Allocation of Resources for HARQ Retransmission in Mobile Networks based on C-RAN

Mohammed Elfiky, *Graduate Student Member, IEEE*, Zdenek Becvar, *Senior Member, IEEE* and Pavel Mach, *Member, IEEE*

***Abstract*—The radio resources are utilized for both transmissions of new data and for retransmission(s) of erroneous data to handle errors due to transmission over the wireless link. Hence, the more retransmissions occur, the fewer resources remain available for transmission of new data resulting in a lower goodput. This problem is even emphasized in Cloud Radio Access Networks (C-RAN), where Remote Radio Heads (RRHs) cooperate with Base Band Unit (BBU) over fronthaul imposing additional delay and limiting goodput. In this paper, we propose a flexible and dynamic Hybrid Automatic Repeat reQuest (HARQ) resource pre-allocation based on the actual retransmission needs of individual User Equipment (UEs). This is managed by two algorithms, first exploiting information on the error rate experienced by individual UEs and the second using the length of the scheduling period of the UEs. Both algorithms are further integrated together to improve the efficiency of the HARQ retransmissions resource pre-allocation. Simulation results show the proposed solution provides goodput close to the theoretical upper-bound while outperforming existing approaches by up to 39%. Furthermore, the transport block loss rate and mean absolute percentage error of the amount of pre-allocated resources for the HARQ are notably decreased by around 38% and 57%, respectively.**

***Index Terms*—5G; C-RAN; HARQ; hierarchical scheduler; ARIMA; RRH; BBU.**

I. Introduction

The fifth generation (5G) of mobile networks is expected to support various traffic patterns and unlock numerous applications for low latency and reliable communication [1]. However, meeting such stringent requirements is challenging, as it requires an efficient radio resource management. The radio resource management encompasses various functionalities that differ in complexity and operating timescale. Many of these functionalities, such as resource allocation, can be centralized [2] [3] [4]. To this end, the concept of a Cloud Radio Access Network (C-RAN) has been introduced to ensure energy and cost-efficient solution [4] [5].

The C-RAN comprises a Baseband Unit (BBU) and Radio Remote Heads (RRHs) connected to the BBU via a fronthaul. However, the fronthaul introduces challenges related to high goodput and low latency requirements to ensure a swift exchange of the baseband signals over the fronthaul links between the BBU and the RRHs [5]. The challenges imposed by the fronthaul are emphasized even more if the fronthaul is facilitated via a wireless link [6]. The high latency at the fronthaul would negatively impact the data transmissions and time-critical radio resource management protocols, such as error correction via Hybrid Automatic Repeat reQuest (HARQ). The HARQ is responsible for retransmissions and corrections of data that were not received correctly. Nevertheless, such a process introduces an additional delay to the data transmission, including processing delay, propagation delay, and retransmission delay. The whole retransmission process should be accomplished in up to 8 time transmissions intervals (TTIs) [7]. Thus, the TTI duration, which is ranging from 62.5µs to 1 ms [8] [9] in 5G, imposes challenges on the HARQ process in C-RAN [9] [10].

There are several works addressing the problem of the HARQ targeting the aspect of low latency for ultra-reliable and low latency communications (URLLC) in 5G, see, e.g., [11] [12] [13] [14]. In [11], the authors present a semi-persistent scheduling of resources for the UEs' retransmissions. To this end, for any potential retransmissions, a pre-defined amount of resources is shared by a pre-defined group of UEs based on the Block Error Rate (BLER) of the UEs' first transmission. In [12] and its extension in [13], a periodic radio resource allocation is proposed for retransmissions of individual UEs to meet latency and reliability requirements. The solution is based on selecting an optimal modulation and coding scheme (MCS) and subsequent allocation of the required resources. The paper [14] exploits the queuing model to optimize the HARQ resource requirement in URLLC. However, none of the works presented in [11] [12] [13] [14] assume the C-RAN architecture with the realistic fronthaul with non-zero delay for the HARQ and resource allocation.

The resource allocation for C-RAN considering strict HARQ requirements is assumed in [15]. The HARQ itself is, however, not optimized in any way. The optimization of HARQ tailored for C-RAN is assumed in [16], [17], [18], [19]. The authors in [16] propose a centralized low-complexity packet scheduling scheme to reduce communication delay. Nevertheless, the inter-cell interference (ICI) among the deployed UEs is neglected and this work is limited only to URLLC traffic. In [17], the authors consider sharing computing resources among multiple RRHs for the uplink in the C-RAN architecture to improve the HARQ retransmission process. However, the work considers only a single-user scenario, and extension toward a practical multi-user scenario is not straightforward. In [18] and [19], the authors focus on a proactive HARQ, which transmits proactively redundancy versions until the receiver indicates

The authors are with the Department of Telecommunication Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, 166 27 Prague, Czech Republic (e-mail: elfikmoh@fel.cvut.cz; zdenek.becvar@fel.cvut.cz; machp2@fel.cvut.cz). This work has been supported by the grant of Czech Technical University in Prague No. SGS20/169/OHK3/3T/13.

correct reception with ACK. This leads to reduced latency of HARQ, but, at the same time, it also lowers spectral efficiency notably. To this end, the authors in [18] propose a feedback prediction scheme for C-RAN to reduce the redundancy in proactive HARQ. The paper is further extended in [19], where machine-learning-assisted HARQ prediction schemes for C-RAN is proposed in order to decrease the maximum transmission latency. Still, neither [18] nor [19] deals with the pre-allocation of resources for HARQ in C-RAN and, thus, our work can be seen as complementary to these.

To cope with the fronthaul delay between the BBU and RRHs in the C-RAN, we have introduced a concept of a hierarchical scheduler in [20]. The key aspect of the hierarchical scheduler is to allow shifting of scheduling-related functions between the BBU and the RRHs. The scheduling of resources is handled so that the cell edge UEs (CE UEs) are scheduled centrally by the BBU to allow mitigation of ICI, while non-CE UEs (nCE UEs) are scheduled in a distributed way by the RRHs to avoid a negative impact of the fronthaul.

One of the key challenges in the hierarchical scheduler in the C-RAN is to determine the amount of resources to be pre-allocated for the HARQ process. Note that since we allocate the resources for the HARQ in advance for multiple upcoming TTIs (i.e., in relatively long-term compared to traditional scheduling), we use the term "pre-allocation" in the paper. In [20] [21], the number of pre-allocated RBs for the HARQ process is set to a fixed amount regardless of the actually required retransmissions. Unfortunately, this fixed pre-allocation degrades the performance of the scheduler. Therefore, in this paper, we present a dynamic resource pre-allocation scheme to cope with HARQ retransmissions. The paper's objective, in other words, is to determine the required amount of resources to be pre-allocated for HARQ retransmissions depending on the individual UEs' actual needs. Our contributions in this paper are summarized as follows:

- We propose a comprehensive framework for the HARQ resource pre-allocation in the C-RAN, considering the hierarchical scheduling to maximize the goodput of UEs via minimizing the transport block loss rate and maximizing the resource pre-allocation accuracy.
- We develop two distinct approaches to determine the required amount of pre-allocated resources for the HARQ and optimize them jointly to improve the resource pre-allocation accuracy.
- We also show that the idea of resource pre-allocation is not limited to the hierarchical scheduling only, but it is extended to be applicable also to other existing type of schedulers, such as centralized and partially distributed schedulers.
- Via simulations, we demonstrate that the proposed solution enhances the goodput by up to 39%, where the transport block loss rate and mean absolute percentage error of the amount of pre-allocated resources for the HARQ are decreased by up to 38% and 57%, respectively, in comparison to state-of-the-art-works.

The rest of the paper is organized as follows. The system model is described in Section II. In Section III, we formulate the HARQ resource pre-allocation problem. Section IV describes the proposed solution for estimating the proper amount of pre-allocated resources for the HARQ retransmissions. The simulation setup and results are presented in Section V. Major conclusions, and possible future research directions are outlined in Section VI.

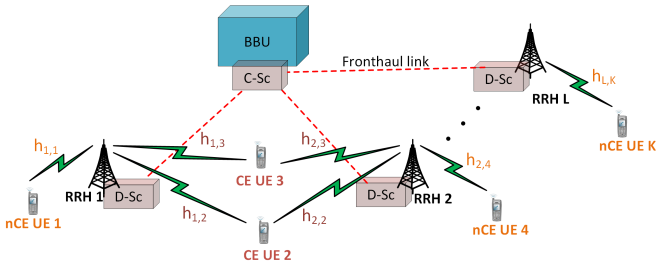


Fig. 1: High-level overview of the hierarchical scheduler

II. System Model

This section describes the system model based on the C-RAN architecture, the background on the hierarchical scheduler, and the HARQ process. Each part is described in the following subsections.

A. C-RAN based architecture

We assume a single BBU interconnected with L RRHs via the fronthaul, as shown in Fig.1. In our study, we consider the fronthaul from the perspective of fronthaul latency for data retransmissions and we do not expect any errors originating at the fronthaul. Furthermore, K UEs are deployed randomly over an area covered by the RRHs. The UEs are individually associated with the RRH providing the highest Signal to Interference and Noise Ratio (SINR). We further classify the UEs into K_{CE} CE UEs and K_{nCE} nCE UEs so that $K = K_{CE} + K_{nCE}$. This classification is based on the experienced SINR via individual UEs. Since, intuitively, the CE UEs experience more substantial inter-cell interference from adjacent RRHs than the nCE UEs, an Inter-Cell Interference Coordination (ICIC) technique [22] is adopted to mitigate such interference.

