Optimal random packet replication policies for IIoT in 5G and Beyond considering different feedback regimes

Salah Eddine Elayoubi, Patrick Brown, Meriem Mhedhbi

Optimal random packet replication policies for IIoT in 5G and Beyond considering different feedback regimes

Salah Eddine Elayoubi[1], Patrick Brown[2] and Meriem Mhedhbi[2]
[1]Université Paris Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, Gif-Sur-Yvette, France
[2]Orange Labs, France

***Abstract*—In the context of Industrial Internet of Thing (IIoT) applications, network administrators must use their available bandwidth to both identify best 5G NR (New Radio) configuration and best transmission schemes. In this paper we derive optimal channel access and packet replication schemes for Ultra Reliable Low Latency Communications (URLLC) traffic in 5G and Beyond networks. Based on typical system configurations, we identify three main regimes, characterized by the delay and the consequent presence or absence of feedback received from the network. In particular, we identify (a) extreme situations where there is only one time slot opportunity for packet transmission with possible replications in the frequency dimension, (b) blind repeated replication with several transmission slots opportunities, with no received ACK before the delay expiration, and (c) far-sighted scenarios where, additionally, some ACKs may be received, acknowledging older replicas of the packet. We propose adapted replication models for each radio scenario and develop corresponding mathematical models from which the transmission schemes may be optimized. The proposed policies are semi-distributed, in the sense that the optimal policy is communicated to each device by the network, that then applies it autonomously in order to respect the stringent timing constraints of URLLC. We then show the corresponding system dimensioning for achieving the target reliability and identify the best NR configurations to be deployed in the controlled industrial environments.**

I. Introduction

Industrial Internet of Things (IIoT) applications with stringent latency and reliability constraints are served in 5G using the URLLC class [2]. In this context, packet re-transmission is regarded as a key enabler for reliability in the uplink and the design of re-transmission schemes has been a hot research topic in the recent years. In order to respect latency requirements, such schemes are generally associated with a "grant-free" approach, under which neither issuing a scheduling request nor waiting for a scheduling grant are required [1]. As a result replicas must contend for shared resources, causing possible collisions between packets, generating losses that are added to the losses due to imperfections in the radio channel. In the context of IIoT applications, network administrators have the choice of radio configurations. These may be dictated by the expected uplink and downlink traffic volumes and the expected number of transmitting devices. Proposed re-transmission policies must be adapted to these configurations.

Driven by the observation that waiting for a negative acknowledgment (NACK) before retransmitting the packet, may introduce unacceptable latency, the 5G standard [1] proposed to send multiple replicas of the same packet without waiting for the acknowledgment (ACK) (which we term *blind retransmissions*). There have been several propositions in this direction. Authors in [10] proposed to send these replicas in a contention-based manner on different frequency resources on consecutive TTI, while in [4] the authors considered a more flexible scheme where replicas can be sent on any of the available time-frequency resources before the delay budget expires. Other sets of works consider the presence of ACKs that cause early stopping before all the replicas are transmitted. For instance, [12] considered a repetition scheme where, once a packet is decoded, an ACK is generated that stops immediately the transmission. The presence of NACKs has also been considered in [6] for 5G networks and in [7] for WIA-FA (Wireless networks for factory automation) considered a cyclic traffic, where a first phase is dedicated to blind replication, followed by a grouped ACK and a second phase dedicated to retransmissions of failed packets.

We show in this paper that both approaches (waiting for a NACK before re-transmitting versus blind re-transmissions), as well as approaches combining the two, should be used for URLLC, depending on the radio and service configuration chosen by the network administrator. We start by analyzing typical 5G New Radio (NR) configurations and show that, for some of them, there is room for receiving one or more ACK before the delay expiration, while in others there is no transmission opportunity after the feedback is received. We evaluate their impact on the achievable URLLC quality of service. We thus address the optimal policy, i.e., the one that minimizes the delay outage, as follows:

- We take into consideration the feedback regime, i.e. the presence of ACKs and the delay before receiving them. Observing the different 5G NR (New Radio) configurations, we classify replication scenarios between "blind" versus "far-sighted", and we identify the possibility of a repeated transmission in the time domain. We also explore the Beyond 5G scenario of early feedback where the receiver implements advanced decoding mechanisms and NACKs are almost instantaneous, as in [11].
- We consider the diversity of policies, including sending a packet with a probability $p \leq 1$ in each available time slot, or multiple replicas per slot.
- We develop mathematical models for the loss probability

for each of the policies. While the model for the blind one-shot case is derived from [4], the models for the other schemes are novel. In addition, these models consider sporadic and asynchronous traffic patterns, dislike other works that consider synchronous and cyclic traffic ([6], [7]) and explicitly integrate the interaction between interfering users in an industrial setting where all devices are configured for following the same transmission policy.
- We identify the optimal policy for each case. These policies are semi-distributed, as the network communicates them to the devices whom apply them autonomously upon packet generation.

