sensors

MDPI

Article

On the Capacity of 5G NR Grant-Free Scheduling with Shared Radio Resources to Support Ultra-Reliable and Low-Latency Communications

M. Carmen Lucas-Estañ *, Javier Gozalvez and Miguel Sepulcre

Department of Communications Engineering, Universidad Miguel Hernández de Elche (UMH), Avda. de la Universidad s/n, 03202 Elche, Spain

* Correspondence: m.lucas@umh.es; Tel.: +34-965-222-424

Received: 14 June 2019; Accepted: 10 August 2019; Published: 16 August 2019

Abstract: 5G and beyond networks are being designed to support the future digital society, where numerous sensors, machinery, vehicles and humans will be connected in the so-called Internet of Things (IoT). The support of time-critical verticals such as Industry 4.0 will be especially challenging, due to the demanding communication requirements of manufacturing applications such as motion control, control-to-control applications and factory automation, which will require the exchange of critical sensing and control information among the factory nodes. To this aim, important changes have been introduced in 5G for Ultra-Reliable and Low-Latency Communications (URLLC). One of these changes is the introduction of grant-free scheduling for uplink transmissions. The objective is to reduce latency by eliminating the need for User Equipments (UEs—sensors, devices or machinery) to request resources and wait until the network grants them. Grant-free scheduling can reserve radio resources for dedicated UEs or for groups of UEs. The latter option is particularly relevant to support applications with aperiodic or sporadic traffic and deterministic low latency requirements. In this case, when a UE has information to transmit, it must contend for the usage of radio resources. This can lead to potential packet collisions between UEs. 5G introduces the possibility of transmitting K replicas of the same packet to combat such collisions. Previous studies have shown that grant-free scheduling with K replicas and shared resources increases the packet delivery. However, relying upon the transmission of K replicas to achieve a target reliability level can result in additional delays, and it is yet unknown whether grant-free scheduling with K replicas and shared resources can guarantee very high reliability levels with very low latency. This is the objective of this study, that identifies the reliability and latency levels that can be achieved by 5G grant-free scheduling with K replicas and shared resources in the presence of aperiodic traffic, and as a function of the number of UEs, reserved radio resources and replicas K. The study demonstrates that current Fifth Generation New Radio (5G NR) grant-free scheduling has limitations to sustain stringent reliability and latency levels for aperiodic traffic.

Keywords: grant-free; scheduling; URLLC; ultra-reliable and low-latency communications; 5G; deterministic; time-critical; reliability; latency; aperiodic traffic; Industry 4.0

1. Introduction

5G networks are being designed with the objective to support a broad range of verticals such as manufacturing, transport, health, energy and entertainment. To this aim, important changes have been introduced to increase data rates (enhanced mobile broadband, or eMBB), efficiently support large amounts of devices (massive machine type communications, or mMTC) and guarantee unprecedented reliability and latency levels (Ultra-Reliable and Low-Latency Communications or URLLC) [1].

Supporting URLLC is particularly relevant for many Industry 4.0 manufacturing applications, such as motion control (requires a maximum latency of 1 ms and a reliability of $1–10^{-6}$ [2]), control-to-control applications (maximum latency of 4 ms and a reliability of $1–10^{-8}$ [1]) and factory automation (maximum latency between 0.25 ms and 2.5 ms and reliability requirements up to $1–10^{-9}$ [3]). These applications require the exchange of information between sensors, actuators and controllers through an industrial sensor and control network. 5G has the potential to provide the connectivity required by the Industry 4.0 to digitalize factories and to support data-intensive services while ubiquitously guaranteeing low latency and reliable connections. This has actually been acknowledged through the establishment of the 5G Alliance for Connected Industries and Automation (5G-ACIA) [4].

5G has introduced significant changes to support URLLC [5]. Some of these changes focus at the Radio Access Network level, since the medium access mechanisms account for an important part of the total end-to-end transmission delay [6]. This is for example the case of the grant-based scheduling process for uplink (UL) transmissions in legacy LTE (Long Term Evolution) 4G networks. Grant-based scheduling requires a User Equipment (UE) and a Base Station (BS) to exchange scheduling requests (SRs) and grant messages before transmitting any data. This process alone already results in an average delay of up to 11.5 ms when considering a Transmission Time Interval (TTI) equal to 1 ms and an SR periodicity of 10 ms [3]. Reducing the slot duration can reduce this delay. However, additional scheduling changes have been necessary to sustain the URLLC requirements that characterize some vertical applications, such as those in Industry 4.0. In particular, Release 15 and 16 of the 3rd Generation Partnership Project (3GPP) standards have introduced the concept of grant-free scheduling (also referred to as Configured Grant for 5G New Radio [7]) to support URLLC.

With grant-free scheduling, the BS reserves resources for UL transmissions and informs the UEs of the reserved resources. When a UE wants to initiate a UL transmission, it directly utilizes the reserved resources, without sending an SR and waiting for the subsequent grant message from the BS. Recent studies have shown that grant-free scheduling in 5G NR considerably reduces the end-to-end latency [8]. The 3GPP standards introduce the possibility for grant-free scheduling to reserve resources to dedicated UEs, or to a group of UEs. In the first case, each resource is reserved for a specific UE, and only this UE can utilize the resource at any time. This approach is adequate for periodic traffic since the resource allocations can be planned, and resources can then be utilized efficiently. Such planning is not possible in the case of aperiodic, sporadic or uncertain traffic. Sharing dedicated resources by a group of UEs is hence an interesting option to optimize the usage of the radio resources in the presence of aperiodic traffic. In this case, UEs have to contend for their usage, and collisions are possible. 5G NR introduces the possibility to transmit K replicas of the same packet in consecutive slots to combat potential collisions. However, relying on the transmission of K replicas to achieve a target reliability level can result in additional delays. It is yet unknown whether 5G NR grant-free scheduling with K-repetitions and shared resources can satisfy critical applications and guarantee very high reliability levels with very low latency. In this context, this study presents an in-depth analysis of the reliability and latency levels that can be achieved with existing 5G NR grant-free scheduling solutions as a function of the number of UEs, the number of reserved radio resources, and the number of replicas K. To this aim, the study analytically quantifies the probability of successfully delivering a packet when using grant-free scheduling with K-repetitions and shared resources. In addition, the study analyzes the impact of self-collisions. Self-collisions occur when a UE has to transmit a new packet, and the transmission of the K replicas of the previous packet has not finished. If this happens, the new packet must be stored, and its transmission is delayed until all replicas of the previous packet have been transmitted. This study demonstrates for the first time that self-collisions have a non-negligible impact upon the capacity of 5G NR grant-free scheduling to support stringent URLLC reliability and latency levels.

2. Related Work

The 5G NR standard introduces the use of grant-free scheduling (also referred to as Configured Grant [7]). With grant-free scheduling, the network pre-configures the radio resources and assigns them to UEs without waiting for UEs to request resources. UEs can utilize the pre-assigned resources as soon as they have data to transmit. This is in contrast to grant-based scheduling, where UEs must request access to radio resources through the transmission of Scheduling Requests (SR). The BS assigns the radio resources to the UEs and notifies them using grant messages. UEs must wait to receive these grant messages before transmitting any data. Grant-free scheduling eliminates all delays introduced by the handshaking present in grant-based scheduling. Grant-free scheduling also improves the energy consumption of the UEs, reduces their complexity, and decreases the signaling overhead compared with grant-based scheduling ([8,9]). Grant-free scheduling can assign dedicated or shared resources to the UEs. The BS decides whether resources are dedicated to specific UEs, or are shared by a group of UEs [10]. Reserving resources to dedicated UEs is an interesting approach when we can plan ahead what is the demand for resources. This is for example the case of periodic traffic. However, reserving resources to dedicated users can be highly inefficient if the traffic demand is uncertain or aperiodic, and it is not possible to anticipate when these resources will be needed. In this case, it is possible to share radio resources by a group of UEs. This option ensures a more efficient utilization of resources, and the possibility to satisfy URLLC communication requirements. However, users must contend for the resources, and collisions can happen if two or more UEs simultaneously contend for the same resources. 5G NR introduces the possibility of transmitting K replicas of the same packet in consecutive slots to combat collisions and thus increase the probability of a correct reception [11,12].

