Sensing-based Grant-Free Scheduling for Ultra Reliable Low Latency and Deterministic Beyond 5G Networks

M.Carmen Lucas-Estañ, *Member, IEEE*, J. Gozalvez, *Senior Member, IEEE*

***Abstract*—5G and beyond networks should efficiently support services with stringent and diverse QoS requirements. This includes services for verticals that demand Ultra Reliable and Low Latency Communications (URLLC). Scheduling strongly impacts the communication latency, and 5G NR introduces grant-free scheduling to reduce the latency at the radio level. Grant-free scheduling can use shared resources and the transmission of *K* replicas per packet to increase the packet delivery ratio and efficiently utilize the spectrum. Previous studies have shown that existing 5G NR grant-free scheduling has limitations to sustain URLLC requirements for aperiodic (or uncertain) and deterministic traffic that is characteristic of verticals such as Industry 4.0 or manufacturing. In this context, this paper proposes and evaluates a novel grant-free scheduling scheme that can efficiently support deterministic and aperiodic uplink traffic. The scheme avoids packet collisions among UEs sharing resources using a priority-based contention resolution process that relies on the transmission of announcement messages in minislots and a local channel sensing process. This study demonstrates that the proposed sensing-based grant-free scheduling scheme outperforms current 5G NR grant-free scheduling implementations, and can support a higher number of UEs with URLLC and deterministic requirements with a considerably lower number of radio resources.**

***Index Terms*—Grant-free, scheduling, configured grant, URLLC, ultra reliable, low latency, deterministic, aperiodic, 5G, NR, Beyond 5G, 6G, Industry 4.0, manufacturing.**

I. Introduction

5G networks and beyond need to support a broad range of applications and services including those required by vertical industries such as manufacturing or transportation. To this aim, 5G and beyond networks must not only increase data rates but also be able to support Ultra-Reliable and Low-Latency Communications (URLLC). URLLC are critical to support verticals with increasing levels of automation [1]. Significant efforts have been devoted in the community and 3GPP (3rd Generation Partnership Project) standardization to reduce the latency in 5G. Physical layer (PHY) and Medium Access Control (MAC) mechanisms can significantly contribute to the total end-to-end transmission delay [2]. 3GPP Release 15 and 16 include several mechanisms to reduce this delay including the use of shorter slot durations (between 1 ms and 0.0625 ms) and the possibility to use minislots in the PHY layer [3]. At the MAC level, 5G introduces the possibility to use grant-free scheduling (referred to as Semi-Persistent Scheduling or SPS in downlink and Configured Grant in uplink) to reduce the transmission latency [4]. In grant-free scheduling, UEs are pre-assigned resources periodically. UEs can use them when they have data to transmit without requesting access to the BS. However, pre-assigning dedicated resources can be inefficient when the uplink traffic is aperiodic or uncertain since many of them might end up under-utilized. Sharing radio resources among a group of UEs can improve their utilization [5]. However, packet collisions can happen when two or more UEs simultaneously contend for the same resources. Collisions can compromise the capacity to sustain stringent reliability and latency requirements, in particular, for deterministic services that require data to be delivered within a maximum latency deadline. In this context, it is still an open issue for Beyond 5G networks how to sustain stringent reliability and latency requirements for aperiodic and deterministic traffic while efficiently utilizing the radio resources. This study progresses the state-of-the art by proposing a novel sensing-based grant-free scheduling scheme designed to support stringent URLLC services with deterministic and aperiodic UL traffic using shared radio resources. The proposed scheduling scheme uses a priority-based contention resolution process that relies on the transmission of announcement messages in minislots and a local channel sensing process. This study demonstrates that the proposed sensing-based grant-free scheduling scheme can achieve higher reliability levels and support a higher number of UEs (with URLLC and deterministic requirements) using a considerably lower number of radio resources compared to existing 5G NR grant-free scheduling implementations with shared resources. The proposed sensing-based grant-free scheduling scheme can be utilized to support deterministic and aperiodic traffic in any vertical. However, the evaluation presented in this study focuses on Industry 4.0 or industrial services with URLLC requirements since 5G and Beyond 5G networks are expected to play a key role in the digital transformation of manufacturing or factory automation [6].

Manuscript received xxx; revised xxx; accepted xxx. Date of publication xxx; date of current version xxx. This work was supported in part by the Ministerio de Ciencia e Innovación, AEI and FEDER funds through the projects TEC2017-88612-R and PID2020-115576RB-I00, and by Generalitat Valenciana under Project GV/2021/044.

M.Carmen Lucas-Estañ an Javier Gozalvez are with the Uwicore Lab of the Universidad Miguel Hernández of Elche, Elche (03202), Spain (e-mails: m.lucas@umh.es, j.gozalvez@umh.es).

The paper is organized as follows. Section II reviews some of the main requirements for deterministic aperiodic and URLLC communications in Industry 4.0 use cases. Section III reviews the state of the art, and Section IV presents our proposed sensing-based grant-free scheduling scheme. Section V derives and validates an analytical expression of the latency achieved with the proposed scheme. Section VI presents a reference grant-free scheduling implementation using shared resources that transmit several replicas of a packet to increase reliability (*K*-repetitions) and that is used as benchmark in this study. Section VII compares the latency, reliability, resource efficiency and energy consumption achieved with the proposed and reference schemes. Section VIII summarizes the main conclusions and outcomes of this study.

II. Deterministic Aperiodic Communications and URLLC Requirements in Industry 4.0

Communication in vertical domains follows certain patterns that can be broadly classified within three traffic classes [1]: non-deterministic communications, deterministic periodic communications, and deterministic aperiodic communications. Non-deterministic communications include all non-time critical traffic. Deterministic communications require data to be delivered within a maximum latency. In this case, reliability accounts for the percentage of data packets that are successfully delivered before the maximum latency tolerated by the application following [7]. Use cases with deterministic communications include, for example, motion control, factory automation, process automation or remote control. Deterministic communication can generate periodic or aperiodic traffic. Periodic traffic is the most usual one in control applications. However, aperiodic traffic is also present in factory automation (for example in closed-loop control applications [1]), and is generally the most difficult traffic to handle without overdimensioning the network since it is not easy to predict when packets will be generated. However, 5G and Beyond must efficiently support deterministic aperiodic traffic to support the digital transformation of manufacturing.

3GPP analyses in [1], [8] and [9] the latency and reliability requirements of Industry 4.0 use cases. 3GPP Release 15 establishes as a general requirement for URLLC services that a packet of 32 bytes must be transmitted with a reliability of $1\text{-}10^{-5}$ and a latency deadline of 1 ms [8]. 3GPP considers use cases with reliability requirements of up to $1\text{-}10^{-6}$ and $1\text{-}10^{-8}$ under Release 16 [9]. Other sources also analyze the requirements of Industry 4.0 applications. For example, NIST establishes in [10] a reliability requirement of $1\text{-}10^{-8}$ and latencies between 0.5 and 4 ms for critical safety applications in factories. [10] also establishes latency requirements between 0.25 and 4 ms for closed-loop and open-loop regulatory control applications, and between 4 and 20 ms for closed-loop supervisory control applications; NIST considers a reliability requirement of $1\text{-}10^{-7}$ for all these applications. ETSI defines some of the most stringent latency and reliability requirements for Industry 4.0 use cases in [11]. For example, ETSI estimates that discrete manufacturing requires a latency between 1 and 12 ms, and motion control a latency between 0.25 and 1 ms; according to [11], both use cases need a reliability of $1\text{-}10^{-9}$. Very stringent requirements are also identified in [12] for automation applications with real-time control of machines. For example, [12] estimates that applications such as printing machines, packaging machines or manufacturing cells require maximum latencies between 0.25 and 5 ms and reliability levels higher than $1\text{-}10^{-9}$.

III. Related Work

5G NR introduces grant-free scheduling, also referred to as Configured Grant [4]. With grant-free scheduling, resources are pre-configured and assigned to UEs in advance. This eliminates the need to exchange the scheduling requests and grants transmitted with grant-based scheduling, and therefore reduces the signaling overhead, the transmission latency and the UEs energy consumption [13][14]. Configured Grant can assign dedicated or shared resources to the UEs [15]. Using dedicated resources can be adequate for periodic traffic, but can be inefficient when the traffic is uncertain or aperiodic and it is not possible to anticipate when resources will be needed. The use of shared resources by a group of UEs is a reasonable alternative to satisfy URLLC requirements while efficiently utilizing scarce spectrum [5]. In this case, UEs contend for the shared resources, and collisions can happen if more than one UE tries to transmit data in the same resource. The performance degradation resulting from packet collisions can be overcome with redundant transmissions (retransmission and/or repetition), but these transmissions introduce additional delays that could compromise URLLC requirements.

Current grant-free scheduling approaches can be classified as reactive, *K*-repetitions, and proactive schemes ([16], [17]). In reactive approaches, the UE only retransmits a packet when the previous transmission is not correctly received. This can improve the utilization of resources, but the latency introduced by the feedback process can compromise the possibility to support stringent latency requirements [16]. An alternative is the use of *K*-repetitions with grant-free scheduling [18]. In this case, a transmitter sends *K* replicas of the same packet in consecutive slots. [19] showed that the probability of correct reception increases with the number of replicas. However, this is achieved at the expense of using more radio resources, which might be unnecessary if one of the previous replicas was correctly received. The transmission of *K* replicas per packet may also result in that packets have to be queued while the *K* replicas of the previous packet are being transmitted. This effect is referred to as self-collisions. [20] demonstrated that queuing delays resulting from self-collisions can notably impact the capacity of *K*-repetitions to satisfy latency requirements. An alternative to mitigate the negative effects of transmitting *K* replicas per packet are proactive schemes. In proactive schemes, the BS notifies the UE when a packet is received correctly. The UE then stops the transmission of the remaining replicas in order to reduce the probability of packet collisions. [16] analytically analyses the reliability and latency performance that can be achieved with reactive, *K*-repetitions and proactive approaches. This study demonstrated that proactive schemes are more adequate when latency

requirements are more stringent, while *K*-repetitions becomes the best option when the latency requirement is more relaxed.

