HARQ-Based Grant-Free NOMA for mMTC Uplink

Faramarz Jabbarvaziri, *Member, IEEE,* Naveen Mysore Balasubramanya, *Member, IEEE,* and Lutz Lampe, *Senior Member, IEEE*

***Abstract*—Massive machine-type communication (mMTC) is one of the most rewarding and at the same time challenging components in the fifth-generation (5G) cellular solutions supporting the Internet of things (IoT). The 5G mMTC is considering the use of a combination of two key mMTC enabling technologies —grant-free (GF) transmission and non-orthogonal multiple-access (NOMA), called GF-NOMA, which can potentially exploit the advantages of both schemes. A primary challenge in GF-NOMA is to reduce the packet drop rate. Owing to the decentralized nature of the GF schemes and the lack of control over user equipment, only hybrid automatic repeat request (HARQ) Type I has been employed for enhancing the reliability of GF-NOMA so far. In this paper, uplink GF-NOMA transmission schemes using HARQ Type III are proposed. Two types of packet combining —a) chase combining and b) incremental redundancy combining are considered. Moreover, we introduce a grant-free single-transmission (GFST) scheme where all redundancy versions of the packet are transmitted in one shot. We present a comprehensive evaluation of both the GF and the conventional grant-based methods in mMTC scenarios, and demonstrate the superiority of our proposed methods over the existing ones.**

***Index Terms*—Massive machine-type communication (mMTC), Grant-free (GF) transmission, Non-orthogonal multiple access (NOMA), Hybrid automatic repeat request (HARQ), Internet of things (IoT)**

I. Introduction

Massive connectivity is a critical requirement to support massive machine-type communication (mMTC) applications in future networks hosting the Internet of things (IoT). Typically, the mMTC devices are designed to stay dormant unless there is an event to be reported, resulting in sporadic user activity [1]. Moreover, multiple mMTC applications require low data rate communications, where the transmitted packets are assumed to be small (200 bytes) and the battery life of the mMTC devices is expected to be between 10 and 15 years [2]. However, the number of devices can reach up to 10^6 per square kilometers [3].

In the fourth generation (4G) cellular systems, like long term evolution (LTE) and LTE-advanced (LTE-A), user equipments (UEs) use a random access (RA) procedure to connect to the network and request time-frequency resources for data transmission. While the RA process is contention-based, devices retain their connection in a contention-free manner by following the scheduling information sent by the serving base station. This scheme is called grant-based (GB) transmission.

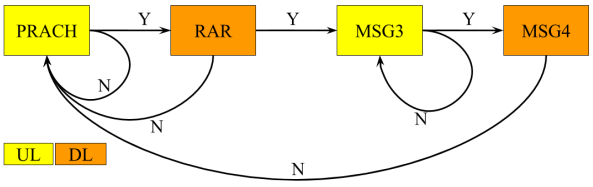


Fig. 1: RA process of the LTE

The four-step RA process in GB transmission (referred to as the conventional random access) includes physical random access channel (PRACH) preamble transmission, random access response (RAR), radio resource control (RRC) connection request (MSG3) and the RRC connection setup (MSG4) [4]. After completing the four-step RA process, a UE is in "connected mode", where it constantly receives resource allocation information from the base station for data transmission. Fig. 1 illustrates the four-step RA process from a state transition viewpoint associated with uplink (UL) and downlink (DL) messages. The current MTC standards, such as, the narrowband-IoT (NB-IoT) and enhanced MTC (eMTC) also adopt similar procedures for GB transmission.

Recently, in the fifth generation (5G) new radio (NR), a two-step RA procedure is introduced in order to reduce the latency resulting from the conventional four-step RA procedure [5]. In this method, preamble transmission and data transmission from the conventional four-step RA procedure are combined into one message called MSGA, which corresponds to the first step. The second step corresponds to a combination of the random access response and the data acknowledgment of the four-step RA procedure, denoted as MSGB. The preamble transmission is similar to the four-step RA (albeit different preamble sequences and transmission occasions may be used). That is, preambles from different users are transmitted in a contention-based manner and occupy a predetermined set of time-frequency resources. Therefore, in terms of probability of preamble collision, its performance is similar to the conventional four-step RA. A detailed analysis of this scheme is presented in Section IV-B.

Although GB transmission was very useful in LTE/LTE-A, its control mechanism is already not suitable for mMTC. This is because the scheduling of a massive number of devices incurs a considerable amount of signalling overhead, and the back-and-forth controlling messages not only drain the UEs' battery, but also lead to increased latency [1], [6]. This motivated academia [1], [6]–[20] and industry [21] (more than 15 companies) to explore grant-free (GF) transmission schemes for mMTC and ultra-reliable low-latency communi-

F. Jabbarvaziri and L. Lampe are with the Department of Electrical and Computer Engineering, University of British Columbia, BC V6T 1Z4, Canada (e-mail: jabbarva@ece.ubc.ca, lampe@ece.ubc.ca).

N. M. Balasubramanya is with the Department of Electrical Engineering, Indian Institute of Technology Dharwad, Karnataka, India (e-mail: naveenmb@iitdh.ac.in).

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

TABLE I: Messages in GB and GF transmissions.

Message	GB	GF
UL: RACH Preamble	✓	✓
DL: RAR	✓	✗
UL: MSG3 = RRC CR	✓	✗
DL: MSG4 = RRC CS	✓	✗
UL: Scheduling Request	✓	✗
DL: DCI: UL grant	✓	✗
UL: RRC CS complete	✓	✗
UL: Data	✓	✓

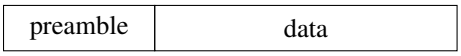


Fig. 2: Two-phase transmission scheme

cation (URLLC) in 5G. The UEs following the GF transmission scheme adopt a "transmit when data is ready" strategy, without requesting a grant from the base station [6]. Table I summarizes the messages involved in GB and GF transmission methods. Moreover, GF transmission is being considered for standardization in Release 16 of NB-IoT.

The main challenge in GF transmission is to handle collisions resulting from the contention-based access. Traditionally, this was handled by mechanisms, such as, tree algorithms and successive interference cancellation assisted tree algorithms introduced in ALOHA systems [22], [23]. Notwithstanding the benefits of such mechanisms in a lightly loaded network, they have been shown to be inefficient in mMTC scenarios [24]. As an alternative, NOMA methods have been found to support massive connectivity through the employment of efficient multi-user detection strategies [10], [11], [14], [25]. They have also been shown to be better than the conventional orthogonal multiple access (OMA) methods used in current wireless networks [10]. Therefore, there has been a trend in industry and academia to shift from OMA to NOMA in recent years.

GF-NOMA, an integration of both GF transmission scheme and NOMA, which can potentially combine the advantages of both the schemes, is also being evaluated for 5G mMTC [1], [7]–[9], [11], [12], [15], [17], [25]–[27]. The majority of GF-NOMA schemes propose a two-phase transmission method that includes a preamble transmission in the first phase, followed by the data transmission in the second phase, as illustrated in Fig. 2 [1], [7]–[9], [11], [12], [15], [25], [26]. The preamble is used for activity detection, UL synchronization and channel estimation. Preambles are generally assumed to be non-orthogonal, so that the system can serve a massive number of users. In addition, preambles are also used to identify NOMA signatures through a predefined mapping [9], [11], [15], [26] and it is assumed that users are free to randomly choose their preambles. Other GF-NOMA schemes consider joint user activity and data detection without the aid of the preambles [17], [27]. While these schemes exploit the sparsity of the user activity, the base station will have to blindly decode the NOMA signals, which is computationally costly. In this paper, we adopt the two-phase GF-NOMA transmission method using a preamble.

A. Motivation

A primary problem in GF-NOMA is the high packet drop rate, which is mainly caused by the interference introduced by non-orthogonality of the different users' signals and the collisions introduced by unsupervised contention-based transmissions. As shown in conventional GB transmission schemes such as LTE and LTE-A, hybrid automatic repeat request (HARQ) is an effective means for enhancing the packet reception reliability. HARQ Type I, i.e., repetition of forward error correction (FEC) coded packets, has already been employed in GF-NOMA to improve transmission reliability [9], [28].

Further benefits can be expected from HARQ with packet combining, i.e., chase combining (CC) or incremental redundancy (IR) combining. For CC, i.e., maximum-ratio combining (MRC) of repeated packets, the base station requires the UE identifier (ID) or the data transaction ID. For IR combining, the base station needs to know the redundancy version (RV) of each received packet as well as the UE or the data transaction ID. Each RV is a predefined subset of the FEC coded data. If the packet cannot be decoded with one RV, the remaining RVs are transmitted and IR combining is employed to improve the probability of data decoding. If each RV is self-decodable, it corresponds to HARQ Type III [29], otherwise, it corresponds to HARQ Type II [30]. For example, the NB-IoT and eMTC standards use HARQ Type III with turbo-coded data. The coded data is divided into four RVs, where RV-0 is used for the first transmission and the remaining RVs are used for re-transmissions.

Prior works in HARQ for conventional NOMA (grant-based power-domain NOMA) analyze the performance of various HARQ schemes for a system with two users [31]–[33] in the downlink. For instance, [31] determines an upper bound for outage probability of an IR based HARQ scheme along with the power allocation. A similar scenario is considered in [32], where closed-form expressions of the outage probability and diversity gain are derived for a CC based HARQ mechanism. In [33], the diversity order (asymptotic scaling law of the outage probability with respect to the transmit power) is analyzed for HARQ Type I, CC-HARQ and IR-HARQ with IR. However, due to the lack of DL control channels for carrying the HARQ meta-data (e.g. UE ID, RV information) in GF scenarios, implementing Type II and III CC-HARQ and IR-HARQ for GF-NOMA is challenging. To the best of our knowledge, such schemes have not been proposed for GF systems yet (especially in a multi-user uplink scenario). In this paper, we propose decentralized GF HARQ Type III schemes, where the HARQ meta-data of a UE is embedded in its UL preamble through predefined mappings.

B. Main Contributions

The contributions of this paper are summarized as follows.

- We propose a physical layer solution for encoding the HARQ meta-data in UL preambles and enable Type III CC-HARQ and Type III IR-HARQ in the GF-NOMA systems.
- We introduce a GF single transmission (GFST) method, where all the RVs are transmitted in one shot. We

quantitatively show that this method is the most energy-efficient one in scenarios with high channel code rate or high network load.
- We present probability-based models facilitating the analysis of the conventional GB IR HARQ mechanism and the proposed GF HARQ schemes, and adopt the same to analyze the probability of successful packet reception.
- Using the probability of successful packet reception, we analyze the key performance indicators (KPIs) of the system, including the UE energy efficiency, DL overhead and packet delay. We show that practical aspects, such as, the network load and channel code rate play significant roles in determining the most energy-efficient method in different scenarios.

C. Organization

The rest of this paper is organized as follows. Section II describes the system model. The proposed and the benchmark HARQ schemes are introduced in Section III and their analytical models are presented in Section IV. Section V presents the analysis of the probability of successful packet reception and Section VI is dedicated to the analytical characterization of the KPIs of the system. Section VII provides the numerical results, both from analytical expressions and system simulations. Finally, conclusions are drawn in Section VIII.

D. Notation

Scalars are denoted by both lower-case and upper-case italic letters, and vectors are denoted by bold-face letters. The letter T in the sub-script or super-script of the variables denotes the term "total" and $(.)^{\mathrm{H}}$ stands for matrix Hermitian operation. The complex Normal distribution with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu, \sigma^2)$. $\mathrm{Bernoulli}(X, \epsilon)$ denotes a Bernoulli distribution of rate ϵ and $\mathrm{Binomial}(X, M, \epsilon)$ denotes a Binomial distribution of rate ϵ, number of components M and random variable X. Also, expected values are denoted by $\bar{(.)}$ and logical not by $\tilde{(.)}$.