The study in [13] analyzes the performance of the K replicas scheme. The authors propose transmitting the first copy of a packet using dedicated resources, and the following replicas using shared resources. The proposal also exploits shared diversity and advanced receiver processing techniques to reduce the impact of packet collisions. The proposal achieves adequate reliability levels and reduces the number of reserved (shared) radio resources, compared to a configuration that reserves resources to dedicated UEs. The study in [14] also transmits the first copy of a packet using dedicated resources. However, it does not consider the transmission of K replicas of a packet. Instead, the authors propose to retransmit the original packet in a shared resource only if the first transmission is not successful. This requires a handshaking between the UEs and the BS to exchange acknowledgement messages. This handshaking increases the latency, and can compromise the capability to adequately support URLLC applications with stringent latency requirements. In [15], the authors study the optimum number of replicas (K) necessary to achieve a target reliability level within a deterministic latency deadline. The study focuses upon aperiodic traffic and the case in which a group of UEs share resources. The authors show that randomly choosing the resource for each replica increases the probability of correctly delivering a packet. However, the study focuses on reliability levels up to $1–10^{-5}$ while some critical Industry 4.0 applications require higher reliability levels.

Previous studies have shown that transmitting K-repetitions of a packet increases the reception rate. However, this can be done at the expense of an inefficient use of the radio resources due to packet collisions or the unnecessary reservation of resources when the first replicas are correctly delivered. Latency requirements may also impose restrictions on the number of replicas that can be transmitted, and consequently on the reliability levels that may be achieved. In this context, several recent contributions have analyzed slight modifications to the K-repetitions scheme. For example, [16] proposed adaptively configuring the number of replicas transmitted based on the channel conditions. The objective is to utilize the radio resources efficiently by avoiding unnecessary retransmissions when the channel quality is good. A similar objective is sought in [17] where authors propose conditions to stop the transmission of replicas. Other interesting proposals in 3GPP standardization working groups include: the transmission of replicas within mini-slots (to reduce the latency) [18], the possibility for transmitting replicas across the slot border, or the concept of periodicity boundary [19]. These studies propose interesting variants of the K-repetitions scheme. However, it is yet unknown

whether 5G NR grant-free scheduling with K-repetitions and shared resources can really support URLLC communications with strict reliability and latency requirements under the presence of aperiodic or sporadic traffic. This traffic is critical in many verticals, for example in Industry 4.0. In this context, this study conducts an in-depth evaluation of 5G NR grant-free scheduling with K-repetitions and shared resources in the presence of aperiodic or sporadic traffic. The study identifies the reliability and latency levels that can be achieved with 5G NR grant-free scheduling, and identifies its current limitations. The study analyzes the impact of the number of UEs in the network, the number of reserved radio resources, and the number of replicas K. The study also analyzes for the first time the impact of self-collisions. The conducted analysis helps to identify the reliability and latency levels that can be achieved based on network deployments and configuration options for 5G NR grant-free scheduling.

It should be noted that 3GPP standards define the possibility of utilizing grant-free scheduling and transmitting K replicas, but do not define a specific scheme to be implemented. This study is based on the implementation of 5G NR grant-free scheduling with K replicas and shared resources proposed in [15]. This implementation is chosen because it has been specifically designed to guarantee stringent URLLC latency and reliability requirements. To this aim, the implementation transmits original packets and all of the replicas using grant-free scheduling on shared radio resources. A different approach is proposed in [13] where dedicated resources are used to transmit the original packets, and shared resources are used for the following replicas. This approach can increase the delay compared to [15] if grant-based scheduling is utilized to allocate the dedicated resources. The efficient utilization of resources could also be compromised if dedicated resources were reserved for each UE when supporting applications with aperiodic traffic. The implementation of 5G NR grant-free scheduling with K–repetitions and shared resources proposed in [15] is therefore better suited to support URLLC applications with aperiodic or sporadic traffic.

3. Grant-Free Scheduling

This paper uses grant-free scheduling with K-repetitions and shared resources to evaluate the reliability and latency levels that can be achieved in the presence of aperiodic traffic. Following [20], reliability for URLLC services is defined as the percentage of data packets that are successfully delivered before the latency deadline L established by the service or application. Following 3GPP standards [11], UEs transmit the same data packet in K consecutive transmission slots with a duration T_{slot}. The UE randomly selects an RB (Resource Block) for each transmission from the U RBs available per T_{slot}. This is illustrated in Figure 1 that represents the time/frequency resource grid map in 5G NR, where the unit is an RB. In 5G NR, a wideband channel is divided into sub-frames, slots and RBs. An RB is the smallest unit of frequency resources that can be allocated to a UE. Without loss of generality, this study considers a numerology μ equal to 3 with a subcarrier spacing of 120 kHz [21]. An RB is then 1440 kHz (Δf) wide in frequency (12 sub-carriers of 120 kHz) and lasts for one time slot with the duration T_{slot} equal to 0.125 ms.

The reliability at the medium access level that can be achieved with grant-free scheduling with K-repetitions and shared resources depends upon two main factors. The first factor is the possibility that a packet is not correctly received due to the collision of all its K replicas with other transmissions; this is due to the random selection of the RB for the transmission of each replica. The study in [15] showed that the possibility to successfully deliver a packet increases with the number K of replicas. The second factor is the effect of self-collisions. A self-collision occurs when a UE has to transmit a new packet, and the transmission of the K replicas of the previous packet has not finished. If this happens, the new packet must be stored, and its transmission is delayed until all the replicas of the previous packet have been transmitted. This delay can result in the case that the new packet cannot be delivered within the latency limit, and hence self-collisions can impact the reliability of URLLC services. It is important then that the reliability (or probability that a packet is correctly received before the latency deadline) of grant-free scheduling with K-repetitions and shared radio resources is computed considering both the effect of collisions from other UEs, and the effect of self-collisions.

In this case, the reliability or probability P_{rel} that a packet is correctly received by the BS must consider the probability P_{sc} that the transmission of the K replicas of a packet is not completed before the latency deadline L due to the effect of self-collisions. For the packets that are not affected by the effect of self-collisions, it must be considered the probability P_c that a packet is not correctly received due to the collision of all its K replicas with other transmissions. Hence, P_{rel} can be expressed as:

$$P_{rel} = 1 - (P_{sc} + (1 - P_{sc}) \cdot P_c) \quad (1)$$

In [15], its authors presented an expression to approximate the probability P_c of the collision of the K replicas of a packet with the transmission of other UEs. The expression was derived in scenarios where N UEs share the same pool of RBs. However, [15] did not analyze the impact of self-collisions, since the study only considered low values of K (equal to or lower than 4). For these low values, self-collisions might not have an impact upon the reliability, as will be later shown. In this paper, we analytically derive the exact probability of any collision of the K replicas of a packet with packets transmitted by other UEs (P_c). We also quantify the impact of self-collisions (P_{sc}), and analytically compute the reliability that can be achieved by grant-free scheduling with K-repetitions and shared resources (P_{rel}). These analytical expressions are a valuable contribution to the community since they can be easily utilized to evaluate 5G NR grant-free scheduling. The availability of these exact analytical expressions is particularly useful when considering applications with very demanding reliability and latency URLLC requirements. This is the case of certain Industry 4.0 applications. For example, motion control requires a maximum latency of 1 ms and a reliability of $1–10^{-6}$. Control-to-control applications require a maximum latency of 4 ms and a reliability of $1–10^{-8}$. Factory automation applications usually demand maximum latency values in the range 0.25–2.5 ms and reliability levels up to $1–10^{-9}$. In this case, simulations can be very computationally expensive if we want to compute the packet reception rate $(1 - P_c)$ with reliability demands in the order of $1–10^{-6}$ to $1–10^{-9}$. In these scenarios, errors are very rare, and we need long and computationally expensive simulations to achieve accurate results. The analytical methodology utilized in this study is then an adequate and efficient tool for scenarios with demanding URLLC communication requirements.

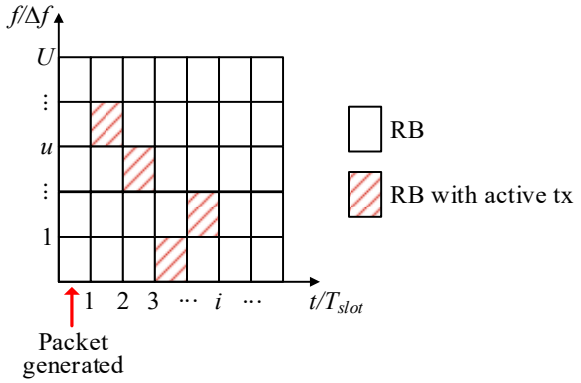

Figure 1. Illustration of the Fifth Generation New Radio (5G NR) resource grid map: Transmission of a data packet with four repetitions and a random selection of Resource Blocks (RBs) per slot.

3.1. Collisions with Other UEs

First, we focus on the probability P_c that a packet is not correctly received due to the collisions of its K replicas with the packets transmitted by other UEs. To this end, we consider UL transmissions and N UEs within a single cell with aperiodic traffic. Packets are generated by each UE following a Poisson distribution with exponential inter-arrival time. The average packet inter-arrival time is equal to $1/\lambda$, where λ is the average number of packets generated per second. We consider the transmission of small packets with a size of 32 bytes [22], and we assume without loss of generality that each packet requires only one RB.

