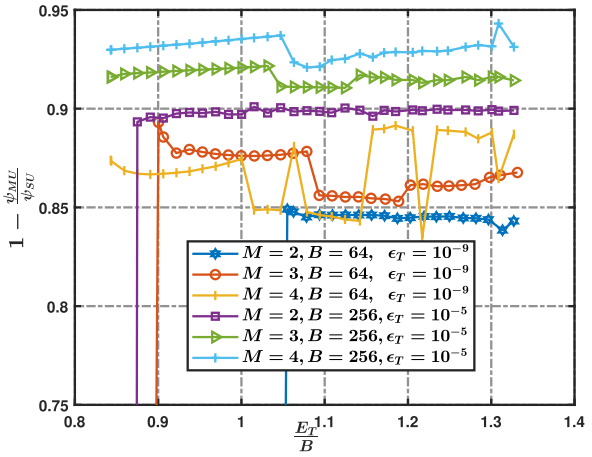


FIGURE 9. Wasted resources reduction by switching from *SU* to MU-HARQ for $G = 10$.

optimal MSU-HARQ average latency [10], even for different values of M.

The latency reduction $1 - \frac{t_{MU}}{t_{SU}}$ is represented in the y-axis of Fig.8, for $\{B = 64, \epsilon_T = 10^{-9}\}$ and $\{B = 256, \epsilon_T = 10^{-5}\}$. We can see that switching from MSU-HARQ to MU-HARQ entails a latency reduction that can go as high as 24.5% for $\{B = 64, \epsilon_T = 10^{-9}\}$ and 13% for $\{B = 256, \epsilon_T = 10^{-5}\}$ when considering the same range of target energy per information bit B. This difference in results is justified by the fact that a higher B and ϵ_T the optimal MSU-HARQ solution spends most of the delay budget on the first transmission, leaving a smaller margin for MU-HARQ to improve. It should be emphasized that these results assume zero latency feedback. Indeed, as the feedback latency increases with respect to the HARQ round duration, one can expect $1 - \frac{t_{MU}}{t_{SU}}$ to diminish while $t_{SU} - t_{MU}$ stays constant. Nevertheless, for use cases like factory automation, the propagation times will be extremely low, being the remaining feedback latency dependent on the processing time at the BS. As the BS processing power is a centralized resource, it can be increased in order to keep the feedback latency under control. The latency reduction means that it is possible to allocate the original bandwidth w_T to a smaller time interval. Therefore, the amount of preallocated resources is also reduced by the same percentage as the latency reduction depicted in Fig. 8. For this case however, this reduction is independent of the feedback latency.

Since freed resources are used by active group members, the group allocated resources are only wasted when no group member is active. On other hand, the SU-HARQ wastes resources every time an UE does not use all the M pre-allocated transmissions. As seen in Fig. 9, switching from SU-HARQ to MU-HARQ can entail a wasted resources reduction in the interval [90%, 93%], for $\{B = 256, \epsilon_T = 10^{-5}\}$ and [85%, 88%] for $\{B = 64, \epsilon_T = 10^{-9}\}$. Given a target M, there are values of $E_T < E_{min}$ to which there are no possible MU-HARQ solutions in given range, justifying the format of some of the curves seen in both Fig. 8 and Fig. 9.

One crucial consideration lies in the practicality of the optimal parameterization. For SU-HARQ, the optimization reveals that in a wide range of scenarios, approximately 90% of the delay budget is allocated to the first transmission round. Consequently, if success is not achieved in this initial round, there is a very limited time left, and power in the subsequent rounds has to increase significantly. In contrast, with MU-HARQ, the resources in the frequency domain are more likely to expand after the first round, providing greater degrees of freedom for optimization and eliminating the need for extreme values in peak power. Therefore if constant power or peak limitations are required, the advantages of MU-HARQ become considerably more evident.

VII. CONCLUSION

In this paper a URLLC distributed cooperative scheme was devised. The proposed scheme combines group-based preallocation and IR-HARQ in order to: eliminate dynamic signalling overhead without collisions and provide improved latency and resource efficiency performance. In terms of signalling, these benefits are achievable using only the standard IR-HARQ signalling. We proved that this group-scheme, by implicitly making use of the law of large numbers, is able to reduce the latency to a value as low as the average latency of any SU-HARQ scheme. For this reason, we formulated a latency minimization problem, and devised both an optimal an sub-optimal algorithm to solve it. The optimal algorithm is suitable for two transmissions since its complexity, that increases exponentially with the number of transmission, is still affordable with two transmissions. To cope with this increase in complexity, we designed a sub-optimal low-complexity algorithm, suitable for practical implementations involving a higher number of MU-HARQ transmissions. The results showed that the proposed MU-HARQ scheme is able to reduce the latency down to the average SU-HARQ latency. The amount of preallocated resources is reduced to the average used resources in the MSU-HARQ. The probability

that these resources are left unused is also reduced in the MU-HARQ.

