# Uplink Contention-based Transmission Schemes for URLLC Services

Matha Deghel, Patrick Brown, Salah Eddine Elayoubi, Ana Galindo-Serrano, Ana Galindo

# Uplink Contention-based Transmission Schemes for URLLC Services

Matha Deghel
Orange Labs
Chatillon, France
matha.deghel@orange.com

Patrick Brown
Orange Labs
Sophia Antipolis, France
patrick.brown@orange.com

Salah Eddine Elayoubi
CentraleSupelec, L2S
Gif-Sur-Yvette, France
salaheddine.elayoubi@centralesupelec.fr

Ana Galindo-Serrano
Orange Labs
Chatillon, France
anamaria.galindoserrano@orange.com

## ABSTRACT

We investigate in this paper uplink multiple transmission schemes for 5G Ultra-Reliable Low Latency Communications (URLLC) traffic. The URLLC class of services has been defined for applications requiring extremely stringent latency and reliability. We show that, in systems with episodic traffic and many users compared with the number of transmission resources, randomly transmitting multiple copies of a packet allows to meet the URLLC requirements. We develop analytical models for the packet loss rate for two contention based multiple transmission schemes and show that one dominates the other in the parameter range for which the URLLC requirements are met. We then show on a possible radio setting for 5G, an example of radio resource dimensioning for different user traffic levels and we illustrate how the latency constraint may limit the allowable traffic for a given radio bandwidth.

## KEYWORDS

URLLC, grant-free transmissions, contention-based access, packet replicas, collision

**ACM Reference format:**
Matha Deghel, Patrick Brown, Salah Eddine Elayoubi, and Ana Galindo-Serrano. 2019. Uplink Contention-based Transmission Schemes for URLLC Services. In *Proceedings of Valuetools conference, Palma de Mallorca, Spain, March 2019 (Valuetools'19),* 8 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

In 5G networks, *Ultra-Reliable Low-Latency Communication (URLLC)* is the class of services with the most stringent latency and reliability requirements [7]. In the 3GPP (3rd Generation Partnership Project) standard, one set of URLLC requirements is a 99.999 % target reliability with a 1 ms (two-way) user-plane latency [2]. Decreasing the Transmission Time Interval (TTI) length is one efficient way to shorten the latency in the system [1, 6] but is not enough to attain those latency targets. This class of services is arguably the most challenging because guaranteeing low latency is conflicting with achieving ultra-high reliability using current Long-Term Evolution (LTE) solutions.

Two mechanisms contribute importantly in achieving reliability objectives for uplink transmissions in LTE systems: grant-based scheduling [9] to avoid collisions between User Equipments (UEs) transmitting to the same base-station (BS) and Hybrid Automatic Repeat Request (HARQ) retransmission procedures [9] for packets which have not been received correctly. In LTE, a User Equipment (UE) wishing to transmit a packet sends a scheduling request to the BS in dedicated resources which are available periodically, typically every 5 to 10 ms. The BS then determines an uplink schedule and returns a transmission grant to the UE. This mechanism does not allow achieving the latency required for URLLC. The HARQ retransmission procedure requires the BS attempting to decode the packet and then sending HARQ message to the UE in case the packet has not been decoded correctly. The latency introduced greatly depends on the decoding time which is variable and typically much longer when decoding fails.

In this context, to ensure fast uplink access, *grant-free* scheduling must be used, under which neither issuing a scheduling request nor waiting a scheduling grant are required [8]. Under this latter approach, two possible access schemes can be adopted depending on the nature of traffic to be carried. Specifically, if the traffic pattern is almost periodic and the number of users is less than the amount of resources in the system, a semi-persistent scheduling is the most suitable scheme. Under this scheme, each user has preallocated resources that repeat according to a predefined periodicity [4]. If, however, the packet arrivals are sporadic and/or the number of users exceeds the amount of resources, then contention-based access is the appropriate scheme to be exploited. In this case, the users contend to access a set of shared time and frequency resources which are preallocated for the contention procedures [1].

In this paper, we are interested in URLLC uplink transmissions with sporadic packet arrivals and a larger number of users than the number of resources for URLLC. In this context we consider grant-free scheduling with contention-based access. As the transmissions are grant-free and users do not have dedicated resources, collisions may occur. To still meet latency and reliability requirements, we


*Valuetools'19, March 2019, Palma de Mallorca, Spain*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

consider access schemes with preventive packet repetitions. This consists in sending multiple copies of the same packet to achieve the required reliability objective despite the occurrence of collisions. The transmission of multiple packets also allows reaching reliability objectives even when the radio conditions and the packet encoding are not sufficient to guarantee a correct reception probability without relying on the HARQ retransmissions.

We propose the use of a contention based access with repetitions randomly placed on a limited and predetermined number of resources. We derive the performance (in terms of collision probability and latency) of such a procedure and evaluate its performance gain compared to a more constrained procedure [10]. We also derive the exact performance of [10] where an approximate expression was presented. We show, for a typical set of system parameters, what resources would be required to respect a stringent collision probability and what would be the maximum offered load as a function of different latency constraints. Finally we analyse the reception success probability in presence of transmissions errors and show that multiple transmissions allow achieving URLLC reliability even in presence of contention and of packet transmission error rates several orders of magnitude higher than required for single transmissions.