Several studies have proposed mechanisms to improve the performance of *K*-repetitions and proactive schemes. For example, [21] proposes a *K*-repetitions approach where resources for the transmission of the consecutive replicas are only shared by a limited group of UEs. [22] and [5] showed that the effect of collisions can be reduced using advanced receivers and multi-user detection schemes for decoding collided packets. [22] also proposed that UEs randomly choose the shared resource (among the available ones) to transmit each replica to increase the probability of correctly delivering a packet. [23] and [17] highlight the importance of selecting the number of reserved radio resources as a function of the traffic load to achieve satisfactory performance levels. To this aim, [23] proposes to dynamically adapt the number of resource blocks or RBs allocated for URLLC at each subframe based on estimations of the network load carried out by the BS. In [17], authors propose to dynamically adapt the number of slots within a subframe that are assigned for high and low priority traffic based on the estimation of the traffic load. Despite current advances, the 5G Infrastructure Association identifies in its 6G vision [24] the need for more advanced access schemes that can support high number of nodes with sporadic data. To this aim, [24] highlights the potential and need for grant-free schemes that limit or avoid retransmissions.

IV. Sensing-based Grant-Free Scheduling

The previous section has discussed existing scheduling schemes to support URLLC services and has discussed their limitations for supporting deterministic aperiodic traffic while efficiently utilizing scarce radio resources. This paper addresses these limitations and proposes a novel grant-free scheduling scheme designed to guarantee stringent latency and reliability requirements of deterministic aperiodic traffic while efficiently utilizing the radio resources. The proposed scheme is sensing-based and assigns shared resources to a group of UEs. The proposal avoids collisions among UEs sharing resources using a priority-based contention resolution process that relies on the transmission of announcement (AN) messages in minislots and a local channel sensing process.

To describe the proposed scheduling scheme, we consider the following scenario without loss of generality. The scenario has a cell with N_{UE} UEs that generate aperiodic traffic. Each packet needs to be transmitted with a maximum latency L and a reliability target equal to P_{rel}. We consider the transmission of small packets that only require one resource block or RB in the frequency domain (an RB corresponds to 12 subcarriers in 5G NR) and one slot in the time domain (with duration T_{slot}). 5G NR supports multiple OFDM numerologies, and each numerology is characterized by a subcarrier spacing and different slot durations [3]. Without loss of generality, we consider that the total bandwidth assigned to the cell is divided into N_F RBs for a given numerology.

The proposed scheme divides the total number of UEs in the cell in $U \in \mathbb{N}$ groups, and $U \leq N_F$. Each group S_u of UEs (with $u = 1,\ldots,U$) is assigned one RB per slot. All UEs in S_u share the resources allocated to the group. The number of UEs in S_u is given by $|S_u|$ and $\sum_{u=1}^{U}|S_u| = N_{UE}$. The BS manages the groups of UEs. When a UE starts a new session, the BS decides which group the user joins and informs the UE about the allocated shared resources. Collisions can happen if more than one UE in S_u wants to transmit in the same slot. To avoid collisions, the proposed scheme includes a contention resolution process based on the transmission and sensing of AN messages. These messages are transmitted in announcement or AN minislots. We propose utilizing minislots for the transmission of AN messages to optimize the usage of the spectrum. AN minislots are located just before the shared resource for data transmission as depicted in Fig. 1 (this figure shows an example with 3 AN minislots). The figure represents the organization of AN minislots and slots for data transmission (referred to as data slots in the rest of the paper). n_{AN} denotes the number of AN minislots prior to a data slot. To avoid collisions when accessing radio resources, the maximum number of UEs in a group sharing resources must not exceed $2^{n_{AN}}$ (i.e. $|S_u| \leq 2^{n_{AN}}$). A UE that wants to transmit a packet in the next data slot using the shared resource must previously transmit an AN message in the AN minislots, and sense these minislots to detect whether other UEs also want to transmit in the same data slot. The time between the start of the AN period i-1 (t_i-T_G) and the start of the AN period i (t_i) is referred to as Generation period i or G_i in Fig. 1. A generation period has a time duration of $T_G = n_{AN} \cdot T_{AN} + T_{slot}$, where T_{AN} and T_{slot} are the duration of an AN minislot and a data slot respectively. UEs that generate a new data packet in G_i will contend for data slot i during the AN period i.

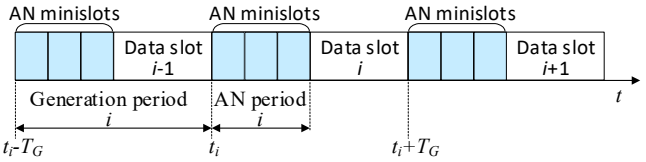


Fig. 1. Announcement minislots for contention resolution.

The transmission of AN messages within a group S_u is organized based on priorities. The BS assigns to each UE in S_u a different priority between 1 (highest priority) and p_{min} (lowest priority). The maximum number of priorities that can be managed with n_{AN} minislots is equal to $2^{n_{AN}}$ ($p_{min} = 2^{n_{AN}}$). This is why the maximum number of UEs in a group must be limited to $2^{n_{AN}}$ in order to avoid collisions. The priority of a UE$_j$ is represented by p_j. A UE$_j$ in S_u that wants to transmit data in the next shared data slot i will first execute the contention resolution process described in Algorithm I. Algorithm I determines if UE$_j$ has to transmit an AN message or sense the channel at each AN minislot s of the AN period i based on its priority p_j. UE$_j$ will transmit an AN message in the first AN minislot if $\left\lfloor (p_j\text{-}1)/(2^{n_{AN}}/2) \right\rfloor$ (where $\lfloor\cdot\rfloor$ represents the integer part) is an even number (lines 2 and 3 of Algorithm I with s=1). If $\left\lfloor (p_j\text{-}1)/(2^{n_{AN}}/2) \right\rfloor$ is an odd number, UE$_j$ will

sense the channel in the first AN minislot to determine if another UE has transmitted an AN message in the first AN minislot (lines 4 and 5). If UE_j senses the channel as busy, it will postpone its transmission (line 7) because a UE with higher priority has requested to transmit in the same data slot i. In this case, UE_j will not continue with the contention resolution process (line 8) and will compete instead for the following shared data slot i+1. If UE_j does not sense any AN message in the first AN minislot, it continues with the contention resolution process in the second AN minislot (s=2, line 1). UE_j transmits an AN message in the second AN minislot if $\left\lfloor (p_j\text{-}1)/(2^{n_{AN}}/4) \right\rfloor$ is an even number. If not, the UE senses the channel to detect whether other UEs are transmitting AN messages. The process finishes when the UE senses the channel as busy or when the process is completed for all AN minislots. The UE postpones its transmission to the next data slot i+1 if it senses the channel as busy due to the request to transmit in data slot i by another UE with higher priority. UE_j transmits its packet in data slot i if it completes the sensing process for all AN minislots without sensing a request to transmit from a higher priority UE (line 12). This process organizes access to the shared resource based on the priorities of UEs and ensures a collision-free access.

ALGORITHM I: CONTENTION RESOLUTION PROCESS FOR UE_j

1. **For** all AN minislots (s ⇠1 to n_{AN})
2. **If** $\left\lfloor (p_j\text{-}1)/(2^{n_{AN}}/2^{s}) \right\rfloor$ is an even number
3. UE_j transmits AN message
4. **Else**
5. UE_j senses the channel
6. **If** channel is busy
7. UE_j postpones its transmission
8. **End process**
9. **End If**
10. **End If**
11. **End For**
12. UE_j transmits data in data slot i

Fig. 2 illustrates an example of the contention process with 3 AN minislots. UE_3, UE_4 and UE_7 with priorities 3, 4, and 7 belong to the same group S_u. UE_3 is the UE with highest priority among the three UEs, and UE_7 is the one with the lowest priority. All UEs want to transmit a data packet and contend for the same data slot i. We represent in blue and stripped pattern the AN minislots where UEs transmit AN messages. The dashed AN minislots depict when a UE senses the channel. All UEs use Algorithm I to determine (based on their priorities) if they have to transmit an AN message or sense the channel at each AN minislot s of an AN period i. The execution of Algorithm I results in that UE_3 transmits two AN messages in AN minislots 1 and 3, and senses the channel in AN minislot 2. UE_4 transmits one AN message in AN minislot 1 and senses the channel in AN minislots 2 and 3. Finally, UE_7 senses the channel in AN minislots 1 and 2, and transmits one AN message in AN minislot 3. This organization results in that UE_7 detects the transmissions from UE_3 and UE_4 in the first AN minislot. UE_7 finishes then its contention process and postpones its transmission. UE_7 will contend to transmit in data slot i+1. UE_3 and UE_4 do not sense any AN message in AN minislot 2 and continue their contention processes. UE_3 transmits then an AN message in AN minislot 3 that is sensed by UE_4. UE_4 postpones its contention and potential transmission to the next data slot as well. UE_3 transmits its data packet in the data slot i. A similar contention process is run by UE_4 and UE_7 for data slot i+1, and UE_4 transmits its data packet in data slot i+1 since it has higher priority than UE_7. UE_7 transmits its data packet in data slot i+2 since the highest priority UEs (UE_3 and UE_4) do not contend for this data slot in this example.

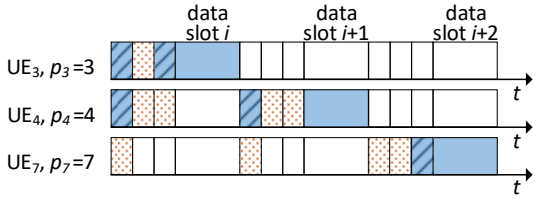


Fig. 2. Illustration of the contention process.

V. ANALYTICAL LATENCY MODELING

This section derives an analytical expression of the latency achieved with the proposed sensing-based grant-free scheduling scheme. This modeling provides a valuable tool to evaluate the performance when events are rare, and the computational cost of simulations significantly increase to achieve accurate statistical results. This is particularly relevant in this study that considers URLLC applications with aperiodic traffic and stringent latency and reliability requirements. Table I compiles all variables and functions used in this section and throughout the paper.

A. Modeling

Without loss of generality, we derive the latency achieved with the proposed scheme in a scenario where UEs generate packets following a Poisson process with exponentially distributed inter-arrival time ([22], [25])[1]. The average packet inter-arrival time is equal to $1/\lambda$ with λ equal to the average number of packet arrivals per second. We consider the transmission of 32 bytes packets following 3GPP standards in [8] that specify that a packet of 32 bytes must be transmitted in less than 1 ms with a reliability of $1\text{-}10^{-5}$. With the proposed scheme, UEs generating new data packets during G_i with duration T_G compete for the shared data slot i. The probability $\mathbb{P}_g$ that a UE generates one or more packets in T_G is equal to:

$$\mathbb{P}_g(T_G)=1\text{-}\exp(-T_G\cdot\lambda). \quad (1)$$

Each group S_u of UEs is assigned a different RB, so the scheduling process is independent for each S_u. We establish a different priority for each UE in S_u with 1 being the highest priority and p_{min} the minimum one. We analyse next different scenarios that help understand how to derive the probability

[1] The latency can be derived for other traffic models following the process explained in this section.