In short, the proposed scheme is able to keep all the benefits of having persistent scheduled resources while mitigating its drawbacks and reducing the latency. These results showed that the proposed scheme can be leveraged in order to comply with strict URLLC requirements of 5G and beyond communications.

APPENDIX A
PROOF OF BINOMIAL EXPRESSION

Let us define $S^{(m)} = 1$ and $S^{(m)} = 0$ as the event of transmission sucess and transmission failure of an UE at the mth round, respectively. Therefore we know that $P(S^{(1)} = 0) = \epsilon^{(1)}$ and $P(S^{(1)} = 1) = 1 - \epsilon^{(1)}$ where $\epsilon^{(m)}$ is the probability of error on the mth round. Likewise, we know that $P(S^{(2)} = 0, S^{(1)} = 0) = \epsilon^{(2)}$ which generalizes for $P(S^{(m)} = 0, S^{(m-1)} = 0, \cdots, S^{(1)} = 0) = \epsilon^{(m)}$. Hence, we know that

$$P(S^{(2)} = 0 | S^{(1)} = 0) = \frac{P(S^{(2)} = 0, S^{(1)} = 0)}{P(S^{(1)} = 0)} = \frac{\epsilon^{(2)}}{\epsilon^{(1)}}, \tag{36}$$

which we generalize as follows

$$P(S^{(m)} = 0 | S^{(m-1)} = 0, \cdots S^{(1)} = 0) = \frac{P(S^{(m)} = 0, S^{(m-1)} = 0, \cdots, S^{(1)} = 0)}{P(S^{(m-1)} = 0, \cdots, S^{(1)} = 0)} = \frac{\epsilon^{(m)}}{\epsilon^{(m-1)}}. \tag{37}$$

When group of size G is at a state $X^{(m)} = x^{(m)}$, then it can only transit to a state $X^{(m+1)} = x^{(m+1)}$ such that $x^{(m)} \geq x^{(m+1)}$. This state transition happens when $x^{(m)}$ UEs failed the first m transmissions and from these $x^{(m)}$ still active UEs, $x^{(m+1)}$ fail the $(m+1)$th transmission round. The probability one UE failing the $(m+1)$th round knowing that it already failed the previous m rounds, is equal to $\frac{\epsilon^{(m+1)}}{\epsilon^{(m)}}$ (37). Therefore, the transition from state $X^{(m)} = x^{(m+1)}$ to $X^{(m+1)}$ follows a binomial distribution

$$X^{(m+1)} \sim B\left(x^{(m)}, \frac{\epsilon^{(m+1)}}{\epsilon^{(m)}}\right) \tag{38}$$

or more generally, the conditional binomial distribution of $X^{(m+1)}$ given $X^{(m)}$

$$X^{(m+1)} | X^{(m)} \sim B\left(X^{(m)}, \frac{\epsilon^{(m+1)}}{\epsilon^{(m)}}\right). \tag{39}$$

APPENDIX B
MU-HARQ ASYMPTOTIC BEHAVIOR

We start this proof for a MU-HARQ with $M = 2$ which is then generalized for a general M, The Chernoff bound provides a upper bound on the cumulative distribution of a binomial RV $X = B(n, p)$,

$$P(X \leq x) \leq e^{-nD\left(\frac{x}{n} \| p\right)} \tag{40}$$

being $D(a\|b)$ the Kullback-Leiber divergence,

$$D(a\|b) = a \log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right). \tag{41}$$

To show $\lim_{G\to\infty} P(X^{(2)} \geq \frac{G}{R} + 1) = 0$ where $R = \frac{G}{x^{(2)}} = \frac{w^{(2)}}{w^{(1)}}$, we introduce the random variable $\tilde{X}^{(2)}$ which represents the number of non-active group members. Hence, $\tilde{X}^{(2)} = G - X^{(2)}$ and $\lim_{G\to\infty} P(X^{(2)} \geq \frac{G}{R} + 1) = 0 \Leftrightarrow \lim_{G\to\infty} P(\tilde{X}^{(2)} \leq G - \frac{G}{R} - 1) = 0$ and we get

$$P(\tilde{X}^{(2)} \leq G - \frac{G}{R} - 1) \leq e^{-GD\left(\frac{G-\frac{G}{R}-1}{G} \| 1-\epsilon^{(1)}\right)}, \tag{42}$$

as G increases to infinity.