The remainder of this paper is organized as follows. Section II presents some 5G NR configurations and their impact on the feedback availability. Section III derives the loss probabilities and the corresponding optimal policies for the different configurations. Section IV evaluates numerically the performances of the different policies and compares their performances. Section V concludes the paper.

II. System model

A. Traffic model

We consider a typical uplink IIoT scenario where a large number of sensors ($n >> 1$), distributed over a plant, forward measurement data to process controllers. This use case requires a reliability (e.g. 99.99%), an end-to-end latency τ ranging between 1 ms and 10 ms and small data packets.

Time is slotted and each transmitter is active during a time slot with probability $q < 1$. For instance, if each user generates 1 packet each y ms, and the time is slotted with a slot length of T, a user generates a packet in a slot with a probability $q = T/y$, with $T << y$.

B. Resources and feedback regimes in 5G NR

We next present in Table I six representative 5G NR configurations of industrial environments [8]. These will be used throughout the paper and in the numerical applications. The configurations are summarized in column 2. They are characterized by different numerologies (subcarrier spacing and mini-slot duration), duplexing modes (FDD and TDD) and frame designs. In all these configurations, we consider a total bandwidth for the URLLC traffic of $B = 10$ MHz, a fixed MCS with a spectral efficiency of $e = 2$ bit/s/Hz and application packets of fixed size equal to $a = 32$ Bytes. For instance, for a subcarrier spacing of $b = 30$ KHz, the OFDMA symbol duration is half that of 15 KHz, leading to a lower number of packets per slot for the same number of symbols per slot (e.g. configurations 1 and 3). The next columns show the resulting slot lengths, number of packets per slot (K), number of time transmission opportunities (D), the ACK delay dictating if an acknowledgment is possible (next column), transmission pattern. We next overview how these are derived.

1) Maximal number of packets per slot: Let the spectral efficiency of the considered Modulation and Coding Scheme (MCS) be equal to e bit/s/Hz. For an application packet of size a bits, a bandwidth per RB of b Hz and a time slot length of T, the number of physical RBs, R, for transmitting an application packet is $R = \lceil a/(eTb) \rceil$ ($\lceil x \rceil$ being the smallest integer greater than or equal to x). While a depends on the application, b and T depend on the radio configuration, and e on the chosen MCS. For a total available spectrum for the uplink URLLC transmission of B (in Hz), the total number of RBs is equal to B/b (usually an integer), and the number of packets that can be multiplexed per slot is given by:

$$K = \left\lfloor \frac{B/b}{R} \right\rfloor = \left\lfloor \frac{B/b}{\lceil a/(eTb) \rceil} \right\rfloor \tag{1}$$

($\lfloor x \rfloor$ being the largest integer smaller than or equal to x).

2) Number of time transmission opportunities: Let τ be the delay budget for the application (expressed in ms). In Table I and the rest of the paper, we consider the typical case of $\tau = 1$ ms. For FDD configurations, the number of transmission opportunities in the time domain is easily computed by dividing the delay budget τ by the mini-slot duration: $D = \left\lfloor \frac{\tau}{T} \right\rfloor$. However, for TDD frames, one has also to account for the downlink slots that cannot be used for uplink transmissions, leading to a lower value of time transmission opportunities, as illustrated by column 5.

3) Acknowledgement regime: We now compute the ACK delay in slots (column 6). Based on this ACK delay and on the value of D, we indicate in column 7 if this delay enables at least one NACK-based retransmission (i.e. retransmission knowing that a replica is not well received). The ACK delay corresponding to a correctly decoded packet is composed of the UE processing delay ($T_{UE} = 1$ slot), the packet alignment time ($T_{Alg} = 1$ slot in FDD while in TDD it depends on the configuration), the packet transmission time ($T_{Tr} = 1$ slot), the gNodeB processing delay (T_{gNB}) that depends on the performance of the equipment (that we approximate here to 0.1 ms), and the ACK transmission time ($T_{ACK} = 1$ slot).

4) Example of transmission patterns: The last column of Table I provides a typical transmission pattern for each configuration, indicating the slots where replicas could possibly be placed (i.e. slots where uplink transmissions are possible). It also indicates the feedback opportunities, based on the previous delay computation. For example, for configuration 1, the delay budget allows for placing replicas in 7 consecutive slots, and the feedback (f_1) for replica sent on the first slot (u_1) arrives on the fifth slot, meaning that u_5 could be used for a NACK-based retransmission in the absence of ACK reception.

5) Extension to Beyond 5G: In addition to these 5G configurations, our schemes can be extended to Beyond 5G configurations. For instance, when considering 6G extreme URLLC scenario, with a very stringent delay constraint of the order of 0.1 ms [13], there is no possibility for repeated replication, but the performance model of proposition 1 below will still be valid.