II. System Model

Consider a network consisting of N_{UE} users[1], that are randomly distributed within a circle of radius R, as shown in Fig. 3. Assume that each user transmits a new, fixed-size packet in a grant-free manner with probability ϵ. After transmission, UEs wait for an acknowledgement (ACK) message from the base station (gNB). If a UE does not receive an ACK[2], it re-transmits the packet. The maximum number of transmissions allowed is denoted by n^{HARQ}. If a packet is not decoded successfully after n^{HARQ} transmissions, it will be considered dropped.

In a conventional grant-based NOMA in the uplink, either open loop or closed loop power control can be adopted for data transmission. Typically, closed loop power control is used

[1]We use the terms user, device and UE synonymously to refer to mMTC entities participating in the UL of the system model.

[2]Negative acknowledgment (NACK) messages are not considered in this GF scheme, since they mitigate the efficiency of the system.

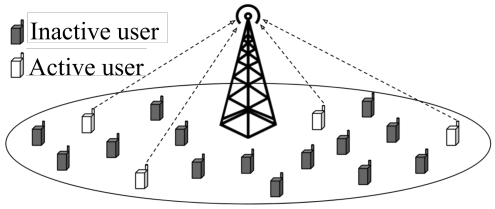


Fig. 3: System model of mMTC

after the users have successfully obtained access to the gNB and want to transmit data. Then, the gNB conveys the transmit power for each user along with the time-frequency resource grant for data transmission through the downlink control channels. However, in GF-NOMA, there are no dedicated grants of resources for each user's data transmission and the data of different users are transmitted in a contention-based manner along with their random access preambles. Hence, open-loop power control is more suitable for GF-NOMA scenarios and thus adopted in this paper (consistent with the power control mechanism used in [9], [12], [34]). Specifically, users regularly listen to the primary synchronization signal (PSS) transmitted by the gNB for downlink synchronization in order to determine their transmit power. We assume perfect time and frequency synchronization, accurate power control and an ideal ACK reception in the DL for simplicity.

Following the notations of LTE, we refer to the smallest block of time in which a UE can transmit as the transmission time interval (TTI) and assume that the channel coherence time is larger than the TTI. Accordingly, we consider a block-fading channel model, where each coherence interval is considered as a separate block and channel realizations are independent across blocks [7]. In each block, we can write the complex-valued channel coefficient of the u^{th} user as

$$h'_u = \sqrt{\beta_u} h_u, \tag{1}$$

where β_u denotes the path-loss component, and h_u denotes the small-scale fading coefficient assumed to be complex Gaussian distributed, i.e., Rayleigh fading.

Using the open-loop power control mechanism, each UE first estimates β_u from the received PSS and then compensates for it by adjusting the transmit power to a certain level so that the received power of all users be the same at gNB. At each TTI, active users transmit a preamble of length L_{p} followed by a NOMA-coded data signal that is spread over orthogonal resources using non-orthogonal Zadoff-Chu sequences of length L_{s}[3]. We consider one-to-one mapping between the preamble and NOMA spreading sequence. We assume that, after applying the adjustment of the power control mechanism, the average received signal powers of the preamble sequence and symbols of data are the same for all users and denoted by ρ_{p} and ρ_{d}, respectively. After detecting preambles and UL synchronization, the gNB estimates the channel and learns the NOMA sequence used for data spreading owing to the unique

[3]We consider the case where L_{s} is a prime number, since it limits the maximum correlation between two Zadoff-Chu sequences to $\frac{1}{\sqrt{L_{\mathrm{s}}}}$ [10].

mapping between the preambles and the NOMA spreading sequences. The gNB then de-spreads and decodes the data.

Following previous works in massive connectivity scenarios [7], [8], we adopt randomly generated preambles and the approximate message passing (AMP) algorithm for user activity detection and channel estimation. Accordingly, the preamble $\boldsymbol{p}_u = [p_u^1, p_u^2, ..., p_u^{L_\mathrm{p}}]$ selected by user u^{th} is modelled as a sequence of complex Gaussian random variables. Denoting the set of active users by $\mathcal{K}$, the received preamble sequence of the u^{th} user at the gNB is given by

$$\boldsymbol{y}_u^{\mathrm{p}} = h_u \boldsymbol{p}_u + \sum_{n \in \mathcal{K}, n \neq u} h_n \boldsymbol{p}_n + \boldsymbol{n}_{\mathrm{p}}. \tag{2}$$

In (2), $\boldsymbol{n}_{\mathrm{p}} = [n_{\mathrm{p}}^1, n_{\mathrm{p}}^2, ..., n_{\mathrm{p}}^{L_\mathrm{p}}]$ denotes the additive white Gaussian noise (AWGN) vector (i.e., $n_{\mathrm{p}}^k \overset{i.i.d}{\sim} \mathcal{CN}(0, \sigma_{\mathrm{p}}^2)$) affecting the preamble part. The preamble is used to obtain the channel estimation $\hat{h}_u$ with the estimation error $\Delta h_u = h_u - \hat{h}_u$. According to [8], $\hat{h}_u \overset{i.i.d}{\sim} \mathcal{CN}(0, \sigma_{\mathrm{F}}^2)$ and $\Delta h_u \overset{i.i.d}{\sim} \mathcal{CN}(0, \sigma_{\mathrm{E}}^2)$, where σ_{F}^2 denotes the variance of the fading channel gain and σ_{E}^2 denotes the variance of the channel estimation error of a minimum mean-squared error (MMSE)-based AMP receiver after convergence of the algorithm:

$$\sigma_{\mathrm{E}}^2 = \frac{\tau_{t\to\infty}^2}{1 + \tau_{t\to\infty}^2}, \tag{3}$$

with

$$\tau_{t\to\infty}^2 = \frac{1}{\gamma^{\mathrm{p}}(1 - \kappa\mu)}. \tag{4}$$

In equation (4), $\gamma^{\mathrm{p}} = \frac{\rho_{\mathrm{p}}}{\sigma_{\mathrm{p}}^2}$ denotes the signal-to-noise ratio (SNR) of the preamble, κ and μ are determined using

$$\kappa = \frac{N_{\mathrm{p}}}{L_{\mathrm{p}}}, \ \mu = \frac{N_{\mathrm{a}}}{N_{\mathrm{UE}}}. \tag{5}$$

where N_{p} and N_{a} denote the total number of preambles and the number of active users, respectively.

For the data part, let d_u and $\boldsymbol{s}_u$ denote a complex-valued transmit symbol and the NOMA spreading sequence of the user u, respectively. We can write the received sample after de-spreading as

$$\begin{aligned} y_u^{\mathrm{d}} &= h_u d_u \boldsymbol{s}_u^{\mathrm{H}} \boldsymbol{s}_u + \sum_{m \in \mathcal{K}, m \neq u} h_m d_m \boldsymbol{s}_u^{\mathrm{H}} \boldsymbol{s}_m + n_{\mathrm{d}} \\ &= \hat{h}_u d_u + \Delta h_u d_u + \sum_{m \in \mathcal{K}, m \neq u} h_m d_m \boldsymbol{s}_u^{\mathrm{H}} \boldsymbol{s}_m + n_{\mathrm{d}}, \end{aligned} \tag{6}$$

where, $n_{\mathrm{d}} \sim \mathcal{CN}(0, \sigma_{\mathrm{d}}^2)$ is the AWGN affecting the data part.

Next, we introduce the GF-NOMA HARQ schemes covered in this paper.

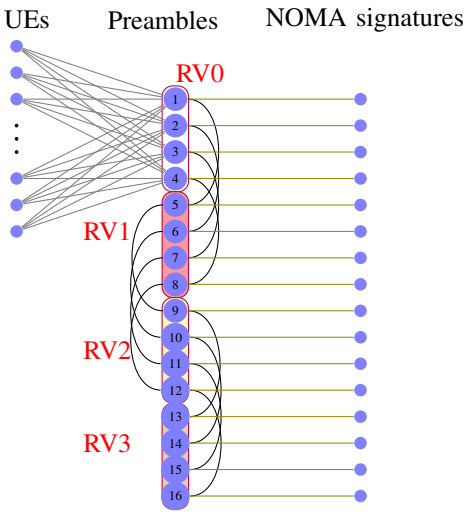
III. GF-NOMA HARQ Schemes

We propose and analyze three new UL transmission methods for GF-NOMA in this work, which are described in the following. We will compare these methods with four existing schemes as benchmarks, which we explain first.

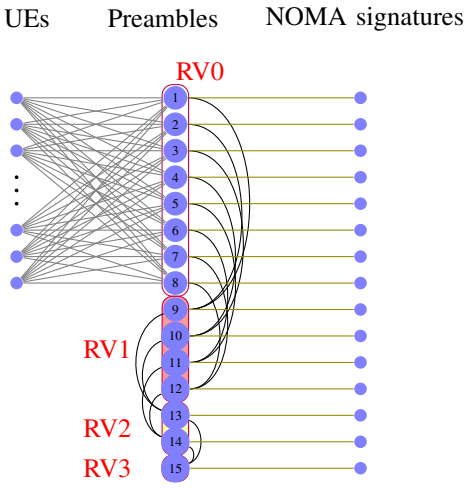
- Conventional GB IR-HARQ: This corresponds to the HARQ Type III method followed in the current 4G standards (LTE/NB-IoT/eMTC), where each re-transmission corresponds to a different RV and the gNB employs the IR combining technique to decode the packets.
- Two-step RA: This corresponds to the GF transmission method proposed in 3GPP Release 16 for 5G new radio (NR) [5], where data transmission can be done in a contention-based manner through the two-step RA procedure.
- GF-NOMA HARQ Type I (GF HARQ-I): This corresponds to the state-of-the-art method used in GF uplink [9], [28]. In this method, users transmit a preamble followed by data transmission and wait a certain amount of time for an ACK message in DL. In case a UE does not receive an ACK, it re-transmits the same preamble and the same RV. The gNB decodes each transmission independently and does not combine the packets before decoding (HARQ Type-I).
- Randomized GF HARQ-I: This corresponds to a modified version of the GF HARQ-I in which each user is authorized to change its preamble during the HARQ process to prevent consecutive preamble collisions.
- GF-NOMA single transmission (GFST): In this first proposed method, each user transmits all RVs in one shot. The gNB combines the RVs using IR combining in order to decode the data. Essentially, this method does not have ARQ and the IoT devices need not keep track of the ACK message for every RV, thereby resulting in a low-complexity GF transmission mechanism.[4]
- GF-NOMA HARQ Type III with chase combining (GF CC-HARQ): In our second proposed method, the transmission (and re-transmission) procedure is similar to that of GF HARQ-I, but the received packets get combined at the gNB using MRC. This process continues until either the gNB decodes the combined packet or the number of re-transmissions surpasses the maximum number of attempts allowed.
- GF-NOMA HARQ Type III with incremental redundancy (GF IR-HARQ): The third proposed method is similar to GF CC-HARQ, but each re-transmission from a user corresponds to a different RV. The gNB combines multiple RVs in the bit domain using IR combining and tries to decode the packet at a lower code rate, in order to improve the probability of successful decoding.

In a centralized network, the base station is in charge of allocating the physical and virtual resources, such as physical resource blocks (PRBs), NOMA signatures, RV indices of the HARQ, etc. However, in mMTC scenarios, a centralized structure with one management entity is infeasible owing to the large overhead, increased UE power consumption and large packet delay. To this end, we introduce a de-centralized system for the proposed GF-NOMA HARQ schemes, where UEs are free to choose their resources (preamble, NOMA sequence, RV index) based on predefined mappings.