$\mathbb{P}_{S_u}(l \geq L)$ that a UE ϵ S_u experiences a latency l equal to or higher than a deadline L. We first consider the scenario where a UE$_j$ that has generated a new packet in G_i gets access to the next shared data slot i. In this case, the packet is generated before t_i (Fig. 1) and its transmission is completed at t_i+T_G. The minimum transmission latency is then equal to T_G, and $\mathbb{P}_{S_u}(l \geq T_G)$ is equal to one. We now analyse the case for $\mathbb{P}_{S_u}(l \geq 2 \cdot T_G)$. When UE$_j$ does not gain access to the shared data slot i, it will contend for the shared data slot i+1. In this case, the minimum transmission latency experienced by UE$_j$ will be equal to $2 \cdot T_G$. $\mathbb{P}_{S_u}(l \geq 2 \cdot T_G)$ is then given by the probability that a UE that has generated a new packet in G_i does not gain access to the shared data slot i. Let's consider $R_i = S_u$-{UE$_j$} is the set of UEs that do not have any pending packet to transmit at the beginning of G_i. Let's also consider that n_i UEs ϵ R_i generate new packets in G_i. UE$_j$ and the n_i UEs ϵ R_i contend to transmit in data slot i. UE$_j$ will not gain access to the data slot i if at least one of the n_i UEs has higher priority than UE$_j$. $\mathbb{P}_{S_u}(l \geq 2 \cdot T_G)$ depends then on the probability $\mathbb{P}_{\mathrm{n}}(n_i,R_i,T_G)$ that n_i UEs from R_i generate new packets during G_i multiplied by the probability $\overline{\mathbb{P}_{\mathrm{p}}}(n_i,p_{min})$ that at least one of these n_i UEs has higher priority than UE$_j$. $\mathbb{P}_{S_u}(l \geq 2 \cdot T_G)$ can then be expressed as:

$$\mathbb{P}_{S_u}(l \geq 2 \cdot T_G) = \sum_{n_i=1}^{|R_i|} \left(\mathbb{P}_{\mathrm{n}}(n_i,R_i,T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(n_i,p_{min}) \right) \quad (2)$$

with $|R_i| = |S_u| - 1$. The probability $\mathbb{P}_{\mathrm{n}}(n_i,R_i,T_G)$ that n_i UEs from R_i generate new packets during G_i is equal to the multiplication of the probability that n_i UEs generate at least one packet during T_G ($\mathrm{P_g}(T_G)^{n_i}$), the probability that $|R_i|$-n_i UEs do not generate a packet during T_G $((1-\mathbb{P}_{\mathrm{g}}(T_G))^{|R_i|-n_i})$, and the number of possible combination of n_i UEs in a set R_i. This is expressed as:

$$\mathbb{P}_{\mathrm{n}}(n_i, R_i, T_G) = \binom{|R_i|}{n_i} \cdot \mathbb{P}_{\mathrm{g}}(T_G)^{n_i} \cdot (1-\mathbb{P}_{\mathrm{g}}(T_G))^{|R_i|-n_i}. \quad (3)$$

To compute $\overline{\mathbb{P}_{\mathrm{p}}}(n_i,p_{min})$, we compute first the probability that n_i UEs within S_u have lower priority than UE$_j$ with priority p_j:

$$\mathbb{P}_{\mathrm{p}}(n_i, p_j, p_{min}) = \frac{1}{p_{min}} \prod_{j=1}^{n_i} \frac{p_{min} - p_j - (j-1)}{p_{min} - j}. \quad (4)$$

Following (4), the probability that n_i UEs ϵ S_u-{UE$_j$} have lower priority than UE$_j$ is given by:

$$\mathbb{P}_{\mathrm{p}}(n_i, p_{min}) = \sum_{p_i=1}^{p_{min}} \mathbb{P}_{\mathrm{p}}(n_i,p_i,p_{min}) = \sum_{p_i=1}^{p_{min}-n_i} \mathbb{P}_{\mathrm{p}}(n_i,p_i,p_{min}). \quad (5)$$

We can then compute the probability that at least one of the n_i UEs has higher priority than UE$_j$ as:

$$\overline{\mathbb{P}_{\mathrm{p}}}(n_i, p_{min}) = 1-\mathbb{P}_{\mathrm{p}}(n_i, p_{min}). \quad (6)$$

The process followed to compute $\mathbb{P}_{S_u}(l \geq 2 \cdot T_G)$ can be used to compute $\mathbb{P}_{S_u}(l \geq m \cdot T_G)$ with $m \in \mathbb{N}$. Appendix A details, as an example, how to compute the probability $\mathbb{P}_{S_u}(l \geq 3 \cdot T_G)$. The process is here generalized to derive the probability $\mathbb{P}_{S_u}(l \geq m \cdot T_G)$ that a UE experiences a latency higher than $m \cdot T_G$ for any value of $m \in \mathbb{N}$. We denote as C_{i+q} the amount of UEs that contend with UE$_j$ to access data slot $i+q$. C_{i+q} is equal to n_{i+q} if q=0, and equal to $\max(C_{i+q-1}-1, 0)+n_{i+q}$ if q>0. n_{i+q} is the number of UEs ϵ R_{i+q} that generate new packets in G_{i+q}. R_{i+q} is the set of UEs that do not have any pending packet to transmit at the beginning of G_{i+q}. $|R_{i+q}|$ is equal to $|S_u|$-1 if q=0, and equal to $|S_u|$-1-C_{i+q-1} if q>0. $\mathbb{P}_{S_u}(l \geq m \cdot T_G)$ is shown in (7) as a function of the auxiliary function $f(q,n_{i+m-q})$ defined in (8). The function $f(q,n_{i+m-q})$ is a recursive function defined for $2 \leq q \leq m$. $f(q,n_{i+m-q})$ considers the probability $\mathbb{P}_{\mathrm{n}}(n_{i+q},R_{i+q},T_G)$ that other UEs generate new packets in G_{i+q}, and the probability $\overline{\mathbb{P}_{\mathrm{p}}}(C_{i+q},p_{min})$ that at least one of the C_{i+q} UEs contending with UE$_j$ for the data slot $i+q$ has higher priority than UE$_j$. $f(q,n_{i+m-q})$ also depends on $f(q\text{-}1,n_{i+m-q})$ for all possible values of n_{i+q} between $n_{i+q,min}$ and $n_{i+q,max}$ defined in (9). Using (8), $\mathbb{P}_{S_u}(l \geq m \cdot T_G)$ in (7) is defined as the sum of $f(m,n_i)$ for all possible values of n_i between $n_{i,min}$=1 and $n_{i,max}$=$|S_u|$-1.

$$\mathbb{P}_{S_u}(l \geq m \cdot T_G) = \sum_{n_i=n_{i,min}}^{n_{i,max}} f(m,n_i) \quad (7)$$

$$f(q,n_{i+m-q}) = \begin{cases} \mathbb{P}_{\mathrm{n}}(n_{i+m-q},R_{i+m-q},T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(C_{i+m-q},p_{min}) \cdot \\ \quad \sum_{n_{i+m-q-1}=n_{i+m-q-1,min}}^{n_{i+m-q-1,max}} f(q\text{-}1,n_{i+m-q-1}) & \text{if } 2 < q \leq m \\ \mathbb{P}_{\mathrm{n}}(n_{i+m-q},R_{i+m-q},T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(C_{i+m-q},p_{min}) & \text{if } q = 2 \end{cases} \quad (8)$$

$$n_{i+m-q,min} = \begin{cases} 1 & \text{if } q = m \text{ or } (q < m \ \&\ |R_{i+m-q}| = |S_u|-2) \\ 0 & \text{if } q < m \ \&\ |R_{i+m-q}| > |S_u|-2 \end{cases} \quad (9)$$

$$n_{i+m-q,max} = |R_{i+m-q}| \quad (10)$$

Once $\mathbb{P}_{S_u}(l \geq m \cdot T_G)$ is computed for all S_u ($u \in [1, U]$), the probability $\mathbb{P}(l \geq m \cdot T_G)$ that any of the N_{UE} UEs in the cell experiences a latency l equal to or higher than $m \cdot T_G$ can be expressed as:

$$\mathbb{P}(l \geq m \cdot T_G) = \frac{\sum_{u=1}^{U} \mathbb{P}_{S_u}(l \geq m \cdot T_G) \cdot \lambda \cdot |S_u|}{\sum_{u=1}^{U} \lambda \cdot |S_u|}. \quad (11)$$

where $\lambda \cdot |S_u|$ is the expected value for the generation of new packets for a set S_u of UEs.

B. Validation

This section validates the derived model by comparing the latency obtained with the analytical expressions with that obtained through simulations. To this aim, we have developed a system level simulator in Matlab™ that accurately implements the proposed sensing-based grant-free scheduling scheme. The simulator emulates a single cell with N_{UE} UEs that generate aperiodic traffic following a Poisson process with exponentially distributed inter-arrival packet time. UEs