On the other hand, early feedback mechanisms are being proposed for Beyond 5G [11], leading to almost instantaneous feedbacks. We will consider this scenario in our performance models too.

TABLE I: System configurations for different numerologies and duplexing (FDD or TDD with a pattern for switching between D=Downlink and U=Uplink slots). Example transmission patterns are also provided: u_i=transmission slot number i with possible multiple transmissisons in the frequency domain; d=reserved slot for downlink in TDD mode; f_i=feedback sent on the downlink channel for slot i packet.

Conf.	Description	Slot length	K packets/slot	D	ACK delay slots	ACK?	Example UL transmission pattern & corresponding DL feedback
1	(15 KHz;2 sym/slot;FDD)	0.144 ms	11	7	5	Yes	$u_1u_2u_3u_4u_5u_6u_7$ 0 0 0 0 f_1 f_2 f_3
2	(15 KHz;4 sym/slot;FDD)	0.288 ms	22	3	5	No	$u_1u_2u_3$ 0 0 0
3	(30 KHz;2 sym/slot;DDUU)	0.072 ms	5	6	7	Yes	$du_1u_2ddu_3u_4d\ d\ u_5u_6\ d\ d$ 0 0 0 00 0 0 0f_1 0 0 $f_2$0
4	(30 KHz;4 sym/slot;DDUU)	0.144 ms	11	4	7	No	$du_1u_2ddu_3u_4d$ 0 0 0 00 0 0 0
5	(30 KHz;2 sym/slot;DDDU)	0.072 ms	5	3	7	Yes	$ddu_1dddu_2dd\ d\ u_3dd$ 00 0 000 0 00f_1 0 00
6	(30 KHz;4 sym/slot;DDDU)	0.144 ms	11	1	7	No	ddu_1ddd 00 0 000

III. Performance of replication policies

We describe the possible replication policies and the system configurations in which they may be applied. In the case of closed expressions we present the best replication factors to obtain optimal performance for the policy depending on system parameters. Otherwise best replication factors, p, must be obtained numerically. All the following results are new except for equation (2) from [4].

A. The blind one-shot scenario

We start by a very challenging scenario from the reliability perspective, that is the case where there is just one time slot for transmission opportunities as in configuration 6 in Table I. This model also applies to 6G extreme URLLC where the very stringent delay does not allow for time-repeated replication.

In this case the natural scheme is a random selection of the K possible replica positions in the frequency dimension for each replica. Let n be the number of users and q their activity (probability of generating a packet in a slot). Let $p > 0$ be the number of replicas in the frequency dimension.

Proposition 1. *For the one-shot blind case and a transmission scheme that sends p replicas per slot, on resources selected randomly, the loss probability is computed by:*

$$l(p) = 1 - \sum_{j=1}^{p} (-1)^{j+1} C_p^j \left((1-q) + q \frac{C_{K-j}^p}{C_K^p} \right)^{n-1} \quad (2)$$

where C_p^j is the number of combinations of j among p.

Proof. Define $\mathcal{A}_i$ to be the event that the i-th resource is free, i.e. no (other) active user chooses this resource for its packet transmissions. We would like to express the probability that one of the p resources is free, i.e. $\mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_p\}$. To this end, we determine the probability that a subset of j resources is free. Note that in a set containing p resources there are C_p^j subsets of size j. All j resources will be collision-free if all other users are either not transmitting or none of their p resources fall in the j slots. For a given user, this happens with probability $1 - q + q\frac{C_{K-j}^p}{C_K^p}$, where q is the probability that a user is active. Since there are $n-1$ other users, the probability that all j slots of this subset are collision-free:

$$\mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_j\} = \left(1 - q + q\frac{C_{K-j}^p}{C_K^p}\right)^{n-1}.$$

Equation (8) may then be derived from the expression:

$$\mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_p\} = \textstyle\sum_{j=1}^{p} (-1)^{j+1} C_p^j \, \mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_j\} \quad \square$$

B. The blind repeated scenario

In scenarios as in configurations 2 and 4 of Table I, there may be $D > 1$ time slots where replicas can be placed, but still no feedback may be be received before the expiration of the delay budget. We refer to this case as ”blind repeated”, as considered in several works. [10] advocated to place one replica per available slot within the latency budget, as also proposed by 3GPP [1]. [5] observed that there is no need for sending exactly one replica per slot, and the UE may decide to place a replica or not following a given probability p, that has to be uniform over all slots. This corresponds to a p-persistent Aloha scheme [3], [9]. On the contrary, [6] proposed a scheme where an integer number $p \geq 1$ of replicas is placed randomly in the available time frequency resources. In the following, we propose a unifying framework where the policy p depends on the traffic regime as will be illustrated in the numerical section.