[4]We note that the error-rate performance of GFST is similar to that of the conventional GF-NOMA scheme without HARQ, which transmits only one RV (say RV-0), if the coding rate of RV-0 is chosen to be the same as the net coding rate of the four RVs in the proposed GFST.



(a) Equal-size preamble groups



(b) Variable-size preamble groups

Fig. 4: Illustration of preamble to preamble and preamble to NOMA signature mappings for the case of $n^{\text{HARQ}} = 4$ RVs.

A. GF-NOMA HARQ Process Design

Let us consider the GF IR-HARQ scheme with a maximum of n^{HARQ} transmissions. We construct an HARQ process as {RV-0's (preamble, NOMA sequence), RV-1's (preamble, NOMA sequence),..., RV-($n^{\text{HARQ}} - 1$)'s (preamble, NOMA sequence)}. Basically, the HARQ process determines the preamble and the associated NOMA sequence that the UE must use if it has to re-transmit. For example, consider a GF IR-HARQ process for $n^{\text{HARQ}} = 4$, denoted by $\{(P_1, S_1), (P_5, S_5), (P_9, S_9), (P_{13}, S_{13})\}$. This means that if the UE uses (P_1, S_1) for its initial transmission, it is bound to use (P_5, S_5) for the first re-transmission, (P_9, S_9) for the next re-transmission and (P_{13}, S_{13}) for the last re-transmission. Note that the RV index can be determined by identifying the preamble. Such a process can be designed by partitioning the pool of preambles and the associated NOMA sequences across different RVs. We propose two types of HARQ processes for GF IR-HARQ based on the number of preambles per RV —i) equal size preamble groups and ii) variable size preamble groups.

1) Equal-size preamble groups: In this case, each group contains the same number of preambles. As illustrated in Fig. 4a, each preamble (indexed by $\{1, 2, \cdots, 16\}$) is uniquely mapped to a NOMA spreading sequence. Considering four RVs, the sixteen preambles are partitioned into groups of four preambles, with each group indicating a different RV index. For example, the preamble indices $\{1, 2, 3, 4\}$ correspond to RV-0. Also, each (preamble, NOMA-sequence) pair in a group is mapped to exactly one (preamble, NOMA-sequence) pair in the next group. The set of ordered (preamble, NOMA-sequence) pairs that are connected through this mapping form an HARQ process. For example, the set $\{1, 5, 9, 13\}$ is associated with one HARQ process.

It is intuitive that evenly distributing preambles between RV indices is not optimal. The reason is that all UEs use the preambles of the first group (RV-0) for their new packet transmissions, while the preambles of the RV-1, RV-2 and RV-3 groups are used only in cases of re-transmissions. Since the number of packets in re-transmission stages is less than the initial stage, equal-size groups cause congestion for earlier stages and under-utilization for later stages of the HARQ process.

2) Variable-size preamble groups: Here, we propose to unequally divide the set of preambles, such that the number of preambles assigned to the earlier stages are more than those assigned to the later stages. To this end, we define HARQ processes with many-to-one mappings between the (preamble, NOMA-sequence) pairs of consecutive groups. Fig. 4b illustrates the structure of this process. This design distributes the collisions between all the re-transmission stages, thereby alleviating the high collision rate in the early stages. In Section VII-B, we show that the overall successful packet reception rate (which strongly depends on the collision rate) can be improved by carefully choosing the size of the preamble groups.

We conclude this section by noting the preamble partitioning is required only for GF IR-HARQ, since it employs multiple RVs. The GF HARQ-I and the GF CC-HARQ schemes use one RV and grouping of preambles per RV is not necessary. Therefore, the HARQ process for these schemes contains one set of (preamble, NOMA sequence) pairs only. The same holds for the GFST scheme, since all the RVs are transmitted in a single shot.

In the following, we present the analytical models for the aforementioned UL transmission schemes.

IV. Analytical Models for the GB HARQ, Two-step RA, and the GF-NOMA HARQ Schemes

In this section, we describe the probability-based models used to analyze the GB HARQ (Section IV-A), two-step RA (Section IV-B), and the GF-NOMA HARQ (Section IV-C) schemes. For the modelling, it is useful to define the following

random events for any active user at the r^{th} re-transmission stage of the GB/GF-NOMA HARQ process:

$$\begin{aligned} E_1^r &= \{\text{no preamble collision at stage } r\},\\ E_2^r &= \{\text{successful preamble detection at stage } r\},\\ E_3^{r,k} &= \{\text{successful data decoding after combining}\\ &\qquad k \text{ packets at stage } r\}, \end{aligned}$$

where $1 \leq k \leq r$ and $1 \leq r \leq n^{\text{HARQ}}$, and $r = 1$ is the first packet transmission. Accordingly, we can express the probabilities of successful and failed packet reception, respectively, after combining k packets for the r^{th} stage and the probability of missed-detection of the preamble as

$$P_{r,k}^{\text{s}} = \Pr(E_1^r, E_2^r, E_3^{r,k}), \tag{7}$$

$$P_{r,k}^{\text{f}} = \Pr(E_1^r, E_2^r, \tilde{E}_3^{r,k}), \tag{8}$$

$$P_r^{\text{m}} = \Pr(E_1^r, \tilde{E}_2^r), \tag{9}$$

respectively.

A. Modelling the Conventional GB IR-HARQ Scheme

As mentioned in Section I, the GB transmission method is a two-stage routine, applying four-step RA in the first stage, followed by data transmission in the second stage. Note that a collision occurs at the RA stage when two or more users pick the same preamble. In the case of a collision, each user backs off for a random duration of time before attempting the next RA. If there is no collision, the gNB detects the preambles transmitted by the users. Upon successful preamble detection, it grants the time-frequency resources for uplink data transmission. The users then transmit their data through a HARQ process, which is decoded by the gNB.

For modelling the GB scheme, we assume an ideal four-step RA procedure, which means that the RA procedure is collision-free and all the back-and-forth messages of this procedure are delivered free of error. Hence, one RA attempt is enough for attaching UEs to the network. Then, $k = r$ at the r^{th} HARQ stage and the probabilities in (7)-(9) simplify to

$$P_{r,r}^{\text{s}} = \Pr(E_3^{r,r}), \tag{10}$$

$$P_{r,r}^{\text{f}} = \Pr(\tilde{E}_3^{r,r}) = 1 - \Pr(E_3^{r,r}), \tag{11}$$

$$P_r^{\text{m}} = 0. \tag{12}$$

Hence, we only need to consider the event $E_3^{r,r}$ to model the HARQ process. This idealistic assumption for RA means that gains attained by the proposed GF methods against the GB scheme are conservative estimates.

Furthermore, in line with mMTC scenarios, we consider that the number of available orthogonal resources in the system is always less than the total number of active users at any given time. Thus, only a part of the access requests can be granted, and we denote the probability of receiving an access grant by P_{AG}. Then, using (10) and (11), the probability of successful packet reception in the r^{th} stage is given by

$$\Omega_r = P_{\text{AG}} \prod_{k=1}^{r-1} P_{k,k}^{\text{f}} P_{r,r}^{\text{s}}, \tag{13}$$

and the probability of successful decoding within a maximum of n^{HARQ} HARQ transmissions can be calculated as

$$\Psi_{n^{\text{HARQ}}} = \sum_{r=1}^{n^{\text{HARQ}}} \Omega_r. \tag{14}$$

B. Modelling the Two-step RA Scheme

Assume that the number of PRACH preambles mapped to a PRB is N_{PRACH} and that the probability of no data (payload) collision is P_{NDC}. There are three possible outcomes.

1) The base station successfully detects the preamble upon reception of MSGA and decodes the payload accompanying it through the physical uplink shared channel (PUSCH). In this case, the base station transmits a success message via RAR to the UE containing the contention resolution ID of MSGA. If the success message is detected free of error, the two-step RACH procedure ends.
2) The base station detects MSGA preamble successfully, but the data on PUSCH is not decoded. In this situation, the base station transmits a fallback message via RAR to the UE with the random-access preamble ID and an uplink grant for the MSGA PUSCH re-transmission. Upon receiving the RAR, the UE resorts to the four-step RACH, starting with MSG3 transmission (re-transmission of the MSGA PUSCH).
3) The base station cannot detect the RACH preamble and hence does not send any feedback. The UE resorts to the conventional four-step RA procedure.

Similar to our analysis of the conventional four-step RA procedure, for simplicity, we assume ideal collision-free transmission and ideal detection for PRACH preambles. Therefore, Scenario 3 is not considered. The probability of occurrence of Scenario 1 is

$$\Omega_1 = P_{\text{NDC}} \Pr(E_3^{1,1}), \tag{15}$$

where $P_{\text{NDC}} \triangleq \Pr(\text{no data collision})$. The probability that Scenario 2 occurs is $1 - \Omega_1$, in which case the UE re-transmits MSGA in a grant-based manner. Therefore, ideally there should be no data collision. However, since the number of available PRBs is limited, the re-transmission of MSGA cannot be granted for all users in Scenario 2. Let P_{AG} denote the probability that gNB grants a MSGA re-transmission. Thus, assuming that the re-transmitted MSGA can be IR-combined with the initial MSGA, the probability of successful reception in Scenario 2 is

$$\Omega_2 = (1 - \Omega_1) P_{\text{AG}} \Pr(E_3^{2,2}), \tag{16}$$

and the overall probability of successful packet reception is $\Psi = \Omega_1 + \Omega_2$.

C. Modelling the GF-NOMA HARQ Methods

In GF scenarios, the gNB combines packets based on the detected preamble IDs. Consider a scenario where UE-1 has picked preamble-1 and is transmitting a packet. Another user, UE-2 picks the same preamble and starts transmitting its packet before UE-1's HARQ process has been completed.

$$\Pr(S_{r,t,i,j}) = \begin{cases} \Pr(S_{r-1,t-1,i,j})P_r^{\mathrm{c}} + \Pr(S_{r-1,t,i-1,j})P_r^{\mathrm{m}} + \Pr(S_{r-1,t,i,j-1})P_{r,j}^{\mathrm{f}} & t+i+j=r \\ \Pr(S_{r-1,t,i,j})P_{r,j+1}^{\mathrm{s}} & t+i+j=r-1 \\ 0 & \text{else} \end{cases} \tag{17}$$

Then, the gNB may combine UE-1's packet with UE-2's packet, since they had the same preamble ID, resulting in a decoding failure. This is termed as packet combining confusion, which occurs due to the random delay between re-transmissions. To avoid such scenarios, the delay between re-transmissions is set to a fixed value. This way, the gNB does not combine any two packets unless the time difference between them is a certain value. In LTE, UL re-transmissions are 8 ms apart. In [9], although the timing of re-transmissions is random, there is a fixed 4 ms waiting time before starting the random re-transmission process and on average, the time difference between re-transmissions is 8 ms. This motivates us to set a fixed delay of 8 ms between UL re-transmissions in all proposed methods.

Unlike the GB IR-HARQ and two-step RA schemes with idealistic RA, we consider the possibility of preamble collision and preamble-detection failure in our analysis of the GF NOMA HARQ schemes. To model the HARQ process, we define the state variables $S_{r,t,i,j}$ corresponding to the state of the r^{th} stage of the HARQ (i.e., after r transmissions), with t, i, j, $0 \leq t, i, j \leq r$, being the number of preamble collisions, missed preambles and failed decodes, respectively. For example, $S_{2,0,0,2}$ refers to the second stage of the HARQ process, where the preambles were successfully detected in both the first and second transmissions, but data was not decoded in either case. The probabilities associated with the states follow the recursion given in (17), where $1 \leq r < n^{\mathrm{HARQ}}$ and $\Pr(S_{0,0,0,0}) = 1$.