The probability P_g that one or more packets are generated for a UE in a time period T_{slot} is equal to:

$$P_g = 1 - \exp(-T_{slot} \cdot \lambda) \tag{2}$$

We define R_i as the set of UEs for which a new packet could be generated in a slot s_i (the slot has a time duration equal to T_{slot}). Here, n_i is the number of UEs that do have a new packet to transmit in s_i. This n_i can then take any value between 0 and the cardinality of R_i. The probability $P_{tx}(n_i, R_i)$ that n_i UEs from the set R_i of UEs have new packets to be transmitted in s_i with duration T_{slot} is equal to:

$$P_{tx}(n_i, R_i) = \begin{pmatrix} |R_i| \\ n_i \end{pmatrix} \cdot {P_g}^{n_i} \cdot \left(1 - P_g\right)^{|R_i| - n_i} \tag{3}$$

where $|R_i|$ represents the number of elements or the cardinality of the set R_i.

A packet will not be successfully delivered to the BS if all its K replicas collide with the trans-missions of other UEs. A UE has an active transmission in s_i if it generated a new data packet in the previous slots $s_{i-(K-1)}, \ldots, s_{i-1}$, and s_i. If this is the case, then the UE would be transmitting one of the K replicas in s_i. We denote as n_i^{act} the number of UEs with active transmissions in s_i. The probability $P_{nrc}(n_i^{act}, U)$ that n_i^{act} UEs do not collide with a given UE is equal to the probability that they do not select the same RB at a given slot for their next transmission as the UE under study. $P_{nrc}(n_i^{act}, U)$ is given by:

$$P_{nrc}(n_i^{act}, U) = \begin{pmatrix} U - 1 \\ U \end{pmatrix}^{n_i^{act}} \tag{4}$$

Equations (2)–(4) are necessary to compute the probability P_c that a packet is not correctly received at the BS due to the collision of all its K replicas with the transmissions of other UEs. To compute P_c, let us consider the case of a particular UE_1 that has to transmit the K replicas of a packet in slots s_i, $s_{i+1}, \ldots, s_{i+K-1}$. For the sake of clarity, we consider an example with $K = 4$, and s_i corresponding to s_3. P_c is then equal to the probability of collision of the 4 replicas transmitted in s_3, s_4, s_5, and s_6, which is represented by $P_{rc}(s_3, s_4, s_5, s_6)$:

$$P_c = P_{rc}(s_3, s_4, s_5, s_6) \tag{5}$$

To determine $P_{rc}(s_3, s_4, s_5, s_6)$, we first study the probability $P_{rc}(s_3)$ that the replica of the packet transmitted in s_3 collides with a transmission from any other UE. $P_{rc}(s_3)$ is given by the probability that one or more UEs (in addition to UE_1) have an active transmission in s_3 (i.e., $n_3^{act} \geq 1$), and that one or more of the n_3^{act} UEs select the same RB as UE_1 for their transmission. n_3^{act} is equal to $n_0 + n_1 + n_2 + n_3$, and the probability $P_{rc}(s_3)$ has to consider all possible combinations of n_0, n_1, n_2 and n_3 that result in $n_3^{act} \geq 1$. The probability $P(n_3^{act} \geq 1)$ can then be expressed as:

$$P(n_3^{act} \geq 1) = \sum_{n_0 = n_0^{min}}^{n_0^{max}} \left\{ P_{tx}(n_0, R_0) \cdot \sum_{n_1 = n_1^{min}}^{n_1^{max}} \left\langle P_{tx}(n_1, R_1) \cdot \sum_{n_2 = n_2^{min}}^{n_2^{max}} \left[P_{tx}(n_2, R_2) \cdot \sum_{n_3 = n_3^{min}}^{n_3^{max}} P_{tx}(n_3, R_3) \right] \right\rangle \right\} \tag{6}$$

where n_i^{max} and n_i^{min} represent the maximum and minimum possible values of n_i in each slot, and are equal to:

$$n_i^{max} = |R_i|, \ \forall i \leq 3 \tag{7}$$

$$n_i^{min} = \begin{cases} 1 & \text{if } i = 3 \,\&\, |R_i| = N - 1 \\ 0 & \text{otherwise} \end{cases}, \ i \leq 3 \tag{8}$$

where R_i is the set of UEs that could have a new packet to be transmitted in s_i. R_i is equal to the total number of UEs (N) minus UE_1 and all active UEs in the slot previous to s_i. The cardinality of R_i is then equal to:

$$|R_i| = N - 1 - \sum\nolimits_{j = \max\{i-3, 0\}}^{i-1} n_j, \ i \leq 3 \tag{9}$$

It should be noted that n_i^{min} is equal to 0 or 1 in order to guarantee that n_3^{act} is equal to or higher than one. n_i^{act} can be expressed as:

$$n_i^{act} = \sum_{j=\max\{i-3,0\}}^{i} n_j,\ i \leq 3 \tag{10}$$

To achieve finally the expression of $P_{rc}(s_3)$, we need to incorporate to the expression of $P(n_3^{act} \geq 1)$ in (6) the probability that one or more of the n_3^{act} UEs select the same RB as UE_1 for their transmissions. This probability is equal to $1 - P_{nrc}(n_3^{act}, U)$. $P_{rc}(s_3)$ is then calculated as:

$$P_{rc}(s_3) = \sum_{n_0=n_0^{min}}^{n_0^{max}} \left\{ P_{tx}(n_0, R_0) \cdot \sum_{n_1=n_1^{min}}^{n_1^{max}} \left\langle P_{tx}(n_1, R_1) \cdot \sum_{n_2=n_2^{min}}^{n_2^{max}} \left[P_{tx}(n_2, R_2) \cdot \sum_{n_3=n_3^{min}}^{n_3^{max}} \left\{ P_{tx}(n_3, R_3) \cdot (1 - P_{nrc}(n_3^{act}, U)) \right\} \right] \right\rangle \right\} \tag{11}$$

The probability of collision of the replica transmitted in s_4 depends upon the number n_4^{act} of UEs with active transmissions in s_4. This n_4^{act} depends on the number n_1, n_2, n_3 and n_4 of UEs that have new packets to transmit in s_1, s_2, s_3, and s_4, respectively. The probability that UEs have new packets to transmit in s_1, s_2, and s_3 is already included in (11) ($P_{tx}(n_1, R_1)$, $P_{tx}(n_2, R_2)$, and $P_{tx}(n_3, R_3)$ respectively). In this context, $P_{rc}(s_3)$ and $P_{rc}(s_4)$ are not independent, and they must be calculated jointly. We then compute the joint probability $P_{rc}(s_3, s_4)$ that the replicas transmitted in s_3 and s_4 collide with transmissions from other UEs. Computing $P_{rc}(s_3, s_4)$ only requires including in (11) the probability that there are UEs with new packets to be transmitted in s_4 (i.e., $P_{tx}(n_4, R_4)$), and the probability that one or more of the active n_4^{act} UEs in s_4 select the same RB for their transmission than UE_1. $P_{rc}(s_3, s_4)$ can then be expressed as:

$$\begin{aligned} P_{rc}(s_3, s_4) = \sum_{n_0=n_0^{min}}^{n_0^{max}} \Bigg\{ P_{tx}(n_0, R_0) \cdot \sum_{n_1=n_1^{min}}^{n_1^{max}} \Bigg\langle P_{tx}(n_1, R_1) \cdot \sum_{n_2=n_2^{min}}^{n_2^{max}} [P_{tx}(n_2, R_2) \cdot \\ \sum_{n_3=n_3^{min}}^{n_3^{max}} \Bigg\langle \Big\{ P_{tx}(n_3, R_3) \cdot (1 - P_{nrc}(n_3^{act}, U)) \Big\} \cdot \sum_{n_4=n_4^{min}}^{n_4^{max}} \Big\{ P_{tx}(n_4, R_4) \cdot (1 - P_{nrc}(n_4^{act}, U)) \Big\} \Bigg\rangle \Bigg] \Bigg\rangle \Bigg\} \end{aligned} \tag{12}$$

where n_4^{act}, $|R_4|$, n_4^{max} and n_4^{min} are defined as:

$$n_4^{act} = \sum_{j=1}^{4} n_j \tag{13}$$

$$|R_4| = N - 1 - \sum_{j=1}^{3} n_j \tag{14}$$

$$n_4^{max} = |R_4| \tag{15}$$

$$n_4^{min} = \begin{cases} 1 & \text{if } |R_4| = N - 1 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