In the ICIC, a set of RRHs (i.e., $ICIC$ set, L_k^{ICIC}) cooperates together to mitigate the ICI of individual CE UEs. More specifically, each CE UE is served via orthogonal resource blocks (RBs) with respect to the transmission of other RRHs in the same ICIC set. The ICIC set encompasses L_k^{ICIC} RRHs that are involved in the cooperation to improve the k-th CE UE's SINR. The new RRH is added to the ICIC set for the k-th CE UE (i.e., to the L_k^{ICIC}) if and only if the CE UE communication goodput is improved by adding such RRH(s). Hence, the RRH is included in the UE's ICIC set if the RRH satisfies the following condition:

$$\Gamma_k < \alpha_C \frac{n_k^S}{n_k^{ICIC}} \tag{1}$$

where Γ_k defines the number of RRHs in L_k^{ICIC} (i.e, $\Gamma_k > 1$), n_k^S represents the number of RBs required for the transmission of data to the k-th CE UE without ICIC (i.e., single RRH), n_k^{ICIC} corresponds to the number of RBs allocated for the

data transmission to the k-th CE UE from each RRH in the ICIC set, and α_C is the ICIC threshold (i.e., α_C>1). The parameter α_C indicates the transmission efficiency with the ICIC utilization so that the RRH is added to the ICIC set if the ratio of the number of required RBs without the ICIC to the amount of the required RBs with the ICIC is α_C times higher than the number of RRHs in the ICIC set. This way, we can guarantee that the ICIC is exploited if and only if the ICIC increases the communication goodput of the CE UE.

Considering the ICIC is used, the CE UE's SINR between the l-th RRH and the k-th CE UE ($\gamma_{k,l}^{CE}$) is calculated as:

$$\gamma_{l,k}^{CE} = \frac{p_l.h_{l,k}}{\sigma_n + \sum_{i \notin L_k^{ICIC}} p_i g_{i,k}} \quad (2)$$

where p_l is the transmission power of l-th RRH, $h_{l,k}$ is the channel gain between the l-th RRH and the k-th CE UE, σ_n is the noise power, and the term $\sum_{i \notin L_k^{ICIC}} p_i.h_{i,k}$ represents the inter-cell interference from all the RRHs except those included in L_k^{ICIC}. Likewise, since the UEs' classification depends on individual UEs' SINR, the nCE UE's SINR between the l-th RRH and the k-th nCE UE is calculated as:

$$\gamma_{l,k}^{nCE} = \frac{p_l.h_{l,k}}{\sigma_n + \sum_{i=1, i \neq l}^{i=L} p_i h_{i,k}} \quad (3)$$

where the term $\sum_{i=1, i \neq l}^{i=L} p_i h_{i,k}$ represents the inter-cell interference from all but the serving RRHs.

B. The hierarchical scheduler

The hierarchical scheduler, the basis of this work, splits the scheduling process into two tiers: a centralized scheduler (C-Sc) and a distributed scheduler (D-Sc), as described in [20]. The C-Sc runs in the BBU and D-Sc in the RRH (see Fig.1). The D-Sc handles data transmission and manages the allocation of resources for the nCE UEs, as these UEs suffer less from inter-cell interference. However, the C-Sc schedules data transmission for the CE UEs, enabling a high level and long-term scheduling, reinforced by an awareness of the mutual interference among individual RRHs. The long-term scheduling is understood as a scheduling decision not only for a single TTI but for N consecutive TTI (i.e., $NxTTI$). The N is set individually for each k-th CE UE (i.e., N_k) as considered in our previous work [21] to maximize the sum goodput of the UEs. The value of the scheduling period, N_k, is set based on: *i*) the individual predicted future CSI of the individual CE UEs' radio channel dynamicity and *ii*) the fronthaul delay. By one of the channel prediction tools, we predict the future UE CSI based on the UE CSI history record. The UE CSI history record is understood as CSI values in the time interval just before the time of the prediction and is presented as the input of the channel prediction tool, as explained in [21].

C. HARQ process

The idea behind the HARQ is to model a system that detects the received erroneous data transport block and then requests the needed retransmissions in case of the erroneous data transport block. The retransmissions can be classified as adaptive and non-adaptive, where we have two ways to implement HARQ for downlink; synchronous HARQ and asynchronous HARQ [23]. In the adaptive HARQ, the MCS and other transmission attributes (such as the redundancy version and sub-carrier) have the option to be updated for each retransmission, where the transmission attributes are fixed or pre-defined in the non-adaptive context.

In practice, the HARQ course of action for the CE UEs and nCE UEs in the hierarchical scheduler are quite different. For the nCE UEs, the HARQ process is handled in a standard way at the RRH, as the nCE UEs are scheduled directly by the D-Sc in the RRH, and the fronthaul does not have any direct negative impact on the HARQ process. The standard way is understood so that the resources for the HARQ retransmissions are scheduled directly by the D-Sc at the RRH based on ACKs/NACKs received from individual nCE UEs served by the given RRH. In the case of the NACK, the D-Sc allocates any available RB(s) that are not dedicated to CE UEs at the moment of the retransmission (see [21] for more details). The HARQ process is being more complicated for the CE UEs due to the fronthaul delay intervention. The HARQ process would be significantly prolonged due to the transmissions taking place over the fronthaul in case the HARQ would be processed in the BBU.

To understand our proposal for pre-allocating resources for CE UEs' HARQ, described in the following sections, let us first define $R_{k,n}^*$ and $R_{k,n}$ as the estimated and actual numbers of the required RBs for the k-th CE UE's potential retransmissions in the n-th TTI within N_k, respectively. Based on that, we introduce a new performance evaluation parameter; the resource pre-allocation efficiency of individual CE UEs, ζ_k. This parameter is defined as the mean absolute percentage error (MAPE) of the amount of pre-allocated resources for the HARQ to estimate how far the $R_{k,n}^*$ value from $R_{k,n}$ value and is expressed as:

$$\zeta_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{|R_{k,n}^* - R_{k,n}|}{R_{k,n}} * 100 \quad (4)$$

Then, considering also retransmissions due to HARQ, we can define the goodput experienced by the k-th UE as:

$$G_k = N_{Sym} N_{SC} \left(\sum_{l=1}^{RB_k} CR_{l,k} log_2 M_{l,k} - \right. \\ \left. - \sum_{r=1}^{R_k} CR_{r,k} log_2 M_{r,k} \right) (1 - OH_k) \quad (5)$$

where N_{Sym} represents the number of OFDM symbols per one RB, N_{SC} stands for the number of subcarriers per RB, $CR_{l,k}$ is the coding rate applied at the l-th RB allocated to the k-th UE, $M_{l,k}$ corresponds to the number of possible modulation states based on the modulation used for data transmission at the l-th RB allocated for the k-th UE, RB_k is the number of all RBs allocated to the k-th UE per second, R_k represents the number of RBs allocated only for retransmissions of the k-th UE per one second, and finally OH_k stands for the overhead due to various signaling and control messages to serve the k-the UE (expressed as a ratio between the amount of resources allocated for signaling to the resources allocated for data transmissions).

III. Problem formulation

This paper aims to maximize the sum goodput of the CE UEs in the hierarchical scheduler architecture based on the C-RAN. This objective can be attained by optimizing the resource scheduling efficiency for the CE UEs' potential retransmissions. The main challenge in such optimization is estimating the required scheduling resources for any potential CE UEs' data retransmissions. One way to address this challenge is to perform a dynamic adjustment of the pre-allocated amount of RBs instead of pre-allocating a fixed amount of resources for the CE UEs' HARQ requirements, which is suggested in our earlier works [20] and [21].

Basically, the individual CE UEs' HARQ resource requirements depend on many factors, including *i*) the CE UE radio channel condition, *ii*) the fronthaul delay, and *iii*) the scheduling period, N_k. The motivation behind presenting this work is to fulfill the varied resources needed to be pre-allocated for individual CE UEs' HARQ. Thus, the problem is formally formulated as:

$$\begin{aligned} \max_{R_k^*} \quad & \sum_{k=1}^{K_{CE}} G_k \\ \text{s.t.} \quad & \text{a) } 0 \leq R_k^* \leq R_{max} \\ & \text{b) } 0 \leq N_k \leq N_{max} \end{aligned} \tag{6}$$

where R_k^* is the amount of RBs pre-allocated for the HARQ of the k-th CE UE, R_{max} represents the maximum affordable RBs depending on the system bandwidth and the number of served UEs, and N_{max} stands for the maximum length of centralized period set up dynamically by the network operator based on the overall average CE UEs' radio channel dynamicity. The constraint a) in (6) limits the possible values of R_k^* to be pre-allocated to each k-th CE UE while constraint b) gives the lower and upper limits on the scheduling period N_k. Note that the calculation of the R_k^* is made independently for each CE UE to maximize the individual CE UEs' goodput and, hence, also to maximize the sum goodput of all UEs. The dynamic setting of the amount of the HARQ pre-allocated RBs also minimizes the unexploited pre-allocated resource in case of free-error delivery data.