The probability of successful packet reception at the r^{th} stage can now be computed by summing the probabilities of all states that satisfy $t+i+j = r-1$ (packet reception is successful only when $t+i+j=r-1$) as follows:

$$\Omega_r = \sum_{t=0}^{r} \sum_{i=0}^{r} \Pr(S_{r,t,i,r-t-i-1}), \tag{18}$$

and the probability of successful packet reception after n^{HARQ} attempts is achieved from (14).

1) GF HARQ-I: As mentioned in Section II, in this method, UEs re-transmit the same preamble and RV for each re-transmission and the gNB does not employ any combining techniques. Thus, preamble collision persists once it happens and leads to packet drop[5], i.e., (17) is zero for all $t > 0$. Therefore, this method is modelled by setting $t = 0$ in (18) and $k = 1$ while computing $P_{r,k}^{\mathrm{s}}$ and $P_{r,k}^{\mathrm{f}}$.

2) Randomized GF HARQ-I: In this method, since UEs are allowed to re-transmit each time with a different preamble, collisions do not necessarily repeat. Thus, (17) has non-zero values for $t \geq 0$. However, similar to the GF HARQ-I method, the receiver cannot combine the received packets. Hence, this method is modelled by setting $k = 1$ while computing $P_{r,k}^{\mathrm{s}}$ and $P_{r,k}^{\mathrm{f}}$.

3) GF IR-HARQ and GF CC-HARQ: In these two methods, similar to the GF HARQ-I, preamble collision leads to a packet drop. However, since receiver can combine packets upon reception, k varies from 0 to r. Therefore, to model these methods, we have to set $t = 0$ in (18).

4) GFST: In this method, since all the RVs are transmitted in one shot, we have $t = 0$, $r = 1$ and $k = n^{\mathrm{HARQ}}$. Note that the TTI of this method is n^{HARQ} times larger than that of the previous methods. Consequently, the number of arrivals in GFST method is $n^{\mathrm{HARQ}}\epsilon$ in one TTI.

V. Analysis of the Probability of Successful Packet Reception

The probability of successful packet reception is the primary metric for evaluation of the system under consideration. Other metrics such as power efficiency, delay, and overhead directly depend on this metric. Therefore, we first derive this metric for each UL transmission method and then derive other metrics based on it in Section VI.

As probabilities of preamble collision and successful data decoding for one user depend on the simultaneous use of resources by other users, the evaluation of probabilities (7)-(9) is facilitated by the conditioning of events on the number of active users at the n^{HARQ} stages of the HARQ process denoted by $\boldsymbol{N}_{\mathrm{a}} = [N_{\mathrm{a}}^1, N_{\mathrm{a}}^2, ..., N_{\mathrm{a}}^{n^{\mathrm{HARQ}}}]$. In particular, we can write

$$\Pr(E_1^r, E_2^r, E_3^{r,k}) = \sum_{\boldsymbol{N}_{\mathrm{a}}} \Pr(\boldsymbol{N}_{\mathrm{a}}) \Pr(E_1^r, E_2^r, E_3^{r,k} | \boldsymbol{N}_{\mathrm{a}}). \tag{19}$$

Since the numbers of active user at different stage r, N_{a}^r, is independent of N_{a}^k, $1 \leq k \leq r-2$, given N_{a}^{r-1}, we can write

$$\Pr(\boldsymbol{N}_{\mathrm{a}}) = \Pr(N_{\mathrm{a}}^1) \prod_{r=2}^{n^{\mathrm{HARQ}}} \Pr(N_{\mathrm{a}}^r | N_{\mathrm{a}}^{r-1}). \tag{20}$$

For all the UL transmission methods considered in this paper we can show that

$$\Pr(N_{\mathrm{a}}^1) = \mathrm{Binomial}(N_{\mathrm{a}}^1, N_{\mathrm{UE}}, \epsilon), \tag{21}$$

$$\Pr(N_{\mathrm{a}}^r | N_{\mathrm{a}}^{r-1}) = \mathrm{Binomial}(N_{\mathrm{a}}^r, N_{\mathrm{UE}}, \epsilon_r), \tag{22}$$

where

$$\epsilon_r = (1 - \Omega_{r-1}) \frac{N_{\mathrm{a}}^{r-1}}{N_{\mathrm{UE}}} \tag{23}$$

and $2 \leq r \leq n^{\mathrm{HARQ}}$. The proof of (21) and (22) is relegated to the Appendix. Moreover, for typical parameters considered in mMTC scenarios we observe that (21) and (22) are highly concentrated around their expected values especially when the

[5]When two or more users select the same preamble, the base-station may be able to detect the preamble. But the associated channel estimate is likely far from any of the individual user channels, and thus the probability of decoding failure is high. Therefore, we assume that a preamble collision always results in data decoding failure. This assumption is also consistent with that used in [9, Section V.A]. For further details, please refer to Appendix B.

number of active users is large. The expected values can be calculated as

$$\bar{N}_{\mathrm{a}}^{r} = N_{\mathrm{UE}}\epsilon \prod_{j=1}^{r-1}(1-\Omega_j), \quad 1 \le r \le n^{\mathrm{HARQ}}. \tag{24}$$

Exploiting this concentration property for the ease of analysis[6], (19) simplifies to

$$\Pr(E_1^r, E_2^r, E_3^{r,k}) \approx \Pr(E_1^r, E_2^r, E_3^{r,k} | \bar{\boldsymbol{N}}_{\mathrm{a}}), \tag{25}$$

where $\bar{\boldsymbol{N}}_{\mathrm{a}} = [\bar{N}_{\mathrm{a}}^1, \bar{N}_{\mathrm{a}}^2, \ldots, \bar{N}_{\mathrm{a}}^{n^{\mathrm{HARQ}}}]$. Likewise, we approximate the total number of active users $N_{\mathrm{a}} = \sum_{r=1}^{n^{\mathrm{HARQ}}} N_{\mathrm{a}}^r$ by

$$\bar{N}_{\mathrm{a}} = \sum_{r=1}^{n^{\mathrm{HARQ}}} \bar{N}_{\mathrm{a}}^r = N_{\mathrm{UE}}\epsilon \sum_{r=1}^{n^{\mathrm{HARQ}}} \prod_{j=1}^{r-1}(1-\Omega_j). \tag{26}$$

Writing

$$\begin{aligned}\Pr(E_1^r, E_2^r, E_3^{r,k} | \bar{\boldsymbol{N}}_{\mathrm{a}}) = &\underbrace{\Pr(E_1^r | \bar{\boldsymbol{N}}_{\mathrm{a}})}_{\triangleq P_r^{E_1}} \\ &\underbrace{\Pr(E_2^r | E_1^r, \bar{\boldsymbol{N}}_{\mathrm{a}})}_{\triangleq P_r^{E_2}} \\ &\underbrace{\Pr(E_3^{r,k} | E_1^r, E_2^r, \bar{\boldsymbol{N}}_{\mathrm{a}})}_{\triangleq P_{r,k}^{E_3}}\end{aligned}$$

we can express (7)-(9) as

$$P_{r,k}^{\mathrm{s}} = P_r^{E_1} P_r^{E_2} P_{r,k}^{E_3}, \tag{27}$$

$$P_{r,k}^{\mathrm{f}} = P_r^{E_1} P_r^{E_2} (1 - P_{r,k}^{E_3}), \tag{28}$$

$$P_r^{\mathrm{m}} = P_r^{E_1} (1 - P_r^{E_2}). \tag{29}$$

Next, we derive expressions for $P_r^{E_1}$, $P_r^{E_2}$, and $P_{r,k}^{E_3}$. We note that for GB IR-HARQ, $P_r^{E_1} = 1$ and $P_{r,k}^{E_3}$ can be calculated using $k = r$. Also, for GFST, we can calculate the probabilities by setting $r = 1$ and $k = n^{\mathrm{HARQ}}$.

A. Conventional GB IR-HARQ

For the GB scheme, we require expressions for the probability of access P_{AG} and the decoding probability $P_{r,r}^{E_3}$. For the former, we note that in the mMTC scenario we can assume that at any given time, the number of active users ($\bar{N}_{\mathrm{a}}$) exceeds the number of available orthogonal resources (N_{PRB}) in the system. Hence, the probability of an access request being granted is given by

$$P_{\mathrm{AG}} = \frac{N_{\mathrm{PRB}}}{\bar{N}_{\mathrm{a}}}. \tag{30}$$

where $\bar{N}_{\mathrm{a}}$ depends on the probabilities of successful packet reception Ω_r as indicated by (26).

Denoting the code rate and the length of each RV after encoding by R and L, respectively, the decoding probability at the r^{th} HARQ stage is computed using the total number of rL transmitted coded bits and the effective code rate of R/r. Then, applying the results from [35] for coding with finite block-length, the probability of successfully decoding a data packet given a channel realization $\hat{h}_u$ is

$$\Pr(E_3^{r,r} | \hat{h}_u) = 1 - Q\left(\sqrt{\frac{rL}{V}}\left(\log_2(1+\hat{h}_u\gamma^{\mathrm{d}}) - \frac{R}{r}\right)\right), \tag{31}$$

where $\gamma^{\mathrm{d}} = \frac{\rho_{\mathrm{d}}}{\sigma_{\mathrm{d}}^2}$ is the SNR of the data, Q denotes the Gaussian Q-function, and $V = (\log_2 \mathrm{e})^2 \left(1 - \frac{1}{(1+\gamma_0^{\mathrm{d}})^2}\right)$. The marginalization over $\hat{h}_u$ to obtain $P_{r,r}^{E_3}$ needs to be done numerically, where we note that only (31) depends on the absolute value of $\hat{h}_u$.

Using the result for $P_{r,r}^{E_3}$ in (10), (11) and (13) we obtain the probability of successful packet reception $\Psi_{n^{\mathrm{HARQ}}}$ using (14). Note that Ω_r in (13) depends on P_{AG} and hence on $\bar{N}_{\mathrm{a}}$. Since $\bar{N}_{\mathrm{a}}$ itself depends on Ω_r, we adopt the numerical method introduced in [34] to solve (26) recursively using

$$\bar{N}_{\mathrm{a}}^{(i+1)} = N_{\mathrm{UE}}\epsilon \sum_{r=1}^{n^{\mathrm{HARQ}}} \prod_{j=1}^{r-1}(1-\Omega_j(\bar{N}_{\mathrm{a}}^{(i)})), \tag{32}$$

starting with $\bar{N}_{\mathrm{a}}^{(1)} = N_{\mathrm{UE}}\epsilon$.

B. Two-step RA

In this method, the probability of the MSGA re-transmission being granted is obtained from (30). To calculate P_{NDC}, note that since $\bar{N}_{\mathrm{a}} \gg N_{\mathrm{PRB}}$, the two-step RACH method must support $\bar{N}_{\mathrm{a}}$-to-N_{PRB} RACH preamble to PRB mapping. Each group of preambles that is mapped to the same PRB has $\frac{\bar{N}_{\mathrm{a}}}{N_{\mathrm{PRB}}}$ members. Any two users who pick their preambles from the same group will experience a data collision. The probability of no data collision for a user can be calculated as

$$P_{\mathrm{NDC}} = \left(1 - \frac{N_{\mathrm{PRB}}}{\bar{N}_{\mathrm{a}}}\right)^{\bar{N}_{\mathrm{a}}-1}. \tag{33}$$

C. GF-NOMA HARQ

For the GF-NOMA cases, we need to evaluate the probability for preamble collision and consider the presence of multiuser interference during preamble detection and data decoding.