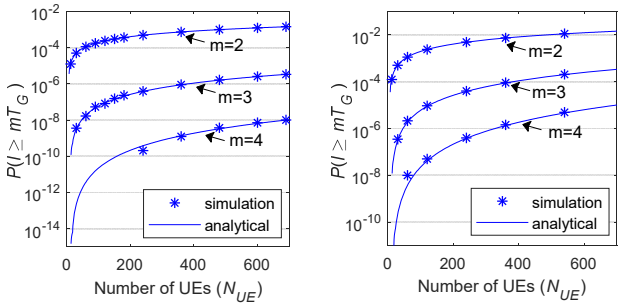
TABLE I
VARIABLES AND FUNCTIONS

Variable/Function	Definition	Variable/Function	Definition
N_{UE}	Number of UEs in a cell	C_i	# of UEs that contend with UE_j to access data slot i
L	Latency requirement	$f(q,n_{i+m-q})$	Auxiliary recursive function
P_{rel}	Reliability requirement	K	# of replicas transmitted per packet by the reference scheme
T_{slot}	Time duration of a slot	$\mathbb{P}_{sc}$	Probability of self-collisions
n_{AN}	Number of AN minislots	$\mathbb{P}_{c}$	Probability that a packet is not correctly received due to collisions of the transmitted replicas with transmissions from other UEs
T_{AN}	Time duration of an AN minislot	$\mathbb{P}_{c,K}$	Probability that a packet is not correctly received due to the collision of all its K replicas
N_F	# of RBs in frequency	$\mathbb{P}_{c,k}$	Probability that a packet is not correctly received due to the collision of the first k replicas (with $k<K$)
λ	Average packet arrivals per second	n_i^{act}	# of UEs that have a replica to transmit in the same slot i as UE_j
G_i	Generation period i	Q_i	Set of UEs (excluding UE_j) that do not have a packet to transmit at the beginning of slot i
T_G	Time duration of a generation period	$\overline{\mathbb{P}_{nrc}}(n_i^{act},N_F)$	Probability that one or more of the n_i^{act} UEs select the same RB as UE_j from the N_F RBs available in slot i
U	# of UEs groups	$\mathbb{P}_{nrc}(n_i^{act},N_F)$	Probability that n_i^{act} UEs select a different RB at slot i than UE_j
S_u	Group of UEs (u =1,…,U)	$h_{i,k}$	Auxiliary recursive function
p_j	Priority of UE_j	$\mathbb{P}_r(n_i^{act},N_F)$	Probability that a replica transmitted by UE_j in slot i is correctly received when there are n_i^{act} UEs also transmitting a replica in the same slot i
p_{min}	Maximum number of priorities	$\mathbb{P}_{rc}(t_i,n_i^{act},N_F)$	Probability that t_i UEs from the n_i^{act} UEs with active transmissions in slot i select the same RB as UE_j
n_i	# of UEs ϵ R_i that generate new packets in G_i	$\mathbb{P}_d(t_i)$	Probability of successfully decoding a replica when it has collided with the replicas transmitted by other t_i UEs
$n_{i,min}$, $n_{i,max}$	Minimum and maximum value of n_i	$g_{i,k}$	Auxiliary recursive function
R_i	Set of UEs that do not have any pending packet to transmit at the beginning of G_i	$\overline{E}$	Average energy consumption of a UE
$\mathbb{P}_g(T)$	Probability that a UE generates one or more packets in a time period T	E_{slot}	Energy consumed by a UE transmitting during T_{slot}
$\mathbb{P}_{S_u}(l \geq L)$	Probability that a UE ϵ S_u experiences a latency l equal to or higher than a deadline L	E_{AN}	Energy consumed in a contention process
$\mathbb{P}(l \geq L)$	Probability that any of the N_{UE} UEs in the cell experiences a latency l equal to or higher than L	E_p	Energy consumed in the transmission of a packet
$\mathbb{P}_n(n_i,R_i,T_G)$	Probability that n_i UEs from R_i generate new packets during G_i	N_{RB}	# of RBs used by a scheduling scheme to support N_{UE} UEs
$\mathbb{P}_p(n_i, p_j, p_{min})$	Probability that n_i UEs have lower priority than UE_j with priority p_j		
$\mathbb{P}_p(n_i, p_{min})$	Probability that n_i UEs have lower priority than UE_j		
$\overline{\mathbb{P}_p}(n_i,p_{min})$	Probability that at least one of n_i UEs has higher priority than UE_j		

transmit small packets of 32 bytes. The simulator implements the time/frequency resource grid map of 5G NR. The time and frequency duration of RBs is configurable based on the numerology. The validation presented in this section corresponds to the 5G NR numerology 3 with T_{slot} equal to 0.125 ms. We consider there are 7 AN minislots per data slot and the AN minislots have a duration of 2 symbols. The simulator can configure the number N_F of available RBs, and the results corresponds to N_F=6.

Fig. 3 compares the value of $\mathbb{P}(l \geq m \cdot T_G)$ obtained analytically and through simulations when considering different number of UEs and values of λ equal to 0.1 and 1 packet/s. The figure clearly shows that the latency performance obtained analytically precisely matches that obtained through the simulations for different values of m and λ[2]. Fig. 3 shows that $\mathbb{P}(l \geq m \cdot T_G)$ decreases several orders of magnitude when m increases and λ decreases. For example, $\mathbb{P}(l \geq 4 \cdot T_G)$ is lower than 10^{-8} when λ=0.1 packet/s, and can even be lower than 10^{-10} when there are less than 157 UEs. This is hence a very rare event, and the computational cost of simulations significantly increases if we want to achieve accurate statistical results for rare events. This explains why simulation results are not shown for less than 240 UEs when m=4 and λ=0.1 packet/s, and highlights the value of the analytical expressions to estimate the latency in the presence of rare events with aperiodic traffic and URLLC requirements.

[2] Different numerologies and values of T_{slot} have been evaluated, and in all cases the analytical results precisely match those achieved with simulations.



a) λ=0.1 packet/s. b) λ=1 packet/s.

Fig. 3. Comparison of analytical and simulation $\mathbb{P}(l \geq m \cdot T_G)$ results for different values of m and λ.

VI. REFERENCE SCHEME

The performance achieved with the proposed sensing-based grant-free scheduling scheme is compared against a reference grant-free scheduling implementation using K-repetitions. 3GPP standards introduce grant-free scheduling with dedicated or shared resources and K-repetitions [18]. With K-repetitions, it is possible to transmit K replicas[3] of the same packet to increase the reliability. We should note that 3GPP standards do not specify a particular implementation of grant-free scheduling with K-repetitions. Some proposals (e.g. [26]) propose implementations that use dedicated resources for the first transmission of a packet and shared resources for the following replicas of the same packet. Using dedicated

[3] K accounts for the original packet and subsequent transmitted copies.

resources can increase the delay and inefficiently utilize resources when supporting applications with aperiodic traffic since it is not possible to predict when packets will be generated. We therefore implement the state-of-art proposal in [22] that transmits the K replicas of a packet using shared resources. Following 3GPP standards [18], UEs transmit the same data packet in K consecutive slots (with a duration T_{slot}) in the implemented reference scheme. Following [22], UEs randomly select an RB for each transmission from the N_F RBs available per T_{slot} to increase the probability to successfully deliver a packet. [20] evaluated the reliability and latency levels that can be achieved with the reference scheme implemented following [22]. This paper compares the performance that can be achieved with the proposed sensing-based grant-free scheduling scheme to that obtained with the reference scheme. To estimate the reliability that can be achieved with the reference scheme, we use the analytical expression derived in [20]. Reliability for URLLC services is defined as the percentage of packets that are successfully delivered before the latency deadline L. It can be expressed as P_{rel}=1-$\mathbb{P}(l \geq L)$ where $\mathbb{P}(l \geq L)$ is the probability that a packet is received after L. For the reference scheme, the probability $\mathbb{P}(l \geq L)$ that a UE experiences a latency l equal to or higher than a deadline L depends on two factors: 1) the probability $\mathbb{P}_c$ that a packet is not correctly received due to the collision of the transmitted replicas with transmissions from other UEs, and 2) the probability $\mathbb{P}_{sc}$ of self-collisions. A self-collision can occur when a new packet is generated but the transmission of the K replicas of the previous packet has not finished. In this case, the transmission of the new packet has to be delayed until the transmission of all the replicas of the previous packet is completed. This delay can result in that the packet cannot be delivered within the established deadline L. We derived $\mathbb{P}(l \geq L)$ in [20] as:

$$\mathbb{P}(l \geq L) = \mathbb{P}_{sc}+(1-\mathbb{P}_{sc}) \cdot \mathbb{P}_c \tag{12}$$

$\mathbb{P}_{sc}$ can be computed as the probability that the time difference between two consecutive packets generated by a UE is between 0 and Δt, where Δt=2$\cdot K \cdot T_{slot}$-L-t_{p_1} and t_{p_1} represents the time at which the first of the two consecutive packets is generated. $\mathbb{P}_{sc}$ is expressed in [20] as:

$$\mathbb{P}_{sc}= \int_0^{\Delta t} \lambda \cdot e^{-t \cdot \lambda} \cdot dt \tag{13}$$

The estimation of $\mathbb{P}_c$ depends on whether L is higher or smaller than $K \cdot T_{slot}$. If $L \geq K \cdot T_{slot}$, $\mathbb{P}_c$ will be equal to the probability $\mathbb{P}_{c,K}$ that a packet is not correctly received due to the collision of all its K replicas. If $L < K \cdot T_{slot}$, $\mathbb{P}_c$ will be equal to the probability $\mathbb{P}_{c,k}$ that a packet is not correctly received due to the collision of the first k replicas (with $k < K$) transmitted before the maximum latency L.

$\mathbb{P}_{c,K}$ was derived in [20][4]. We compute first $\mathbb{P}_{c,k}$ when $k<K$ using the same procedure followed in [20] to derive $\mathbb{P}_{c,K}$. $\mathbb{P}_{c,k}$ is the probability that the first k replicas of a packet transmitted by UE$_j$ in consecutive slots i, i+1, …, i+k-1 are not correctly received due to collisions with replicas transmitted by other UEs. The probability of collision of a replica on slot i depends on:

1) The probability that n_i^{act} UEs also have a replica to transmit in the same slot i as UE$_j$, with $n_i^{act} \geq 1$. n_i^{act} is equal to n_{i-K}+…+n_{i-1}, where n_{i-K}, …, n_{i-1} represent the number of UEs with a new packet generated in slot i-K,…, i-1, respectively. The probability that n_i^{act} UEs have a replica to transmit in slot i is then given by $\mathbb{P}_n(n_{i-K},Q_{i-K},T_{slot}) \cdot \ldots \cdot \mathbb{P}_n(n_{i-1},Q_{i-1},T_{slot})$, where $\mathbb{P}_n(n_t,Q_t,T_{slot})$[5] is the probability that n_t UEs from the set Q_t generated new packets during slot t for t=i-K,…, i-1 ($\mathbb{P}_n(n_t,Q_t,T_{slot})$ was presented in (3)). Q_t is the set of UEs (excluding UE$_j$) that do not have a packet to transmit at the beginning of slot t, and is calculated as follows:

$$|Q_t| = N_{UE}\text{-}1\text{-}\sum_{z=\max\{t-(K-1),\,0\}}^{t-1} n_z \tag{14}$$

2) The probability that one or more of the n_i^{act} UEs with an active transmission in slot i select the same RB as UE$_j$ from the N_F RBs available in slot i. This probability is given by $\overline{\mathbb{P}_{nrc}}(n_i^{act},N_F)$=1-$\mathbb{P}_{nrc}(n_i^{act},N_F)$, where $\mathbb{P}_{nrc}(n_i^{act},N_F)$ is the probability that n_i^{act} UEs select a different RB at slot i than UE$_j$. $\mathbb{P}_{nrc}(n_i^{act},N_F)$ is computed as:

$$\mathbb{P}_{nrc}(n_i^{act},N_F)=\left(\frac{N_F\text{-}1}{N_F}\right)^{n_i^{act}} \tag{15}$$

To estimate the probability of collision of each of the first k replicas of a packet, we need to consider all possible combinations of n_i^{act} (in the range [1, N_{UE}-1]) and n_{i-K}, …, n_{i-1}.