The formulated problem can be classified as a nonlinear integer programming problem. The reason is that the dependence of goodput on the amount of pre-allocated resources is non-linear with respect to the channel quality experienced by individual UEs. Moreover, both the objective function and the constraints are integer (discreet) variables as: *i*) the goodput in objective function strictly depends on selected coding rate and modulation and is limited to several discreet values (see (5)), and *ii*) constraints on the scheduling period N and the number of pre-allocated resource blocks for HARQ are also integer variables.

In general, the resource allocation formulated as nonlinear integer problem is usually solved by various deterministic algorithms [24] [25] [26] [27] [28] or evolutionary algorithms [9] [29] [30] [31]. The main limitation of the deterministic algorithms is a huge complexity, since finding the optimal solution in a large search space is infeasible while limiting the search to only a subset of the search space results in a poor and far from optimal solution [32]. Moreover, the uncertainty in the channel quality in the future (several TTIs for which the pre-allocation of resources for HARQ is done) adds another dimension to the complexity. Hence, utilizing such deterministic algorithms would make our proposal computationally complex, time-consuming, and impractical, especially for large-scale problems with multiple UEs. Therefore, the deterministic algorithms would not be a good fit for our problem, which demands swift and instantaneous pre-allocation scheduling decisions for the available resources in a horizon of milliseconds. Along similar lines, the evolutionary algorithms are not suitable for our problem as these are known to suffer from slow convergence [33].

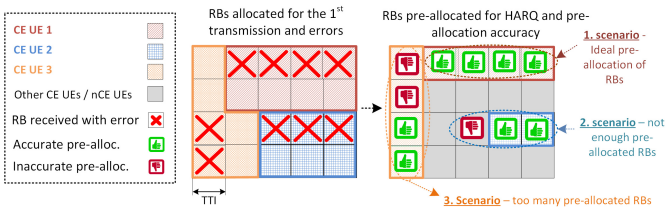


Fig. 2: High-level overview of the HARQ resource pre-allocation when the number of pre-allocated RBs is equal to, lower than, and larger than the actual number of RBs required by the HARQ process.

In contrast to these traditional tools, the heuristic algorithms can be designed to be fast, because they do not require a complete search in the search space [32]. Thus, the heuristic algorithms are practical, serving as fast and feasible solutions for planning and scheduling problems (i.e., see [32]) as targeted in our work (i.e., finding the R_k^*). Therefore, we adopt the heuristic approach to solve the defined optimization problem. The proposal is described in detail in the following section.

IV. Dynamic resource pre-allocation for the HARQ

This section describes the proposed approach for the HARQ resource pre-allocation in the C-RAN based on the hierarchical scheduler. We tackle in this work the resource pre-allocation problem specifically for the CE UEs, since the retransmissions for the nCE UEs are handled directly by the RRHs, as explained in [20] [21]. Furthermore, the resource scheduling decision for the nCE UEs is not negatively impacted by the fronthaul delay. We first outline a high-level principle of the resource pre-allocation for the CE UEs' HARQ retransmissions. Then, we describe two proposed approaches for determining the number of pre-allocated RBs for the CE UEs' HARQ retransmissions.

A. High-level principle

Let us first illustrate the possible scenarios that can occur during the pre-allocation of resources for the CE UEs' HARQ and discuss the motivation behind the proposed work. To cope with the fronthaul delay affecting the HARQ process of the CE UEs, a part of RBs is pre-allocated in the BBU for any retransmission needs of all CE UEs transmitting data at any given TTI. Based on the amount of pre-allocated RBs

in comparison with the actual required RBs for the HARQ retransmissions, we can distinguish three scenarios (see Fig.2):

- *1. scenario*: The pre-allocated amount of RBs is precisely equal to the amount of actually required RBs. Thus, there are no further actions to be taken since the number of pre-allocated RBs exactly matches the required RBs.
- *2. scenario*: The C-Sc at the BBU pre-allocates an insufficient number of RBs for retransmitting all erroneous data transport blocks. Consequently, some of the retransmitted data is delivered with an additional delay due to the fronthaul, as the HARQ process, in this scenario, is performed in the C-Sc at the BBU instead of the D-Sc in the RRHs. This additional delay postpones the retransmission process to later TTIs and increases the overall delay. The situation is getting even more critical for delay-sensitive services, such as URLLC in 5G mobile networks [34]. For such services, the retransmitted data transport blocks might even get rejected once the data retransmission deadline is expired due to the introduced additional delay in the HARQ process.
- *3. scenario*: The C-Sc pre-allocates too many RBs for CE UEs' HARQ retransmissions. This scenario alleviates the bottleneck of additional HARQ delay and reduces the probability that some retransmitted data transport blocks are not delivered in time. However, relatively, a lower number of RBs remains available for the new data transmission of UEs (both CE UEs and nCE UEs) due to the overbooking of scheduling resources for the CE UEs' HARQ retransmissions.

In order to rectify the problem of the HARQ resource pre-allocation and estimate a proper number of RBs in the C-RAN with the hierarchical scheduler, we propose a flexible and dynamic resource pre-allocation approach based on the estimated retransmission requirements of individual CE UEs. The following subsection illustrates the proposed framework and details the resource pre-allocation principle.

B. The proposed pre-allocation of resources for HARQ retransmissions

This section describes the proposed solution for estimating the number of pre-allocated RBs for any possible retransmissions of the k-th CE UE, R_k^*. The value of R_k^* is set independently for each CE UE based on the individual CE UEs' retransmissions needs. The number of retransmissions is set up to a pre-defined maximum limit, δ_{max}. Note that the value of R_k^* is not set only for a single TTI but for N_k consecutive TTIs (i.e., for the whole scheduling period of the k-th CE UE).

In order to estimate the R_k^*, part of our proposal is to predict the evolution of individual CE UEs' CSI. This prediction is exploited via the Auto-regressive Integrated Moving Average model (ARIMA) [35]. Compared with other statistical models, such as the exponential smoothing model and the moving average algorithm, the ARIMA makes the prediction process more reliable and flexible [36]. Fundamentally, the ARIMA model is defined by a combination of coefficients p, d, and q representing the order of the autoregressive model, the degree of differencing, and the order of the moving-average model, respectively. This combination of ARIMA coefficients is adjusted individually for each CE UE based on extensive experiments on CE UE's CSI (see [21] for more details). To assess the combinations of coefficients for individual CE UEs (i.e., p_k, d_k, and q_k), we exploit the Bayesian information criterion (BIC) [37]. The combination of coefficients achieving the lowest BIC is selected for the CE UE's CSI prediction process [37]. Note that the selected BIC (lowest BIC) contains the maximum likelihood estimation, which penalizes free parameters to combat overfitting. After individual CE UEs' CSI is predicted, the block error rate (BLER) and then the required resources for any retransmission at any given TTI (i.e., R_k^*) can be estimated.

The value of the R_k^* depends on three distinct parameters: *i*) the length of the scheduling period of the k-th CE UE, N_k, *ii*) the number of required retransmissions for the k-th CE UEs, δ_k, where $0 \leq \delta_k \leq \delta_{max}$, and *iii*) the number of pre-allocated RBs for each retransmission (i.e., τ retransmission) at the n-th TTI, $R_{k,n}^{\tau}$, where $R_{k,n}^{*} = \sum_{\tau=1}^{\delta_k} R_{k,n}^{\tau} \ \forall \ \tau \in (1, 2, \ldots, \delta_k)$. Let us investigate these three parameters in more detail. First is the scheduling period's length N_k, which is determined according to the individual CE UEs' channel quality information, as introduced in our prior work [21]. Second, the number of required retransmissions, δ_k, which depends on the loss rate of the data transport blocks. The loss rate is calculated based on the CE UE's MCS, which is set to keep the BLER below a certain threshold. Third and lastly, the number of pre-allocated RBs for each retransmission, $R_{k,n}^{\tau}$. The $R_{k,n}^{\tau}$ can be variant for each retransmission since the adaptive HARQ retransmission is assumed in our proposal, as explained in the system model.

Based on the parameters mentioned above for estimating the R_k^* (i.e., N_k, δ_k, and $R_{k,n}^{\tau}$), we can categorize our proposed solution into two distinct aspects: 1) the CE UE's error rate, and 2) the CE UE's scheduling period. Both aspects are described vividly in the following sub-sections. Then, both aspects are combined to make the HARQ resource pre-allocation estimation more precise.