1) High traffic regime: In the high traffic regime (nq large and K small), it is convenient not to load excessively the channel. We consider in this case the p-persistent Aloha with $p < 1$. In p-persistent Aloha, starting from its generation time, the packet is set on the first slot following its generation, and then a replica is transmitted in each of the consecutive $D-1$ slots with probability $p \leq 1$ in each slot, to be optimized. We suppose that a replica is lost if it collides with the packet of another transmitter. The objective is to derive p^* that maximizes the probability of success.

Proposition 2. *In a scheme that sends a replica in each slot with probability p, the packet loss is:*

$$l(p) = \left[1 - \left(1 - \frac{q + (1-(1-q)^{D-1})p}{K}\right)^{n-1}\right] \times \left[1 - p\left(1 - \frac{q + (1-(1-q)^{D-1})p}{K}\right)^{n-1}\right]^{D-1} \quad (3)$$

Proof. Consider a user who transmits a packet in the slot following its generation with probability 1, and then transmits a replica in the subsequent $D-1$ slots with probability p.

The probability the first transmission is lost is:

$$l_1(p) = 1 - \left(1 - \frac{q}{K} - \frac{(1-(1-q)^{D-1})p}{K}\right)^{n-1} \quad (4)$$

Indeed, this packet is correctly received if none of the other $n-1$ transmitters (if active) selects the same frequency resource. An interfering user may have just generated a packet (with probability q), or sent a replica with probability p of a packet originating from one of the previous $(D-1)$ slots (with probability $(1-(1-q)^{D-1})p)$[1]. The probability of collision with our given user transmission is: $\frac{q}{K} + \frac{(1-(1-q)^{D-1})p}{K}$, where $1/K$ is the probability for two users to choose one of the K positions in a slot. As all $n-1$ users have the same behaviour, we obtain the loss for the first transmission in equation (4).

For replicas sent on subsequent slots, the collision probability is conditioned on the user deciding to send a replica p:

$$p\left(1 - \frac{q}{K} - \frac{(1-(1-q)^{D-1})p}{K}\right)^{n-1}.$$

As all the subsequent $D-1$ slots are similar, we obtain (3). □

Proposition 3. *An upper bound of the loss in equation (3) is:*

$$\hat{l}(p) = \left[1 - p\left(1 - \frac{q + (1-(1-q)^{D-1})p}{K}\right)^{n-1}\right]^{D} \quad (5)$$

and the value of p that maximizes it is:

$$p^* = \begin{cases} \frac{K-q}{n(1-(1-q)^{D-1})}, & \text{if } n > n^* = \frac{K-q}{(1-(1-q)^{D-1})} \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

Proof. First note that as $p \leq 1$, $\hat{l}(p)$ is an upper bound for expression 3. We first define the function

$$f(p) = 1 - p\left(1 - \frac{q + (1-(1-q)^{D-1})p}{K}\right)^{n-1}$$

As $D > 1$, the transmission probability p that minimizes this bound $\hat{l}(p)$ minimizes also function $f(p)$, whose derivative is:

$$f'(p) = -\left(1 - \frac{q}{K} - \frac{\beta p}{K}\right)^{n-2}\left(1 - \frac{q}{K} - \frac{\beta p n}{K}\right) \quad (7)$$

with $\beta = (1-(1-q)^{D-1})$. As the amount of resources per slot $K > 1$ and we assume small q: $(1 - \frac{q}{K} - \frac{\beta p}{K}) > 0$. The loss starts by decreasing and reaches its minimum for $p = \frac{K-q}{\beta n}$. If however, $\frac{K-q}{\beta n}$ is larger than 1, the best probabilistic policy corresponds to $p = 1$, which gives the optimum of (6). □

[1]The probability for a packet being carried for the past δ slots is $q(1-q)^{\delta-1}$. The probability of carrying an old packet in the buffer is then $\sum_{\delta=1}^{D-1} q(1-q)^{\delta-1} = 1-(1-q)^{D-1}$

2) Low traffic regime: When traffic is low, it may be interesting to send more than one replica per slot as in the blind scenario case. Let $p \in [1, K]$ be an integer value that represents the number of replicas per slot. We prove the following result:

Proposition 4. *For the blind repeated case and a transmission scheme that sends p replicas per slot, on resources selected randomly, the loss probability is computed by:*

$$l(p) = \left[1 - \sum_{j=1}^{p} (-1)^{j+1} C_p^j \left((1-\bar{q}) + \bar{q}\frac{C_{K-j}^p}{C_K^p}\right)^{n-1}\right]^D \quad (8)$$

where $\bar{q}$ is the probability that a user is active (has a packet to transmit) computed by:

$$\bar{q} = 1 - (1-q)^D \quad (9)$$

Proof. Probability $\bar{q}$ in (9) is the probability a packet has been generated in one of the previous D slots (we assume that once a packet is generated by a user, no other packet generation occurs before it is received as q is small).