$\mathbb{P}_{c,k}$ finally depends on the number k of replicas transmitted before L, the total number K of replicas transmitted for each packet, the number N_{UE} of UEs sharing radio resources, the number N_F of RBs, and the time duration of a slot T_{slot}. $\mathbb{P}_{c,k}$ can be expressed as shown in (16) and is defined as a function of the auxiliary recursive function $h_{i,k}(K,N_{UE},N_F,T_{slot})$[6] in (17).

$$\mathbb{P}_{c,k}(K, N_{UE}, N_F,T_{slot}) =h_{0,k}(K, N_{UE}, N_F,T_{slot})= h_{0,k} \tag{16}$$

$$h_{q,k}=\begin{cases} \sum_{n_{i+q}=n_{i+q,min}}^{n_{i+q,max}} \left[\mathbb{P}_n\left(n_{i+q},Q_{i+q},T_{slot}\right)\cdot h_{q+1,k}\right] & \text{if } q \in [0, K\text{-}1) \\ \sum_{n_{i+q}=n_{i+q,min}}^{n_{i+q,max}} \left[\mathbb{P}_n(n_{i+q},Q_{i+q},T_{slot})\cdot \left(\overline{\mathbb{P}_{nrc}}\left(n_{i+q}^{act},N_F\right)\right)\cdot h_{q+1,k}\right] & \text{if } q \in [K\text{-}1, K\text{+}k\text{-}2) \\ \sum_{n_{i+q}=n_{i+q,min}}^{n_{i+q,max}} \left[\mathbb{P}_n(n_{i+q},Q_{i+q},T_{slot})\cdot \overline{\mathbb{P}_{nrc}}\left(n_{i+q}^{act},N_F\right)\right] & \text{if } q = K\text{+}k\text{-}2 \end{cases} \tag{17}$$

with $n_{i+q,min}=\begin{cases}1 & \text{if } q \geq K\text{-}2 \;\&\; |Q_{i+q}| = N_{UE}\text{-}1 \\ 0 & \text{otherwise}\end{cases}$,

$n_{i+q,max}=|Q_{i+q}|$.

[4] In [20], $\mathbb{P}_c$ is equal to $\mathbb{P}_{c,K}$ since [20] did not consider the scenario where $L < K \cdot T_{slot}$.

[5] $\mathbb{P}_n(\cdot)$ is denoted as $P_{tx}(\cdot)$ in [20].

[6] [20] derived the expression of $h_{q,K}$. In [20], h_i is equal to $h_{q,K}$ since [20] only considers the scenario where $L \geq K \cdot T_{slot}$.

$\mathbb{P}_{c,K}$ is obtained from (16) and (17) with k=K[7], and it is then possible to estimate $\mathbb{P}_c$.

The expression of $\mathbb{P}_{c,k}$ in (16) is obtained assuming that a collision cannot be decoded. However, advanced receivers with multi-user detection capability can resolve positively collisions. In this paper, we extend the expression of $\mathbb{P}_{c,k}$ in (16) to account for scenarios where advanced receivers are used. In these scenarios, the probability $\mathbb{P}_r(n_i^{act},N_F)$ that a replica transmitted by UE$_j$ in slot i is correctly received when there are n_i^{act} UEs also transmitting a replica in the same slot i is given by:

$$\mathbb{P}_r(n_i^{act},N_F)=\mathbb{P}_{nrc}(n_i^{act},N_F)+\sum_{t_i=1}^{n_i^{act}}[\mathbb{P}_{rc}(t_i, n_i^{act},N_F)\cdot\mathbb{P}_d(t_i)] \quad (18)$$

$\mathbb{P}_{nrc}(n_i^{act},N_F)$ is the probability that any of the n_i^{act} UEs select the same RB at slot i than UE$_j$, and there is no collision. $\mathbb{P}_{nrc}(n_i^{act},N_F)$ was defined in (15). The second term of the expression in (18) represents the probability of successfully decoding the replica even when a collision happened. In (18), $\mathbb{P}_{rc}(t_i,n_i^{act},N_F)$ defined in (19) represents the probability that t_i UEs from the n_i^{act} UEs with active transmissions in slot i select the same RB as UE$_j$, and $\mathbb{P}_d(t_i)$ represents the probability of successfully decoding a replica when it has collided with the replicas transmitted by other t_i UEs.

$$\mathbb{P}_{rc}(t_i, n_i^{act},N_F)=\binom{n_i^{act}}{t_i}\cdot\frac{(N_F\text{-}1)^{(n_i^{act}-t_i)}}{N_F{}^{n_i^{act}}} \quad (19)$$

We can then derive the expression of $\mathbb{P}_{c,k}$ considering the impact of using advanced receivers with multi-user detection capability as follows:

$$\mathbb{P}_{c,k}(K,N_{UE},N_F,T_{slot}) =g_{0,k}(K, N_{UE}, N_F,T_{slot})= g_{0,k} \quad (20)$$

where $g_{i,k}$ is equal to:

$$g_{q,k}=\begin{cases}\sum_{n_{i+q}=n_{i+q,min}}^{n_{i,max}}\left[\mathbb{P}_n(n_{i+q},Q_{i+q},T_{slot})\cdot g_{q+1}\right] & \text{if } q\in[0,K\text{-}1)\\ \sum_{n_{i+q}=n_{i+q,min}}^{n_{i,max}}\left[\mathbb{P}_n(n_{i+q},Q_{i+q},T_{slot})\cdot\left(1\text{-}\mathbb{P}_r(n_{i+q}^{act},N_F)\right)\cdot g_{q+1}\right] & \text{if } i\in[K\text{-}1,K\text{+}k\text{-}2)\\ \sum_{n_{i+q}=n_{i+q,min}}^{n_{i,max}}\left[\mathbb{P}_n(n_{i+q},Q_{i+q},T_{slot})\cdot\left(1\text{-}\mathbb{P}_r(n_{i+q}^{act},N_F)\right)\right] & \text{if } i=K\text{+}k\text{-}2\end{cases} \quad (21)$$

Similar to the scenario without advanced receivers, $\mathbb{P}_{c,K}$ can be derived from (20) and (21) with k=K. We can then estimate $\mathbb{P}_c$ using advanced receivers as follows: 1) if $L \geq K\cdot T_{slot}$, $\mathbb{P}_c$=$\mathbb{P}_{c,K}$; 2) if $L < K\cdot T_{slot}$, $\mathbb{P}_c$=$\mathbb{P}_{c,k}$. A summary of all variables and functions is included in Table I.

[7] We should note that $\mathbb{P}_{c,K}$ was already introduced in [20] but we include the process to derive it for a better understanding of the reader.

VII. Evaluation

This section evaluates the proposed sensing-based grant-free scheduling scheme, and compares its performance to that achieved with the reference state-of-art grant-free scheme with shared resources and K-repetitions. The numerical evaluation is conducted using the validated analytical expressions since they can accurately quantify the performance even under the presence of sporadic and rare events. The evaluation considers first the 3GPP Release 15 requirements for URLLC services identified in [8]. [8] establishes that a packet of 32 bytes must be transmitted with a reliability of P_{rel}=1-10^{-5} within a maximum latency L of 1 ms. The reliability is defined as the percentage of data packets that are successfully delivered before the latency deadline L. The evaluation is then extended to different reliability and latency requirements. Finally, the section compares the efficiency in the utilization of the radio resources and the UE energy consumption achieved with the proposed and reference scheduling schemes.

A. Scenario

The evaluation considers a single cell scenario with N_{UE} UEs that generate aperiodic traffic following a Poisson process with exponentially distributed inter-arrival packet time with λ equal to 0.1 and 1 packet/s. UEs transmit small packets of 32 bytes following [8]. We consider numerology 3 with a subcarrier spacing of 120 kHz and T_{slot} equal to 0.125 ms[8]. There are 7 AN minislots per slot (i.e., n_{AN} = 7 and T_{AN} = T_{slot}/7), and the AN minislots have a duration of 2 symbols. We consider 6 RBs in frequency (i.e., N_F=6). We utilize the values of $\mathbb{P}_d$ derived in [22] when using advanced receivers with multi-user detection capability for the reference scheme. $\mathbb{P}_d$ was calculated in [22] using simulations considered a single cell network deployed in a factory hall. [22] as the percentage of collided packets that are at least 5 dB stronger than the other packets (transmitted by other UEs) that are simultaneously received.

B. Latency and Reliability Performance

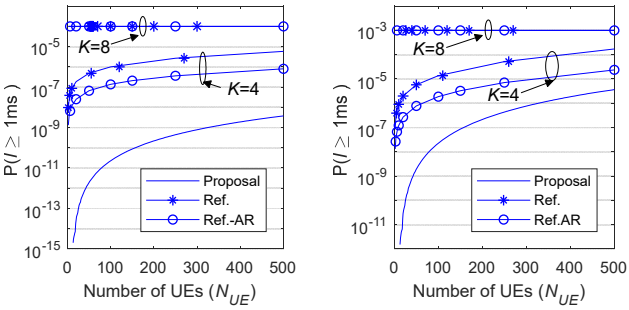
We first evaluate the capacity to meet the latency and reliability requirements established by 3GPP in [8] for URLLC services (i.e. P_{rel}=1-10^{-5} and L=1 ms for packets of 32 bytes). We should note that 1 ms corresponds to m=4 generation periods T_G with the proposed sensing-based grant-free scheduling under the considered scenario. This is the case because T_G is equal to 0.25 ms (Section IV) since T_{slot}=0.125 ms and $n_{AN}\cdot T_{AN}$=0.125 ms. We should also note that the reference scheme can transmit up to 8 replicas of a packet within 1 ms. This is the case because each replica is transmitted in a slot with T_{slot}=0.125 ms.