The authors in [10] propose to place replicas in consecutive TTIs, where the resources used for each replica are randomly selected in each TTI. We show in the following that placing the replicas freely among the total set of available resources improves the performance. In addition it allows choosing the number of replicas to send, independently of the number of TTIs, and best adapted to the system parameters. Authors in [5] adopt a different approach from ours and evaluate the probability of decoding an uplink transmission when the resources are split into shared and dedicated parts. They make use of advanced receiver processing in order to satisfy the URLLC constraints. The approach of [5] mainly intends to combat radio errors as it allocates dedicated resources for each user in addition to shared resources where several users place their replicas. However this may not be feasible for a large number of users with sporadic traffic as in the scenario considered here. We also introduce the impact of radio errors and show how they impact our contention-based scheme. For satellite communications, [3] considers diversity transmission, by transmitting multiple copies of each generated packet, to improve the throughput of the slotted ALOHA random access scheme by randomly transmitting copies in future slots. In contrast to [3], we consider a bounded transmission period due to stringent delay constraints and we focus on the problem of how to efficiently transmit multiple copies on this time period.

The rest of the paper is structured as follows. In Section 2, we present the system model. The contention-based schemes are presented in Section 3, and the analysis is conducted in Section 4. The numerical results are given in Section 5. We finally draw conclusions in Section 6.

## 2 SYSTEM MODEL

We focus on uplink transmissions for 5G URLLC services. A system with a total of $N$ independent users having sporadic packet arrivals is considered. When a packet is generated, it has to be received within a delay constraint of $T$ ms, otherwise it is to be discarded, and the objective is a packet loss ratio less than a small value $\theta$. The packet has thus to carry on a timestamp and there is no need for packet reordering. In addition, a tight latency constraint would not allow waiting for ACKs. Let $\tau$ be the TTI length[1] and $p$ the probability that a user has at least one packet to transmit in a TTI, which we refer to as *activation probability*. A set of RBs are supposed to be available for uplink transmissions in each TTI. For the sake of simplicity, we assume that each packet transmission requires the same number of Resource Blocks (RBs), and we define the so-called Resource Unit (RU) composed from several RBs so that one packet occupies 1 RU. The amount of RBs per RU depends on the numerology (i.e. the size of the TTI and the subcarrier spacing), the spectral efficiency and the packet size. We will give in the numerical applications an example calculation of the number of RBs per RU. In the remainder of this paper, we consider that at set of $K$ RUs is reserved for URLLC transmissions per TTI.

We are interested in the case where the amount of resources reserved for URLLC uplink transmissions is less than the number of users in the system and where the traffic is sporadic (i.e. not periodic). As explained earlier, for these reasons we consider grant-free contention-based schemes. Grant-based access schemes are not suitable due to the tight latency constraint, while grant-free semi-persistent scheduling is not suitable for random packet arrivals and a number of users exceeding the amount of resources.

## 3 PROPOSED CONTENTION-BASED SCHEMES

The contention-based access solution is by definition a non-orthogonal solution, thus involving possible collisions among the data transmissions occurring in the same contention period. This may result in poor reliability performance, which implies an inability to support URLLC services. Increasing the resource pool reduces the collision probability very slowly, for example by a mere factor two if the number of resources are doubled. A more efficient approach to achieve high reliability in this case is by transmitting a packet multiple times instead of only once, i.e. sending multiple packet replicas. In this work, we adopt this approach and aim at proposing an efficient contention-based scheme that achieves high reliability while respecting tight latency constraints.

It should be noted that the impact of the channel will not be taken into consideration in the analysis: we assume that a collision among a number of packets results in the loss of these packets. This assumption can be seen as a worst-case scenario regarding the impact of packet collisions on the system performance. Indeed, in practice, even when colliding with other packets, there is a non-zero probability that the packet can be decoded correctly.

Define $\delta$ to be the number of TTIs during which the multiple packet transmissions can be done; $\tau\delta$ can be then seen as the contention interval (in ms) for a given packet. In this section, the following general assumption is made: if for a user there is a packet arrival in a given TTI, then the first transmission of this packet can be done in the next TTI.

In the following, we describe two approaches for the contention-based scheme with repetitions.

[1] We adopt here a general terminology for the TTI, including that of short TTI where only a subset of the TTI symbols is used for transmissions.

### 3.1 One Transmission per TTI

This approach is proposed in [10], and we will refer to it as OT. It consists in transmitting each data packet $\delta$ times, where each transmission is done in a different TTI. In detail, when a packet arrives in a given TTI, the corresponding user sends a replica of this packet in each of the next $\delta$ TTIs. For each of these transmissions, a random RU selection is adopted, i.e. the RU used is chosen randomly from the set of $K$ RUs that are available in each TTI; recall that a packet transmission requires only one RU.

The above approach uses multiple transmissions in time as a way to increase the reliability performance.

### 3.2 Random Transmissions

We propose an approach that provides better reliability performance than OT, according to our numerical results, when very low collision probabilities are targeted. Let us denote this approach by RT for *Random Transmissions*. Under RT, the packet is transmitted a number of times denoted $\beta$. The subset of $\beta$ RUs is chosen randomly from the set of $K\delta$ RUs available during the next $\delta$ TTIs following the packet arrival. Unlike for OT, it may be possible for the active user not to transmit a copy of the packet in one or more of the $\delta$ TTIs. Obviously, this will depend on the (random) choice of the $\beta$ RUs.

The main reasoning behind the approach here is to ensure better resource-selection flexibility than OT, so that greater reliability performance can be achieved. As explained above, this is accomplished by not constraining transmissions to be done in each of the $\delta$ TTIs.

## 4 ANALYTICAL MODELING

In this section, we derive the collision probability under each of the transmission strategies that we adopt. This probability is used as a measure of reliability performance of these strategies.

Note that, in order to satisfy reliability targets for URLLC, users are assigned a robust Modulation and Coding Scheme (MCS) that ensures a low Block Error Rate (BLER). This does not eliminate completely the errors. However, for the ease of presentation, we neglect in this section the impact of radio errors and focus on errors due to collisions only. We will introduce the impact of errors in section 5.3.2.