We know that $D(a\|b) \geq 0$. When $\frac{1}{R} = \epsilon^{(1)}$, $D\left(1 - \frac{1}{R} \| 1 - \epsilon^{(1)}\right) = 0$, which is its minimum value. On all the remaining possible values of $R \in [1, \infty] \setminus \{\frac{1}{\epsilon^{(1)}}\}$, $Z = D\left(1 - \frac{1}{R} \| 1 - \epsilon^{(1)}\right)$ being $Z \in]0, \infty[$,

$$\lim_{G\to\infty} e^{-GD\left(\frac{G-\frac{G}{R}-1}{G} \| 1-\epsilon^{(1)}\right)} = e^{-\infty Z} = 0. \tag{43}$$

This implies that, asymptotically $P(\tilde{X}^{(2)} \leq G - \frac{G}{R} - 1)$ is upper bounded by zero (42)(43). Being $P(\tilde{X}^{(2)} \leq G - \frac{G}{R} - 1)$ a probability, implies $P(\tilde{X}^{(2)} \leq G - \frac{G}{R} - 1) = 0$. Another thing to note is the following relationship

$$\lim_{G\to\infty} P(X^{(2)} \leq \frac{G}{R} - 1) = \lim_{G\to\infty} P(X^{(2)} \geq \frac{G}{R} + 1). \tag{44}$$

Therefore, as the number of group members G grows, the amount of bandwidth available per UE will be, with certainty, equal to $w^{(2)} = Rw^{(1)} = \frac{w^{(1)}}{\epsilon^{(1)}}$. Hence,

$$\lim_{G\to\infty} P(|W^{(2)} - \frac{w^{(1)}}{\epsilon^{(1)}}| < \zeta) = 1. \tag{45}$$

We can generalize this proof to a general number of transmissions M by noting that, the state probabilities at the $m-1$th round, only depend on both $\epsilon^{(m-1)}$ and the group size G (39). Hence, one can apply the same two transmission proof to any transmission round, by assuming that the current transmission round m is the second transmission of a two round IR-HARQ, where the first transmission had a probability of error $\epsilon^{(m-1)}$. We can apply this proof to successive transmissions, since the parameters $\mathcal{P}_{x^{[m-1]}}$, $\mathcal{N}_{x^{[m-1]}}$ and $\mathcal{E}_{x^{[m-1]}}$ are kept constant as the proof goes on. Hence, in general

$$\lim_{G\to\infty} P(|W^{(m)} - \frac{w^{(1)}}{\epsilon^{(m-1)}}| < \zeta) = 1 \ \ \forall m \in [1, M] \tag{46}$$

APPENDIX C
PROOF THAT $E_T = E^{\star}_{\Theta}(T^{\star}_{\Theta}(E_T))$

In this section we want to prove that $E^{\star}_{\Theta}(t_T) = T^{\star\,-1}_{\Theta}(t_T)$. Following *Conjecture 2* we know that $E^{\star}_{\Theta}(t_T) > E^{\star}(t_T + 1)$, being $E_0 = E^{\star}_{\Theta}(t_T)$ the minimum average energy for delay budget t_T. Similarly, let $t_0 = T^{\star}_{\Theta}(E_0)$, be the minimum

achievable latency given the energy budget E_0. We prove that t_0 equals to t_T by showing that the proposition $t_0 \neq t_T$ leads to contradictions. For this consider consider a solution Θ where $E^{\star}_{\Theta}(t_T) = E_0$, a solution Θ' such that $t_{\Theta'} = T^{\star}_{\Theta'}(E_0) = t_0$ and a solution Θ'' that is optimal energy wise for the delay budget t_0 i.e., $E_{\Theta''} = E^{\star}_{\Theta''}(t_0)$. For the case where $t_0 < t_T$ it follows from *Conjecture 2* that $E^{\star}_{\Theta}(t_T) < E^{\star}_{\Theta''}(t_0) \Leftrightarrow E_0 < E^{\star}_{\Theta''}(t_0)$, however $T^{\star}_{\Theta'}(E_0) = t_0$ states that the solution Θ' complies with a delay budget t_0 by using a lower energy budget, implying that Θ'' is not optimal, which is a contradiction. Likewise, if $t_T < t_0$ then Θ' is not optimal as Θ would comply to a lower delay budget with the same energy constraint, which is also a contradiction. Therefore, t_0 has to be equal to t_T, meaning that $\Theta'' = \Theta' = \Theta$ and $E^{\star}_{\Theta}(t_T) = T^{\star}_{\Theta}{}^{-1}(t_T) \Leftrightarrow E_T = E^{\star}_{\Theta}(T^{\star}_{\Theta}(E_T))$.