If we consider now one of the D slots, the probability of the replica being lost is computed as for the one-shot blind case in equation (2), except that the activity factor is larger due to the repeated transmission. Knowing that there are D opportunities for transmission, the loss probability is as in equation (8). □

C. Instantaneous feedback scenario

Before dealing with the delayed feedback case, we consider the case of instantaneous feedback where the receiver knows, before the next transmission opportunity, if the packet is well received. While this is not realistic in current 5G NR configurations, advances in fast packet decoding and early feedback [11] make this perfect case possible in some TDD configurations (see, e.g., configuration 7 of Table I).

1) High traffic regime: We consider the same probabilistic policy as in section III-B1. However, if a packet is well received, the transmitter knows immediately about it and stops transmissions.

Proposition 5. *Under the instantaneous feedback regime, the loss probability is computed as a solution of the following fixed point equation:*

$$l(p) = l_1(p) l_2(p)^{D-1} \quad (10)$$

with

$$l_1(p) = 1 - (1-\beta(p))^{n-1} \quad (11)$$

$$l_2(p) = 1 - p(1-\beta(p))^{n-1} \quad (12)$$

$$\beta(p) = \frac{q}{K} + \frac{p}{K} \frac{q l_1(p)\left(1-((1-q)(1-p+pl_2(p))^{D-1}\right)}{1-(1-q)(1-p+pl_2(p))}. \quad (13)$$

Proof. As for the blind case, we first compute the probability of collision, $\beta(p)$, with a given other user, that depends on the probability sending a replica, p. This latter is now dependent on the loss probabilities, $l_1(p)$ and $l_2(p)$, as a packet that is successfully received is immediately removed for the transmission buffer. We assume the probabilities (resp. $l_1(p)$ and $l_2(p)$) seen by different users at their first and following transmissions are stationary. The collision probability with a given user is then:

$$\beta(p)=\frac{q}{K}+ql_1(p)\sum_{\delta=1}^{D-1}(1-q)^{\delta-1}(1-p+pl_2(p))^{\delta-1}\frac{p}{K} \quad (14)$$

resulting in (13). Equation (14) reads as follows: a collision will occur with the packet sent by another user if a) the other user receives a packet to send, q, and chooses the same position in the slot, $1/K$, or b) she previously had a packet to send $\delta \in [1, D-1]$ slots before, q, which was not acknowledged, $l_1(p)$, did not receive a new packet since, $(1-q)^{\delta-1}$, and each time either did chose to replicate, $1-p$, or the packet was lost, $(1-p+pl_2(p))^{\delta-1}$, and she decided to transmit in the current slot, p, in the same position in the slot, $1/K$.

As there are $n-1$ other users, and the $D-1$ slots after the generation of the packet are all similar, we obtain (10-12).

The loss probability (10) depends on β, while β (13) depends on loss, we thus obtain a fixed point equation depending on the given replication probability p. □

Remark: The collision probability (14) obtained reduces to the blind case (see proof of proposition 3) if retransmissions do not cease even in case of correct transmission (equivalent to setting $l_1(p) = l_2(p) = 1$): $\beta(p) = \frac{q}{K} + \frac{p(1-(1-q)^{D-1})}{K}$.

2) Low traffic regime: For low traffic, one can afford to send more than one replica per slot, leading to the following:

Proposition 6. *Under the instantaneous feedback regime and when $p \geq 1$ replicas are sent randomly per slot resources, the loss probability is computed as for the repeated blind case (equation (8)), where $\bar{q}$ is the probability that a user is active computed by solving the following fixed point equation:*

$$l_1(p) = 1 - \sum_{j=1}^{p}(-1)^{j+1}C_p^j\left((1-\bar{q}) + \bar{q}\frac{C_{K-j}^p}{C_K^p}\right)^{n-1} \quad (15)$$

$$\textit{and} \qquad \bar{q}(p) = 1-(1-q)\prod_{i=1}^{D-1}(1-q+q(1-l^i)) \quad (16)$$

Proof. The success probability, for a given number of replicas and a known activity factor, is computed as for the blind case. For computing the activity factor, we remark that a user is active if it has generated a packet in one of the previous slots, and did not receive any feedback for it, which gives (16). □

D. Far-sighted scenarios

We now move to the situations where one or more ACK/NACKs may be received before the delay budget expires, but the feedback is not instantaneous (see e.g. configurations 1, 3 and 5 of Table I). The transmission decision can thus exploit delayed information, i.e. information about the old replicas and not the latest sent ones. We call policies associated with such scenarios "far-sighted" policies, in opposition with the completely blind ones where no ACK is received (as discussed in the next paragraph). Intuitively, these policies have to change dynamically with time, as the reception of an ACK stops the transmissions.