To compare both strategies we consider intervals of $\delta$ consecutiveTTIs each. We assume that any packet arriving in one of these intervals cannot be transmitted before the beginning of the subsequent interval. (We will see in section 5.2.2 that this assumptions has little impact on collision probabilities.) In other words, each interval can be seen as a contention-based access cycle of the packets that arrived in the preceding interval.

Note that since in each TTI there are $K$ RUs, then $K\delta$ RUs are available for each contention cycle. However, in our mathematical analysis, we assume that any active user participates in the contention cycle with only one packet, even if it has more than one packet arrival in the preceding $\delta$ TTIs. This latter assumption has negligible impact on the analysis since the probability of having more than one packet arrival for a user is very low. This observation will be validated by the numerical results.

We next provide the collision probability for each of the adopted schemes. Note that when multiple copies of a packet are sent, a collision occurs if all these copies collide with other transmissions. The collision rate is measured from a predefined-user perspective, given that this user has data to transmit. It is worth recalling that any user having packet arrival(s) in a time interval is an *active* user in the next interval.

### 4.1 Collision Probability under OT

Here, we provide the collision probability under OT. It is worth noting that in [10] (Equation (9) therein) only an approximation of this probability is given, as their formulae expresses the probability all $K$ transmissions collide with all those of one other user.

Let $P_{\mathrm{c}}^{\mathrm{OT}}$ denote the collision probabilities under OT. Define $\mathcal{E}_n$ to be the event of having $n$ active users, other than the predefined user.

Using the above definitions and considerations, $P_{\mathrm{c}}^{\mathrm{OT}}$ can be expressed as follows

$$\begin{aligned} P_{\mathrm{c}}^{\mathrm{OT}} &= \mathbb{P}\{\text{collision with predefined user}\} \\ &= \sum_{n=1}^{N-1} \mathbb{P}\{\text{collision with predefined user} \mid \mathcal{E}_n\}\mathbb{P}\{\mathcal{E}_n\}. \end{aligned} \quad (1)$$

Let $\binom{N}{n}$ be defined as the number of $n$-combinations of a set containing $N$ elements, i.e. $\binom{N}{n} = \frac{N!}{n!(N-n)!}$. The explicit expression of $P_{\mathrm{c}}^{\mathrm{OT}}$ is provided in following proposition.

PROPOSITION 4.1. *The collision probability under approach OT, where each packet is sent once in each of the $\delta$ TTIs, can be given as*

$$P_{\mathrm{c}}^{\mathrm{OT}} = \sum_{n=1}^{N-1} \left(1 - \left(1 - \frac{1}{K}\right)^n\right)^{\delta} \binom{N-1}{n} \times (1 - (1-p)^{\delta})^n \, (1-p)^{\delta(N-1-n)}. \quad (2)$$

PROOF. The probability that, in one TTI, an active user does not choose the same RU as the predefined user is $1 - \frac{1}{K}$. Suppose that event $\mathcal{E}_n$ occurs, i.e. there are exactly $n$ other active users. The probability that, in one TTI, there is *at least* one collision between the $n$ active users and the predefined user can then be given as

$$1 - \left(1 - \frac{1}{K}\right)^n. \quad (3)$$

Conditioned on the number of users, $\mathcal{E}_n$, these last probabilities are independent on different TTIs. From the above, the probability that in each of the $\delta$ TTIs there is at least one collision between any of the other $n$ active users and the predefined user can be expressed as

$$\mathbb{P}\{\text{collision with predefined user} \mid \mathcal{E}_n\} = \left(1 - \left(1 - \frac{1}{K}\right)^n\right)^{\delta}. \quad (4)$$

We now want to find $\mathbb{P}\{\mathcal{E}_n\}$. Event $\mathcal{E}_n$ occurs if exactly $n$ users are active, meaning that (i) $n$ users are active so that each has at least a packet arrival in the previous interval of $\delta$ TTIs, and (ii) $N-1-n$ other users have zero packet arrival in this time interval. Note that there are $\binom{N-1}{n}$ choices of $n$ active users among the other $N-1$ users. Recalling that $p$ is the probability that there is at least one packet arrival in a TTI, and there are $\delta$ TTIs before the corresponding contention cycle starts, the probabilities associated with (i) and (ii) are respectively $(1-(1-p)^{\delta})^n$ and $(1-p)^{\delta(N-1-n)}$. Hence, the above yields

$$\mathbb{P}\{\mathcal{E}_n\} = \binom{N-1}{n} (1 - (1-p)^{\delta})^n \, (1-p)^{\delta(N-1-n)}. \quad (5)$$

Combining (1), (4) and (5), the desired result holds. □

## 4.2 Collision Probability under RT

Let $P_c^{\mathrm{RT}}$ denote the collision probability under approach RT. Recall that under this approach, each active user selects $\beta$ RUs randomly from the set of $K\delta$ RUs that are available in the contention cycle. In a similar way to OT, the collision probability here is measured from a predefined-user perspective who is assumed to be active. The explicit expression of this probability is given in the following proposition.