APPENDIX D
PROOF OF CONVEXITY

For a two transmission MU-HARQ the average transmission energy of one UE can be defined as

$$E = n_{x[1]} p_{x[1]} + \sum_{x \in \mathcal{S}} n_{x[2]} p_{x[2]} P_{bin}(x^{(2)}, G, \epsilon_{x[1]}) \frac{x^{(2)}}{G} \tag{47}$$

where $p_{x[1]}$ is a function of $\epsilon_{x[1]}$ and $n_{x[1]}$, $p_{x[1]}\left(n_{x[1]}, \epsilon_{x[1]}\right)$, however $p_{x[2]}$ is a function of $\epsilon_{x[2]}, n_{x[1]}$, and $\epsilon_{x[1]}$ since $\mathcal{T}_{\mathcal{S}[2]} = t_T - t_1$. The derivative of (47) relative to $\epsilon_{x[2]}$ $\forall x \in \mathcal{S}$

$$\begin{aligned}\frac{\partial E}{\partial \epsilon_{x[2]}} &= P_{bin}(x^{(2)}, G, \epsilon_{x[1]}) \frac{x^{(2)}}{G} n_{x[2]} \frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}} \\ &= P_{bin}(x^{(2)}, G, \epsilon_{x[1]}) \frac{(t_T - n_{x[1]})G}{x^{(2)}} \frac{x^{(2)}}{G} \frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}}. \end{aligned} \tag{48}$$

Since $\frac{\partial E}{\partial \epsilon_{x[2]}}$ does not depend on any $\epsilon_{x'[2]}$ $\forall x' \in \mathcal{S} \setminus \{x\}$ means that the if $n_{x[1]}$ and $\epsilon_{x[1]}$ are kept constant, the Hessian matrix is diagonal as the cross-derivatives are zero. Since it is not possible to define the function $p_{x[2]}(\epsilon_{x[1]}, n_{x[1]}, \epsilon_{x[2]})$ [12], we follow the same approach in [12] and obtain $\frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}}$ through the implicit formula theorem [27]. From the PPV bound [6], one can define $F_{x[m]}(\epsilon_{x[1]}, \ldots, \epsilon_{x[m]}, n_{x[1]}, \ldots, n_{x[m]})$ [12]

$$\begin{aligned} &F_{x[m]}\left(\Theta_{x[m]}\right) \\ &= \sum_{i=1}^{m} n_{x[i]} \log(1 + p_{x[i]}) \\ &\quad - Q^{-1}\left(\epsilon_{x[m]}\right) \sqrt{\sum_{i=1}^{m} n_{x[i]} p_{x[i]} \frac{2 + p_{x[i]}}{(1 + p_{x[i]})^2}} - B \log(2) \end{aligned} \tag{49}$$

meaning that $F_{x[m]} = 0 \forall x$. Using the implicit function theorem and $F_{x[2]}(\epsilon_{x[1]}, \epsilon_{x[2]}, n_{x[1]}, n_{x[2]})$ one can obtain $\frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}}$ with

$$\frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}} = -\frac{\partial C_{x[2]}}{\partial \epsilon_{x[2]}} \frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}}, \tag{50}$$

where $C_{x[2]} = Q^{-1}\left(\epsilon_{x[2]}\right)$. Knowing that

$$\frac{\partial \epsilon_{x[2]}}{\partial C_{x[2]}} = \lim_{\Delta \to 0} \frac{-c}{\Delta} \int_{C_{x[2]}}^{C_{x[2]}+\Delta} e^{-\frac{t^2}{2}} dt = -\frac{e^{-\frac{C_{x[2]}{}^2}{2}}}{\sqrt{2\pi}} < 0, \tag{51}$$