Fig. 4 compares the probability $\mathbb{P}(l\geq 1$ ms) that a packet experiences a latency equal to or higher than 1 ms with the

[8] We also evaluated the performance with different numerologies and values of T_{slot}. As expected, the evaluation showed that varying these two parameters impacts the performance, e.g. the latency increases when using lower numerologies and higher values of T_{slot}. However, similar trends regarding the comparison of the proposed and reference scheduling schemes have been obtained for all the analysed numerologies and values of T_{slot}.

proposed scheme and with the reference scheme when configured to transmit K=4 or K=8 replicas of each packet[9]. Results are shown for the reference scheme with (Ref.-AR) and without (Ref.) advanced receivers. The figure represents the performance achieved when λ is equal to 0.1 or 1 packet/s. Fig. 4 shows that the proposed grant-free scheme and the reference scheme with K=4 can meet the reliability requirement P_{rel}=1-10^{-5} (i.e. $\mathbb{P}(l \geq 1\,\text{ms}) \leq 10^{-5}$) for up to 500 UEs when λ is equal to 0.1 packets/s. Fig. 4 shows that increasing K to 8 degrades the performance of the reference scheme as it can only guarantee a reliability of P_{rel}=1-10^{-4}. [22] showed that transmitting more replicas per packet reduces the probability of collision and improves the performance of the reference scheme. However, the evaluation in [22] was limited to a maximum of K=4. The authors demonstrated in [20] that the performance of the reference scheme can degrade with higher values of K due to the effect 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 generates a queueing delay that ultimately impacts the latency and reliability performance. The probability $\mathbb{P}_{sc}$ of self-collisions was derived in [20]. We consider in this study the scenario where self-collisions are less probable, i.e. the best-case scenario for the reference scheme. This best-case scenario occurs when the first of two consecutive packets is generated just before the beginning of the slot where the first replica of the packet is transmitted. In this case, $\mathbb{P}_{sc}$ is equal to 9.99·10^{-5} and 9.99·10^{-4} when K=8 and λ is equal to 0.1 packet/s and 1 packet/s, respectively (see (13)). When K=4, $\mathbb{P}_{sc}$ is null, and self-collisions do not impact the latency (and hence the reliability). Fig. 4 shows that self-collisions limit the reliability of the reference scheme only when K=8 (independently of whether using ARs or not) and $\mathbb{P}(l \geq 1\text{ms}) \backsimeq \mathbb{P}_{sc}$. On the other hand, self-collisions does not affect the reference scheme when K=4 and $\mathbb{P}(l \geq 1\text{ms}) \backsimeq \mathbb{P}_{c}$. Fig. 4 also shows that the proposed grant-free scheduling scheme is still able to support 500 UEs with the established latency and reliability requirements when λ (and the traffic load) increases to 1 packet/s. This is not the case for the reference scheme with K=4 that can only meet the latency and reliability requirements for 312 UEs if ARs are used, and for 82 UEs if ARs are not used.

Table II reports the number of UEs that each scheme can support when considering different reliability requirements and traffic loads (λ) with a latency deadline L of 1 ms. Results are only reported for K=4 since Fig. 4 already showed that an implementation of the reference scheme with K=8 cannot even meet the lowest reliability requirement in Table II (i.e. P_{rel}=1-10^{-5}). The table shows that the proposed sensing-based grant-free scheduling scheme always outperforms the reference scheme and supports more UEs than the reference scheme independently of whether ARs are used or not. For example, the proposed scheme guarantees a maximum latency equal to 1 ms with a reliability of 1-10^{-9} to 325 and 42 UEs when λ is equal to 0.1 and 1 packets/s respectively. However, the reference scheme is not able to guarantee reliability levels higher than 1-10^{-7} for more than 1 UE even when using ARs because of collisions when K=4 and self-collisions when K=8. For lower reliability levels, the use of ARs increases the number of UEs that the reference scheme can support.



a) λ=0.1 packet/s. b) λ=1 packet/s.

Fig. 4. Probability that UEs experience a latency higher than 1 ms (N_F=6).

TABLE II
NUMBER OF UES SUPPORTED WITH A RELIABILITY OF P_{rel} AND A LATENCY LOWER THAN 1 MS (N_F=6, K=4 FOR THE REFERENCE SCHEME)

P_{rel}	λ = 0.1 packets/s			λ = 1 packets/s		
	Proposal	Ref.	Ref.-AR	Proposal	Ref.	Ref.-AR
1-10^{-5}	>500	>500	>500	>500	82	312
1-10^{-7}	>500	12	78	158	2	8
1-10^{-9}	325	1	1	42	1	1
1-10^{-11}	78	1	1	18	1	1

Previous results have shown that the proposed sensing-based grant-free scheduling scheme can support a large number of UEs with latency requirements as low as 1 ms. However, many URLLC applications have more relaxed latency requirements, e.g., discrete manufacturing demands latencies between 1 and 12 ms and a reliability of 1-10^{-9} [11]. Fig. 5 evaluates then the probability $\mathbb{P}(l \geq L)$ for the proposed scheme when L is set equal to 1, 1,5 and 2 ms. The figure also shows $\mathbb{P}(l \geq L)$ for the reference scheme with and without AR when configured to transmit K=8 replicas of each packet and L=2 ms. When L=2 ms, the impact of self-collisions is null. However, $\mathbb{P}(l \geq L)$ increases due to the effect of self-collisions when L<2 ms. Fig. 5 shows that the proposed and reference schemes significantly reduce $\mathbb{P}(l \geq L)$ and increase the

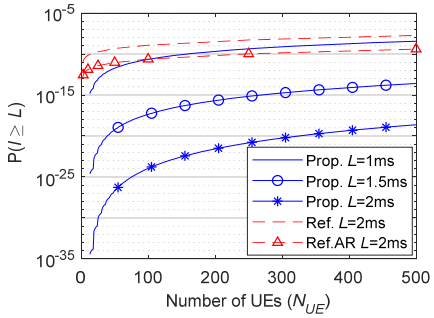


Fig. 5. $\mathbb{P}(l \geq L)$ as a function of the number of UEs for different values of the latency requirement L. Results are shown for the proposed and reference scheme with K=8 (λ=0.1 packet/s).

[9] 3GPP Release 15 consider values of K equal to 1, 2, 4 or 8 [27].

reliability levels (P_{rel}=1- $\mathbb{P}(l \geq L)$) when relaxing the latency requirements. Higher gains are achieved with the proposed scheme that can sustain reliability levels as high as 1-10^{-18} for more than 500 UEs when L is set equal to 2 ms. The proposed scheme can then support most URLLC applications (that demand lower reliability levels) for a large number of UEs.

C. Radio Resource Efficiency

Achieving high reliability levels should not be done at the expense of an inefficient use of radio resources. This section evaluates then the efficiency in the utilization of radio resources achieved by the proposed and reference scheduling schemes. To this end, we compute first the number N_{RB} of RBs used by each scheme to serve a number of UEs. Without loss of generality, we consider $L>K\cdot T_G$ for K equal to 4 and 8 for the reference scheme, and that UEs can contend for a maximum of M data slots before the maximum latency L with the proposed scheme, i.e. $L \geq M\cdot T_G$. For the reference scheme, N_{RB} depends on the mathematical expectation of the number of packets generated for all UEs in a given time period Δt ($\lambda\cdot N_{UE}\cdot\Delta t$) multiplied by the number K of replicas transmitted for each packet; Δt is selected higher than L. In addition, we need to consider the probability of collision of a replica which is given by $\mathbb{P}_{c,1}$ computed using (16) with k=1[10]. $\mathbb{P}_{c,1}$ calculates the probability that two or more UEs have a replica of a packet to transmit in the same slot, and that two or more of these UEs select the same RB for their transmissions. The number of RBs used with the reference scheme to support N_{UE} UEs is then equal to:

$$N_{RB} = \lambda\cdot K\cdot\Delta t + \lambda\cdot K\cdot\Delta t\cdot(N_{UE}\text{-}1)\cdot(1\text{-}\mathbb{P}_{c,1}). \quad (22)$$

The number N_{RB} of RBs used by the proposed sensing-based grant-free scheduling scheme is given by the mathematical expectation of the number of data packets generated by all UEs in a given time period Δt ($\lambda\cdot N_{UE}\cdot\Delta t$) multiplied by the probability that a UE gains access to a data slot to transmit its packet before the latency deadline L expires. This probability is given by 1-$\mathbb{P}(l\geq(M+1)\cdot T_G)$. Considering that the proposed scheme requires one RB for the contention process and one RB for the transmission of a data packet, we can compute N_{RB} as follows:

$$N_{RB} = 2\cdot\lambda\cdot\Delta t\cdot N_{UE}\cdot[1\text{-}\mathbb{P}(l\geq(M+1)\cdot T_G)] \quad (23)$$

Fig. 6 compares the number of RBs used per frame[11] (Δt=10 ms) by the proposed and reference schemes. We consider L=1 ms, and M=4 for the proposed scheme. The reference scheme is configured with K=4 and with/without ARs. The results are normalized to the average traffic arrival rate (λ). Fig. 6 shows that the reference scheme requires a higher number of RBs to serve a given number of UEs than the proposed sensing-based grant-free scheduling scheme. The number of RBs used by the proposed scheme is approximately equal to $2\cdot\lambda\cdot\Delta t\cdot N_{UE}$ for the values of λ and N_{UE} evaluated; $\mathbb{P}(l\geq 5\cdot T_G)$ is approximately equal to zero for all values of λ and N_{UE} evaluated. The number of RBs used by the reference scheme is approximately equal to $K\cdot\lambda\cdot\Delta t\cdot N_{UE}$ when λ=0.1 packet/s, but it decreases when λ increases to 1 packet/s for large number of UEs due to the increase of the probability $\mathbb{P}_{c,1}$ of collision of a replica. The proposed scheme reduces then the use of RBs by a factor of 1.99 and 1.92 compared to the reference scheme when N_{UE} is equal to 500 and λ is equal to 0.1 and 1 packet/s respectively. The reduction tends to K/2 when N_{UE} and/or λ decrease.

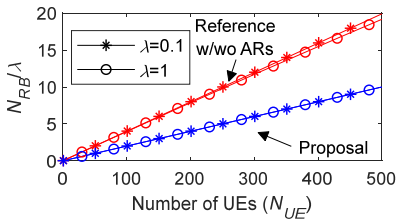


Fig. 6. Number of RBs used per frame (Δt=10 ms) normalized to λ with the proposed grant-free (M=4) and reference (K=4) scheduling schemes (L=1 ms).