1) GF CC-HARQ: At each stage of GF CC-HARQ, $\bar{N}_{\mathrm{a}}$ users contribute to collisions and interference. Hence, the probability of no collision is

$$P_r^{E_1} = \left(1 - \frac{1}{N_{\mathrm{p}}}\right)^{\bar{N}_{\mathrm{a}}-1}. \tag{34}$$

The probability of successful preamble detection if there was no collision can be obtained, using the result from [7, Eq. 34] for MMSE-based AMP detection, as

$$P_r^{E_2} = 1 - \frac{\underline{\Gamma}\left(1, \tau_{t\to\infty}^2 \log_2(1 + \frac{1}{\tau_{t\to\infty}^2})\right)}{\Gamma(1)}, \tag{35}$$

where Γ and $\underline{\Gamma}$ denote the Gamma function and the lower incomplete Gamma function, respectively. Also, (35) depends

[6]In Section VII, for massive connectivity scenarios, we evaluate the probability of successful packet reception without this approximation through Monte-Carlo simulations, and show that it aligns well with those obtained using this approximation.

on $N_{\rm p}$ through (4) and (5). We note that the expressions (34) and (35) are independent of the HARQ transmission stage r.

To evaluate the probability of successful decoding for GF-NOMA, we need to consider the signal-to-interference-plus-noise ratio (SINR) of the $u^{\rm th}$ UE obtained after combining k packets. Let $\mathcal{K}^i$ denote the set of active UEs excluding the $u^{\rm th}$ UE in the $i^{\rm th}$ HARQ stage. Conditioned on the fading realizations for all active users during the transmission of k RVs from UE u, i.e., $\boldsymbol{h}_u = [h_1, \ldots, h_{u-1}, \hat{h}_u, h_{u+1}, \ldots, h_{|\bigcup_{i=1}^{k}\mathcal{K}^i|}]$, the probability of decoding success for MRC is

$$\Pr(E_3^{r,k}|\boldsymbol{h}_u) = 1 - Q\left(\sqrt{\frac{L}{V}}\left(\log_2(1+\gamma_u^{\rm MRC}) - R\right)\right), \quad (36)$$

where

$$\gamma_u^{\rm MRC} = \frac{k|\hat{h}_u|^2}{\frac{1}{\gamma^{\rm d}} + \frac{1}{k}\sum_{i=1}^{k}\sum_{n\in\mathcal{K}^i, n\neq u}|h_n|^2|\boldsymbol{s}_u^{\rm H}\boldsymbol{s}_n|^2 + \sigma_{\rm E}^2}. \quad (37)$$

is the SINR for user u. We note that (36) only depends on the number of combined packets (k) and not the number of transmissions (r). The marginalization with respect to $\boldsymbol{h}_u$ needs to be performed numerically using the Monte-Carlo method, so that we obtain $P_{r,k}^{E_3}$. We found empirically that on the order of 10,000 realizations are sufficient for stable results.

Substituting the results for $P_r^{E_1}$, $P_r^{E_2}$ and $P_{r,k}^{E_3}$ into (27)-(29), we can evaluate Ω_r (18)[7] and finally $\Psi_{n^{\rm HARQ}}$ (14). Similar as for the GB case, Ω_r and $\bar{N}_{\rm a}$ are inter-dependent and we need to recursively solve for $\bar{N}_{\rm a}$ using the numerical method from [34].

2) GF HARQ-I: This method is a special case of GF CC-HARQ setting $k = 1$ in (36).

3) Randomized GF HARQ-I: This method can be modeled similar to the GF CC-HARQ method. However, t is not zero in (18) and we have to set $k = 1$ in (36).

4) GF IR-HARQ: The analysis for GF IR-HARQ is somewhat more complicated because of the preamble partitioning schemes introduced in Section III-A. At each HARQ transmission stage, $\bar{N}_{\rm a}$ users contribute to interference but only $\bar{N}_{\rm a}^r$ compete for preambles and thus contribute to collisions at stage r. Furthermore, the number of preambles available is a function of r, i.e., $N_{\rm p}^r$. Accordingly, the probability of no collision is

$$P_r^{E_1} = \left(1 - \frac{1}{N_{\rm p}^r}\right)^{\bar{N}_{\rm a}^r - 1}. \quad (38)$$

Furthermore, we can reuse expression for $P_r^{E_2}$ (35) from the GF CC-HARQ case, with using $N_{\rm p}^r$ in (5). The probability of successful decoding after IR combining of k packets and conditioned on the fading realizations for all active users is

$$\Pr(E_3^{r,k}|\boldsymbol{h}_u) = 1 - Q\left(\sqrt{\frac{kL}{V}}\left(\log_2(1+\gamma_u) - \frac{R}{k}\right)\right), \quad (39)$$

where the SINR is calculated as

$$\gamma_u = \frac{|\hat{h}_u|^2}{\frac{1}{\gamma^{\rm d}} + \sum_{n\in\mathcal{K}^r, n\neq u}|h_n|^2|\boldsymbol{s}_u^{\rm H}\boldsymbol{s}_n|^2 + \sigma_E^2}. \quad (40)$$

[7]Note that $t = 0$ in (18).

TABLE II: Comparison of UE energy consumption.

Method	UE energy consumption
GB IR-HARQ	$W = \Lambda_{\rm RA} + (\Lambda_{\rm DCI} + \Lambda_{\rm data} + \Lambda_{\rm ACK})\left(\sum_{r=1}^{n^{\rm HARQ}} r\Omega_r + (1-\Psi_{n^{\rm HARQ}})n^{\rm HARQ}\right)$
Two-step RA	$W = (\Lambda_{\rm MSGA} + \Lambda_{\rm MSGB})\Omega_1 + (\Lambda_{\rm MSGA} + \Lambda_{\rm MSGB} + \Lambda_{\rm MSG3} + \Lambda_{\rm data} + \Lambda_{\rm MSG4})\Omega_2 + (\Lambda_{\rm MSGA} + \Lambda_{\rm MSGB} + \Lambda_{\rm MSG3} + \Lambda_{\rm payload})(1-\Psi)$,
GF NOMA with HARQ[8]	$W = (\Lambda_{\rm preamble} + \Lambda_{\rm data} + \Lambda_{\rm ACK})\left(\sum_{r=1}^{n^{\rm HARQ}} r\Omega_r + (1-\Psi_{n^{\rm HARQ}})n^{\rm HARQ}\right)$
GFST	$W = \Lambda_{\rm preamble} + n^{\rm HARQ}\Lambda_{\rm data} + \Lambda_{\rm ACK}$

The remaining steps are same as for GF CC-HARQ, including the recursive method to solve for $\bar{N}_{\rm a}^r$, $1 \leq r \leq n^{\rm HARQ}$.

5) GFST: GFST can be analyzed as a special case of GF IR-HARQ with $t = 0$, $r = 1$, $k = n^{\rm HARQ}$, packet arrival rate equal to $n^{\rm HARQ}\epsilon$ and $N_{\rm p}^1 = N_{\rm p}$.

VI. Analysis of UE Energy Efficiency, DL Overhead and Packet Delay

In this section, we use the statistical analysis of the probability of successful packet reception developed in Section V to derive other important KPIs, such as, energy efficiency, packet delay, and DL overhead.

A. UE Energy Efficiency

As mentioned in Section I, UE energy efficiency is a critical metric for mMTC scenarios. Since we model the UL scenario, a major portion of the energy is consumed for transmitting uplink data. In order to assist uplink data transmission, the UE typically needs to decode the relevant downlink control information (DCI) and process the ACK (or ACK/NACK) messages received in the downlink. Let $\Lambda_{\rm q}$ denote the energy consumed by the UE for executing the operation q.

For the GB transmission scheme, one must account for the energy consumed by the UE for the four-step RA process (see Fig. 1) in addition to that consumed by the DCI decoding and ACK processing. Based on the ideal RA process assumption, the amount of energy consumed in this process is

$$\Lambda_{\rm RA} = \Lambda_{\rm PRACH} + \Lambda_{\rm RAR} + \Lambda_{\rm MSG3} + \Lambda_{\rm MSG4}. \quad (41)$$

where the PRACH involves the preamble transmission.

For the two-step RA method, similar to the four-step RA scheme, one needs to account for the energy consumption for MSGA transmission ($\Lambda_{\rm MSGA} = \Lambda_{\rm PRACH} + \Lambda_{\rm data} + \Lambda_{\rm MSG3}$) and MSGB reception ($\Lambda_{\rm MSGB} = \Lambda_{\rm RAR} + \Lambda_{\rm MSG4}$) in addition to the energy consumption for payload transmission.

The proposed GF-NOMA HARQ methods only require preamble transmission, data transmission and ACK processing. The DCI decoding is not required since the access mechanism is grant-free. The expected values of the energy consumption (denoted as W) corresponding to the different UL transmission methods presented in this paper are summarized in Table II.

[8]GF NOMA with HARQ represents GF HARQ-I, GF randomized HARQ-I, GF CC-HARQ, and GF IR-HARQ.

Further, we define the energy efficiency of each method as the ratio of the number of successfully delivered bits to the energy consumed, given by

$$\eta = \frac{\Psi_{n^{\mathrm{HARQ}}} L_{\mathrm{uncoded}}}{W} \quad [\mathrm{bits/Joule}], \tag{42}$$

where L_{uncoded} is the number of bits of data before coding and $\Psi_{n^{\mathrm{HARQ}}}$ can be obtained using the analysis in Sections IV and V.

B. Packet Delay

We define the packet delay as the time difference between the instant of successful packet reception by the gNB and the first transmission with any missed packet being ignored. The expected value of the packet delay can be expressed as

$$\tau = \tau_{\mathrm{RA}} + n \cdot \tau_{\mathrm{TTI}} + w \frac{\sum_{r=1}^{n^{\mathrm{HARQ}}} r\Omega_r}{\Psi_{n^{\mathrm{HARQ}}}}, \tag{43}$$

where τ_{RA} is the delay due to the four-step RA process, τ_{TTI} is the duration of the TTI, n is the number of RVs in each transmission and w is the fixed delay between re-transmissions. Evidently, $\tau_{\mathrm{RA}} = 0$ for all GF-NOMA methods. Also, $n = 1$ and $w = 8$ ms for all the methods except the two-step RA and GFST schemes. For the two-step RA we have $n = 0$ and $w = 0$ and for the GFST we have $n = 4$ and $w = 0$, since four RVs are transmitted in a single shot.

In order to determine the delay caused by the RA process, we assume that the gNB handles excessive access requests by running a queue on a first-come, first-served basis. The length of the queue is $\bar{N}_{\mathrm{a}}$ and the number of available resources is N_{PRB}. Therefore, it takes $\bar{N}_{\mathrm{a}}\tau_{\mathrm{TTI}}/N_{\mathrm{PRB}}$ longer for users to start transmitting than in the GF case. Considering an ideal RA process as per Fig. 1, we obtain

$$\tau_{\mathrm{RA}} = \begin{cases} \tau_{\mathrm{PRACH}} + \tau_{\mathrm{RAR}} + \tau_{\mathrm{MSG3}} + \tau_{\mathrm{A-RP}} + \\ \qquad \tau_{\mathrm{MSG4}} + \dfrac{\bar{N}_{\mathrm{a}}}{N_{\mathrm{PRB}}}\tau_{\mathrm{TTI}} & , 4-\mathrm{step} \\ \dfrac{\tau_{\mathrm{MSGA}}\Omega_1 + (2\tau_{\mathrm{MSGA}} + \tau_{\mathrm{MSGB}})\Omega_2}{\Psi} + \dfrac{\bar{N}_{\mathrm{a}}}{N_{\mathrm{PRB}}}\tau_{\mathrm{TTI}} & , 2-\mathrm{step} \end{cases} \tag{44}$$

where $\tau_{\mathrm{A-RP}}$ is the delay between the acknowledgment of MSG3 by the gNB and acceptance of the connection request of the UE in the four-step RA procedure and $\tau_{\mathrm{MSGA}} = \tau_{\mathrm{PRACH}} + \tau_{\mathrm{MSG3}}$ and $\tau_{\mathrm{MSGB}} = \tau_{\mathrm{RAR}} + \tau_{\mathrm{MSG4}}$.