Let Δ be the delay (in slots) before the feedback is received. The transmitter then remains active (willing to transmit) for at least Δ slots, before a possible stop due to a positive ACK. The activity of the users is thus higher than their activity in the instantaneous feedback case. Due to lack of space, we show here how to derive the loss probability for $p \leq 1$ replicas per slot. A similar approach can be followed for the $p \leq 1$ case.

Proposition 7. *In the far-sighted case with D slots and a delayed feedback of $\Delta < D$ slots, the loss probability is computed as in proposition 6 for the multiple replicas per slot case, with the activity probability computed by:*

$$\bar{q} = 1-(1-q)^{\Delta+1}\prod_{i=1}^{D-\Delta-1}(1-q+q(1-(1-s)^i)) \quad (17)$$

Proof. As no feedback is received during the first Δ slots following the packet generation, a UE that generated a packet in the past Δ slots is still active. Nevertheless, a packet generated between $-D$ and $-\Delta$ may have already received an ACK. This leads to the activity probability of (17). □

IV. Numerical Experiments

A. Evaluation of the replication schemes

1) Blind one-shot replication: We start by the blind one-shot case of proposition 1, considering configuration 6 of Table I. Figure 1 shows the impact of the number of replicas on the loss probability, for a constant number of users ($n = 15$) and different activity factors that impact the load (average number of packets per slot). We observe that the number of replicas impacts the loss rate and that the optimal number of replicas decreases when the traffic load increases (4 replicas for $q = 0.01$ and 3 replicas for $q = 0.1$). We then vary the number of users with a constant activity factor and show in Figure 2 the loss probability and the corresponding optimal number of replicas. When the load increases, the loss logically increases and the optimal number of replicas decreases.

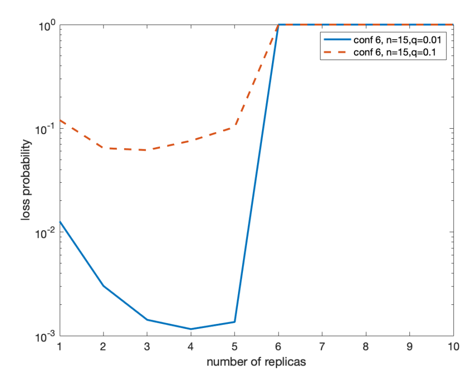

Fig. 1: Loss probability for the blind one shot case (conf 6).

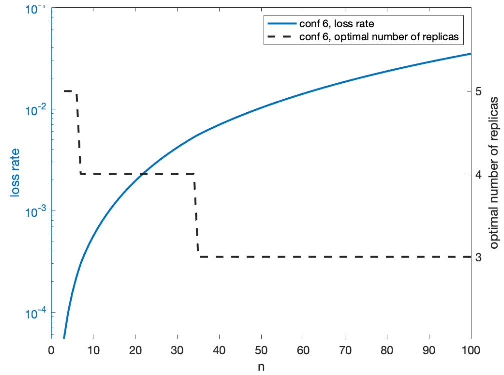

Fig. 2: Loss probability and the corresponding optimal number of replicas for the blind one shot case (conf 6, $q = 0.01$).

2) Repeated blind replication: In this case, we have proposed two schemes: p-persistent replication with $p \leq 1$ (proposition 2) or multiple replicas per slot (proposition 4).

Figure 3 illustrates the optimal policy for a system that corresponds to configuration 2. We concatenate the results of equations (3) and (8), for the cases of $p < 1$ and $p \in \mathcal{N}^*$, respectively and illustrate in Figure 3 the loss probability of three values of the number of concurrent devices n, equal to 20, 50 and 100 users, and a fixed activity factor of $q = 0.1$. We first observe that the two systems are equivalent when $p = 1$, as this corresponds to exactly one replica per slot. Second, for high traffic, p-persistent scheme with $p < 1$ is preferred while more aggressive policies ($p \in \mathcal{N}^*$) are better for low traffic.

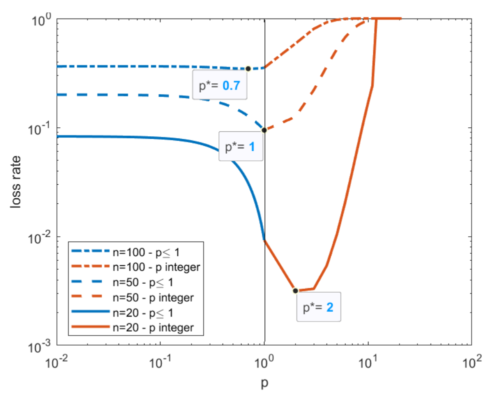

Fig. 3: Loss rate for the blind repeated case (conf 2, $q = 0.1$) as a function of p .