1) The error rate aspect:

In this subsection, we determine the amount of the pre-allocated resource for the HARQ based on the individual CE UEs' error rate. The individual CE UEs' error rate is one of the primary data transmission metrics influencing the average number of retransmissions for a successful data transport block delivery. Each transmission/retransmission is defined by a delivery state S so that $S = 1$ for the data packet received correctly and $S = 0$ for delivery with error(s). Hence, the resulting amount of pre-allocated resources in this aspect, R_k^{er}, can be derived via the Bernoulli random variable and the Poisson Binomial Distribution [38]. Since the received data transport block of the k-th CE UE has a delivery state at every TTI along the scheduling period N_k, it can be written as a vector V_k in such that: $V_k = \{S_1, ..., S_{N_k}\}$. We define all possible combinations of delivery states (i.e., vectors) of the k-th CE UE with the scheduling period N_k as a sample space

(SS_{N_k}) in such that:

$$SS_{N_k} = \{V_k^1, ..., V_k^e, ..., V_k^{Q_k}\} = \begin{pmatrix} S_1^1 & S_2^1 & \cdots & S_{N_k}^1 \\ S_1^2 & S_2^2 & \cdots & S_{N_k}^2 \\ \vdots & \vdots & \ddots & \vdots \\ S_1^{Q_k} & S_2^{Q_k} & \cdots & S_{N_k}^{Q_k} \end{pmatrix} \quad (7)$$

where $e \in \{1, 2, ..., Q_k\}$ is defined as a single vector outcome out of $Q_k = 2^{N_k}$ possible vectors for N_k TTIs. Note that the selected vector of the delivery states of the k-th CE UE, V_k^s, over other outcome vectors in the SS_{N_k} depends on the predicted BLER of individual CE UEs at every TTI along the scheduling period, N_k. Hence, the amount of HARQ pre-allocated resources for the selected V_k^s is estimated as:

$$R[V_k^s] = \sum\nolimits_{n=1}^{N_k} V_{k,n}^s R_{k,n}^\tau \quad (8)$$

where $R[V_k^s]$ represents the number of pre-allocated RBs for HARQ along N_k TTI in case part, or all of data transport blocks of the k-th CE UE are received with an error. Note that the $R_{k,n}^\tau$ is estimated based on the predicted future evolution of the individual CE UEs' CSI at each TTI along the scheduling period, N_k.

Let us discuss the way to estimate R_k^{er}. Since the probability of error of individual CE UEs' is randomly distributed along the scheduling period N_k, we exploit the Poisson Binomial Distribution (PBD) to estimate the error probability distribution of the vector V_k^s for the k-th CE UE. Fundamentally in the PBD, two subsets of vectors are defined from the SS_{N_k}; a_k and a_k^c. The subset a_k is understood as a collection of vectors that occur for the k-th CE UE with the scheduling period equal to N_k. The subset a_k^c (i.e., $a_k^c = \{SS_{N_k} - a_k\}$) is complementary to a_k and includes vectors that are not occurring for the k-th CE UE with the same scheduling period, N_k. The occurred and not occurred vectors in the respective subsets a_k and a_k^c depend on the error rate for each vector in the SS_{N_k}, $\eta_{k,e}$, and a pre-defined error rate threshold, ψ. The value of the $\eta_{k,e}$ is expressed as $\eta_{k,e} = \prod_{n=1}^{N_k} \varphi_{k,e,n}$, where $\varphi_{k,e,n}$ is the predicted BLER for the vector e at every TTI along the scheduling period N_k of the k-th CE UE. Notice that each vector in the SS_{N_k} is indicated by an index: e. Therefore, we can define subset A_k and subset A_k^c as they correspond to the vectors' indices in the subset a_k and the subset a_k^c, respectively. Based on that, the classification of the vectors' indices either belonging to subset A_k or subset A_k^c subset as:

$$\eta_{k,e} = \begin{cases} \eta_{k,i} & \text{if } \eta_{k,e} \geq \psi_{N_k} \quad \forall\ i \in A_k \\ \eta_{k,j} & \text{if } \eta_{k,e} < \psi_{N_k} \quad \forall\ j \in A_k^c \end{cases} \quad (9)$$

where the i, and j refer to the individual vectors' indices within the subset A_k and subset A_k^c, respectively. The value of the ψ_{N_k} for the scheduling period N_k is determined based on the average of the experienced vectors' error rate over a long period of time so that:

$$\psi_{N_k} = \overline{\eta_{k,e}}(T_s | N_k) \quad (10)$$

where $T_s | N_k$ is the communication session period when the scheduling period length is equal to N_k. The communication session period, i.e., T_s, is defined as the period of time (in seconds) for a series of interactions between two communication endpoints (i.e., UE and RRH/BBU) that occur during the span of a single connection. Note that the ψ_{N_k} is calculated independently for each scheduling period, and then it varies depending on the N_k. Hence, the error probability distribution of the vector V_k^s for the k-th CE UE is written as follows:

$$P_k[V == V_k^s] = \sum\nolimits_{A_k} \prod\nolimits_{i \in A_k} \eta_{k,i} \prod\nolimits_{j \in A_k^c} (1 - \eta_{k,j}) \quad (11)$$

Then, the number of pre-allocated scheduling resources for the HARQ over N_k TTI is calculated as:

$$R_k^{er} = P_k[V_k^s]\ R[V_k^s] \quad (12)$$

The proposed pre-allocation based on the error rate aspect is summarized in Algorithm 1. Note that the algorithm is illustrated for any k-th CE UE. The algorithm starts with definition of SS_{N_k} matrix in line with (7) giving all possible combinations of delivery states (i.e., vectors) of the k-th CE UE with the scheduling period N_k (see line 1 in Algorithm 1). Based on SS_{N_k} matrix, the amount of HARQ pre-allocated resources for the selected vector V_k^s (i.e., $R[V_k^s]$) is calculated according to (8) (line 2). Then, $\eta_{k,e}$ (i.e., error rate for V_k^e) and ψ_{N_k} (i.e., pre-defined error rate threshold of the scheduling period N_k) is estimated according to (9) and (10), respectively (line 3). In the next step, the error probability distribution for any e-th combination out of Q_k is estimated via the Poisson Binomial Distribution (lines 4-9). Finally, the number of pre-allocated scheduling resources for the HARQ, R_k^{er}, is calculated via (12) (line 12).

2) The scheduling period aspect:

Now, let us turn our attention to the scheduling period aspect for estimating the required pre-allocated resources for HARQ retransmission(s), R_k^{sp}. The importance of this aspect comes from the fact that the CE UE's scheduling period reflects two factors in its estimation: 1) the CE UE's radio channel dynamicity and 2) the fronthaul delay (see [21] for more details). Both factors are essential in the way for achieving network reliability and fulfilling retransmission requirements.

To estimate the number of pre-allocated resources for the HARQ retransmission of CE UEs, we reformulate the sample space in (7) into the number of error states, ρ, for each vector.

Algorithm 1 The Error Rate Aspect

1: Define matrix SS_{N_k} in line with (7)
2: Calculate $R[V_k^s]$ according to (8)
3: Estimate $\eta_{k,e}$ and ψ_{N_k} via (9) and (10), resp.
4: **for** $e = 1 : Q_k$ **do**
5: **if** $\eta_{k,e} \geq \psi_{N_k}$ **then**
6: $e \in A_k$
7: **else if** $\eta_{k,e} < \psi_{N_k}$ **then**
8: $e \in A_k^c$
9: **end if**
10: **end for**
11: Estimate $P_k[V_k^s]$ according to (11)
12: Calculate R_k^{er} according to (12)

In other words, the error state ρ indicates the number of TTIs in which the errors occur for the individual CE UEs along N_k TTIs. It means the sample space of error states, ES, is expressed as: $ES = \{0, 1, ..., \rho, ..., N_k\}$. Hence, the case $\rho = 0$ indicates an error-free transmission event(s) over the N_k, and the case $\rho = N_k$ refers to the event(s) with an error in each TTI within the scheduling period N_k. Then, the probability mass function of the error for each ρ in the ES with N_k scheduling period of the k-th CE UE is:

$$P_{k,N_k}(\rho) = \prod_{\vartheta_\rho} P(\eta_{k,e}) \tag{13}$$

where ϑ_ρ represents the group of vectors that have the same number of errors ρ in such that $\vartheta_\rho = SS_{N_k}[ES == \rho]$, and $P(\eta_{k,e})$ is the probability of the vector's error rate, which is calculated as:

$$P(\eta_{k,e}) = \prod_{n=1}^{N_k} P(\varphi_{k,e,n}) \tag{14}$$

where $P(\varphi_{k,e,n})$ is the probability of BLER at every n TTI within N_k for the vector e. The size of the ϑ_ρ is indicated by Υ_ρ, and represents the number of vectors that have the same number of erroneous TTI, i.e., ρ. Each vector in the ϑ_ρ is indicated by V^q_{k,ϑ_ρ}, where q is the vector index in the ϑ_ρ in such that: $\vartheta_\rho = \{V^1_{k,\vartheta_\rho}, ..., V^q_{k,\vartheta_\rho}, ..., V^{\Upsilon_\rho}_{k,\vartheta_\rho}\}$. Each V^q_{k,ϑ_ρ} requires a number of pre-allocated RBs for the HARQ retransmissions as estimated in (8). Hence, the number of pre-allocated RBs required for the HARQ of the k-th CE UEs with the scheduling period, N_k and ρ erroneous TTI is written as follows:

$$R_{k,\Upsilon_\rho} = \overline{R}\{V^1_{k,\vartheta_\rho}, ..., V^q_{k,\vartheta_\rho}, ..., V^{\Upsilon_\rho}_{k,\vartheta_\rho}\} \tag{15}$$