PROPOSITION 4.2. *The collision probability under approach RT can be expressed as follows*

$$P_c^{\mathrm{RT}} = 1 - \sum_{l=1}^{\beta} (-1)^{l+1} \binom{\beta}{l} \left( 1 + (1-(1-p)^{\delta})(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1}. \quad (6)$$

PROOF. Here, we use the term 'slot' to denote any RU the predefined user selects. Define $\mathcal{A}_i$ to be the event that the $i$-th slot used by the predefined user is free, i.e. no (other) active user chooses this RU for its packet transmissions. We would like to express the probability that one of the $\beta$ slots is free, i.e. $\mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_\beta\}$. To this end, we determine the probability that a subset of $l$ slots is free. Note that in a set containing $\beta$ slots there are $\binom{\beta}{l}$ subsets of size $l$. All $l$ slots will be free if all other users are either not transmitting or non of their $\beta$ RUs fall in the $l$ slots. For a given user, this happens with probability

$$1 - (1-(1-p)^{\delta}) + (1-(1-p)^{\delta}) \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}, \quad (7)$$

where $1-(1-p)^{\delta}$ represents the probability that a user is active and $\frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}$ is the probability this user has not chosen any of the $l$ slots. The expression in (7) can be re-expressed as

$$1 + (1-(1-p)^{\delta})(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}). \quad (8)$$

Since there are $N-1$ other users, we can write

$$\mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_l\} = \left( 1 + (1-(1-p)^{\delta})(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1}. \quad (9)$$

Using the above, we conclude that

$$\mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_\beta\} = \sum_{l=1}^{\beta} (-1)^{l+1} \binom{\beta}{l} \mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_l\} =$$

$$\sum_{l=1}^{\beta} (-1)^{l+1} \binom{\beta}{l} \left( 1 + (1-(1-p)^{\delta})(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1}. \quad (10)$$

Therefore, the collision probability is

$$1 - \mathbb{P}\{\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_\beta\} =$$

$$1 - \sum_{l=1}^{\beta} (-1)^{l+1} \binom{\beta}{l} \left( 1 + (1-(1-p)^{\delta})(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1}. \quad (11)$$

This concludes the proof. □

## 4.3 Collision Probabilities for small $p$

We present linear approximations at the point $p = 0$ for the collision probabilities of RT and OT. It is difficult to compare equations (6) and (2). However we show that for small values of $p$ the collision probabilities with a given user for RT and OT, are both closely approximated by the probability that a single other user be active and that this other user occupies exactly the same RUs as the given user. As a consequence we show that RT offers a smaller collision probability for small activation probabilities $p$ than OT, when both schemes transmit the same number of copies $\delta$ (so that $\beta = \delta$). In our numerical experiences we find that in the range of collision probabilities acceptable for URLLC, i.e. $10^{-5}$, these linear approximations are sufficiently accurate.

We use the following notations to explicit the dependency on certain system parameters: $P_c^{RT}(p, \beta, N)$ and $P_c^{OT}(p, N)$.

From (6) and (2) we easily deduce, as expected, that collision probabilities are null when $p = 0$: $P_c^{RT}(0, \beta, N) = P_c^{OT}(0, N) = 0$.

To simplify notations we introduce two functions $f(p, \beta, N)$ and $g(p, N)$ to calculate derivatives:

$$f(p, \beta, N) = \sum_{l=0}^{\beta} (-1)^{l} \binom{\beta}{l} \left( 1 + p(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1},$$

$$g(p, N) = \sum_{n=1}^{N-1} \left( 1 - \left( 1 - \frac{1}{K} \right)^{n} \right)^{\delta} \binom{N-1}{n} p^{n} (1-p)^{(N-1-n)},$$

so that $P_c^{RT}(p, \beta, N) = f((1-(1-p)^{\delta}), \beta, N)$ and $P_c^{OT}(p, N) = g((1-(1-p)^{\delta}), N)$. Thus $\frac{dP_c^{RT}}{dp} = \delta(1-p)^{\delta-1} \frac{df}{dp}$ and $\frac{dP_c^{OT}}{dp} = \delta(1-p)^{\delta-1} \frac{dg}{dp}$.

Taking the derivatives of $f(p, \beta, N)$ and $g(p, N)$ in $p$ at the point $p = 0$:

$$\frac{df}{dp}(0, \beta, N) = (N-1) \sum_{l=0}^{\beta} (-1)^{l} \binom{\beta}{l} \left( -1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}} \right)$$

$$= (N-1) \sum_{l=0}^{\beta} (-1)^{l} \binom{\beta}{l} \left( \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}} \right) = (N-1) P_c^{RT}(1, \beta, 2),$$

and

$$\frac{dg}{dp}(0, N) = \left( 1 - \left( 1 - \frac{1}{K} \right)^{1} \right)^{\delta} \binom{N-1}{1} = \frac{N-1}{K^{\delta}} = (N-1) P_c^{OT}(1, 2).$$

We deduce the linear approximations at the point $p = 0$ for the collision probabilities of RT and OT:

$$P_c^{RT}(p, \beta, N) = P_c^{RT}(0, \beta, N) + p \frac{dP_c^{RT}}{dp}(0, \beta, N) + o(p)$$

$$= (N-1) \delta p P_c^{RT}(1, \beta, 2) + o(p),$$

and

$$P_c^{OT}(p, N) = P_c^{OT}(0, N) + p \frac{dP_c^{OT}}{dp}(0, N) + o(p)$$

$$= (N-1) \delta p P_c^{0T}(1, 2) + o(p),$$

where $o(p)$ is a term such that $\lim_{p \to 0} \frac{o(p)}{p} = 0$.

In each case we recognize the (approximate) probability one other user is active, $(N-1)\delta p$, multiplied by the probability of choosing exactly the same RUs, respectively $P_c^{RT}(1,\beta,2)$ and $P_c^{OT}(1,2)$. For example $P_c^{OT}(1,2)$, is the probability of collision in a system with one other user (i.e. $N=2$) which is active (i.e. $p=1$). This is consequently the probability of choosing the same RUs.