C. DL Overhead

The DL overhead analysis assumes that the radio resource configuration (RRC) connection is already established. It includes only the physical and link layer aspects in the calculation *after* the uplink access is attempted. This is because decoding of messages conveying system parameters, such as, the system information block (SIB) happens in the downlink (prior to uplink access). Also, the pre-defined mappings introduced in Section III-A can be treated as system parameters and conveyed through SIB. Hence, conveying the mapping is not included as a part of the overhead calculation. Moreover, SIBs are not as frequent as Ack, DCI and other messages (PRACH transmission, RAR, MSG3, and MSG4) that are exchanged after the uplink access attempt.

The DL overhead for the conventional GB method comprises the resources consumed for transmitting the RAR and MSG3 as well as the DCI and ACK (or ACK/NACK) that are denoted by ν_{RAR}, ν_{MSG4}, ν_{DCI}, and ν_{ACK} respectively. For an ideal RA assumption as per Fig. 1, the expected value of the overhead of GB IR-HARQ method is

$$\begin{aligned} O_{\mathrm{GB}}^{\mathrm{DL}} = \nu_{\mathrm{RAR}} + \nu_{\mathrm{MSG4}} + (\nu_{\mathrm{DCI}} + \nu_{\mathrm{ACK}}) \sum_{r=1}^{n^{\mathrm{HARQ}}} r\Omega_r \\ + n^{\mathrm{HARQ}}(1 - \Psi_{n^{\mathrm{HARQ}}})\,(\nu_{\mathrm{ACK}} + \nu_{\mathrm{DCI}})\,, \end{aligned} \tag{45}$$

and for the two-step RA we have

$$O_{2-\mathrm{step}}^{\mathrm{DL}} = \nu_{\mathrm{MSGB}}\Omega_1 + (\nu_{\mathrm{MSGB}} + \nu_{\mathrm{MSG4}})\Omega_2 + \nu_{\mathrm{MSGB}}(1 - \Psi), \tag{46}$$

where, $\nu_{\mathrm{MSGB}} = \nu_{\mathrm{RAR}} + \nu_{\mathrm{MSG4}}$. As mentioned in Section II, NACK messages are not considered for the GF schemes and an ACK message is sent in the DL if the UL packet is decoded successfully. Hence, the DL overhead of the GF methods only consists of the resources consumed to transmit the ACK message. Therefore, we infer that

$$O_{\mathrm{GF}}^{\mathrm{DL}} \leq \nu_{\mathrm{ACK}}. \tag{47}$$

VII. Numerical Results

In this section, we present quantitative results to evaluate the performance of the proposed GF-NOMA HARQ methods. We consider a scenario with $N_{\mathrm{UE}} = 2000$ users distributed uniformly at random within a cell of radius 1 km. We adopt an LTE system with normal cyclic prefix (CP) with $N_{\mathrm{PRB}} = 6$, which corresponds to a system bandwidth (for the gNB and the UEs) of 1.4 MHz. A PRB pair, which is the minimum unit of resource allocation in LTE spans 12 sub-carriers times 14 symbols = 168 resource elements (REs). For GF transmission, we allocate one PRB for preamble transmission, i.e., $L_{\mathrm{p}} = 168$. The length of the ZC-based NOMA spreading sequence is set to $L_{\mathrm{s}} = 23$, which enables us to generate $23 \times 22 = 506$ unique NOMA spreading sequences. Therefore, we have 506 unique preambles owing to the one-to-one mapping between the NOMA sequences and the preambles. The maximum allowed number of transmissions is set to $n^{\mathrm{HARQ}} = 4$, which is in line with the number of RVs used in conventional GB schemes, such as, LTE/LTE-A, NB-IoT, and eMTC. Similarly, the transmit power is set to 23 dBm for both the preamble and the data signal. The AWGN power is set to -169 dBm and the propagation loss model $\beta_u = -128.1 - 36.7\log_{10}(d_u/1000)$ according to [8] is adopted, where d_u is the distance between user u and the gNB in meters. The variance of the Rayleigh fading channel is set to unity, i.e., $\sigma_{\mathrm{F}}^2 = 1$, and the variance of the channel estimation error given in Eq. (3) in this simulation setup is $\sigma_{\mathrm{E}}^2 = 0.07$.

[9]Each DCI contains 4 CCEs and supports 4 users, thus 1CCE per user [36].

[10]PHICH contains 12 REs and can carry one ACK or NACK [37].

[11]One RAR contains 4 CCE = 144 REs [4] and serves up to three users [38], thus, 48 REs per user.

[12]One MSG4 contains 8 CCEs = 288 REs [4].

TABLE III: Simulation parameters.

parameter	Description	value
ν_{DCI}	overhead of DCI[9]	36 REs
ν_{ACK}	overhead of Ack[10]	12 REs
ν_{RAR}	overhead of RAR[11]	48 REs
ν_{MSG4}	overhead of MSG4[12]	288 REs
W_{RE}	energy consumed for reception of one RE	
W_{RE}	energy consumed for transmission of one RE	
Λ_{DCI}	energy consumed for DCI reception	$36W_{\mathrm{RE}}$
Λ_{data}	energy consumed for payload transmission	$168W_{\mathrm{RE}}$
Λ_{ACK}	energy consumed for ACK reception	$12W_{\mathrm{RE}}$
$\Lambda_{\mathrm{preamble}}$	energy consumed for preamble transmission	$32W_{\mathrm{RE}}$
Λ_{RAR}	energy consumed for RAR reception	$48W_{\mathrm{RE}}$
Λ_{MSG3}	energy consumed for MSG3 transmission	$24W_{\mathrm{RE}}$
Λ_{MSG4}	energy consumed for MSG4 reception	$288W_{\mathrm{RE}}$
τ_{preamble}	processing time required by base station to detect the RACH preamble [38]	2 ms
τ_{RAR}	length of the window of RAR [38]	5 ms
$\tau_{\mathrm{MSG3}}^{\mathrm{HARQ}}$	waiting time for receiving MSG3 HARQ ACK [38]	4 ms
$\tau_{\mathrm{A-RP}}$	gap of monitor connection response message [38]	1 ms
$\tau_{\mathrm{MSG4}}^{\mathrm{HARQ}}$	waiting time for receiving MSG4 HARQ ACK [38]	4 ms
τ_{TTI}	TTI of GF method	4 ms
w	delay of re-transmission	8 ms

Moreover, we consider scenarios with a wide range of network loads with on average $N_{\mathrm{UE}}\epsilon \in \{20, 40, 60, 80, 100\}$ new transmissions [1], [7], and four different data sizes of 16, 40, 72, and 120 bits. These data sizes correspond to the standard transport block size (TBS) configurations in LTE for quaternary phase shift keying (QPSK), denoted by $I_{\mathrm{TBS}} \in \{0, 3, 5, 8\}$. Since there is 24-bit cyclic redundancy checksum (CRC) associated with each transport block and a PRB pair consists of 168 REs, the corresponding code rates can be calculated as $R = \frac{\mathrm{TBS}+24}{2\cdot 168}$. Other relevant simulation parameters are listed in Table III.

A. Probability of Successful Packet Reception

The KPIs considered in this work for comparing the different mechanisms depend on the probability of successful packet reception. Therefore, we first compare the analytical results for the probability of successful packet reception with those from Monte-Carlo simulations. Fig. 5 shows this comparison for the different UL transmission methods. It is evident that the simulation results agree well with analytical results derived in Section V for different initial coding rates and for varying network loads. This also confirms that the probability-based analytical models presented in Section IV are adequate to analyze a wide range of network scenarios.

B. Energy efficiency optimization for GF IR-HARQ

Next, we attempt to maximize the energy efficiency of the proposed GF IR-HARQ method. Recall from Section III-A that we suggest to partition the set of preambles into variable-size groups for GF IR-HARQ. Here, the objective is to find the optimal number of preambles in each group, such that the energy efficiency is maximized. This problem can be

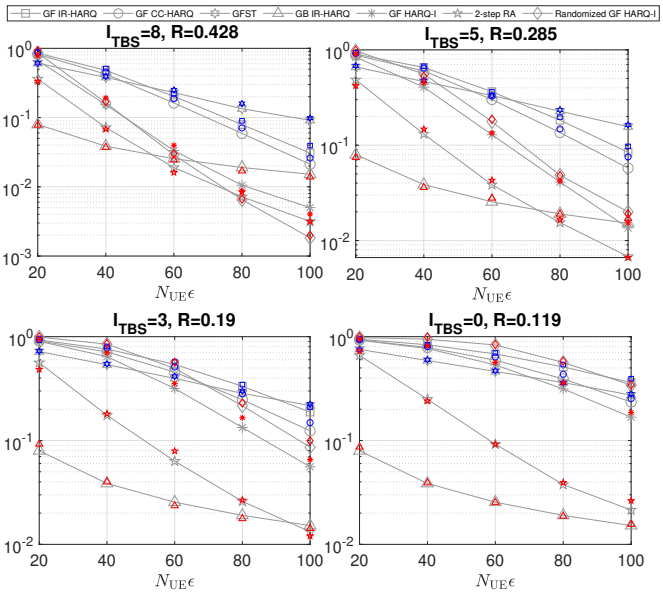


Fig. 5: Comparison of analytical results (solid lines) and simulation results (colored markers) for the probability of successful packet reception.

TABLE IV: GA parameters

Parameter	**Value**
Crossover fraction	0.9
Max stall generations	10
Population size	100
Function tolerance	10^{-3}

formulated as a nonlinear integer programming problem

$$\max_{\boldsymbol{N}_{\mathrm{p}}} \eta$$
$$\text{s.t. } \{\boldsymbol{N}_{\mathrm{p}} \in \mathbb{N}^{n^{\mathrm{HARQ}}} | \left(1 \le N_{\mathrm{p}}^{r} < N_{\mathrm{T}}\right) \cap \left(\textstyle\sum_r N_{\mathrm{p}}^{r} \le N_{\mathrm{T}}\right)\}, \tag{48}$$

where $\boldsymbol{N}_{\mathrm{p}} = [N_{\mathrm{p}}^{1}, N_{\mathrm{p}}^{2}, ..., N_{\mathrm{p}}^{n^{\mathrm{HARQ}}}]$ with each entry N_{p}^{r} indicating the number of preambles required for the r^{th} HARQ stage. The solution to this problem is obtained using a genetic algorithm (GA) with the parameters summarized in Table IV.[13]

Table V shows the results for energy efficiency optimization using GA for $R = 0.119$. It is clear that the use variable-size preamble groups improves the energy efficiency, when compared to the use of equal-size preamble groups, as indicated by the last column of Table V. Moreover, the energy efficiency gain increases with increasing $N_{\mathrm{UE}}\epsilon$.

C. UE Energy Efficiency, Packet Delay and DL Overhead

Finally, we demonstrate the results for the KPIs considered in this work: UE energy efficiency, packet delay and DL overhead. The energy efficiency results are normalized to the energy consumed for the processing of one RE, where for convenience we assume $W_{\mathrm{RE}}^{\mathrm{TX}} = W_{\mathrm{RE}}^{\mathrm{RX}} \triangleq W_{\mathrm{RE}}$ (see Table III), which is reasonable for low-cost UEs.

[13]For small problem instances, optimal results were generated using exhaustive search and it was verified that the results obtained using GA were near-optimal.

TABLE V: Results for preamble group optimization.