where R_{k,Υ_ρ} is the average number of pre-allocated RBs for every V^q_{k,ϑ_ρ} in ϑ_ρ. Based on the selected number of erroneous TTI, ρ^s, the number of pre-allocated RBs is estimated. The selected ρ of the k-th CE UE, ρ^s_k, for a received data transport block depends on the predicted BLER of individual CE UEs at every TTI along the scheduling period, N_k. Then, the number of pre-allocated RBs for the HARQ over N_k TTI with ρ^s_k errors is calculated as:

$$R^{sp}_k = R_{k,\Upsilon_{\rho^s_k}} P_{k,N_k}(\rho^s_k) \tag{16}$$

The proposed solution for the scheduling period aspect is managed by Algorithm 2 as follows. First, the sample space SS_{N_k} defined in (7) is reformulated into the number of error states ρ (line 1 in Algorithm 2). Then, the probability mass function of the error for each ρ (i.e., $P_{k,N_k}(\rho)$) and the probability of the vector's error rate ($P(\eta_{k,e})$) is estimated in (13) and (14), respectively (lines 2-3). Based on that, the number of pre-allocated RBs required for the HARQ of the k-th CE UEs with all possibilities of errors R_{k,Υ_ρ}) is estimated according to (15). Finally, R^{sp}_k is calculated in line with (16).

Algorithm 2 The Scheduling Period Aspect

1: Reformulate the sample space in (7) into ρ
2: Calculate $P_{k,N_k}(\rho)$ according to (13)
3: Estimate $P(\eta_{k,e})$ according to (14)
4: Estimate R_{k,Υ_ρ} according to (15)
5: Calculate R^{sp}_k from (16)

3) Joint optimization of the aspects:

This subsection describes the combination of both HARQ resource's pre-allocation approaches presented in previous subsections (i.e., the error rate aspect and the scheduling period aspect) in order to make the estimation more precise. Following three cases of the HARQ resource pre-allocation can take place: i) $R^{er}_k = R^{sp}_k$, ii) $R^{er}_k > R^{sp}_k$, and iii) $R^{er}_k < R^{sp}_k$.

The first case explains when pre-allocation estimation outcomes are identical in both aspects (i.e., $R^{er}_k = R^{sp}_k$); therefore, no further action is needed since the number of pre-allocated RBs is validated by both aspects. For the second case (i.e., $R^{er}_k > R^{sp}_k$), we pre-allocate the larger of both values (i.e., R^{er}_k) since the amount and placement of erroneous TTI are estimated in advance (i.e., the selected vector, V^s_k). The solution for the third case is quite different since the larger value, i.e., R^{sp}_k, shows only the number of erroneous TTIs and does not contain information on which TTI the erroneous data is placed (i.e., the selected number of erroneous TTI, ρ^s). Therefore, we initially pre-allocate the number of resources indicated in R^{er}_k. Then, in addition, the difference in the number of pre-allocated RBs between R^{er}_k and R^{sp}_k, which is denoted as R^d_k: $R^d_k = R^{sp}_k - R^{er}_k$, is also considered for the HARQ needs. In other words, the number of resources R^d_k is pre-allocated as shared resources for any k-th CE UE retransmission needs. This way, we can fulfill all CE UEs' retransmissions needs and, simultaneously, improve the scheduling resource utilization since the shared pre-allocated resources (i.e., R^d_k) can be fully re-scheduled in case of CE UEs' error-free data delivery.

Thus, the number of pre-allocated RBs for the HARQ in the t-th TTI is determined as follows:

$$R^*_{k,t} = \max_{n=N_k} \{R^{er}_k, R^{sp}_k\} \tag{17}$$

Finally, the total amount of pre-allocated resources for the HARQ retransmissions of all CE UEs over a communication session period, T_S, is:

$$R^* = \sum_{t=1}^{T_s} \sum_{k=1}^{K_{CE}} R^*_{k,t} \tag{18}$$

The integration of both aspects of the proposal is managed by Algorithm 3. At the beginning, the initialization of the algorithm is done by setting N_{max}, R_{max}, δ_{max}, and K_{CE} representing the maximum length of the scheduling period, the maximum number of resources available for pre-allocation, the maximum number of possible retransmissions, and the total number of CE UEs, respectively (see line 1 in Algorithm 3). After that, the centralized scheduling period N_k is estimated for all CE UEs (line 2). Then, Algorithm 1 and Algorithm 2 are executed to obtain R^{er}_k and R^{sp}_k, respectively (lines 4-5). In the sequel, the following two cases of the HARQ resource pre-allocation can take place: i) $R^{er}_k \geq R^{sp}_k$ or ii) $R^{er}_k < R^{sp}_k$. Based on this, R^*_k is calculated for the k-the CE UE (lines 6-10). The steps in lines 3-11 are repeated for each k-th CE UE. Finally, the overall number of pre-allocated resources R^* are calculated based on (18) (line 12).

Algorithm 3 Joint Optimization of Aspects

Initialization: N_{max}, R_{max}, δ_{max}, K_{CE}
Estimate $N_k \ \forall \ k$
for $k = 1 : K_{CE}$ **do**
    Execute Algorithm 1 for k-th CE UE (obtain R_k^{er})
    Execute Algorithm 2 for k-th CE UE (obtain R_k^{sp})
    if $R_k^{er} \geq R_k^{sp}$ **then**
        $R_{k,t}^{*} \Longleftarrow R_k^{er}$
    else if $R_k^{er} < R_k^{sp}$ **then**
        $R_{k,t}^{*} \Longleftarrow R_k^{sp}$
    end if
end for
Calculate R^* according to (18)

C. Discussion on suitability of various types of HARQ

In this section, we discuss a suitability of our proposal for various HARQ types and we outline any potential modifications that need to be done. In general, the HARQ types can be classified according to several criteria:

- *Synchronous vs. asynchronous HARQ* - In the synchronous HARQ, each HARQ process occurs at predefined times relative to the initial transmission. Thus, signaling of the HARQ process number is unnecessary and can be inferred from transmission timing. In the asynchronous HARQ, the retransmissions can occur at any time. Thus, the HARQ process number is necessary to correctly associate each retransmission with the corresponding initial transmission. In other words, the main difference for both HARQs is the retransmission timing. In our work, we adopt the asynchronous HARQ, since it is used in 5G networks [10]. Still, our proposal can be easily adapted also for the synchronous HARQ and only time of individual retransmissions may need to be changed for the synchronous HARQ while the number of pre-allocated resources is unaffected.
- *Adaptive vs. non-adaptive HARQ* – The adaptive HARQ allows to change modulation, coding rate, or number of resource blocks for retransmissions while the non-adaptive HARQ keeps these parameters the same as for the first transmission. In our work, we assume adaptive HARQ process, since it is used in 5G (please see [10]). As the result, the number of pre-allocated resource blocks is modified with respect to initial transmission depending on the current channel quality. In principle, even the non-adaptive HARQ process can be utilized for our proposal. In this case, however, the number of pre-allocated resource blocks for individual retransmissions should be the same as in case of the initial transmission of data.
- *HARQ type I-III* – In HARQ type I (chase combining), the same information and parity bits are retransmitted each time. In HARQ Type II (incremental redundancy), multiple different sets of code bits are generated for the same information bits used in each transmission. The HARQ type III is based on HARQ type II, but each retransmitted packet is self-decodable. In our work, we do not specify HARQ type, as these rather relates to the physical layer and, thus, are not relevant to the proposed pre-allocation of resource targeting higher layers. Hence, all three HARQ types can be used for our proposal.

V. PERFORMANCE EVALUATION

The performance is evaluated in the MATLAB system-level simulator. To this end, the simulation setup, competitive algorithms, and performance metrics are introduced in the following subsections. Then, the simulation findings are presented and comprehensively discussed.

A. Simulation scenario

We assume a square area of 1000x1000 m encompassing a single BBU located in the middle, up to 100 RRHs, and 200 UEs deployed randomly with uniform distribution. Each UE is associated to the RRH providing the highest SINR. In this work, we implement the 3GPP 5G-compliant model described in [39]. The orthogonal frequency division multiple access is assumed for the downlink transmission. The channel between any UE and RRH, including shadowing and fast fading, is modeled according to the Urban Micro-cell model [40] with mixed Line-Of-Sight (LOS) and Non-LOS communication (see [41]). We adopt ICIC, as explained in [22], for interference management.

We assume a realistic fronthaul with the latencies between 0 ms to 30 ms in line with the Small Cell Forum model [43], which is widely adopted by researchers. The proportional fair scheduler [44] is adopted as a basis for resource scheduling among the UEs, as this scheduler provides an adequate trade-off between network goodput and fairness [45]. For the traffic model, we select the full buffer model to examine the performance of our proposal under heavy-load network conditions.