Finally to compare both probabilities when $\beta=\delta$ (one transmission per TTI in average) and $p$ is small, first notice that $P_c^{RT}(1,\delta,2)=\frac{1}{\binom{K\delta}{\delta}}$, the probability of choosing the same $\delta$ RUs out of $K\delta$ possible positions for a system with RT. Thus

$$\lim_{p\to 0}\frac{P_c^{RT}}{P_c^{OT}}=\frac{P_c^{RT}(1,\delta,2)}{P_c^{OT}(1,2)}=\frac{K^\delta}{\binom{K\delta}{\delta}}<1. \tag{12}$$

It thus is preferable to use RT when user activation probabilities, $p$, are small, and the gain of using RT with respect to OT is approximately given by the inverse of the ratio of possible RU positions.

The gain may be approximated, even for moderate $K$, by taking the limit of (12) when $K$ tends to infinity:

$$\frac{K^\delta}{\binom{K\delta}{\delta}}=\frac{K^\delta\delta!}{(K\delta)(K\delta-1)\dots(K\delta-\delta+1)}\xrightarrow[K\to\infty]{}\frac{\delta!}{\delta^\delta} \tag{13}$$

Table 1 shows the asymptotic reduction factor of the collision probability when using RT in comparison to OT when $p$ tends to zero for different values of $K$ and $\delta$. Recall that a lower factor means a larger gain of RT over OT. The factor when $K$ is infinite already gives the order of magnitude for the case where $K=6$. Furthermore we see that the reduction factor enhances when $\delta$ increases, resulting in a factor close to 0.1 when $\delta=4$ (a gain of 10 of RT over OT). Interestingly as we will see in the following section, this order of magnitude is observed for 2 TTIs (i.e. $\delta=2$) and $\beta=4$, resulting in a gain factor close to 0.1 without increasing the latency to 4 TTIs.

**Table 1: Asymptotic collision probability reduction factor of RT vs OT for small $p$ and $\beta=\delta$.**

| $K^\delta/\binom{K\delta}{\delta}$ | $\delta$ | | | |
|---|---|---|---|---|
| $K$ | 2 | 3 | 4 | 5 |
| 6 | 0.55 | 0.26 | 0.12 | 0.05 |
| 10 | 0.53 | 0.25 | 0.11 | 0.05 |
| 20 | 0.51 | 0.23 | 0.10 | 0.04 |
| $\infty$ | 0.50 | 0.22 | 0.09 | 0.04 |

## 5 NUMERICAL RESULTS AND DISCUSSIONS

In this section, we present numerical results to validate the analytical models and to show the performance of the RT and OT schemes. Recall that $p$ stands for the probability that a user has at least one packet arrival in a TTI. Also, note that $\delta$ corresponds to a certain latency budget, which clearly depends on the TTI length. We point out that for all the figures in this section a base-10 log scale.

### 5.1 Validation of the Analytical Model

The analytical models for RT and OT are validated using an ad hoc discrete time simulator. We simulate a single cell serving $N$ URLLC devices. Packets are replicated following the OT or RT strategy. We first perform simulations according to the cyclic transmission scheme as described in section IV. All uplink users accumulate packet transmission requests during a period of $\delta$ TTIs. After this period these packets and their replicas are transmitted during the following $\delta$ TTIs. Thus a cycle lasts $2\times\delta$ TTIs. Cycles overlap, since while packets are being transmitted in the previous cycle, new packets arrive for the next cycle. However cycles do not interfere with each other, as new packets will be transmitted after the previous cycle has finished. In this cyclic scheme the maximum packet delay is $2\times\delta$ TTIs. Replicas contend on the wireless medium and, for each generated packet, we track its replicas and check if they are in collision. We compute the loss probability as the proportion of packets whose replicas are all lost. Recall that in the analytical model we assume that an active user participates in the contention cycle with only one packet, even if it has more than one packet arrival in the preceding $\delta$ TTIs. Nevertheless, in the presented simulations we do not make such an assumption, and the effective number of packets is considered.

Furthermore, to reduce simulation times while obtaining small variances for small activation probabilities $p$, we use importance sampling. In a first step, simulations are performed a large number of times with $n$ active users, with $n$ ranging from 1 to $N$. Then in a second step, the resulting conditional collision probabilities are used to obtain the collision probabilities for different activation probabilities $p$, by multiplying by the probability of having $n-1$ additional active users with respect to the user of interest.

Figure 1 illustrates the analytical and simulated behavior of the collision probability for different values of $p$ under each of RT and OT. We take $N=30$ users, $\delta=2$ TTIs and $K=6$ RUs. For RT, the number of packet replicas is $\beta=4$. The curves in Figure 1 confirm that the above assumption has a negligible impact on the analysis and validate that the derived collision probability expressions, given in (2) and (6), fit with the simulation results.

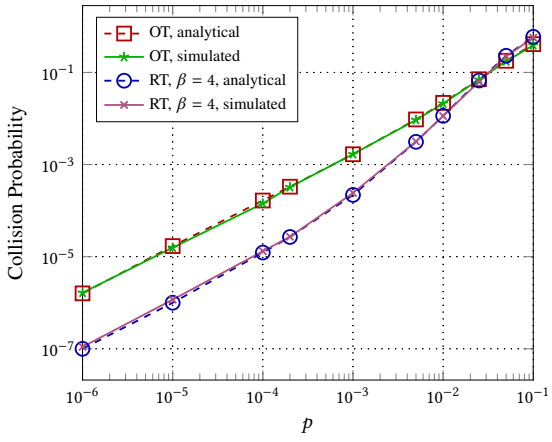


**Figure 1: Collision probability vs $p$ for OT and RT. Here, $N=30$, $\delta=2$, and $K=6$.**

### 5.2 Performance Comparison

Having validated the analytical model, we now turn to the comparison of the proposed scheme with other contention-based schemes.