the first term of (50) can be obtained since $\frac{\partial C_{x[2]}}{\partial \epsilon_{x[2]}} = \frac{1}{\frac{\partial \epsilon_{x[2]}}{\partial C_{x[2]}}} < 0$. From (48) we can see that the cross-derivatives are zero, hence the Hessian H is a diagonal matrix meaning that in order to prove the convexity property it suffices to prove that $H \succeq 0$,

$$\frac{\partial^2 E_n}{\partial \epsilon_{x[2]}{}^2} = \epsilon_{x[1]} P_{bin}(x^{(2)}, G, \epsilon_{x[1]})(t_T - n_1) \frac{\partial^2 p_{x[2]}}{\partial \epsilon_{x[2]}{}^2}. \tag{52}$$

By inspecting (52), one can see that if $\frac{\partial^2 p_{x[2]}}{\partial \epsilon_{x[2]}{}^2} > 0$ then $\frac{\partial^2 E_n}{\partial \epsilon_{x[2]}{}^2} > 0$.

$$\begin{aligned} &\frac{\partial^2 p_{x[2]}}{\partial \epsilon_{x[2]}{}^2} \\ &= \frac{\partial}{\partial \epsilon_{x[2]}} \left(\frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}} \right) \\ &= -\left(\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}} \times \frac{\partial^2 C_{x[2]}}{\partial \epsilon_{x[2]}{}^2} + \left(\frac{\partial C_{x[2]}}{\partial \epsilon_{x[2]}} \right)^2 \times \frac{\partial}{\partial C_{x[2]}} \left(\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}} \right) \right) \end{aligned} \tag{53}$$

We are going to analyze the sign of all the four terms of the expressions (53). Going from left to right we start with $\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}}$. From [11, Lenma 1] we know that $\frac{\partial p_{x[2]}}{\partial \epsilon_{x[2]}} < 0$. Likewise, $\frac{\partial F_{x[2]}}{\partial C_{x[2]}} < 0$ (49) and $\frac{\partial c_{x[2]}}{\partial \epsilon_{x[2]}} < 0$ (51), meaning that $\frac{\partial F_{x[2]}}{\partial p_{x[2]}} > 0$ (50) and $\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}} < 0$. The second expression on (53) is

$$\frac{\partial^2 C_{x[2]}}{\partial \epsilon_{x[2]}{}^2} = -\sqrt{2\pi} C_{x[2]} e^{\frac{C_{x[2]}{}^2}{2}} \frac{\partial C_{x[2]}}{\partial \epsilon_{x[2]}}. \tag{54}$$

From (54) is possible to verify that $\sqrt{2\pi} e^{\frac{C_{x[2]}{}^2}{2}} \frac{\partial C_{x[2]}}{\partial \epsilon_{x[2]}} < 0$. Since $C_{x[2]} \in [-\infty, +\infty]$, means that $\frac{\partial^2 C_{x[2]}}{\partial \epsilon_{x[2]}{}^2} \geq 0 \iff C_{x[2]} \geq 0 \iff \epsilon_{x[2]} \leq 0.5$. The third therm of (53) is a square of a real number, hence is always positive. The fourth element $\frac{\partial}{\partial C_{x[2]}} \left(\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}} \right)$ can be expanded as,

$$\frac{\partial}{\partial C_{x[2]}} \left(\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}} \right)$$

$$= \frac{\frac{\partial^2 F_{x[2]}}{\partial {C_{x[2]}}^2} \times \frac{\partial F_{x[2]}}{\partial p_{x[2]}} - \frac{\partial F_{x[2]}}{\partial C_{x[2]}} \times \frac{\partial}{\partial C_{x[2]}} \left(\frac{\partial F_{x[2]}}{\partial p_{x[2]}}\right)}{\left(\frac{\partial F_{x[2]}}{\partial p_{x[2]}}\right)^2} \tag{55}$$

we verify that $\frac{\partial^2 F_{x[2]}}{\partial {C_{x[2]}}^2} = 0$, hence (55) can be further simplified to

$$\frac{\partial}{\partial C_{x[2]}} \left(\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}}\right) = -\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}} \times \frac{\partial}{\partial C_{x[2]}} \left(\frac{\partial F_{x[2]}}{\partial p_{x[2]}}\right)}{\left(\frac{\partial F_{x[2]}}{\partial p_{x[2]}}\right)^2}, \tag{56}$$