The process followed to account for possible collisions of the replicas transmitted in s_5 and s_6 is similar to that considered for s_4. P_c can then be expressed as follows when K = 4:

$$\begin{aligned} P_c = \sum_{n_0=n_0^{min}}^{n_0^{max}} \Bigg\{ P_{tx}(n_0, R_0) \cdot \sum_{n_1=n_1^{min}}^{n_1^{max}} \Bigg\langle P_{tx}(n_1, R_1) \cdot \sum_{n_2=n_2^{min}}^{n_2^{max}} [P_{tx}(n_2, R_2) \cdot \\ \sum_{n_3=n_3^{min}}^{n_3^{max}} \Bigg\langle \Big\{ P_{tx}(n_3, R_3) \cdot (1 - P_{nrc}(n_3^{act}, U)) \Big\} \cdot \sum_{n_4=n_4^{min}}^{n_4^{max}} \Big[\Big\{ P_{tx}(n_4, R_4) \cdot (1 - P_{nrc}(n_4^{act}, U)) \Big\} \cdot \\ \sum_{n_5=n_5^{min}}^{n_5^{max}} \Bigg\langle \Big\{ P_{tx}(n_5, R_5) \cdot (1 - P_{nrc}(n_5^{act}, U)) \Big\} \cdot \sum_{n_6=n_6^{min}}^{n_6^{max}} \Big\{ P_{tx}(n_6, R_6) \cdot (1 - P_{nrc}(n_6^{act}, U)) \Big\} \Bigg\rangle \Big] \Bigg\rangle \Bigg\rangle \Bigg\} \end{aligned} \tag{17}$$

where n_i^{act}, $|R_i|$, n_i^{max} and n_i^{min} $\forall i \in [0, 2{\cdot}K{-}1]$ are defined as:

$$n_i^{act} = \sum_{j=\max\{i-(K-1),0\}}^{i} n_j \tag{18}$$

$$|R_i| = N - 1 - \sum_{j=\max\{i-(K-1),0\}}^{i-1} n_j \tag{19}$$

$$n_i^{max} = |R_i| \tag{20}$$

$$n_i^{\min} = \begin{cases} 1 & \text{if } i \geq K-1 \ \& \ |R_i| = N-1 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

The process illustrated for $K = 4$ can be followed to compute P_c for any value of K. As shown in (22), P_c can be computed using the auxiliary function $h_i(K, N, U)$ defined in (23) with i equal to cero. To simplify the notation, $h_i(K, N, U)$ is also represented as h_i in (22) and (23). As it can be observed in (23), h_0 depends on h_1, and in general, h_i depends on h_{i+1}, until h_{2K-1}.

$$P_c(K, N, U) = h_0(K, N, U) = h_0 \tag{22}$$

$$h_i = \begin{cases} \sum\limits_{n_i=n_i^{min}}^{n_i^{max}} [P_{tx}(n_i, R_i) \cdot h_{i+1}] & \text{if } i \in [0, K) \\ \sum\limits_{n_i=n_i^{min}}^{n_i^{max}} \left[P_{tx}(n_i, R_i) \cdot (1 - P_{nrc}(n_5^{act}, U)) \cdot h_{i+1}\right] & \text{if } i \in [K, \ 2 \cdot K - 1) \\ \sum\limits_{n_i=n_i^{min}}^{n_i^{max}} \left[P_{tx}(n_i, R_i) \cdot (1 - P_{nrc}(n_5^{act}, U))\right] & \text{if } i = 2 \cdot K - 1 \end{cases} \tag{23}$$

The parameters n_i^{act}, $|R_i|$, n_i^{max} and n_i^{min} in (23) correspond to those expressed in (18)–(21).

3.2. Self-Collisions

The effect of self-collisions is illustrated in Figure 2. We may suppose that a UE starts transmitting a packet p_1 that was generated before t_0. Let us then suppose then that a second packet p_2 is generated before the K replicas of the previous packet p_1 have been transmitted. This is a self-collision. If a self-collision happens, p_2 can be stored, and its transmission will start after the UE has transmitted the K^{th} replica of p_1 (i.e., at t_1 in Figure 2). The transmission of the K replicas of p_2 will finish at t_2 that is equal to:

$$t_2 = 2 \cdot K \cdot T_{slot} + t_0 \tag{24}$$

The transmission of the K replicas of p_2 may finish after the latency deadline L, due to the time p_2 being stored as the K replicas of p_1 are being transmitted. We then analyze the probability P_{sc} that the transmission of K replicas of a packet is not completed before L due to the effect of self-collisions. This probability depends upon the number of replicas K and on the time instant at which p_2 was generated. Figure 2 illustrates how self-collisions affect the probability of completing the transmission of p_2 before L, with L equal to 1 ms. L = 1 ms implies that the maximum number of replicas K that can be transmitted per packet is 8. However, it is possible to transmit less than 8 replicas, and Figure 2 represents the case in which K is set equal to 4, 6 or 8. p_2 can be transmitted before the deadline L if it is generated at any time instant after $t_2 - L$, where t_2 is the time at which the transmission of the K replicas of p_2 is finished (the transmission of p_2 starts when the transmission of the K replicas of p_1 has finished at t_1). If p_2 is generated before $t_2 - L$, it is not possible to complete the transmission of the K replicas of p_2 before the latency deadline L. P_{sc} can then be computed as the probability that the time between the generation of two consecutive packets at a UE falls within the interval $[0, \Delta t]$, where

Δt represents the time difference between $t_2 - L$ and the time t_{p_1} at which p_1 is generated (see (26)). P_{sc} can then be expressed as:

$$P_{sc}(\Delta t) = \int_0^{\Delta t} \lambda \cdot e^{-t \cdot \lambda} \cdot \mathrm{d}t \tag{25}$$

$$\Delta t = t_2 - L - t_{p_1} = 2 \cdot K \cdot T_{slot} - L - t_{p_1} \tag{26}$$

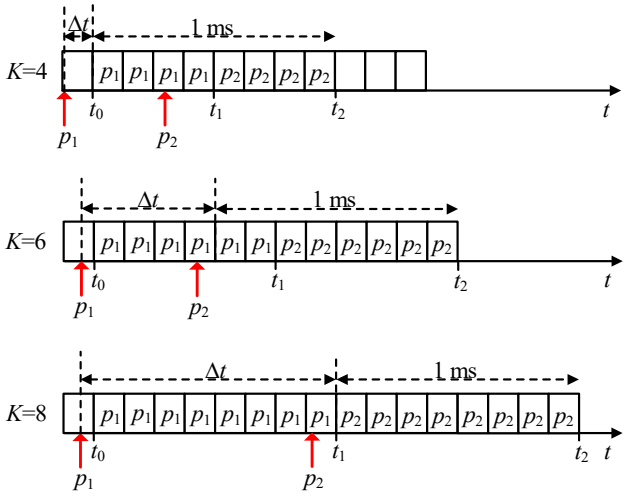

Figure 2. Scenarios with possible self-collisions (L = 1 ms and K = 4, 6 and 8).

As shown in (25) and (26), the negative effect of self-collisions increases with the value of K, since K influences the time a packet might be stored until the transmission of the previous packet is finished. However, increasing the number K of replicas transmitted for each packet is preferred, in order to combat possible collisions with other UEs sharing the same pool of radio resources. The next section will analyze both the effect of collisions from other UEs and the effect of self-collisions to analyze the reliability achievable with the grant-free scheduling with K-repetitions and shared radio resources.

4. Validation

This section validates the analytical expressions derived in Section 3.1 to calculate the probability P_c that a packet is not correctly received due to packet collisions with other UEs. To this aim, we compare the results achieved with the analytical expressions, with that obtained through simulations.

We have implemented a system level simulator in Matlab™ that accurately models the 5G NR grant-free scheduling process with K-repetitions and shared resources. The simulator emulates a single cell with N UEs that generate aperiodic traffic. Each UE models the packet traffic arrival, using a Poisson distribution with exponential inter-arrival time. The average packet inter-arrival time is equal to $1/\lambda$, where λ is the average number of packets generated per second. The simulator implements the time/frequency resource grid map of 5G NR. The time and frequency duration of RBs is configurable based on the considered 5G NR numerology μ. It is possible to also configure the number U of RBs available per time slot. The number K of replicas can also be configured in the simulation platform.

We have conducted a large number of simulations to ensure the accuracy of the simulation results, and compare them to those obtained with our analytical expressions and methodology. Simulations are here shown for K equal to 2, 4 and 8, λ equal to 0.1 packets, μ equal to 3, and U equal to 6 RBs per slot. UEs transmit small packets with a size of 32 bytes [22] that can be transmitted in a single RB. Figure 3 compares the value of P_c achieved analytically and through simulations for a varying number N of users in the cell. The figure shows that the results achieved analytically precisely match those obtained through the simulations. Similar trends have been observed for other values of the parameters. The results achieved clearly validate the proposed methodology and the analytical expressions presented in Section 3.1.