The BLER calculation is based on the Cyclic Redundancy Check (CRC) evaluation, which is attached to transport blocks to detect the error at the receiver side (i.e., UE). The incremental redundancy (IR) HARQ with a 1/3 turbo encoder is considered. The reason behind adopting IR-HARQ is its higher coding gain compared to the chase combining HARQ [46]. The retransmitted data transport blocks are sent with an initial coding rate of 1/2 or 3/4, and the maximum number of simultaneous downlink HARQ processes is limited to 8 [39]. The HARQ adopts the N-channel stop-and-wait protocol, offering low buffering requirements and low acknowledgment (ACK) / negative acknowledgment (NACK) feedback overhead. In particular, the data packet must be delivered with a packet error rate (PER) of less than 10^{-5}, either with or without retransmission(s), as detailed in [41].

The HARQ RTT is scaled by the TTI length, which is assumed as the default time unit in this work. The TTI length depends on the number of OFDMA symbols and the subcarrier spacing of the OFDMA modulation is $t_{TTI} = N_S(1/\Delta_f + t_{CP})$, where N_S is the number of OFDMA symbols per TTI, Δ_f represents subcarrier spacing, and t_{CP} stands for the duration of a cyclic prefix. We adopt a common system configuration with the carrier spacing equal to 15kHz and the normal duration of t_{CP} equal to approximately 4.7 us. Hence, considering 14 OFDMA symbols per one TTI, the length of each TTI is equal to 1 ms. Note that the proposal can be

TABLE I: Parameters and sitting for the paper simulation

Parameters	***Values***
Simulation scenario	3GPP Urban Microcell scenario [40]
Carrier frequency	2 GHz
System bandwidth	20 MHz (200 RBs)
Number of BBU, RRH, UEs	1, up to 100, 200
α_C	1.2
RRH transmit power	27 dBm
Number of retransmissions	up to 3 attempts
TTI length	1 ms
HARQ RTT	8 TTI
Co-channel fading model	Rayleigh and Rician fading [42]
Lognorm shadowing std. dev.	4 dB for LOS, 7.82 dB for NLOS [40]
Scheduler	Proportional fair
Antenna configuration	Single input single output
Fronthaul delay	0; 2; 5; 10; 20; 30 ms [43]
Centralized sched. period	1; 5; 10; 15; 20 ms

adapted for any sub-carrier spacing and any TTI defined for 5G networks (see Table I in [47] with 5G numerologies).

Our proposed resource pre-allocation approach focuses on one part of RTT: decreasing the retransmission delay part. Since we propose HARQ retransmission at the RRH instead of the BBU, our proposal shortens the retransmission delay by at least double fronthaul. Note that the retransmission delay is understood as an additional delay caused by data transport blocks needing retransmission(s). It means the other HARQ RTT components are considered negligible for this purpose. Because the 5G networks have scalable TTIs, we assume 1 ms as a TTI length in this work. Since our proposal pre-allocates part of the scheduling resources for individual CE UEs' HARQ needs at RRH(s) for immediate retransmission(s) without the BBU intervention, any retransmission(s) is admitted and scheduled during individual CE UEs' scheduling period (i.e., N_k). Otherwise, the data packet is assumed to be lost.

Table I lists the simulation scenarios and parameters.

B. Competitive algorithms and performance metrics

To show the gain of the proposal, we compare the results of the proposed scheduling with related competitive approaches. The proposed scheduling settings comprise: *i*) the scheduling period selection, as explained in [20] [21], and *ii*) the dynamic pre-allocation of resources for HARQ retransmission, as proposed in this work. Following approaches are compared:

1) *Centralized scheduler (CS)*: The conventional scheduling process is done only at the BBU for all UEs without any functional split (i.e., split options 6-8 acc. to 3GPP [48]).
2) *Partially–Distributed Scheduler (PDS)*: The conventional scheduling is performed at either the BBU or the partially distributed radio aggregation units (RAU) depending on the individual UEs' fronthaul delay, as proposed in [49]. Since the authors in [49] do not specify any deployment scheme of the RAUs, a realistic case with the RAUs are collocated with the underlying RRH closest to the cluster's center of all underlying RRHs is assumed.
3) *Hierarchical scheduler (HS)*: The conventional hierarchical scheduler based on our previous works [20] [21], where only fixed pre-allocation of resources for the HARQ process is done. This way, we demonstrate the impact of the proposed dynamic pre-allocation.
4) *CS - Proposal*: The conventional CS implemented with our proposed scheduling settings (i.e., dynamic pre-allocation of resources for HARQ retransmissions).
5) *PDS - Proposal*: The conventional PDS implemented with our proposed scheduling settings.
6) *HS - Proposal*: The conventional HS implemented with our proposed scheduling settings.
7) $HS-Optimum$: We also show a theoretical upper bound of the hierarchical scheduler in terms of network goodput and the CE UEs goodput. The scheduling settings are dynamic and optimally adjusted for individual CE UEs. The HS-Optimum comprises two parts: *i*) an estimation of the optimal scheduling period length, N_{opt}, and *ii*) estimation of the optimal amount of pre-allocated resources for the HARQ process R^*_{opt}, while perfect prediction of the future signal characteristic for the monitored period (50 ms in our case) is assumed for estimation of both. The first part, N_{opt}, is determined individually for each CE UE according to its "channel dynamicity". The channel dynamicity is understood as a significance of the changes in CQI per a monitored period of time. In general, the more significant the CQI changes within the monitored period, the lower N is set for the CE UE and vice versa. To find N_{opt}, we subsequently set the scheduling period N from 1 to N_{max} and N yielding the maximum goodput is selected as the optimal individually for each CE UE. Regarding the second part, the HS-Optimum always pre-allocates the exact amount of RBs needed for any retransmission (please see 1. scenario in Fig. 2 representing the ideal pre-allocation). Even though the *HS-Optimum* cannot be determined in real-world networks, it represents the upper bound performance and allows us to demonstrate the efficiency of the proposed solutions.

The performance of the competitive solutions and the proposed algorithms are assessed by four performance metrics:

1) *Network goodput*: Sum goodput over both CE UEs and nCE UEs. The calculation of network goodput is based on (5). Note that the signaling overhead (i.e., OH_k in (5)) is also taken into consideration when the goodput is estimated. Basically, the overhead size usually varies between 7% and 14% of the downlink subframe size.
2) *CE UEs goodput*: Sum goodput only over the CE UEs for whom the proposal is tailored specifically.
3) *Downlink transport block loss rate*: The ratio of data transport blocks not retransmitted to the CE UEs within 8 consecutive TTI due to a lack of scheduling resources divided by the total number of received transport blocks.
4) *MAPE*: The evaluation of the number of pre-allocated RBs accuracy of the paper proposal compared to other counterparts (see (4) for the calculation of the MAPE).

C. Simulation results

This subsection summarizes the paper's findings and contributions made. Let us start with the impact of the fronthaul delay on the goodput of all UEs (i.e., the network goodput)

as Fig. 3a, and also on the goodput of only CE UEs (see Fig. 3b). Disregarding the scheduler type, both the network goodput and the CE UEs goodput gradually decrease with the fronthaul delay increasing. This is because the fronthaul delay postpones the required scheduling information, i.e., channel quality reports and delivery of the scheduling decision to the RRHs. The fronthaul delay impairs the *PDS*, *CS*, and *HS* approaches' goodput (i.e., the CE UEs and the network) more significantly with respect to the proposed ones (*PDS - Proposal*, *CS - Proposal*, and *HS - Proposal*), especially for higher fronthaul delays. More specifically, for the *PDS*, *CS*, and *HS* approaches, the network goodput and the CE UEs goodput are notably decreased compared to the optimum hierarchical scheduler approach (i.e., *HS - Optimum*) by up to 39% if the fronthaul delay is increased from 0 to 30 ms (see Fig. 3). At the same time, the performance gap between *HS - Proposal* and *HS - Optimum* is only by up to 2%. The gain of the hierarchical scheduler is attained via an efficient suppression of the negative fronthaul delay impact as the centralized scheduling decision can be adjusted in the RRHs for the nCE UEs in case the fronthaul delay leads to a notable change in the channel quality. At the same time, the CE UEs can still benefit from ICIC gain, as these are scheduled solely by the BBU.

Another important finding is that the *HS - Proposal* outperforms the *HS* by up to 13% for both the network goodput and the CE UEs goodput (see Fig. 3). Besides, the added value of our proposal is that we can improve the performance of conventional approaches, i.e., *CS* and *PDS*, if they exploit our pre-allocation algorithm. In particular, the CE UEs goodput and the network goodput of *CS-Proposal*, and *PDS-Proposal* approaches outperform the *CS*, and *PDS* approaches' by up to around 11%. This improvement is because the dynamicity of the HARQ resource pre-allocation minimizes the probability of transport blocks getting lost due to the lack of pre-allocated resources. Moreover, the dynamicity of the HARQ resource pre-allocation decreases the amount of unexploited resources since the HARQ pre-allocated resources vary based on the individual CE UEs' actual needs. However, in the *CS*, *PDS*, and *HS* approaches, the amount of the HARQ pre-allocated resources is fixed. Based on that, the probability of data transport blocks getting lost due to a lack of resources is notably increased since there is no flexibility in the amount of HARQ pre-allocated resources.