The number of radio resources to use with each scheduling scheme has to be configured considering the particular requirements in terms of latency, reliability and nodes that need to be supported. Fig. 7 depicts the number of UEs that can be supported by the proposed and reference schemes as a function of the number N_F of RBs available per slot when L=1 ms. The results are depicted for λ=0.1 packet/s, but similar trends have been observed for λ=1 packet/s. Results for the reference scheme are depicted for K=4 with and without ARs since the effect of self-collisions reduces the reliability that can be achieved with K=8. Results are shown for different reliability levels since the reference scheme cannot guarantee P_{rel}=1-10^{-9} independently of whether using ARs or not. Fig. 7 shows that the proposed scheme can support a significantly higher number of UEs with stricter reliability requirements. Fig. 7 shows that the number of UEs supported by the proposed and reference schemes (with and without ARs) reduces when decreasing the number of RBs available per slot (N_F). Higher reduction levels are observed with the reference scheme. This is because the probability of collision $\mathbb{P}_c$ that affects the reference scheme increases when N_F decreases. In addition, the number of UEs that can be supported with the reference scheme varies as a function of 1-((N_F-1)/N_F)$^{n_i^{act}}$ while it linearly varies with N_F for the proposed grant-free scheduling scheme. For example, the reference scheme needs

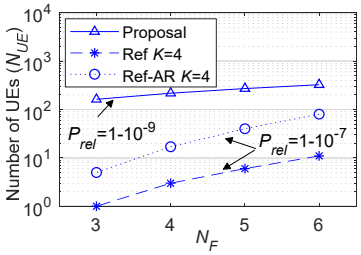


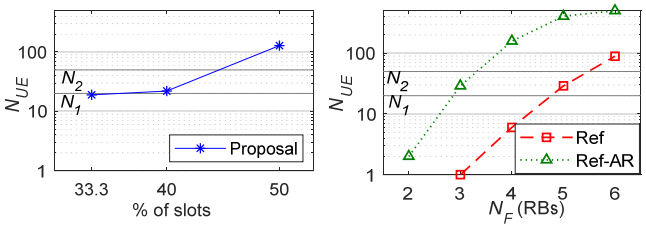
Fig. 7. Number of UEs supported as a function of N_F (L=1 ms and λ=0.1 packet/s).

[10] The same expression is used whether ARs are used or not since we calculate the probability that two or more UEs select the same RB to transmit a replica.

[11] A frame is divided into 80 slots when numerology is equal to 3.

a minimum of 6 RBs to support 80 UEs when using ARs with P_{rel}=1-10^{-7}. The proposed scheme only needs 3 RBs to support 80 UEs with P_{rel}=1-10^{-9}; in fact, it can support up to 162 UEs with N_F=3 RBs.

Fig. 5 in Section VII.B showed that relaxing the latency requirements reduces $\mathbb{P}(l \geq L)$ and increases the reliability P_{rel}. Relaxing the latency requirements resulted in some cases in higher reliability levels than those demanded by most URLLC applications. We then analyse how the proposed scheme can be configured to reduce the usage of RBs when relaxing the latency requirements. To this aim, we consider the case of a packaging machines application in factory automation. Following [12], this application tolerates a maximum latency L of 2.5 ms and requires a reliability equal to 1-10^{-9}. We consider a scenario where we have to support N_1=25 UEs and N_2=50 UEs, the generation period T_G is equal to $4 \cdot T_{slot}$, and there are 6 available RBs per slot in the cell (N_F=6). T_{slot} is set equal to 0.125 ms so there are 20 slots within a time window of 2.5 ms. We consider that all UEs are included in a single group G_1 (i.e. U=1) in the case of the proposed scheme since N_1 and N_2 are lower than the maximum number of UEs that can be supported in a single group. This number is equal to 2^7=128 since we can assign $2^{n_{AN}}$ different priorities with n_{AN} equal to 7 AN minislots in this analysis. We then analyze the minimum percentage of slots within the time window defined by L=2.5 ms during which the RB assigned to G_1 should be used in order to satisfy the application requirements of N_1 and N_2 UEs. We consider that the proposed scheme reserves the RB for data transmissions every x slots. We must also reserve the RB in the slot prior to a data slot for the transmission and sensing of AN messages. Fig. 8.a shows the number of UEs supported by the proposed scheme when varying the percentage of slots during which the RB is used by the proposed scheme. The figure shows that we must reserve the RB for 40% of slots within the time window defined by L=2.5ms in order to satisfy the application requirements of N_1 UEs. This percentage increases to 50% for N_2 UEs.



a) Proposed scheme. b) Reference scheme (K=8).

Fig. 8. Number of UEs supported as a function of the reserved resources (L=2.5ms, P_{rel}=1-10^{-9}, λ=0.1 packet/s).

We also analyze the minimum number of resources that must be reserved for the reference scheme in order to satisfy the application requirements for N_1 and N_2 UEs. The reference scheme requires UEs to transmit K replicas of a packet in consecutive slots. The UEs start the transmission of a packet in the first slot after the generation of the packet. The following replicas are transmitted in consecutive slots, and UEs randomly select an RB from the available N_F RBs for transmitting each replica according to [22]. Fig. 8.b shows the number of UEs supported with the reference scheme with and without using ARs when N_F varies between 3 and 6. The analysis is conducted with K=8 since this configuration can reach reliability levels of 1-10^{-9} because the probability P_{sc} of self-collisions is null when the latency deadline L increases to 2.5 ms; a configuration with K=4 cannot reach the requested reliability level. We should note that the reference scheme cannot satisfy the application requirements for N_1 or N_2 UEs with less than 3 RBs in frequency, while it can always satisfy them with more than 6. Fig. 8.b shows that the reference scheme requires 5 RBs per slot to support N_1 UEs when ARs are not used and only 3 RBs per slot when ARs are used. When the number of UEs to be supported increases to N_2, the reference scheme requires 6 and 4 RBs per slot without and with ARs to satisfy the application requirements. Using ARs reduces the number of required RBs because ARs reduce the impact of collisions thanks to the use of multi-user detection schemes.

To compare how efficiently both schemes utilize the radio resources, Table III reports the average number of UEs that each scheme can support[12] per reserved resource in the scenario under study. A resource corresponds to one RB in one slot. The results are reported for scenarios where we reserve the minimum number of resources needed by each scheme to satisfy the application requirements of N_1 and N_2 UEs (based on Fig. 8). Table III shows that the proposed scheme can support a significantly higher number of UEs per resource compared to the reference scheme (with or without ARs) for the two scenarios. This is the case because the proposed scheme can avoid collisions between UEs thanks to its contention process. The reference scheme cannot avoid such collisions but reduces their impact with the transmission of K replicas per packet. However, transmitting multiple replicas per packet consumes more resources per packet and utilizes less efficiently the resources. Table III shows that the reference scheme improves its efficiency when using ARs. This is the case because multi-user detection schemes reduce the impact of collisions, and hence the reference scheme can use less resources to support a number of UEs than when not using ARs. Table III also shows that it is possible to support more UEs per resource in the scenario where we reserve resources to satisfy N_2 UEs compared to the scenario where we reserve resources to satisfy N_1 UEs. This is the case because we need to reserve more resources to serve N_2 UEs than N_1 UEs (Fig. 8). In this case, the reference scheme experiences a lower probability of packet collisions (see (15)), and it can support more UEs per resource. The proposed scheme increases the percentage of slots reserved when configured to serve N_2 UEs compared to when configured to serve N_1 UEs (Fig. 8). This reduces the time between contention slots and the number of UEs that compete for the next data slot, increases the probability to access a data slot, and decreases the time during which UEs need to contend for

[12] A UE is effectively supported if its packets can be transmitted in less than L=2.5ms.

data slots. This increases the number of UEs that can transmit their packet before the maximum latency L and augments the number of UEs that can be supported per UE.

TABLE III
AVERAGE NUMBER OF UES SUPPORTED PER RESOURCE
(L=2.5 MS, P_{REL}=1-10^{-9}, λ=0.1 PACKET/S).

N_1=20 UEs			N_2=50 UEs		
Proposal	Ref.	Ref.-AR	Proposal	Ref.	Ref.-AR
2.75	0.29	0.48	12.8	0.74	1.98

D. Energy Consumption

Reducing the energy consumption is a key objective for beyond 5G networks. This section compares then the energy consumed by a UE in the transmission of a data packet with the proposed and the reference schemes.

The energy consumed by a UE per data packet with the reference scheme is equal to the energy consumed by the transmission of the K replicas of the packet. If we denote E_{slot} the energy consumed by a UE for the transmission of a replica of a packet in a slot of duration T_{slot}, the average energy consumption of a UE for the reference scheme is expressed as:

$$\overline{E} = K \cdot E_{slot}. \tag{24}$$

With the proposed grant-free scheduling scheme, a UE must first contend and gain access to a data slot before transmitting the packet. In this case, we must then compute the energy E_{AN} consumed in the contention processes in which the UE participates, and the energy E_p consumed in the transmission of the packet. The energy consumed by a UE in a contention process during a generation period depends on the UEs' priority since higher priority UEs will transmit less AN messages before getting access to a data slot. To compute the energy consumed in the contention process, we consider the worst-case scenario in terms of energy consumption, i.e. the case in which a UE transmits an AN message at each AN minislot. Since the duration of the contention process is equal to $T_{slot}=n_{AN} \cdot T_{AN}$, the energy consumed by a UE in the contention process is equal to E_{slot}. We need to calculate the average number of contention processes in which a UE participates to estimate the average energy $\overline{E_{AN}}$ consumed by the UE. To this aim, let us consider a UE$_j$ in a S_u that generated a new packet in G_i. UE$_j$ contends for the next shared data slot i with probability equal to 1. UE$_j$ will contend for shared data slot $i+m$ (with $m \geq 1$) if it cannot gain access to any of the previous slots $i,\ldots, i+m-1$. As presented in Section V.A, the probability that UE$_j$ contends for the shared data slot $i+m$ is given by $\mathbb{P}_{S_u}(l \geq (m+1) \cdot T_G)$. Moreover, UE$_j$ will drop the data packet if it has not been transmitted after the latency deadline L. As presented in Section IV, the time period between the beginning of a contention process and the end of the next data slot is equal to T_G. We can define M as the maximum number of shared data slots for which UE$_j$ can contend before the latency deadline L, i.e. $M \cdot T_G \leq L$. When L=1 ms, the packet will be dropped if UE$_j$ does not get access to any of the M=4 data slots (i, i+1, i+2, i+3) after the packet is generated (this corresponds to a latency equal to or higher than $4 \cdot T_G$ =1 ms). The average number of contention processes in which a UE participates is then equal to $1+\sum_{m=1}^{M-1} \mathbb{P}_{S_u}(l \geq (m+1) \cdot T_G)$, and the average energy consumed by a UE due to contention processes is then given by:

$$\overline{E_{AN}} = \left[1+\sum_{m=1}^{M-1} \mathbb{P}_{S_u}(l \geq (m+1) \cdot T_G)\right] \cdot E_{slot} \tag{25}$$

The probability that a UE cannot finally transmit its packet after participating in M contention processes is $\mathbb{P}_{S_u}(l \geq (M+1) \cdot T_G)$. The probability that a UE is able to transmit its data packet is then $1-\mathbb{P}_{S_u}(l \geq (M+1) \cdot T_G)$, and the average energy $\overline{E_p}$ consumed by a UE in the transmission of a data packet is computed as:

$$\overline{E_p} = \left[1-\mathbb{P}_{S_u}(l \geq (M+1) \cdot T_G)\right] \cdot E_{slot}. \tag{26}$$

The average energy consumed by a UE per data packet with the proposed sensing-based grant-free scheduling is then:

$$\overline{E} = \left[1+\sum_{m=1}^{M-1} \mathbb{P}_{S_u}(l \geq (m+1) \cdot T_G)\right] \cdot E_{slot} + \left[1-\mathbb{P}_{S_u}(l \geq (M+1) \cdot T_G)\right] \cdot E_{slot} \tag{27}$$

In (27), $\mathbb{P}_{S_u}(l \geq (m+1) \cdot T_G) \ll 1$ for all $m \geq 1$, so the energy consumed by a UE per data packet with the proposed scheme is approximately equal to:

$$\overline{E} = 2 \cdot E_{slot}. \tag{28}$$

(24) and (28) show that the proposed sensing-based grant-free scheduling scheme reduces the energy consumed by a UE compared to the reference scheme. The reduction is approximately equal to 50% when K=4 for the reference scheme, and equal to 75% when K=8.