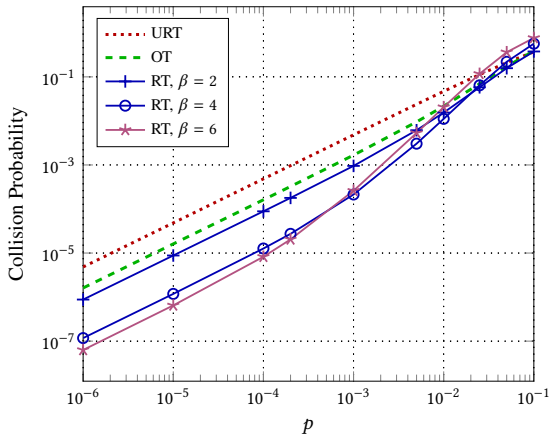


**Figure 2: Collision probability vs $p$ for URT, OT and RT. Here, $N = 30$, $\delta = 2$, and $K = 6$.**

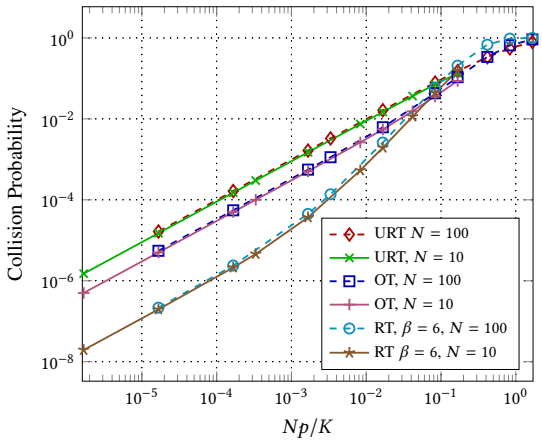


**Figure 3: Collision probability vs $Np/K$ for URT, OT and RT. Here, $\delta = 4$, $K = 6$, and $p = 10^{-6} : 10^{-1}$.**

*5.2.1 Comparing different resource allocation strategies.* Define URT (for Unique Random Transmission) to be the baseline scheme where only one packet replica is sent. This scheme can then be seen as a special case of RT for $\beta = 1$.

In Figure 2, we illustrate the reliability performance achieved by URT, OT and RT, by plotting the variation of the collision probability with respect to $p$. We consider the same setting as in the previous subsection. It can be seen that for $p < 10^{-2}$ and a number of packet replicas $\beta = 2$, 4 or 6, our proposed RT scheme is able to reach high reliability levels compared to the other schemes by producing lower collision probabilities. We can notice that the performance is initially enhanced when the number of replicas increases ($\beta = 2$, 4 or 6) but degrades when the transmission probability, $p$, increases beyond a certain value which depends on $\beta$. There is then an inflexion point. However this degradation occurs for transmission probabilities, $p$, for which none of the schemes allow reaching the low collision probabilities required for URLLC services.

We now turn to the sensitivity of the scheme to the load parameters, $N$ and $p$. Figure 3 depicts the collision probability for different values of the system load which is defined here as $Np/K$; i.e. it represents the ratio of the average packet arrivals in a TTI to the amount of resources in this TTI. Specifically, we consider $N = 10$ and 100, $p = 10^{-6} : 10^{-1}$, $\delta = 4$, $K = 6$. An important observation here is that, under each of URT, OT and RT, changing $N$ or $p$ has a negligible impact on the reliability performance as long as the offered load, i.e. the product $Np$, is the same. To close this paragraph, we compare the maximum load for which the target performance is achievable with the different schemes. If a collision probability of $10^{-5}$ is targeted, the baseline (URT) scheme can only support a load of $10^{-5}$, OT supports a maximal load of $3 \times 10^{-5}$, while RT can support a load of almost $10^{-3}$ (100 times larger than the baseline).

We now compare the ratio of collision probabilities for RT versus OT with the asymptotic expression (12) for small $p$, developed in section 4.3. Figure 4 compares the calculated ratio with the expression (12), for $K = 6$ and $\delta = \beta = 2$. It can be observed that there is a perfect match for low values of $p$ (smaller than $10^{-4}$) and that RT still stays preferable to OT when the load increases.

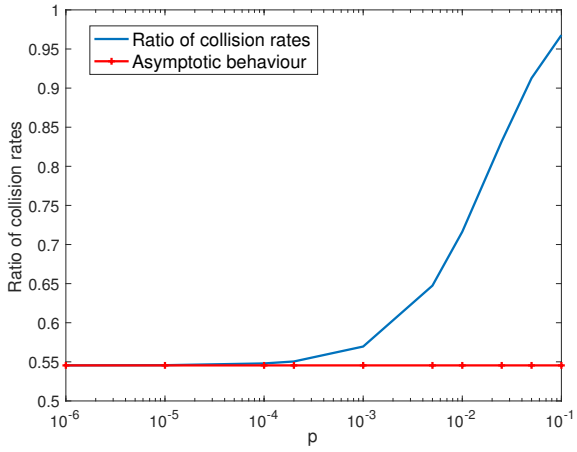


**Figure 4: Asymptotic behaviour of the ratio between collision rates of RT and OT for $K = 6$ and $\delta = \beta = 2$ as a function of $p$.**

*5.2.2 Cyclic vs Acyclic Scheme.* First, note that the simulations we have done until now are based on the scheme explained in Section 3, which we refer to as cyclic scheme. Recall that under this scheme the time is decomposed in intervals of $\delta$ TTIs each, and the packets arriving in one interval cannot be transmitted before the beginning of next interval. The acyclic scheme is defined as the scheme under which if there is a packet arrival in a TTI, the user can transmit this packet starting from the next TTI. This scheme will result in smaller latency as compared to the cyclic one.