where $\frac{\partial F_{x[2]}}{\partial C_{x[2]}} < 0$ and $\frac{\partial}{\partial C_{x[2]}} \left(\frac{\partial F_{x[2]}}{\partial p_{x[2]}}\right) < 0$

$$\frac{\partial}{\partial C_{x[2]}} \left(\frac{\partial F_{x[2]}}{\partial p_{x[2]}}\right) = -\frac{n_{x[2]}}{(1+p_{x[2]})^3 \sqrt{n_{x[1]} p_{x[1]} \frac{2+p_{x[1]}}{(1+p_{x[1]})^2} + n_{x[2]} p_{x[2]} \frac{2+p_{x[2]}}{(1+p_{x[2]})^2}}}, \tag{57}$$

meaning that

$$\frac{\partial}{\partial C_{x[2]}} \left(\frac{\frac{\partial F_{x[2]}}{\partial C_{x[2]}}}{\frac{\partial F_{x[2]}}{\partial p_{x[2]}}}\right) < 0. \tag{58}$$

Knowing the signs of all the terms in (53) allow us to determine that

$$\frac{\partial^2 p_{x[2]}}{\partial {\epsilon_{x[2]}}^2} > 0 \tag{59}$$

and that

$$\frac{\partial^2 E}{\partial {\epsilon_{x[2]}}^2} > 0 \quad \forall x \in \mathcal{S}. \tag{60}$$

Therefore, the Hessian is a positive definite matrix as it is a diagonal matrix positive diagonal elements, proving that the optimization problem is indeed convex as long as $\epsilon_{x[2]} < 0.5 \forall x \in \mathcal{S}$.

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] *Scenarios, Requirements and KPIs for 5G Mobile and Wireless System, Deliverable D1.1*, document ICT-317669, METIS Project, New York, NY, USA, 2013.

[3] 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.

[4] 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.

[5] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8–15, Mar. 2018.

[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] S. Hwan Kim, D. Keun Sung, and T. Le-Ngoc, "Performance analysis of incremental redundancy type hybrid ARQ for finite-length packets in AWGN channel," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA USA, Dec. 2013, pp. 2063–2068.

[8] 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.

[9] 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.

[10] 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)*, Shanghai, China, May 2019, pp. 1–6.

[11] 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.

[12] 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.

[13] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Proc. Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Sep. 2007, pp. 2861–2864.

[14] 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.

[15] 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. (WCNC)*, Mar. 2017, pp. 1–5.

[16] 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. (VTC-Spring)*, Jun. 2022, pp. 1–5.

[17] 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.

[18] 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 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.

[19] 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. (VTC Spring)*, Jun. 2018, pp. 1–5.

[20] 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 (ICC Workshops)*, Jun. 2021, pp. 1–6.

[21] 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.

[22] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[23] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

[24] A. K. Bairagi, Md. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.

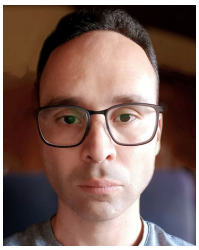
[25] T.-K. Le, U. Salim, and F. Kaltenberger, "Enhancing URLLC uplink configured-grant transmissions," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–5.

[26] M. S. Bazaraa, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[27] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem*. New York, NY, USA: Birkhäuser, 2002.

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 University of Aveiro, under the MAP-Tele Doctoral Program.

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 Assistant Professor with the Department of Electronics, Telecommunication and Informatics, University of Aveiro, and a Researcher at the Instituto de Telecomunicações, Pólo de Aveiro. 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 FP7 ICT. His research interests include signal processing techniques for digital communications and radar, with emphasis on 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 with the national level. He has published more than 150 technical papers in international journals and conference proceedings. His research interests include multicarrier-based systems, cooperative networks, precoding, multiuser detection, and massive MIMO. He served as a member of TPC for several international conferences. He is an Associate Editor of IEEE Access and *IET Signal Processing*.

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 Departamento de Electrónica, Telecomunicações e Informática, University of Aveiro, and a Researcher with the 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 BT Laboratories and one year with NKT Elektronik. He has been involved and has led IT or the University of Aveiro, participation in more than 20 national and European projects. He has published more than 200 technical papers in international journals and conferences. His main research interests lie in signal processing techniques for digital communications and communication protocols, and within this research line, he has done work for optical and mobile communications, either at the theoretical or experimental level. His current research interests include space-time-frequency algorithms for broadband wireless systems and cross-layer design.