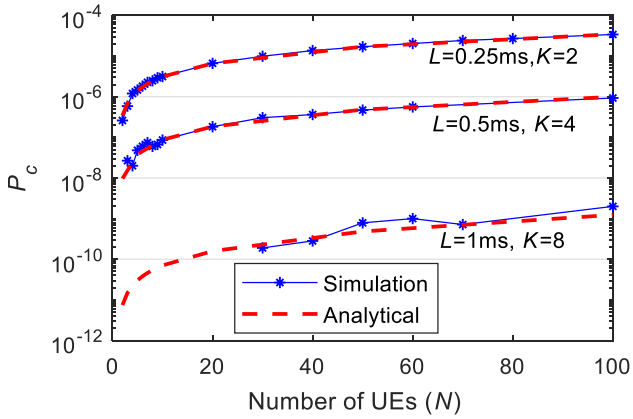

Figure 3. Comparison of analytical and simulation results for different latency requirements L and number of repetitions K ($U = 6$, $\lambda = 0.1$ packets).

It is important to highlight that this study focuses on URLLC applications that demand very high reliability levels. In simulations, we compute the number of packets for which the K replicas have collided with those packets transmitted by other UEs, and then compute the achieved reliability ($P_{rel} = 1 - P_c$). It is rare that all K replicas of a packet collide with transmissions from other UEs for low values of N. This is particularly the case when K increases. In this context, the computational cost of simulations significantly increases if we want to achieve accurate statistical results. This explains why simulation results are not shown for values of N below 30 when $K = 8$. It also highlights the value of our analytical expressions and methodology to estimate the performance of 5G NR grant-free scheduling for demanding URLLC applications and aperiodic traffic.

5. Performance Evaluation

This section evaluates the capacity of 5G NR grant-free scheduling with K-repetitions and shared resources to meet the reliability and latency requirements of URLLC services. To this aim, we use the analytical expressions that are derived in Section 3 and were validated in the previous section. Reliability for URLLC services is defined as the percentage P_{rel} of data packets that are successfully received by the BS before the latency deadline established by the service or application. In this study, we analyze first the reliability, considering only the effect of collisions from other UEs. This study analyzes then the impact of self-collisions on the capacity of 5G NR grant-free scheduling with K-repetitions and shared resources to achieve the reliability levels demanded by URLLC services. This is particularly relevant, as this study extends the state of the art by evaluating the capacity of 5G NR grant-free scheduling to sustain reliability levels even higher than 1–10^{-9}. This study also evaluates the performance of 5G NR grant-free scheduling as a function of the number of UEs, the number of reserved radio resources, and the number K of replicas.

The performance of 5G NR grant-free scheduling is evaluated considering a single cell with N UEs. Packets are generated by each UE following a Poisson process with exponentially inter-arrival time. The average packet inter-arrival time is equal to $1/\lambda$, where λ is the average number of packets generated per second. UEs transmit small packets with a size of 32 bytes [22]. Radio resources are divided in 6×12 subcarriers (i.e., $U = 6$) with a subcarrier spacing of 120 kHz (i.e., $T_{slot} = 0.125$ ms). Figure 4 shows the probability P_c that a packet is not correctly received at the BS due to the collisions from other UEs experienced by all of the replicas of a packet (This would correspond to the reliability achieved with 5G NR grant-free scheduling if there were no self-collisions, i.e., $P_{sc} = 0$ and $P_{rel} = 1 - P_c$). The figure shows the value of P_c that can be achieved as a function of the number of UEs for latency requirements (L) of 0.25, 0.5, 0.75 and 1 ms. We focus on services with the most stringent latency requirements, given the challenge to satisfy high reliability levels when latency decreases [23]. For each value of L, the grant-free scheduling scheme is executed with the maximum possible number of replicas

K that can be transmitted within the required latency. For example, if the maximum latency L that can be tolerated is equal to 1 ms, the maximum number of replicas K that can be transmitted within 1 ms is equal to 8 (L = 1 ms corresponds to $8 \cdot T_{slot}$ when T_{slot} = 0.125 ms). Figure 4 also shows the performance achieved for two values of λ (0.1 and 1 packet(s)). The results depicted in Figure 4 clearly show that reducing the probability P_c of not receiving a packet to values as low as 10^{-9}, (and hence reaching reliability levels of 1–10^{-9} when the effect of self-collisions is not considered), can only be achieved with high values of K and values of L equal to 0.75 or 1 ms. Figure 4 also shows that the probability P_c increases with the number of UEs, since the risk of collision is higher. As a result, the capacity of 5G NR grant-free scheduling to support high reliability levels is significantly decreased as the number of UEs to be supported increases. Figure 4 also shows that the difficulty in supporting high reliability levels increases with λ, since the probability P_c increases as a result of a higher risk of collision between UEs.

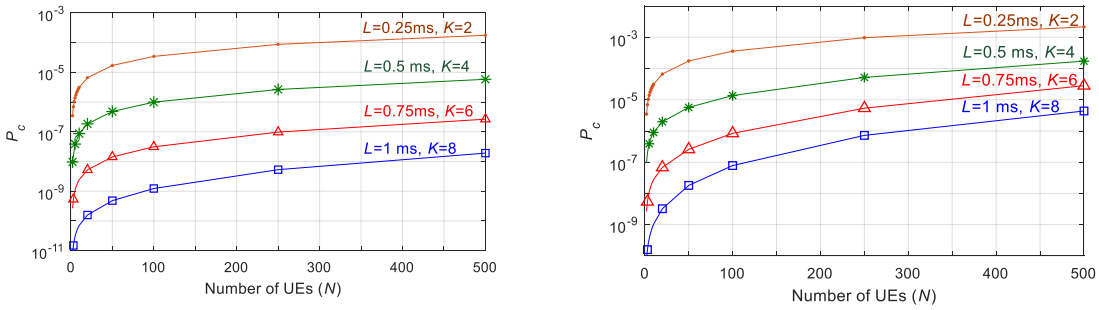

Figure 4. P_c as function of the number of User Equipments (UEs) and for different latency requirements L: (**a**) λ = 0.1 packets; (**b**) λ = 1 packet.

Figure 5 depicts the number of UEs that can be supported with a given latency requirement (L) and a reliability of $P_{rel} = 1 - P_c$ when $P_{sc} = 0$. It is important to remember that L establishes the maximum number of replicas K that can be transmitted. The results (the number of supported UEs) for each value of L in Figure 5 have been obtained for the maximum value of K permitted by L (K equal to 2, 4, 6 and 8 for L equal to 0.25, 0.5, 0.75 and 1 ms, respectively). The Release 15 of the 3GPP standards [22] establishes URLLC requirements with a latency of L = 1 ms and a reliability target of 1–10^{-5}. Figure 5 shows that grant-free scheduling with K-repetitions and shared resources can achieve a reliability equal to 1–10^{-5} with only K = 2 if we do not consider self-collisions. Grant-free scheduling with K = 2 can also guarantee a latency as low as 0.25 ms. For low values of the packet generation rate (i.e., λ = 0.1 packets), grant-free scheduling with 2 repetitions can support up to 34 UEs with a reliability of 1–10^{-5} and L = 0.25 ms if we do not consider self-collisions. The number of UEs that can be supported decreases with λ, since the risk of collision with other UEs increases when each UE transmits more packets per second. For example, only 4 UEs can be supported with L = 0.25 ms and a reliability of 1–10^{-5} when λ = 1 packet. If the latency requirement is relaxed to 0.5 ms or even higher, grant-free scheduling can support more than 500 UEs with only K = 4 when λ = 0.1 packets. If λ increases, grant-free scheduling can only guarantee the required reliability for 500 UEs if the latency requirement is 1 ms, and each UE can transmit 8 replicas of the same packet. These results show that the reliability and latency levels that can be achieved with grant-free scheduling depend upon configuration parameters (e.g., K), the traffic (e.g., λ) and the number of UEs supported. An adequate configuration and optimization of grant-free scheduling based on the network conditions could help support stringent reliability and latency levels. However, it is important to note that these results are achieved without considering self-collisions. The impact of self-collisions might be non-negligible when, for example, K and/or λ increase.