In Figure 5, we present simulation results for the two considered schemes by depicting the collision probability as a function of $p$. We can see that the acyclic scheme achieves better performance in terms of collision probability than the cyclic one, but the difference is small. The analytical model developed in this paper can thus be used as a good approximation for both schemes.

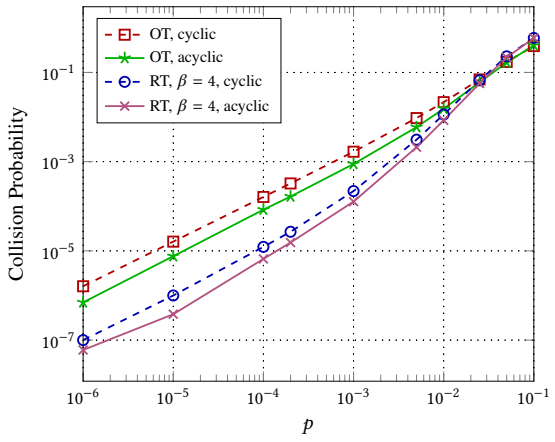


**Figure 5: Collision probability vs $p$ for OT and RT. Here, $N = 30$, $\delta = 2$, and $K = 6$.**

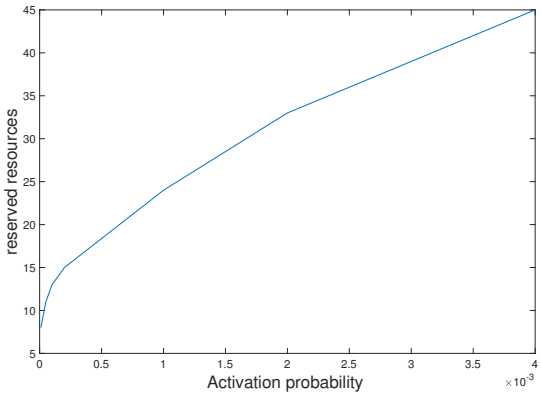


**Figure 6: Reserved resource units $Q$ for a given reliability target of $1 - 10^{-5}$.**

## 5.3 Detailed analysis of the performance of the proposed scheme

*5.3.1 Impact of the available spectrum and the target latency.* In the previous sections, we performed the evaluation based on the number of required RUs, supposing that the latency target is achievable. However, this is not always the case because of spectrum limitations. For a given amount of spectral resources, a tighter latency target limits the number of available TTIs $\delta$ for the transmission of the generated packet (leading to a smaller total number of resources $K\delta$). This makes meeting the reliability target more difficult for a larger number of users or a larger traffic density. We now show how to relate spectral resources, latency targets and feasible configurations (number of RUs).

For an application packet of size $b$ bits, a spectral efficiency of the used MCS of $\eta$ *bit/s/Hz*, a bandwidth per RB of $\omega$ and a TTI $\tau$, the resulting number of physical RBs, $R$, required for transmitting an application packet is:

$$R = \lceil \frac{b}{\eta \tau \omega} \rceil \tag{14}$$

One RU occupies then $R$ RBs. Recall that $K$ is the amount of resource allocation units per TTI; it is obtained by dividing the amount of available spectrum $W$ by the available amount of spectral resources per unit:

$$K = \lfloor \frac{W}{R\omega} \rfloor \tag{15}$$

When $Q$ resources are calculated to be needed for ensuring the target reliability, the latency constraint can be expressed as:

$$\lceil \frac{Q}{K} \rceil \leq \frac{T}{\tau} \tag{16}$$

This constraint is not achievable for all configurations $(W, \tau)$ and constraints (latency constraint $T$ in equation (16) and reliability constraint $\theta$ that determines $Q$).

In order to illustrate the performance of the proposed scheme in a realistic setting, we consider a 5G system with the parameters of table 2. We plot in Figure 6 the amount of RUs to be reserved for the contention-based pool for different values of the activation probability, $p$, to ensure a target reliability of $1 - 10^{-5}$. This is done by using the performance model of equation (6) and varying $K\delta$ until reaching a collision probability of $10^{-5}$ ($Q$ is the minimum $K\delta$ so that this constraint is satisfied). We can observe that, the more intense the traffic, $p$, the greater the amount of resources to be reserved. Starting from smaller values of $p$, the amount of RUs is first lower than the number of users $N$, showing a gain with respect to a deterministic allocation scheme where a RU is reserved for each user. However, when $p$ increases, this gain vanishes, showing that a grant-free contention-based scheme is efficient only for scenarios with low activity patterns.

We now introduce the latency constraint as the amount of reserved resources is constrained by the target latency due to (16). We plot in Figure 7 the maximum activation probability $p_{\max}$ that can be supported by the system as a function of the latency for a target reliability of $1 - 10^{-5}$. For each target latency, $p_{\max}$ corresponds to the maximum value of $p$ in Figure 7 that gives a number of resource units $Q$ that can fit within the corresponding number of TTIs. It is observed that a more stringent latency constraint (small $T$) reduces the supported load for the same amount of available spectrum.