$N_{\mathrm{UE}}\epsilon$	Solution	$\frac{\eta_{\mathrm{variable-size}}-\eta_{\mathrm{equal-size}}}{\eta_{\mathrm{equal-size}}}$
20	$\boldsymbol{N}_{\mathrm{p}} = [439, 42, 17, 8]$	34%
40	$\boldsymbol{N}_{\mathrm{p}} = [347, 103, 38, 18]$	54%
60	$\boldsymbol{N}_{\mathrm{p}} = [318, 135, 48, 5]$	64%
80	$\boldsymbol{N}_{\mathrm{p}} = [279, 162, 54, 11]$	77%
100	$\boldsymbol{N}_{\mathrm{p}} = [269, 166, 62, 9]$	89%

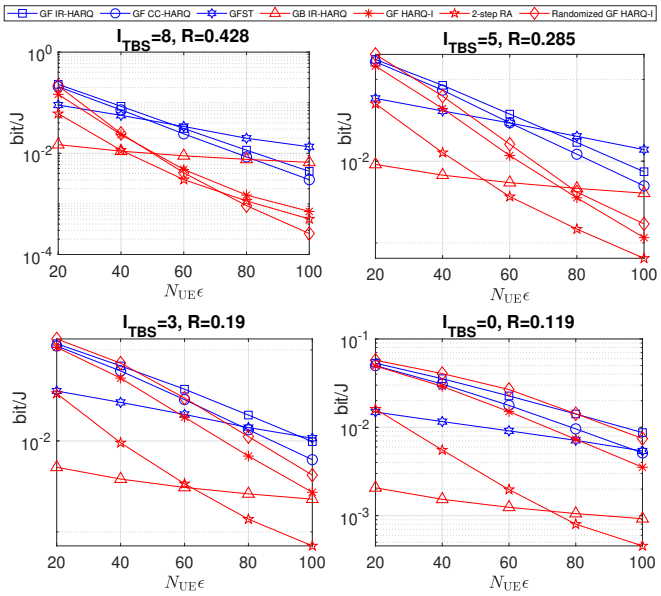


Fig. 6: UE energy efficiency.

Fig. 6 demonstrates the UE energy efficiency for the different methods. We observe that the GFST method outperforms all other methods when the code rate is high ($R \in \{0.428, 0.285\}$) and the network is heavily loaded ($N_{\mathrm{UE}}\epsilon \in \{80, 100\}$). This is because in all the methods except GFST, the probability of successful reception of a packet with a high code rate, in the early stages of HARQ, is small. Hence, most UEs resort to re-transmission of packets, thereby reducing the energy efficiency. But in the GFST method, all n^{HARQ} RVs are sent together, which reduces the effective code rate and increases the probability of successful reception, thereby improving the energy efficiency. Thus, although the GFST method appears to be a fairly straightforward scheme of sending all RVs together, it is indeed the most energy efficient UL transmission scheme in scenarios with high network loads and high initial coding rates.

At lower code rates ($R \in \{0.119, 0.19\}$), the HARQ-assisted methods demonstrate better energy efficiency than the GFST method because the probability of successful reception of a packet in the early stages of HARQ increases when the code rate is low. It is seen from Fig. 7 that Randomized GF HARQ-I performs better than the proposed methods when the initial code rate is very low (i.e., 0.119) in terms of the energy efficiency. But the proposed GF IR-HARQ method is superior in performance when the initial code rate is higher than 0.119 or in scenarios with very heavy network load ($N_{\mathrm{UE}}\epsilon = 100$) because it incorporates packet combining. Hence, the proposed GF IR-HARQ method along with the optimal preamble partitioning is very much applicable to mMTC, where scenarios with low coding rate are not uncommon.

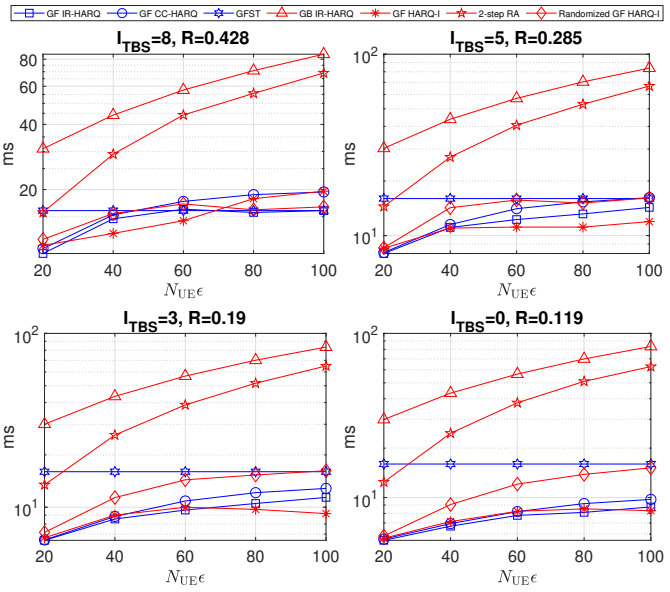


Fig. 7: Packet delay.

Fig. 7 illustrates the results obtained for packet delay using the different methods. We see that the delay incurred by the use of the conventional GB IR-HARQ and two-step RA schemes are much higher that the rest, due to the back-and-forth exchange of control messages related to RA processes. As expected, the GFST scheme has a constant packet delay across different code rates and network loads, which is $n\tau_{\mathrm{TTI}} = 16$ ms. The packet delay for the other GF-HARQ schemes is lower than that of GFST, especially for low code rates. The reasoning is similar to that stated for UE energy efficiency. Furthermore, the packet delay incurred by the usage of the Randomized GF HARQ-I method is inferior to that incurred by our proposed methods and the GF HARQ-I method for coding rates greater than or equal to 0.19. With respect to the DL overhead across all code rates and network loads, it was observed that $O^{\mathrm{DL}}_{\mathrm{GB}}$ and $O^{\mathrm{DL}}_{2-\mathrm{step}}$ are at least 30 times and 28.4 times larger that the maximum amount of DL overhead of the GF-HARQ methods ($= \nu^{\mathrm{ACK}}$) respectively.

VIII. Conclusion

Generally, grant-free transmission and NOMA have been thoroughly studied as separate solutions to enhance the number of devices supported in mMTC networks. A comprehensive analysis of their combination, GF-NOMA, which can potentially exploit the advantages of both the schemes, is presented in this paper. In particular, a GF-NOMA system along with different HARQ mechanisms is considered for mMTC uplink. Three new methods for GF-NOMA HARQ have been proposed, namely, the GF IR-HARQ (using IR combining), the GF CC-HARQ (using CC combining) and the GFST scheme, where all the RVs are transmitted in a single shot. All these methods employ preamble followed by NOMA-coded data transmission. Unlike current grant-based systems, the proposed schemes do not require additional

exchanges from the base-station for the transfer of HARQ-related information, since the preambles are not only used for user activity detection and channel estimation, but also carry the HARQ meta-data (for example, RV index information). This renders the proposed GF-NOMA HARQ schemes to be de-centralized. In terms of performance, the proposed methods have been compared against the conventional GB-IR HARQ and two-step RA mechanisms and the state-of-the-art GF HARQ-I mechanism (which does not incorporate packet combining) in terms of the probability of successful packet reception, followed by the KPIs corresponding to UE energy efficiency, packet delay and DL overhead. Numerical results indicate that the proposed GF IR-HARQ method along with the optimal preamble partitioning strategy presented in this paper, is superior in scenarios with relatively low initial coding rates or low network loads. Moreover, in scenarios with relatively high initial coding rates and heavy network loads, the GFST scheme demonstrates the best performance. Considering their distinguished performance and their decentralized nature, it can be concluded that the proposed methods can be readily employed to improve the operational efficiency of grant-free uplink mechanisms in mMTC.

Appendix A
Proof of (21, 22)

Assume that, α_u^r is the binary activity indicator for user u at state r, where 1 means active and 0 means inactive. The number of active users at HARQ stage r is

$$N_{\mathrm{a}}^r = \sum_{u=1}^{N_{\mathrm{UE}}} \alpha_u^r \,. \tag{49}$$

We assume that α_u^1 follows a Bernoulli distribution with parameter ϵ, thus N_{a}^1 which is the sum of N_{UE} i.i.d. Binomial variables, follows a Binomial distribution as follows

$$\Pr(N_{\mathrm{a}}^1) = \mathrm{Binomial}(N_{\mathrm{a}}^1, N_{\mathrm{UE}}, \epsilon), \tag{50}$$

which proves (21). To prove (22), we first consider the probability $\Pr(\alpha_u^r = 1|N_a^{r-1})$. Defining the binary variable S_u^r indicating the success of packet reception for user u at stage r, we can write

$$\begin{aligned}
\Pr(\alpha_u^r = 1|N_{\mathrm{a}}^{r-1}) &= \sum_{\alpha_u^{r-1}, S_u^{r-1}} \Pr(\alpha_u^r = 1, \alpha_u^{r-1}, S_u^{r-1}|N_{\mathrm{a}}^{r-1}) \\
&\overset{(a)}{=} \Pr(\alpha_u^r = 1, \alpha_u^{r-1} = 1, S_u^{r-1} = 0|N_{\mathrm{a}}^{r-1}) \\
&\overset{(b)}{=} \Pr(\alpha_u^r = 1|\alpha_u^{r-1} = 1, S_u^{r-1} = 0, N_{\mathrm{a}}^{r-1}) \\
&\quad \times \Pr(S_u^{r-1} = 0|\alpha_u^{r-1} = 1, N_{\mathrm{a}}^{r-1}) \\
&\quad \times \Pr(\alpha_u^{r-1} = 1|N_{\mathrm{a}}^{r-1}), \\
&\overset{(c)}{=} \Pr(S_u^{r-1} = 0|\alpha_u^{r-1} = 1, N_{\mathrm{a}}^{r-1}) \\
&\quad \times \Pr(\alpha_u^{r-1} = 1|N_{\mathrm{a}}^{r-1}), \\
&\overset{(d)}{=} (1 - \Omega_{r-1}) \Pr(\alpha_u^{r-1} = 1|N_{\mathrm{a}}^{r-1}) \\
&\overset{(e)}{=} (1 - \Omega_{r-1}) \frac{N_{\mathrm{a}}^{r-1}}{N_{\mathrm{UE}}},
\end{aligned} \tag{51}$$

where we used (a) the fact that transmission at stage r is conditioned on a preceding unsuccessful transmission at stage $r-1$, (b) the chain rule of probability, (c) the fact that a re-transmission is scheduled at fixed time interval, (d) the definition of Ω_r as the probability of successful packet reception in the r^{th} stage, and (e) the fact that

$$\begin{aligned}
\Pr(\alpha_u^{r-1} = 1|N_{\mathrm{a}}^{r-1}) &= \frac{\Pr(N_{\mathrm{a}}^{r-1}|\alpha_u^{r-1} = 1) \Pr(\alpha_u^{r-1} = 1)}{\Pr(N_{\mathrm{a}}^{r-1})} \\
&= \frac{\binom{N_{\mathrm{UE}}-1}{N_a^{r-1}-1} \epsilon^{N_{\mathrm{a}}^{r-1}-1} (1-\epsilon)^{N_{\mathrm{UE}}-N_{\mathrm{a}}^{r-1}} \epsilon}{\binom{N_{\mathrm{UE}}}{N_a^{r-1}} \epsilon^{N_{\mathrm{a}}^{r-1}} (1-\epsilon)^{N_{\mathrm{UE}}-N_{\mathrm{a}}^{r-1}}} \\
&= \frac{N_{\mathrm{a}}^{r-1}}{N_{\mathrm{UE}}}.
\end{aligned} \tag{52}$$

Hence,

$$\Pr(\alpha_u^r|N_{\mathrm{a}}^{r-1}) = \mathrm{Bernoulli}(\alpha_u^r, \epsilon_r), \tag{53}$$

where

$$\epsilon_r = (1 - \Omega_{r-1}) \frac{N_{\mathrm{a}}^{r-1}}{N_{\mathrm{UE}}}. \tag{54}$$

Finally, since N_{a}^r in (49) is the sum of N_{UE} i.i.d. Bernoulli variables, we obtain

$$\Pr(N_{\mathrm{a}}^r|N_{\mathrm{a}}^{r-1}) = \mathrm{Binomial}(N_{\mathrm{a}}^r, N_{\mathrm{UE}}, \epsilon_r). \tag{55}$$

Appendix B

In this appendix, we provide justification for the assumption that a preamble collision results in data decoding failure.