VIII. CONCLUSIONS

This paper has presented and evaluated a novel sensing-based grant-free scheduling scheme for 5G and Beyond networks. The scheme is able to satisfy stringent reliability and latency requirements of deterministic and aperiodic UL traffic characteristic of verticals such as Industry 4.0 or manufacturing. The scheme uses shared resources and avoids packet collisions among UEs using a priority-based contention resolution process that relies on the transmission of announcement messages in minislots and a local channel sensing process. This study demonstrates that the proposed sensing-based grant-free scheduling scheme significantly outperforms current 5G NR grant-free scheduling with K-repetitions and shared resources even when using advanced receivers with multi-user detection capabilities to combat the effect of packet collisions. The proposed scheme reduces the energy consumption and can support a higher number of UEs with URLLC and deterministic requirements with a considerably lower number of radio resources compared to current 5G NR grant-free scheduling implementations with K-repetitions and shared resources.

APPENDIX A: $\mathbb{P}_{S_U}(l \geq 3 \cdot T_G)$

We derive the probability $\mathbb{P}_{S_u}(l \geq 3 \cdot T_G)$ that UE$_j$ experiences a latency equal to or higher than $3 \cdot T_G$ following the same process described in Section V.A to compute

$\mathbb{P}_{S_u}(l \geq 2 \cdot T_G)$. UE$_j$ experiences a latency equal to or higher than $3 \cdot T_G$ if it cannot gain access to data slot i and i+1. $\mathbb{P}_{S_u}(l \geq 3 \cdot T_G)$ is then computed considering the probability that other UEs generate new packets in G_i ($\mathbb{P}_{\mathrm{n}}(n_i,R_i,T_G)$) and G_{i+1} ($\mathbb{P}_{\mathrm{n}}(n_{i+1},R_{i+1},T_G)$), and the probabilities $\overline{\mathbb{P}_{\mathrm{p}}}(C_i,p_{min})$ and $\overline{\mathbb{P}_{\mathrm{p}}}(C_{i+1},p_{min})$ that at least one of these UEs have higher priority than UE$_j$. $\mathbb{P}_{S_u}(l \geq 3 \cdot T_G)$ can then be expressed as follows:

$$\begin{aligned}\mathbb{P}_{S_u}(l \geq 3 \cdot T_G) = & \mathbb{P}_{\mathrm{n}}(1,R_i,T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(1,p_{min}) \\ & \cdot \sum_{n_{i+1}=1}^{|R_{i+1}|} \left(\mathbb{P}_{\mathrm{n}}(n_{i+1},R_{i+1},T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(C_{i+1},p_{min})\right) + \\ & \sum_{n_i=2}^{|R_i|} \left[\mathbb{P}_{\mathrm{n}}(n_i,R_i,T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(C_i,p_{min}) \right. \\ & \left. \cdot \sum_{n_{i+1}=0}^{|R_{i+1}|} \left(\mathbb{P}_{\mathrm{n}}(n_{i+1},R_{i+1},T_G) \cdot \overline{\mathbb{P}_{\mathrm{r}}}(C_{i+1},p_{min})\right) \right]\end{aligned} \tag{29}$$

(29) considers all possible values of n_i and n_{i+1}. The first term of the sum in (29) considers the case for n_i=1. If n_i=1, n_{i+1} should be equal to or higher than 1. The second term of the sum represents the scenarios where n_i is higher than 1, and n_{i+1} can then be equal to or higher than 0. We define variables $n_{i,min}$, $n_{i,max}$ and $n_{i+1,min}$, $n_{i+1,max}$ as the minimum and maximum values of variables n_i and n_{i+1} respectively. If we consider that $n_{i,min}$=1, $n_{i,max}$=$|R_i|$, $n_{i+1,min}$=1 if $|R_i|$=$|S_u|$-2 (this is the case when n_i=1), $n_{i+1,min}$=0 if $|R_i|$>$|S_u|$-2 (this is the case when n_i>1), and $n_{i+1,max}$=$|R_{i+1}|$, (29) can also be expressed as (30).

$$\begin{aligned}\mathbb{P}_{S_u}(l \geq 3 \cdot T_G) = & \sum_{n_i=n_{i,min}}^{n_{i,max}} \left[\mathbb{P}_{\mathrm{n}}(n_i,R_i,T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(C_i,p_{min}) \right. \\ & \left. \cdot \sum_{n_{i+1}=n_{i+1,min}}^{n_{i+1,max}} \left(\mathbb{P}_{\mathrm{n}}(n_{i+1}, R_{i+1},T_G) \cdot \overline{\mathbb{P}_{\mathrm{p}}}(C_{i+1},p_{min})\right) \right]\end{aligned} \tag{30}$$

REFERENCES

[1] 3GPP, TSG SA; Service requirements for cyber-physical control applications in vertical domains; Stage 1; Rel. 16, 3GPP TS 22.104 V16.4.0, Dec. 2019.

[2] 3GPP, TSG RAN; E-UTRA; Study on latency reduction techniques for LTE, 3GPP TR 36.881 V14.0.0, 2016.

[3] 3GPP, TSG RAN; NR; Physical channels and modulation (Release 15), 3GPP TS 38.211, v15.4.0, Dec. 2018.

[4] 3GPP, TSG RAN; NR; Medium Access Control (MAC) protocol specification; Rel. 15, 3GPP TS 38.321 V15.8.0, Jan. 2020.

[5] N. H. Mahmood, *et al.*, "Uplink Grant-Free Access Solutions for URLLC services in 5G New Radio", in *Proc. of the 16th International Symposium on Wireless Communication Systems*, Oulu, Finland, 2019, pp. 607-612.

[6] 5G-ACIA, *5G for Connected Industries and Automation*, Feb. 2019.

[7] 3GPP, TSG SA; Service Requirements for the 5G System; Stage 1; Rel. 15, 3GPP TS 22.261 V15.3.0, Dec. 2017.

[8] 3GPP, TSG RAN; Study on Scenarios and Requirements for Next Generation Access Technologies; Rel.15, TR 38.913, v15.0.0, 2018.

[9] 3GPP, TSG RAN; Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC); Rel. 16, 3GPP TR 38.824 V1.0.0, Nov. 2018.

[10] K. Montgomery et al, "Wireless User Requirements for the Factory Workcell", *NIST Advanced Manufacturing Series 300-8*, R1, Nov. 2020.

[11] ETSI, "Reconfigurable Radio Systems (RRS); Feasibility study on temporary spectrum access for local high-quality wireless networks", ETSI TR 103 588 V1.1.1, Feb. 2018.

[12] P. Schulz et al., "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture", *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70-78, February 2017.

[13] 3GPP TSG-RAN, R1-1810329, "Discussion on configured grant for NR-U", Meeting #94bis, Oct. 2018.

[14] G. Berardinelli, *et al.*, "Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources", *IEEE Access*, vol. 6, pp. 23602-23611, 2018.

[15] Z. Li et al., "5G URLLC: Design Challenges and System Concepts", in *Proc. of ISWCS2018*, Lisbon, 2018, pp. 1-6.

[16] Y. Liu, et al., "Analyzing Grant-Free Access for URLLC Service", *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 741-755, March 2021.

[17] T. N. Weerasinghe, et al., "Priority Enabled Grant-Free Access with Dynamic Slot Allocation for Heterogeneous mMTC Traffic in 5G NR Networks", *IEEE Trans. on Communications*, vol. 69, no. 5, pp. 3192-3206, May 2021.

[18] 3GPP, TSG RAN; NR; Physical layer procedures for data (Release 16), 3GPP TS 38.214, v16.7.0, Sept. 2021.

[19] C. Wang et al., "Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services", in *Proc. of VTC2017-Spring*, Sydney, NSW, 2017, pp. 1-6.

[20] M.C. Lucas-Estañ et al., "On the Capacity of 5G NR Grant-Free Scheduling with Shared Radio Resources to Support Ultra-Reliable and Low-Latency Communications", *Sensors*, vol. 19 (3575), August 2019.

[21] T.-K. Le, et al, "Enhancing URLLC Uplink Configured-grant Transmissions", in *Proc. of VTC2021-Spring*, Helsinki (Finland), 2021.

[22] B. Singh, et al., "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions", *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182-185, April 2018.

[23] Z. Zhao, Q. Du and L. Sun, "Network-Load Estimation for K-Repetition Grant-Free Access Enabling Adaptive Resource Allocation Towards QoS Enhancement", in *Proc. of 2021 PIMRC*, Helsinki (Finland), 2021, pp. 1073-1078.

[24] 5GIA, *European Vision for the 6G Network Ecosystem*, June 2021.

[25] 3GPP, TSG RAN; Study on New Radio Access Technology Physical Layer Aspects; Rel. 14, 3GPP TR 38.802, v14.2.0, Sept. 2017.

[26] R. Kotaba et al., "Uplink Transmissions in URLLC Systems With Shared Diversity Resources", *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 590-593, Aug. 2018.

[27] 3GPP, TSG RAN; NR; Radio Resource Control (RRC) protocol specification; Rel. 16, 3GPP TS 38.331, v16.3.1, Jan. 2021.