The Release 16 of 3GPP standards for 5G NR [2] defines use cases with higher reliability requirements (up to 1–10^{-6}). Some Industry 4.0 applications (e.g., factory automation) require even

higher reliability levels (up to 1–10^{-9}), as discussed in [3]. It is then important analyzing whether grant-free scheduling with K-repetitions and shared resources can guarantee reliability levels of the order of 1–10^{-9}. Figures 4 and 5 show that grant-free scheduling can only guarantee very high reliability levels with high values of K, which limits the latency requirements (L) that can be satisfied. For example, a probability to correctly receive a packet equal to 1–10^{-7} cannot be guaranteed when $L < 0.5$ ms, even for the lower packet generation rates. If the reliability requirement increases to $P_{rel} = 1–10^{-9}$, grant-free scheduling can only support 5 UEs with $L = 0.75$ ms and $\lambda = 0.1$ packets. It can support 86 UEs if the latency requirement is relaxed to 1 ms. However, if λ increases to 1 packet then grant-free scheduling can only support 10 UEs with a reliability of 1–10^{-9} even if L is equal to 1 ms.

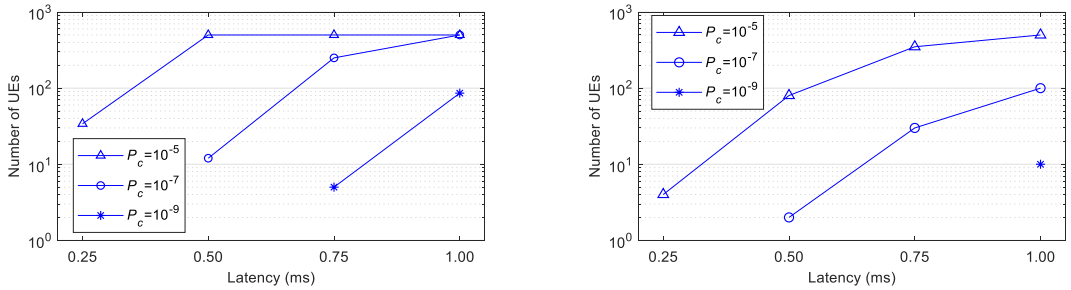

Figure 5. Number of UEs supported with different requirements (L and $P_{rel} = 1 - P_c$, when $P_{sc} = 0$): (**a**) $\lambda = 0.1$ packets; (**b**) $\lambda = 1$ packet.

Figure 6 shows the impact of U upon the performance of the grant-free scheduling scheme with K-repetitions and shared resources. U is the number of available RBs (Resource Blocks) per T_{slot}. In particular, Figure 6 depicts the number of UEs that can be supported with a given reliability and latency L when U decreases and λ is set equal to 0.1 packets (the reliability is equal to $P_{rel} = 1 - P_c$ when the effect of self-collisions is not taken into account, i.e., $P_{sc} = 0$). Figure 6 shows that the number of UEs that grant-free scheduling with K-repetitions can support for a given set of requirements strongly depends upon the number of RBs available. UEs randomly select an RB for each transmission from the U RBs available per slot. The probability that several UEs select the same RB for their transmissions increases when the number of RBs per slot decreases. Consequently, the probability P_c that a packet is not correctly received due to packet collisions, increases. In addition, the number of UEs that can achieve a target reliability level also decreases when the number of RBs per slot decreases. For example, 443 UEs can be supported with $L = 0.5$ ms (and hence $K = 4$) and $P_c = 10^{-5}$ when U is equal to 5 RBs. This number decreases to 69 UEs when U decreases to 3 RBs. This is a significant reduction of 84%. This reduction increases when the reliability demand increases. For example, 86 UEs can be supported with $P_c = 10^{-9}$ and $L = 1$ ms (and hence $K = 8$) when U is equal to 6. However, only 6 UEs can achieve these values of P_c and L if U decreases to 4 (i.e., a 93% reduction).

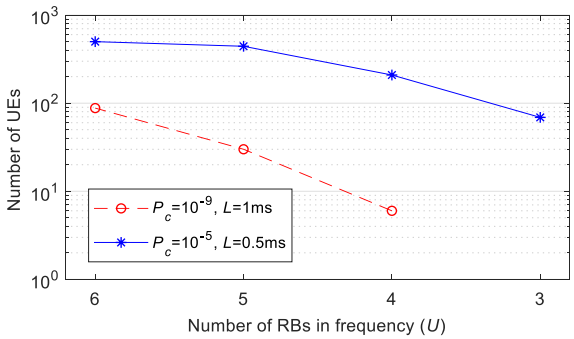

Figure 6. Number of UEs supported for a given L and $P_{rel} = 1 - P_c$ with $P_{sc} = 0$ as a function of the number U of available RBs per T_{slot} ($\lambda = 0.1$ packets).

All previous results have been derived without considering the effect of self-collisions. Self-collisions were illustrated in Figure 2, and the probability of self-collision was derived in Section 3.2. As previously described, if a packet p_2 is generated before the K replicas of the previous packet p_1 have been transmitted, p_2 will be stored and transmitted after completing the transmission of the K replicas of p_1. Due to the time that p_2 is stored, the transmission of its K replicas may finish after the latency deadline L. As presented in Section 3.2, it is not possible to complete the transmission of the K replicas of p_2 before the latency deadline L if p_2 is generated before $t_2 - L$ (t_2 is the time at which the transmission of the K replicas of p_2 is finished as shown in Figure 2). This results in that the probability P_{sc} (the probability that the transmission of K replicas of a packet is not completed before L due to the effect of self-collisions) is equal to the probability that the time between the generation of two consecutive packets at a UE falls within the interval $[0, \Delta t]$, where Δt represents the time difference between $t_2 - L$ and the time t_{p_1} at which p_1 is generated (see (25) and (26)).

We consider that packets are generated following a Poisson process with exponential inter-arrival time. As a result, Δt is homogeneously distributed between Δt_1 and Δt_2. For $K = 4$ in Figure 2, Δt_1 is equal to 0 and Δt_2 is equal to T_{slot}, since p_1 can be homogeneously generated between t_0 and $t_0 - T_{slot}$. When $K = 6$, Δt_1 is equal to $4 \cdot T_{slot}$, and Δt_2 is equal to $(4+1) \cdot T_{slot}$, since p_1 can be homogeneously generated between t_0 and $t_0 - T_{slot}$. Similarly, Δt_1 and Δt_2 are equal to $8 \cdot T_{slot}$ and $(8+1) \cdot T_{slot}$ for $K = 8$. Table 1 shows the value of P_{sc} given in (26) when Δt is equal to Δt_1 or Δt_2 considering $L = 1$ ms and $K = 4$, 6 and 8. $\Delta t = \Delta t_1$ corresponds to the scenario where self-collisions are less probable, while $\Delta t = \Delta t_2$ corresponds to the case in which they are more probable.

The results in Table 1 show that the probability of self-collision is non-negligible. For example, P_{sc} can reach values equal to 1.25×10^{-4} and 9.99×10^{-4} when K is equal to 4 and 8, respectively, and $\lambda = 1$ packet. It is also important to highlight that a comparison of results in Figure 4 and Table 1 shows that P_{sc} can be actually higher than P_c. This is for example the case when $K = 8$: P_c is lower than 10^{-7} and 10^{-5} for λ equal to 0.1 and 1 packet(s), respectively (Figure 4), while P_{sc} is approximately equal to 10^{-4} and 10^{-3} (Table 1). Grant-free scheduling can hence be limited by the effect of self-collisions, in particular when K increases. It is then important that the reliability (or probability that a packet is correctly received before the latency deadline) of grant-free scheduling with K-repetitions and shared radio resources is computed considering both the effect of collisions from other UEs and the effect of self-collisions following (1).

Table 1. P_{sc} for $L = 1$ ms.