**Table 2: System and service parameters**

| | |
|---|---|
| Application packet size, $b$ | 100 bits |
| Number of UEs, $N$ | 50 |
| Reserved bandwidth for URLLC service, $W$ | 2 MHz |
| Subcarrier spacing, $\omega$ | 15 KHz |
| Smallest time scheduling unit (TTI), $\tau$ | 0.144 ms (2 symbols per TTI) |
| Spectral efficiency of the selected MCS, $\eta$ | 1 bit/s/Hz |
| Reliability target, $\theta$ | $1 - 10^{-5}$ |
| Latency constraint, $T$ | 1 to 5 ms |

*5.3.2 Impact of radio errors.* As mentioned earlier, URLLC users are generally assigned a robust MCS that ensures a low error rate. However, some packets will still be lost even without collisions due to radio imperfections. Let $\gamma$ be the probability that a resource is

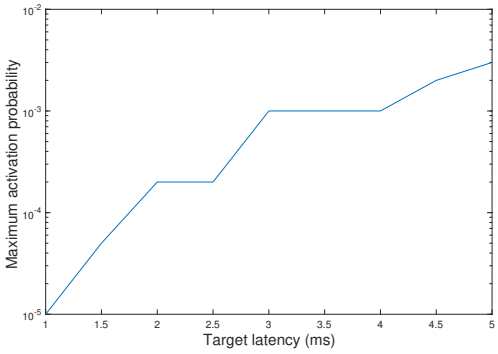


**Figure 7: Maximum traffic as a function of the latency target (reliability target of $1 - 10^{-5}$).**

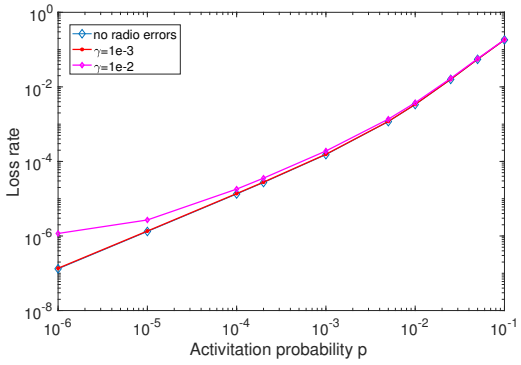


**Figure 8: Impact of radio errors on the loss rates.**

subject to degraded radio condition so that a replica that is transmitted on it would be lost even without collision. We have the following result.

PROPOSITION 5.1. *The loss probability integrating wireless errors for the random transmission scheme can be expressed by:*

$$P_\gamma^{\mathrm{RT}} = 1 - \sum_{l=1}^{\beta} (-1)^{l+1} \binom{\beta}{l} \left( 1 + (1-(1-p)^\delta)(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1} (1-\gamma)^l. \tag{17}$$

PROOF. We now 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 and this resource is not subject to a radio error. These events (occupancy and error) are independent. As before, we determine the probability that a subset of $l$ resources among the $\beta$ resources allocated to the target user is free. Since there are $N-1$ other users and errors are independent, the probability that all $l$ slots of this subset are collision-free and error-free, in the random case, is:

$$\mathbb{P}\{\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_l\} = \left( 1 + (1-(1-p)^\delta)(-1 + \frac{\binom{K\delta - l}{\beta}}{\binom{K\delta}{\beta}}) \right)^{N-1} (1-\gamma)^l.$$

Which leads to the expression (17). □

We now illustrate how the radio errors impact the loss rate. Figure 8 shows the loss performance when introducing radio errors, always with the same configuration (50 users, 24 reserved resources and 3 replicas per packet). However, for the usual target collision probability of $10^{-5}$, there is no significant impact, even if the radio error rate is as high as $\gamma = 10^{-2}$. Interestingly, this example shows that multiple transmissions may allow reaching the target reliability even in mediocre radio conditions, without relying on the HARQ retransmissions mechanism, thus attaining the required reliability while still respecting the latency constraint. We however advocate robust MCS for URLLC services, so that the radio error rate remains around $\gamma = 10^{-3}$ for sustaining reliabilities of $10^{-6}$ or $10^{-7}$.

## 6 CONCLUSIONS

In this paper we consider sporadic uplink transmissions for URLLC services. We combine grant-free contention-based transmissions with packet repetitions as a means to increase the reliability while respecting the latency budget. We explore contention-based schemes and develop an analytical model for the resulting collision probability. We validate this model through simulations and use it to design the transmissions strategies that allow meeting the URLLC requirements. In particular, we find that a strategy that allocates replicas randomly to available resources achieves better performance than a strategy that preallocates a specific replication pattern. We also introduced the impact of radio errors and show how they impact the performance. An important result of this paper is to show that very stringent reliability targets can be achieved with grant-free transmissions, without having to perform hard resource reservation per user.

## REFERENCES


[1] 3GPP. 2016. Study on latency reduction techniques for LTE. (3GPP TR 36.881 v14.0.0, Tech. Rep., June 2016).

[2] 3GPP. 2017. Study on scenarios and requirements for next generation access technologies. (3GPP TR 38.913 v14.2.0, Tech. Rep., June 2017).

[3] G. Choudhury and S. Rappaport. 1983. Diversity ALOHA - A Random Access Scheme for Satellite Communications. *IEEE Transactions on Communications* 31, 3 (March 1983), 450–457.

[4] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala. 2007. Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE System. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*. 2861–2864.

[5] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski. 2018. Uplink Transmissions in URLLC Systems with Shared Diversity Resources. *IEEE Wireless Communications Letters* (2018), 1–1.

[6] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen. 2017. Rethink Hybrid Automatic Repeat reQuest Design for 5G: Five Configurable Enhancements. *IEEE Wireless Communications* PP, 99 (2017), 2–8. https://doi.org/10.1109/MWC.2017.1600319

[7] P. Popovski. 2014. Ultra-reliable communication in 5G wireless systems. In *1st International Conference on 5G for Ubiquitous Connectivity*. 146–151.

[8] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch. 2017. Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture. *IEEE Communications Magazine* 55, 2 (February 2017), 70–78.

[9] Stefania Sesia, Issam Toufik, and Matthew Baker. 2011. *LTE, The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*. Wiley Publishing.

[10] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo. 2018. Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions. *IEEE Wireless Communications Letters* 7, 2 (April 2018), 182–185.