Let us consider two users—user A and user B selecting the same preamble (say P_1) and transmitting at the same time with their channel gains denoted as h_A and h_B, respectively. The received signal at the base station is $h_A P_1 + h_B P_1$. Although the detection of P_1 at the base station is possible through multi-user detection methods like MMSE-based AMP [7], the base station cannot distinguish whether just one UE or several UEs have sent the same preamble. Hence, the estimated channel corresponding to the detected preamble in this situation is $h_A + h_B$. Using $h_A + h_B$ instead of h_A and h_B for equalizing the received signal deteriorates the signal, mostly leading to decoding failure.

Moreover, users with the same preambles use the same NOMA spreading codes. Hence, there is no processing gain for any of the two users over the other one after de-spreading. Therefore, the signal of these two users interfere with each other in a network that already suffers from high interference level owing to the large number of users and non-orthogonality of their signals. Given this scenario, a preamble collision results in decoding failure with a very high probability. Thus, we assume that a preamble collision always results in data decoding failure. Such an assumption was also used in [9, Section V.A].

To further verify this assumption, we perform two simulations to illustrate that preamble collision is indeed detrimental to the data decoding performance. The parameters in these simulations are consistent with the ones adopted in Section VII. Also, for the channel coding technique, we employed the convolution turbo code (CTC) scheme used in LTE standard with a default code rate of $\frac{1}{3}$. Note that this coding rate $\frac{1}{3}$ is well within the four different coding rates considered in the manuscript and is not biased towards one particular setting.

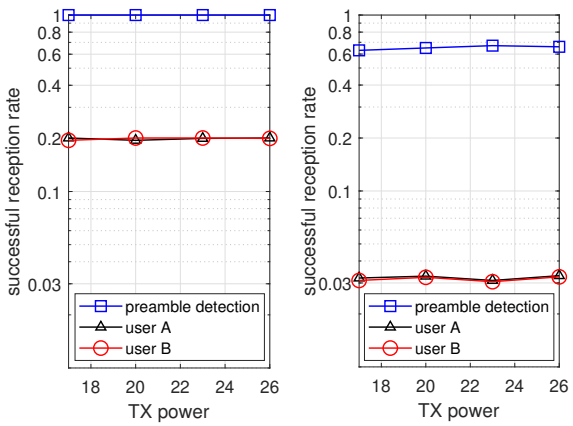


Fig. 8: Monte-Carlo simulation results for Scenario 1 (left) and Scenario 2 (right) with 10000 and 1000 runs, respectively.

Scenario 1 - Simulation of two users: In this simulation we consider the scenario where only two users exist in the network and they pick the same (preamble, NOMA code) pair. The base station receives both signals at the same time with equal power (due to power control) and runs an MMSE-based AMP multi-user detection algorithm to detect the preamble and estimate its channel [7]. Once the preamble is detected, the base station equalizes and de-spreads the signal. The de-spread signal is then passed through the demodulator and turbo decoder. Finally, base station examines whether the data matches that of the first user or the second. Fig. 8 (left) shows the successful decoding rate of this scenario for different transmit powers. As it is seen, there is a 20% chance for successfully decoding either messages. Note that since there are only two users, the interference level is not very high.

Scenario 2 - Simulation of multiple users with one preamble collision: Here, we consider a scenario with 100 active users employing different (preamble, NOMA code) pairs except for two users, say user A and B. We are interested in knowing the decoding success rate of these two users. In this scenario, the SNR of user A and B can be low, owing to the higher interference level caused by multiple active users with non-orthogonal codes. Hence, we expect to see a decoding success rate much lower than 20% (which was observed in the previous scenario with exactly two users). Fig. 8 (right) shows that the decoding rate goes down to about 3% in this scenario. Note that in a typical mMTC scenario where more than two users can have preamble collision, decoding success rate is actually less than 3%. Therefore, we ignore this small probability for simplicity and conservatively assume that a preamble collision is equivalent to packet decoding failure in this work.

REFERENCES

[1] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[2] *Study on Scenarios and Requirements for Next Generation Access Technologies*, ETSI, May 2017, v14.2.0.

[3] S. K. Sharma and X. Wang, "Towards massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, pp. 1–1, 2019.

[4] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, 3GPP, Jun. 2017, v14.3.0.

[5] 5G Americas. The 5G Evolution: 3GPP Releases 16-17. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2020/01/5G-Evolution-3GPP-R16-R17-FINAL.pdf

[6] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Simple semi-grant-free transmission strategies assisted by non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4464–4478, June 2019.

[7] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, June 2018.

[8] ——, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, June 2018.

[9] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. De Carvalho, Y. Ji, C. Stefanović, P. Popovski, Q. Wang, M. Schellmann, E. Kosmatos, P. Demestichas, M. Raceala-Motoc, P. Jung, S. Stanczak, and A. Dekorsy, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28 969–28 992, 2018.

[10] J. Choi, "On the throughput comparison between multi-channel aloha and compressive random access," in *Intl. Conf. on Inform. and Commun. Techn. Convergence (ICTC)*, Oct 2016, pp. 944–948.

[11] A. T. Abebe and C. G. Kang, "Comprehensive grant-free random access for massive low latency communication," in *IEEE Intl. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[12] K. Senel and E. G. Larsson, "Grant-free massive MTC-Enabled Massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec 2018.

[13] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, June 2017.

[14] A. C. Cirik, N. Mysore Balasubramanya, and L. Lampe, "Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 46–49, Feb 2018.

[15] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, March 2019.

[16] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, March 2017.

[17] G. Chen, J. Dai, K. Niu, and C. Dong, "Sparsity-inspired sphere decoding (SI-SD): A novel blind detection algorithm for uplink grant-free sparse code multiple access," *IEEE Access*, vol. 5, pp. 19 983–19 993, 2017.

[18] Y. Du, B. Dong, Z. Chen, X. Wang, Z. Liu, P. Gao, and S. Li, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812–2828, Dec 2017.

[19] F. Wei, W. Chen, Y. Wu, J. Li, and Y. Luo, "Toward 5G wireless interface technology: Enabling nonorthogonal multiple access in the sparse code domain," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 18–27, Dec 2018.

[20] J. Shen, W. Chen, F. Wei, and Y. Wu, "ACK feedback based UE-to-CTU mapping rule for SCMA uplink grant-free transmission," in *Intl. Conf. Wirel.Commun. and Sign. Proc. (WCSP)*, Oct 2017, pp. 1–6.

[21] "Study on new radio access technology physical layer aspects," 3GPP, Tech. Rep. 38.802, 2017.

[22] C. Stefanovic, P. Popovski, and D. Vukobratovic, "Frameless ALOHA protocol for wireless networks," *IEEE Commun. Lett.*, vol. 16, no. 12, pp. 2087–2090, December 2012.

[23] Y. Yu and G. B. Giannakis, "SICTA: a 0.693 contention tree algorithm using successive interference cancellation," in *IEEE Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3, March 2005, pp. 1908–1916 vol. 3.

[24] V. Boljanović, D. Vukobratović, P. Popovski, and C. Stefanović, "User activity detection in massive random access: Compressed sensing vs. coded slotted ALOHA," in *IEEE Intl. Workshop Sign. Proc. Adv. in Wirel. Commun. (SPAWC)*, July 2017, pp. 1–6.

[25] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Intl. Symp. on Wirel. Commun. Syst. (ISWCS)*, Aug 2013, pp. 1–5.

[26] J. Zhang, L. Lu, Y. Sun, Y. Chen, J. Liang, J. Liu, H. Yang, S. Xing, Y. Wu, J. Ma, I. B. F. Murias, and F. J. L. Hernando, "PoC of SCMA-Based Uplink Grant-Free Transmission in UCNC for 5G," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1353–1362, June 2017.

[27] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, July 2016.

[28] G. Berardinelli, N. Huda Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23 602–23 611, 2018.

[29] S. Kallel, "Complementary punctured convolutional (CPC) codes and their applications," *IEEE Trans. Commun.*, vol. 43, no. 6, pp. 2005–2009, June 1995.

[30] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with Chase combining and incremental redundancy for HSDPA," in *IEEE Veh. Techn. Conf. (VTC Fall)*, vol. 3, Oct 2001, pp. 1829–1833 vol.3.

[31] J. Choi, "On harq-ir for downlink noma systems," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3576–3584, 2016.

[32] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the performance of noma with hybrid arq," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10 033–10 038, 2018.

[33] Z. Shi, C. Zhang, Y. Fu, H. Wang, G. Yang, and S. Ma, "Achievable diversity order of harq-aided downlink noma systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 471–487, 2020.

[34] S. Yoon and Y. Bar-Ness, "Packet data communications over coded CDMA - Part II: Throughput bound of CDMA unslotted ALOHA with hybrid type II ARQ using rate compatible punctured turbo codes," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1616–1625, Sep. 2004.

[35] 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.

[36] *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding*, 3GPP, Mar. 2015, v12.4.0.

[37] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*, ETSI, Apr. 2017, v14.2.0.

[38] C. Wei, G. Bianchi, and R. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, April 2015.

Faramarz Jabbarvaziri received his B.A.Sc. and M.A.Sc. degrees in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 2012 and 2015. He is currently a Ph.D. candidate at the Department of Electrical and Computer Engineering, The University of British Columbia. His research interests are massive MIMO communications, non-orthogonal multiple access, and machine learning.

Naveen Mysore Balasubramanya received the M.S. degree in electrical engineering from the University of Colorado, Boulder, USA, in 2010 and the Ph.D. degree in Electrical and Computer Engineering from The University of British Columbia, Vancouver, BC, Canada in 2017. Since September 2018, he is an assistant professor at the Indian Institute of Technology Dharwad, Karnataka, India. He has a rich academic experience as well as six years of industrial R&D experience, having worked as a postdoctoral research associate at Heriot-Watt University, Edinburgh, U.K. and as a Senior Design Engineer for the ADSL, WiMAX, and LTE Systems in leading communication industries, such as, Lantiq Communications (Intel) India Pvt. Ltd., Broadcom Communications India Pvt. Ltd., and Tata Elxsi Ltd., Bangalore, India. His research interests are broadly in theory and practical aspects of wireless communications, with focus on next generation communication technologies and energy efficient Internet of Things.

Lutz Lampe (M'02-SM'08) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the University of Erlangen, Erlangen, Germany, in 1998 and 2002, respectively. Since 2003, he has been with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada, where he is a Full Professor. His research interests are broadly in theory and application of wireless, optical wireless, optical fiber, power line and underwater acoustic communications. He has served as an Associate Editor and a Guest Editor for several IEEE journals, and as a General and Technical Program Committee Co-Chair for IEEE conferences. He has been a Distinguished Lecturer of the IEEE Communications Society and a (co-)recipient of a number of best paper awards. He is a co-editor of the book "Power Line Communications: Principles, Standards and Applications from Multimedia to Smart Grid" (2nd ed.) by John Wiley & Sons.