K	$\Lambda = 0.1$ Packets		$\Lambda = 1$ Packet	
	$\Delta t = \Delta t_1$	$\Delta t = \Delta t_2$	$\Delta t = \Delta t_1$	$\Delta t = \Delta t_2$
4	0	1.25×10^{-5}	0	1.25×10^{-4}
6	5.00×10^{-5}	6.25×10^{-5}	5.00×10^{-4}	6.25×10^{-4}
8	9.99×10^{-5}	1.13×10^{-4}	9.99×10^{-4}	1.13×10^{-3}

Figure 7 plots $1 - P_{rel}$ for different values of K and L when considering both P_c and P_{sc}. The results are plotted considering $\Delta t = \Delta t_1$ for computing P_{sc}. $\Delta t = \Delta t_1$ corresponds to the case where self-collisions are less probable. Figure 4 shows that it is necessary to transmit a high number of replicas K within L to combat collisions from other UEs and correctly receive a packet at the BS. For example, Figure 4 shows that K must be equal to 8 in order to achieve P_{rel} = 1–10^{-9} when $P_{sc} = 0$ and λ is equal to 1 packet. However, Table 1 showed that the effect of self-collisions increases with K even to the point that self-collisions limit the reliability that can be achieved. This is actually shown in Figure 7 when we consider L = 1 ms. In principle, it could be possible to satisfy a 1 ms latency requirement if we transmit 4, 6 or 8 replicas of a packet. Figure 7 shows that if $K = 4$ and $\Delta t = \Delta t_1$ (for computing P_{sc} in (26)), the impact of self-collisions is not relevant, and the reliability levels of 1–10^{-5} can be satisfied for more than 500 UEs and 80 UEs when λ is equal to 0.1 and 1 packet(s), respectively; these results are in line with those observed in Figure 4 for $K = 4$. However, when K is equal to 6 or 8, the effect of self-collisions becomes more relevant (Table 1), and Figure 7 shows that it can actually limit the maximum reliability that can be achieved independently of the number of UEs. In fact, the maximum reliability that can be achieved is approximately equal to $1 - P_{sc}$. In this case, for $K = 8$ and $\lambda = 1$ packet/s, the maximum reliability (when P_{sc} is computed considering $\Delta t = \Delta t_1$) that can be achieved is 1–10^{-3} when the latency requirement L is equal to 1 ms. It should be noted that reliability levels even higher than $1 - P_c$ = 1–10^{-9} were achieved when the effect of self-collisions was not considered (Figure 4). The results discussed so far correspond to the scenario where P_{sc} has been computed considering $\Delta t = \Delta t_1$. This corresponds to the scenario where self-collisions are less probable. Figure 7 also shows the reliability that can be achieved with L = 1 ms and $K = 4$ when $\Delta t = \Delta t_{avg}$. This Δt_{avg} is the average value of Δt. $\Delta t_{avg} = (\Delta t_1 + \Delta t_2)/2$, since Δt is homogeneously distributed between Δt_1 and Δt_2. Figure 7 shows that in this case it is not possible to achieve a reliability higher than 1–6.3×10^{-5} and 1–6.3×10^{-4} when λ is equal to 0.1 and 1 packet(s). Figure 7 also shows that the reliability becomes again nearly independent of the number of UEs that are being supported. The degradation of reliability experienced from $\Delta t = \Delta t_1$ to $\Delta t = \Delta t_{avg}$ is again due to a major relevance of the effect of self-collisions when we compute the reliability.

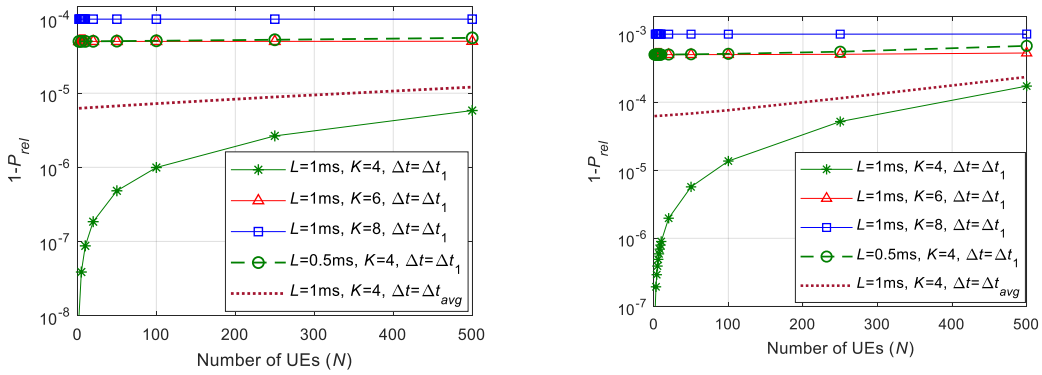

Figure 7. Reliability for different latency requirements L and number of repetitions K ($U = 6$): (**a**) $\lambda = 0.1$ packets; (**b**) $\lambda = 1$ packet.

Expressions in (25) and (26) show that P_{sc} also depends upon the latency requirement L. The effect of self-collisions is more relevant when the latency requirement is stricter. For example, Figure 7 shows that the effect of self-collisions already limits the maximum reliability that can be achieved when $K = 4$ if the latency requirement is equal to 0.5 ms. Latency requirements significantly influence the reliability levels that can be satisfied. This is the case because latency requirements limit the number K of replicas that can be sent for each packet. Figure 4 shows that the maximum reliability level that can be guaranteed depends on the latency requirements when only considering P_c. Figure 7 also shows

that the effect of self-collisions becomes more relevant with stricter latency requirements. These results show that it is a challenge guaranteeing high reliability demands with very low latency levels.

The results in Figure 7 demonstrate that current 5G NR grant-free scheduling with K-repetitions and shared resources cannot guarantee some of the more demanding reliability and latency levels. However, it is important emphasizing that other proposals cannot meet such requirements either, and these actually perform worse than the implementation analyzed in this study. This is actually the case for the proposals that transmit the first copy of a packet in dedicated resources for the UEs. These resources can be reserved using grant-based scheduling (such as in [14]) or semi-persistent scheduling (such as in [13]). Grant-based scheduling requires the UE to send an SR to the BS, and wait for the BS to reply with a grant message. The exchange of these messages between the UE and the BS is illustrated in Figure 8 This handshaking generates a non-negligible T_{total} latency that is equal to:

$$T_{total} = 2\,T_{L1/L2} + T_{align} + 2\,T_{proc} + 3\,T_{tx} = 2.3 \text{ ms} \quad (27)$$

where $T_{L1/L2}$ is the $L1/L2$ processing latency at the BS and the UE, T_{align} is the alignment latency (the alignment latency is the time elapsed from the moment the UE is ready to transmit to the actual time the transmission starts), T_{proc} is the processing latency (this latency represents the latency between the reception of the SR and the transmission of the grant message), and T_{tx} is the time required to transmit the SR and grant messages. Following [24], we consider $T_{L1/L2} = T_{align} = T_{tx} = 1$ TTI, and $T_{proc} = 2.33$ TTI. These values are a best-case scenario, since they represent reduced processing times that can be achieved with 3GPP Release 15 compared to Release 14. Equation (27) shows that the total latency (2.3 ms) introduced by the grant-based scheduling process to assign dedicated resources to UEs is higher than the latency achieved with the 5G NR grant-free scheduling implementation analyzed in this study. For example, Figure 7 shows that this implementation can guarantee latency levels below 1 ms (this latency is guaranteed with a reliability up to $1–10^{-5}$ when $K = 4$, $\lambda = 0.1$ packets, $U = 6$, and $\Delta t = \Delta t_{avg}$).

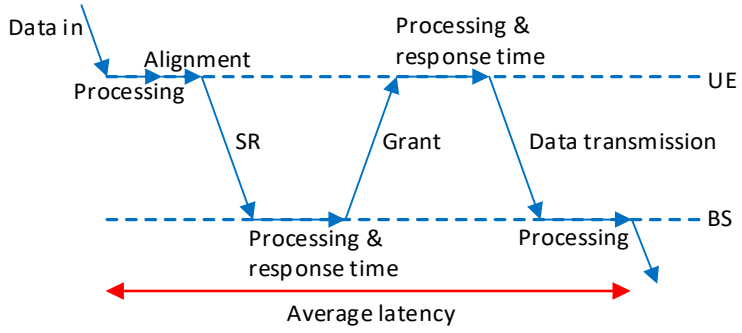

Figure 8. Latency introduced in grant-based scheduling.

The alternative to grant-based scheduling is Semi-Persistent Scheduling (SPS). In this case, UEs are assigned dedicated resources for a period of time. During this period, UEs can utilize the resources without requesting permission from the BS. This avoids the latency introduced by grant-based scheduling. However, semi-persistent scheduling inefficiently utilizes the radio resources when the traffic is aperiodic. This is the case, because it is not possible to predict when UEs will need resources. To illustrate this effect, let us consider a scenario with $N = 300$ users that generate aperiodic traffic ($\lambda = 0.1$ packets). We shall then suppose that users request a maximum latency of 1 ms and a reliability level equal to $1–10^{-5}$. Satisfying this demand requires reserving 300 RBs (one per UE) in a 1 ms time windowA lower number of resources would be necessary if traffic was periodic and we could estimate when each UE would require resources for their transmission. In this case, several UEs could share the same RB if they generate their packets at different time instants. This would reduce the total number of RBs necessary to serve all users. This is not possible in the case of aperiodic traffic, since

we cannot predict when a UE would need radio resources. Figure 7 shows that our implementation of 5G NR grant-free scheduling with 4-repetitions and shared resources can support 300 UEs (with their latency and reliability demands) with only 48 RBs in a time window of 1 ms. This is 84% less radio resources than if we reserve dedicated resources per UE (with aperiodic traffic) for their first transmission using semi-persistent scheduling. These results clearly show that the implemented 5G NR grant-free scheduling with shared resources can better support URLLC applications with aperiodic traffic and stringent communication requirements than other existing proposals. However, the conducted analysis (e.g., Figure 7) has also shown that new solutions will be needed to guarantee very demanding reliability and latency levels such as those foreseen for some URLLC services in 3GPP Release 16.