The impact of the prolonged scheduling period in the C-Sc on the network goodput and the CE UEs goodput is investigated in Fig. 4a and Fig. 4b, respectively. Intuitively, the longer the scheduling period is, the less signaling overhead is required. In Fig. 4, we observe that both the network and the CE UEs goodput increase with the prolonging of N until the maximum goodput is reached at some point. Then, the goodput starts decreasing. In the first phase (i.e., the goodput raising phase from 1 ms to 2 ms scheduling period), the network goodput and the CE UEs goodput are increased as the signaling overhead related to the scheduling is notably reduced and more resources remain for new data transmissions. Moreover, the channel quality for low values of N is generally stable, and the loss in goodput due to a higher error rate

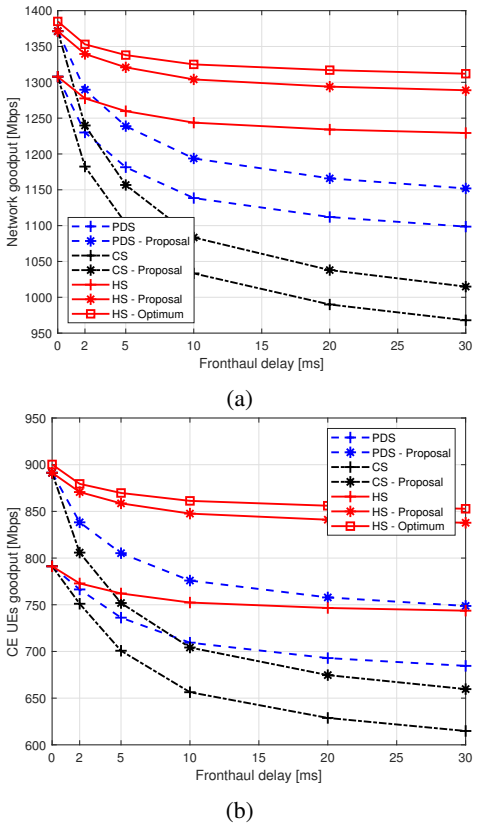
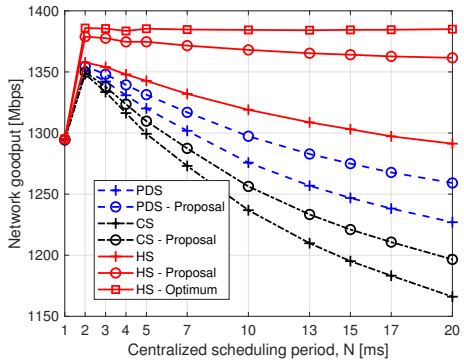


Fig. 3: Impact of fronthaul delay on the network goodput (a) and on the CE UEs goodput (b) for centralized, partially distributed, and hierarchical schedulers (note that N is set dynamically up to 20 ms).

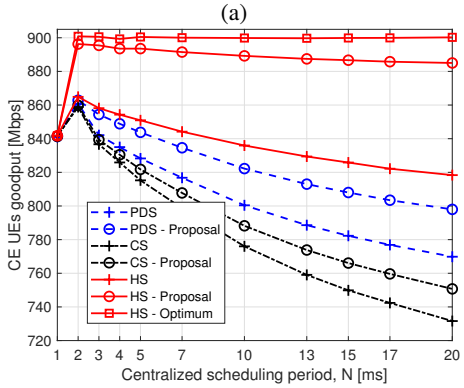
resulting from potentially outdated channel state information for scheduling is negligible. In the second phase (i.e., the goodput declining phase from about 2 ms scheduling period and onward), the goodput starts decreasing gradually with N as the impact of outdated channel state information becomes more significant and dominates the gain introduced by the overhead saving.

Fig. 4 also indicates that the degradation of goodput in case of *HS - Proposal* are suppressed to be below 5% and 2%, if N is set to 1 ms and 20 ms, respectively. This is because the scheduling period and the retransmission of pre-allocated resources are dynamically set based on individual CE UEs CSI and the HARQ resources need on the way to minimize the negative impact of the outdated CE UEs CSI. Note that the hierarchical scheduler does not allow adjusting the scheduling resources for CE UEs at the RRHs, as multiple RRHs serve these resources, and uncoordinated scheduling updates by these RRHs would lead to potentially strong interference. However, the goodput accomplished by *PDS* and *CS* severely decreases to around 15% for the same range of N.

Our results also cast a new light on the HARQ resource's pre-allocation options. Generally, *HS - Proposal*, *CS - Proposal*, *PDS - Proposal* outperform the conventional approaches (i.e., *HS*, *CS*, *PDS*) by nearly up to 9%. The proposal approaches outstanding performance is not surprising as the amount of the HARQ RBs for individual CE UEs is



(a)



(b)

Fig. 4: Impact of scheduling period on the network goodput (a) and on the CE UEs goodput (b) for centralized, partially distributed, and hierarchical scheduler (fronthaul delay=0 ms).

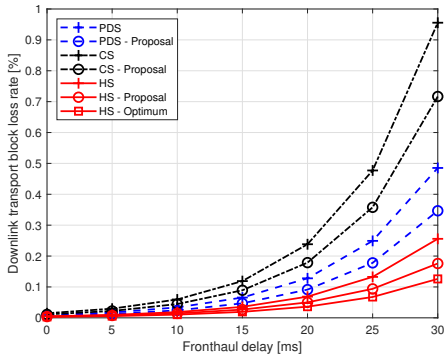


Fig. 5: Impact of the fronthaul delay on the transport block loss rate.

dynamically set according to their actual needs. Overall, the *HS-Proposal* approach is the one that obtained the most robust results in comparison with the *HS - Optimum*. The results show that the *HS-Proposal* is dropped only by around 2% compared to the *HS -Optimum*, while the *HS* is fallen by 10%.

Fig. 5 shows the probability of data transport blocks being lost due to insufficient pre-allocated resources for HARQ retransmissions (i.e., within up to 8 TTI). Fig. 5 reveals that the transport block loss probability increases with the fronthaul delay for all schedulers because the fronthaul delay negatively influences the scheduling decision. Intuitively, a higher error rate is observed for high fronthaul. However, the *HS - Proposal* significantly reduces the loss rate by up to 83%

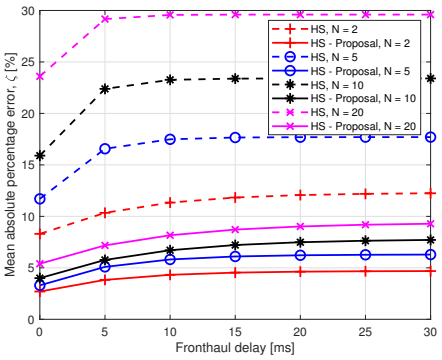


Fig. 6: Impact of the fronthaul delay on the MAPE of the proposed hierarchical scheduler and the conventional hierarchical scheduler for different values of the scheduling period.

and 65%, compared to the *PDS* and *CS*, respectively, due to the dynamicity of the HARQ resource pre-allocation. Moreover, a notable improvement is observed even if we compare the *HS - Proposal* with the *HS*, where our proposal reduces the loss rate by roughly up to 30% for the longer fronthaul delay. The superiority of the dynamic HARQ approach for hierarchical scheduler comes not only from its ability to suppress the negative impact of a fronthaul delay by scheduling part of the UEs (i.e., nCE UEs) at RRHs but also from the dynamic amount of RBs that are assigned for the HARQ process based on individual CE UEs radio channel characteristics. However, even better results are achieved using our proposed scheduling setting on both *CS* and *PDS*, where the loss rate is remarkably reduced by up to 38% and 27%, respectively.

Fig. 6 shows the impact of the fronthaul delay and the scheduling period prolongation on the MAPE of the amount of the HARQ pre-allocated resources, ζ. The ζ for all shown scenarios starts increasing with the fronthaul delay due to the negative impact of outdated CSI. Our approach, *HS - Proposal*, reaches the ζ of up to 57% lower compared to the *HS*, regardless of the length of the scheduling period. This is because the actual value of retransmission pre-allocated resources is adjusted according to the CE UEs' HARQ actual needs rather than keeping it static for all CE UEs. Furthermore, the ζ values increase with the scheduling period (i.e., N) for all presented approaches as the scheduling information is not up to date for the later TTIs within the scheduling period. Still, the proposed dynamic pre-allocation of resources decreases ζ by roughly three times compared to static allocation.

VI. CONCLUSION

This paper proposes the dynamic resource pre-allocation framework for hierarchical scheduling in mobile networks with C-RAN architecture. The dynamic resource pre-allocation calculates the pre-allocated resources for the HARQ by combining two distinct approaches: 1) the error rate and 2) the scheduling period. In both aspects, we derived analytical expressions for estimating the amount of resources needed. Our simulation results illustrate that the proposed dynamic pre-allocation scheduler increases the goodput with other presented schedulers by around 39% and, at the same time,

minimizes the transport block loss rate and mean absolute percentage error of the amount of pre-allocated resources for the HARQ by 38% and 57%, respectively.

The work can be further extended by an energy-efficient resource allocation model to obtain better bandwidth performance and, thus, making it suitable for new scenarios such as the internet of things, big data, and smart city.