We now illustrate the optimal scheme performance for the two cases ($p \leq 1$ and $p \in \mathcal{N}^*$). Figure 4 illustrates the optimal value of p for both cases, We can see that, for p-persistent strategies, $p < 1$ when traffic is low and starts decreasing when traffic increases. In the case of multiple replicas per slot, the optimal number of replicas starts large, and then decreases to reach 1 when the number of UEs n increases. The corresponding loss probability is illustrated in Figure 5. There are clearly two regimes corresponding to when one of the two schemes is preferred, and the scheme with $p < 1$ is slightly better for high traffic regime, even if its performance is unacceptable for URLLC.

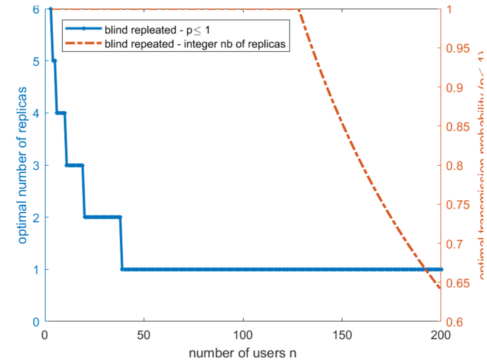

Fig. 4: Optimal policies for the blind repeated case (conf 2, $q = 0.1$), for the p-persistent and integer p cases, respectively.

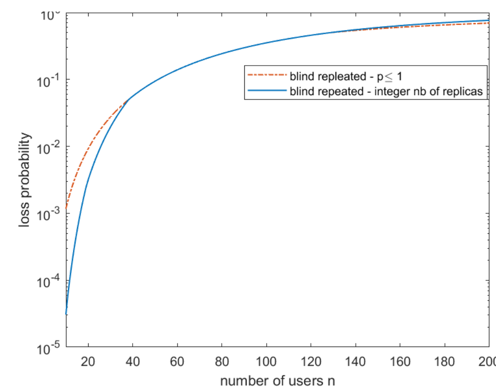

Fig. 5: Loss rate for the blind repeated case (conf 2, $q = 0.1$).

3) Impact of the feedback delay (far-sighted versus instantaneous feedback versus no feedback): We now move to the more general case where there is feedback that is received after some delay. We consider configuration 1 that is the most common in 4G and 5G system. In this case, when a user generates a packet, there are 4 consecutive slots for blind replication, followed by the reception of a feedback about the first replica, another feedback for the second replica, and then for the third replica.

In addition to the nominal case (delayed feedback, studied in proposition 7), two extreme cases are considered: the instant feedback (proposition 5) and the blind case (proposition 2).

We start by the p-persistent case ($p \leq 1$), and compare three cases. In addition to the nominal case (delayed feedback, studied in proposition 7), two extreme cases are considered: the instant feedback (proposition 5) and the blind case (proposition 2). Figure 6 show the loss probability corresponding to the optimal p for each case. The delayed feedback has an intermediate performance between both extreme cases, and having a feedback enhances performance as the devices become less aggressive, thus reducing the traffic load.

We now show in Figure 7 the loss rate for the nominal case (delayed feedback) for the two schemes: $p \leq 1$ and $p \in \mathcal{N}^*$. As expected, putting multiple replicas per slot is better for low traffic, the two schemes are equivalent for medium traffic (as the optimal scheme corresponds $p = 1$ for both) and the p-persistent with $p \leq 1$ is better for high traffic.

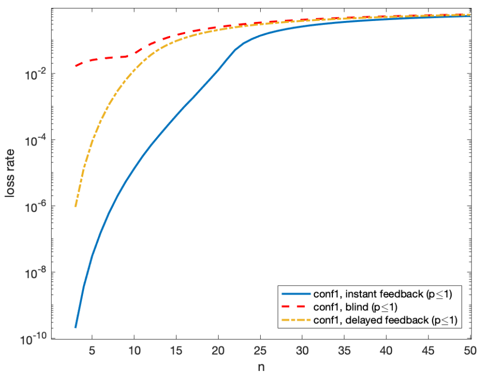

Fig. 6: Loss prob. for a delayed feedback with $p \leq 1$ (conf 1).

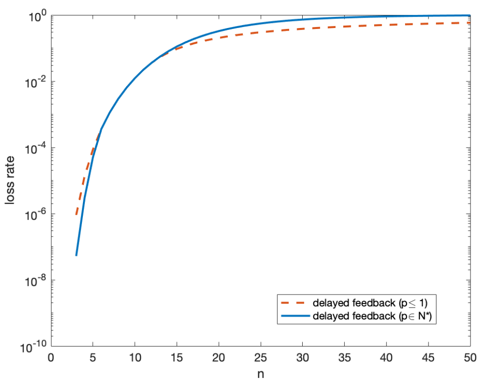

Fig. 7: Loss for a delayed feedback (conf 1).