6. Conclusions

This paper has analyzed the capacity of 5G NR grant-free scheduling to support URLLC services with strict reliability and latency levels such as those demanded by Industry 4.0. The study has focused on aperiodic or sporadic traffic and an implementation of 5G NR grant-free scheduling with K-repetitions and shared radio resources. This implementation has been chosen, since sharing radio resources is an attractive option for aperiodic traffic. In addition, the K-repetitions scheme can combat possible packet collisions between UEs that share radio resources. This study has analyzed the reliability and latency levels that can be achieved with existing 5G NR grant-free scheduling with shared radio resources as a function of the number of UEs, the number of reserved radio resources, and the number of replicas K. To this aim, this study has derived analytical expressions that quantify the exact probability of collision with packets transmitted by other UEs, and the impact of self-collisions. It is important to emphasize that this study is the first one that has evaluated the impact of self-collisions. Packet collisions and self-collisions have then been taken into account to derive analytically the reliability that can be achieved by existing 5G NR grant-free scheduling with shared resources. The derived analytical expressions have been validated against simulations. These expressions are a valuable contribution to the community, since they can be easily utilized to evaluate 5G NR grant-free scheduling.

This study has demonstrated that current 5G NR grant-free scheduling solutions cannot guarantee high reliability levels with strong latency requirements. This is partly due to the fact that strong latency requirements limit the number of replicas K that can be transmitted. In addition, self-collisions have a non-negligible impact that even limits the reliability that can be achieved when K increases. The impact of self-collisions also increases with the latency requirements. The obtained results demonstrate that new solutions are necessary for 5G NR grant-free scheduling to be able to support applications with stringent URLLC latency and reliability requirements under the presence of aperiodic traffic. In particular, the transmission of K replicas per packet might be inadequate to support aperiodic traffic with very low latency levels due to the impact of self-collisions. Consequently, other approaches should be designed to minimize collisions between UEs sharing radio resources. This study has shown that these new solutions cannot be based either on grant-based or semi-persistent scheduling. Grant-based scheduling introduces additional latency due to the exchange of messages between the UEs and the BS for assigning the radio resources. Semi-persistent scheduling with dedicated resources inefficiently utilizes the available resources when considering dedicated resources and aperiodic traffic. Innovative grant-free scheduling solutions are hence necessary to meet the URLLC requirements identified for 3GPP Release 16 and beyond. This could include, for example, the use of sensing mechanisms or full duplex techniques that can reduce packet collisions.

Author Contributions: Conceptualization, M.C.L.-E. and J.G.; methodology, M.C.L.-E. and J.G.; validation, M.C.L.-E. and J.G.; formal analysis, M.C.L.-E., J.G. and M.S.; investigation, M.C.L.-E.; writing—original draft preparation, M.C.L.-E.; writing—review and editing, J.G. and M.S.; funding acquisition, J.G. and M.S.

Funding: This work has been funded by the European Commission through the FoF-RIA Project AUTOWARE: Wireless Autonomous, Reliable and Resilient Production Operation Architecture for Cognitive Manufacturing

(No. 723909), and the Spanish Ministry of Economy, Industry, and Competitiveness, AEI, and FEDER funds (TEC2017-88612-R).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 3GPP. *Technical Specification Group Services and System Aspects*; Study on Communication for Automation in Vertical Domains; 3GPP: Sophia Antipolis, France, 2018; Release 16, 3GPP TR 22.804 V16.2.0.
2. 3GPP. *Technical Specification Group Radio Access Network*; Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC); 3GPP: Sophia Antipolis, France, 2018; Release 16, 3GPP TR 38.824 V1.0.0.
3. Klessig, H.; Ashraf, S.A.; Almeroth, B.; Riedel, I.; Puschmann, A.; Elste, T.; Simsek, M.; Schulz, P.; Matthe, M.; Fettweis, G.; et al. Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture. *IEEE Commun. Mag.* **2017**, *55*, 70–78.
4. 5G Alliance for Connected Industries and Automation (5G-ACIA). *5G for Connected Industries and Automation*, 2nd ed.; 5G Alliance for Connected Industries and Automation (5G-ACIA): Frankfurt, Germany, 2019.
5. Popovski, P.; Nielsen, J.J.; Stefanovic, C.; De Carvalho, E.; Ström, E.; Trillingsgaard, K.F.; Bana, A.-S.; Kim, D.M.; Kotaba, R.; Park, J.; et al. Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks. *IEEE Netw.* **2018**, *32*, 16–23. [CrossRef]
6. 3GPP. *Technical Specification Group Radio Access Network*; Evolved Universal Terrestrial Radio Access (E-UTRA); Study on Latency Reduction Techniques for LTE; 3GPP: Sophia Antipolis, France, 2016; 3GPP TR 36.881 V14.0.0.
7. 3GPP. *Technical Specification Group Radio Access Network*; NR; Medium Access Control (MAC) Protocol Specification; 3GPP: Sophia Antipolis, France, 2018; Release 15, 3GPP TS 38.321 V15.4.0.
8. Berardinelli, G.; Mahmood, N.H.; Abreu, R.; Jacobsen, T.; Pedersen, K.; Kovacs, I.Z.; Mogensen, P. Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources. *IEEE Access* **2018**, *6*, 23602–23611. [CrossRef]
9. 3GPP. *Discussion on Configured Grant for NR-U*; 3GPP: Sophia Antipolis, France, 2018; 3GPP TSG-RAN, R1-1810329.
10. Li, Z.; Uusitalo, M.A.; Shariatmadari, H.; Singh, B. 5G URLLC: Design Challenges and System Concepts. In Proceedings of the 2018 15th International Symposium on Wireless Communication Systems (ISWCS), Lisbon, Portugal, 28–31 August 2018; pp. 1–6.
11. 3GPP. *Technical Specification Group Radio Access Network*; NR; Physical Layer Procedures for Data; 3GPP: Sophia Antipolis, France, 2018; Release 15, 3GPP TS 38.214 V15.4.0.
12. Wu, Y.; Zhang, L.; Wang, C.; Chen, Y. Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services. In Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, Australia, 4–7 June 2017; pp. 1–6.
13. Kotaba, R.; Manchon, C.N.; Balercia, T.; Popovski, P. Uplink Transmissions in URLLC Systems with Shared Diversity Resources. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 590–593. [CrossRef]
14. Abreu, R.; Mogensen, P.; Pedersen, K.I. Pre-Scheduled Resources for Retransmissions in Ultra-Reliable and Low Latency Communications. In Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, 19–22 March 2017; pp. 1–5.
15. Singh, B.; Tirkkonen, O.; Li, Z.; Uusitalo, M.A. Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 182–185. [CrossRef]
16. Jacobsen, T.; Abreu, R.; Berardinelli, G.; Pedersen, K.; Mogensen, P.; Kovacs, I.Z.; Madsen, T.K. System Level Analysis of Uplink Grant-Free Transmission for URLLC. In Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 4–8 December 2017; pp. 1–6.
17. 3GPP. *Grant-Free Transmission for UL URLLC*; 3GPP: Sophia Antipolis, France, 2017; 3GPP TSG-RAN, R1-1706919.
18. 3GPP. *Enhancement of Uplink Grant-free transmission for NR URLLC*; 3GPP: Sophia Antipolis, France, 2018; 3GPP TSG-RAN, R1-1810176.
19. 3GPP. *Enhancement of Configured Grant for NR URLLC*; 3GPP: Sophia Antipolis, France, 2018; 3GPP TSG-RAN, R1-1812162.

20. 3GPP. *Technical Specification Group Radio Access Network*; NR; Service Requirements for the 5G System; 3GPP: Sophia Antipolis, France, 2018; Stage 1, Release 15, 3GPP TR 22.261 V15.7.0.
21. 3GPP. *Technical Specification Group Radio Access Network*; NR; Physical Channels and Modulation; 3GPP: Sophia Antipolis, France, 2018; Release 15, 3GPP TR 38.211 V15.4.0.
22. 3GPP. *Technical Specification Group Radio Access Network*; Study on Scenarios and Requirements for Next Generation Access Technologies; 3GPP: Sophia Antipolis, France, 2018; Release 15, 3GPP TR 38.913 V15.0.0.
23. Bennis, M.; Debbah, M.; Poor, H.V. Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale. *Proc. IEEE* **2018**, *106*, 1834–1853. [CrossRef]
24. 3GPP. *Evaluation of Latency in LTE*; 3GPP: Sophia Antipolis, France, 2017; 3GPP TSG-RAN, R1-1720535.