REFERENCES

[1] O.O. Erunkulu, A. M. Zungeru, C. K. Lebekwe, M. Mosalaosi, and J. M. Chuma, "5G Mobile Communication Applications: A Survey and Comparison of Use Cases," *IEEE Access*, vol. 9, pp. 97251-97295, 2021.

[2] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, "5G RAN: Functional Split Orchestration Optimization," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, pp. 1448-1463, July 2020.

[3] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," in *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94-100, May 2017.

[4] A. Martínez Alba and W. Kellerer, "Dynamic Functional Split Adaptation in Next-Generation Radio Access Networks," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 3, pp. 3239-3263, Sept. 2022.

[5] A. Checko, *et al.* "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, 405-426, 2014.

[6] F. E. Kadan and A. Ö. Yılmaz, "A Theoretical Performance Bound for Joint Beamformer Design of Wireless Fronthaul and Access Links in Downlink C-RAN," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2177-2192, Apr. 2022.

[7] Holma, Harri, Antti Toskala, and Takehiro Nakamura, eds. 5G technology: 3GPP new radio. *John Wiley & Sons*, 2020.

[8] F. Voigtländer, *et al.*, "5G for the factory of the future: Wireless communication in an industrial environment." arXiv:1904.01476, 2019.

[9] W. Ejaz, *et al.*, "A comprehensive survey on resource allocation for CRAN in 5G and beyond networks," *Journal of Network and Computer Applications*, vol. 160, 2020.

[10] A. Ashfaq, *et al.*, "Hybrid automatic repeat request (HARQ) in wireless communications systems and standards: A contemporary survey." *IEEE Commun. Surveys Tuts.*, vol. 23, no.4, pp. 2711-2752, 2021.

[11] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications," In *IEEE WCNC*, 2017.

[12] Y. Han, S. E. Elayoubi, A. G.-Serrano, V. S. Varma, and M. Messai. "Periodic radio resource allocation to meet latency and reliability requirements in 5G networks," In *IEEE 87th VTC Spring*, pp. 1-6. IEEE, 2018.

[13] S. E. Elayoubi, P. Brown, M. Deghel, and A. G.-Serrano, "Radio resource allocation and retransmission schemes for URLLC over 5G networks," *IEEE J. Sel. Areas Commun.* , vol. 37, no. 4, 896-904, 2019.

[14] H. Jang, J. Kim, W. Yoo, and J.-M. Chung, "URLLC mode optimal resource allocation to support HARQ in 5G wireless networks," *IEEE Access*, vol. 8, pp. 126797-126804, 2020.

[15] M. Sharara, S. Hoteit, P. Brown, and V. Veque, "On Coordinated Scheduling of Radio and Computing Resources in Cloud-RAN," to appear in *IEEE Transactions on Network and Service Management*, 2023.

[16] A. Karimi, K. I. Pedersen, and P. Mogensen, "Low-complexity centralized multi-cell radio resource allocation for 5G URLLC." In *IEEE Wireless Communications and Networking Conference (WCNC)*, 2020.

[17] F. Bassi and H. I. Khedher, "HARQ-aware allocation of computing resources in C-RAN." In *IEEE ISCC*, 2020.

[18] B. Goktepe, C. Hellge, T. Schierl, and S. Stanczak, "A hybrid HARQ feedback prediction approach for Single- and Cloud-RANs in the sub-THz regime," *IEEE GLOBECOM Workshop*, 2022.

[19] B. Goktepe, C. Hellge, T. Schierl, and S. Stanczak, "Distributed Machine-Learning for Early HARQ Feedback Prediction in Cloud RANs," to appear in *IEEE Transactions on Wireless Communications*, 2023.

[20] Z. Becvar, P. Mach, M. Elfiky, and M. Sakamoto, "Hierarchical scheduling for suppression of fronthaul delay in C-RAN with dynamic functional split." *IEEE Communications Magazine*, vol. 59, no. 4, pp. 95-101, 2021.

[21] M. Elfiky, Z. Becvar, and P. Mach, "Dynamic Adjustment of Scheduling Period in Mobile Networks Based on C-RAN." In *IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pp. 1-7., 2021.

[22] G. Nardini, G. Stea, A. Virdis, D. Sabella, and M. Caretti, "Practical large-scale coordinated scheduling in LTE-Advanced networks," *Wireless Networks*, vol. 22, pp. 11-31, 2016.

[23] Sesia, Stefania, Issam Toufik, and Matthew Baker. LTE-the UMTS long term evolution: from theory to practice. *John Wiley & Sons*, 2011.

[24] Z. Wang, Y. Wei, and F. Richard Yu, "Utility optimization for resource allocation in multi-access edge network slicing: A twin-actor deep deterministic policy gradient approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 5842-5856, 2022.

[25] Y. Ren, A. Guo, Ch. Song, and Y. Xing, "Dynamic resource allocation scheme and deep deterministic policy gradient-based mobile edge computing slices system," *IEEE Access*, vol. 9, pp. 86062-86073, 2021.

[26] V. Angelakis, I. Avgouleas, N. Pappas, E. Fitzgerald, and D. Yuan, "Allocation of heterogeneous resources of an IoT device to flexible services," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 691–700, 2016.

[27] N. Sharma and K. Krishan, "Resource allocation trends for ultra dense networks in 5G and beyond networks: A classification and comprehensive survey," *Physical Communication*, vol. 48, 2021.

[28] S. Kim, "Asymptotic shapley value-based resource allocation scheme for IoT services," *Comput. Netw.*, vol. 100, pp. 55-63, 2016.

[29] M. Kim and I.Y. Ko, "An efficient resource allocation approach based on a genetic algorithm for composite services in IoT environments," *IEEE International Conference on Web Services*, pp. 543–550, 2015.

[30] P.Y. Yin and J.Y. Wang, "A particle swarm optimization approach to the nonlinear resource allocation problem," *Appl. Math. Comput.*, vol. 183, pp. 232–242, 2006.

[31] A. K. Mohamed, A. W. Mohamed, E. Z. Elfeky, and M. Saleh, "Solving constrained non-linear integer and mixed-integer global optimization problems using enhanced directed differential evolution algorithm," *Machine learning paradigms: Theory and application*, vol. 801, 2019.

[32] A. K. Sangaiah, *et al.*, "IoT resource allocation and optimization based on heuristic algorithm," *Sensors*, vol. 20, no.2, 2020.

[33] J. Plachy, Z. Becvar, P. Mach, R. Marik, and M. Vondra, "Joint Positioning of Flying Base Stations and Association of Users: Evolutionary-Based Approach," *IEEE Access*, vol. 7, 2019.

[34] D. Soldani, *et al.*, "5G for ultra-reliable low-latency communications." *IEEE Network*, vol. 32, no. 2, pp. 6-7, 2018.

[35] Z. Sayeed, E. Grinshpun, D. Faucher and S. Sharma, "Long-term application-level wireless link quality prediction," 2015 36th *IEEE Sarnoff Symposium*, Newark, NJ, USA, 2015, pp. 40-45.

[36] Y. Zheng, *et al.*, "A modified ARIMA model for CQI prediction in LTE-based mobile satellite communications," In *IEEE International Conference on Information Science and Technology*, pp. 822-826. 2012.

[37] Watanabe, Sumio. "A widely applicable Bayesian information criterion." The Journal of Machine Learning Research 14, no. 1 (2013): 867-897.

[38] Wang, Yuan H. "On the number of successes in independent trials," *Statistica Sinica* (1993): 295-312.

[39] 3GPP, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures. TS 36.300 version 17.1.0 Release 17, 2022.

[40] 38.901, G.T., Study on channel model for frequencies from 0.5 to 100 GHz. 3GPP TR 38.901, 2022. 17.0.0.

[41] 3GPP TS 22.261 Service requirements for the 5G systems; Stage 1 (Release 17)," 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, Tech. Rep., 2022

[42] W. Sun, Q. Yu, W. Meng, and V. CM Leung, "Transmission mechanism and performance analysis of multiuser opportunistic beamforming in Rayleigh and Rician fading channels." *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9459-9473, 2018.

[43] Small Cell Forum. Small cell virtualization: Functional splits and use cases. White Paper. Release 6.0. 2016.

[44] R. Kwan, C. Leung, and J. Zhang, "Proportional fair multiuser scheduling in LTE," *IEEE Signal Process. Lett.*, vol. 16, no. 6, 461-464, 2009.

[45] S. Mosleh, L. Liu, and J. Zhang, "Proportional-fair resource allocation for coordinated multi-point transmission in LTE-advanced," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, 5355-5367, 2016.

[46] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with chase combining and incremental redundancy for HSDPA," *IEEE Vehicular Technology Conference. VTC-Fall*, 1829-1833, 2001.

[47] A. B. Kihero, M. S. J. Solaija, and H. Arslan, "Inter-Numerology Interference for Beyond 5G," *IEEE Access*, vol. 7, 146512-146523, 2019.

[48] 3GPP TS 38.801, "Study on new radio access technology: Radio access architecture and interfaces," V. 14.0.0, 2017.

[49] M. Huang and X. Zhang, "Distributed MAC Scheduling Scheme for C-RAN with Non-Ideal Fronthaul in 5G Networks," 2017 *IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1-6.