B. Hints about 5G NR configuration performance

We illustrate in Figure 8 the performances for:

- a delayed feedback: configurations 1, 3 and 5.
- a repeated blind transmission: configurations 2 and 4.
- a one-shot blind transmission: configuration 6.

It is worth noting that these configurations do not offer the same amount of uplink resources to URLLC as they use different duplexing schemes. However, they are pairwise comparable (configurations 1 and 2, configurations 3 and 4, and configurations 5 and 6). For these couples of configurations, the only difference is on the numerology (sTTI size).

Figure 8 shows that the FDD configurations perform better than the TDD ones, as they reserve the whole bandwidth to the uplink. DDUU is genenarly better than DDDU for the TDD scheme, as there are more slots reserved for the uplink. When comparing the FDD configurations, a smaller slot that allows for a delayed feedback (conf.1) is preferred for low traffic. However, for larger traffic, configuration 2 that allows multiplexing more packets per slot achieves a better performance, even if this makes the transmitter blind. For the DDUU case, configuration 4 with a larger number of packets per slot achieves a better performance even if it is blind, because the only feedback in configuration 3 comes too late for the transmitter to be significantly less aggressive. Finally, for the DDDU case, the one-shot case of configuration 6 performs badly compared to the repeated case with delayed feedback, but both conf. 5 and conf. 6 do not achieve URLLC targets.

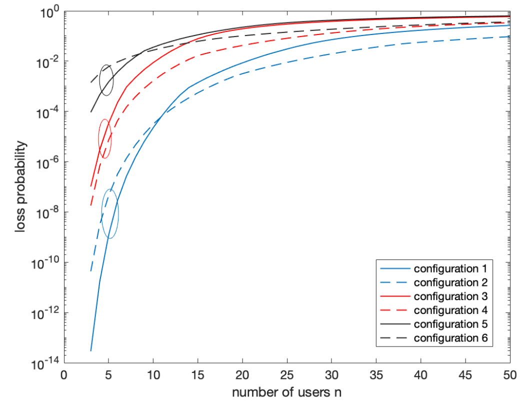

Fig. 8: Comparison of the loss for 5G NR configurations.

V. Conclusion

We study in this paper the problem of designing random packet replication schemes for IIoT, typically in the presence of many transmitters with sporadic traffic. As a function of possible 5G NR radio interface configurations, we show how to design these packet retransmission schemes based on the number of received ACKs before the delay budget expiration. We provide closed-form expressions for the loss probability in different scenarios, including blind replication, far-sighted replication and instant feedback allowing to derive optimal replication factors. We numerically test these policies on different 5G and Beyond configurations showing the impact of radio design to reach IIoT objectives.

References

[1] 3GPP. Physical layer procedures for data. 3GPP TR 38.214 v15.1.0.

[2] 3GPP. *Service requirements for next generation new services and markets*, 3 2018. V16.3.0.

[3] N. Abramson. The throughput of packet broadcasting channels. *IEEE Trans. on Communications*, 25(1):117–128, January 1977.

[4] Patrick Brown, Salah Eddine Elayoubi, Matha Deghel, and Ana Galindo-Serrano. Design and performance evaluation of contention-based transmission schemes for urllc services. *Performance Evaluation*, 143:102132, 2020.

[5] R. Combes, S. Elayoubi, and T. Varela. Optimal retransmission policies for urllc with delayed feedback. In *IEEE GLOBECOM*, 2019.

[6] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano. Radio resource allocation and retransmission schemes for URLLC over 5g networks. *IEEE JSAC*, 37(4):896–904, April 2019.

[7] W. Liang, M. Zheng, J. Zhang, H. Shi, H. Yu, Y. Yang, and S. Liu. Wia-fa and its applications to digital factory: A wireless network solution for factory automation. *Proceedings of the IEEE*, 107(6):1053–1073, 2019.

[8] Meriem Mhedhbi, Mira Morcos, Ana Galindo-Serrano, and Salah Eddine Elayoubi. Performance evaluation of 5g radio configurations for industry 4.0. In *IEEE WiMob*, 2019.

[9] Z. Naor and H. Levy. A centralized dynamic access probability protocol for next generation wireless networks. In *IEEE INFOCOM 2001*.

[10] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo. Contention-based access for ultra-reliable low latency uplink transmissions. *IEEE Wireless Communications Letters*, 7(2):182–185, April 2018.

[11] N. Strodthoff et al. Enhanced machine learning techniques for early harq feedback prediction in 5g. *IEEE JSAC*, 37(11), 2019.

[12] Zhiyi Zhou et al. Resource allocation for uplink grant-free ultra-reliable and low latency communications. In *IEEE VTC Spring*, 2018.

[13] Volker Ziegler and Seppo Yrjola. 6g indicators of value and performance. In *6G wireless summit*. IEEE, 2020.