Analyzing Grant-Free Access for URLLC Service

Yan Liu, Student Member, IEEE, Yansha Deng, Member, IEEE, Maged Elkashlan, Member, IEEE, Arumugam Nallanathan, Fellow, IEEE, and George K. Karagiannidis, Fellow, IEEE

Abstract

5G New Radio (NR) is expected to support new ultra-reliable low-latency communication (URLLC) service targeting at supporting the small packets transmissions with very stringent latency and reliability requirements. Current Long Term Evolution (LTE) system has been designed based on grant-based (GB) random access, which can hardly support the URLLC requirements. Grant-free (GF) access is proposed as a feasible and promising technology to meet such requirements, especially for uplink transmissions, which effectively save the time of requesting/waiting for a grant. While some basic grant-free features have been proposed and standardized in NR Release-15, there is still much space to improve. Being proposed as 3GPP study items, three grant-free access schemes with Hybrid Automatic Repeat reQuest (HARQ) retransmissions including Reactive, K-repetition, and Proactive, are analyzed in this paper. Specifically, we present a spatio-temporal analytical framework for the contention-based GF access analysis. Based on this framework, we define the latent access failure probability to characterize URLLC reliability and latency performances. We propose a tractable approach to derive and analyze the latent access failure probability of the typical UE under three GF HARQ schemes. Our results show that the Reactive scheme and the Proactive scheme can provide the lowest latent access failure probability under shorter latency constraints, whereas the K-repetition scheme achieves the best performance under longer latency constraints, except for the scenario of lower SINR threshold and higher UE density.

Index Terms

URLLC, 5G NR, Grant free access, HARQ

Y. Liu, M. Elkashlan and A. Nallanathan are with School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: {yan.liu, maged.elkashlan, a.nallanathan}@qmul.ac.uk).

Y. Deng is with Department of Engineering, King’s College London, London WC2R 2LS, U.K. (Corresponding author: Yansha Deng (e-mail:yansha.deng@kcl.ac.uk)).

G. K. Karagiannidis is with Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54 124, Greece (e-mail: geokarag@auth.gr).
I. INTRODUCTION

The Fifth Generation (5G) New Radio (NR) considers three new communication service categories: enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC), and Ultra-Reliable Low-Latency Communications (URLLC) [1] [2]. Among them, URLLC service is an essential element for applications, including factory automation [3], automation vehicles [4], remote control [5], and virtual/augmented reality (VR/AR) [6], which has stringent requirements on low latency and high reliability for small packets transmissions. The third Generation Partnership Project (3GPP) has defined a general URLLC requirement: \(1 - 10^{-5}\) reliability within 1ms user plane latency\(^1\) for 32 bytes (0.5ms for both downlink (DL) and uplink (UL)) [1]. More details about the variety of different traffic characteristics and the requirements of some URLLC use cases can be found in [7] [8]. For example, the automation use case requires \(1 - 10^{-5}\) reliability within 10ms for remote motion control; the intelligent transportation use case requires \(1 - 10^{-6}\) reliability within 5ms for cooperative collision avoidance.

Current Long Term Evolution (LTE) system can hardly fulfill the URLLC requirements. Especially in the uplink, current LTE utilizes a scheduling based transmission mode, namely, grant-based (GB) scheduling as specified in [9]. This conventional GB scheduling is initiated by the User Equipment (UE) with an access request to the network in which the Base Station (BS) can respond by issuing an access grant through a four-step random access (RA) procedure as shown in Fig. 1. Such scheduling-request-triggered transmission would take at least 10ms before starting the data transmission, which is far from the URLLC latency requirement. Recently, uplink grant-free (GF) access has been proposed and extensively discussed in 3GPP RAN WG1 [10]–[12] to cope with the URLLC requirement in the uplink transmission. With uplink GF access, a UE with a small packet can transmit data along with required control information in the first step transmission itself. This can greatly reduce the RA and data transmission latency, as the scheduling request and grant issuing step in GB RA are removed as shown in Fig. 1.

In the GF transmission, the frequency resource can be reserved in advance or allocated at the time when there is a request. Preallocation of dedicated resource, known as Semi-Persistent-Scheduling (SPS) [10], is more suitable for periodic traffic with a fixed pattern, whereas contention-based GF transmission over shared resource is more suitable for sporadic small

\(^1\)User plane latency is defined as the one-way latency from the processing of the packet at the transmitter to when the packet has been received successfully, and includes the transmission processing time, transmission time and reception processing time.
Fig. 1. Uplink transmissions for grant-based and grant-free random access

packets, as it is more efficient and flexible in terms of resource utilization. However, contention-based GF transmission is subject to potential collisions with other neighbouring UEs transmitting simultaneously over the shared resource, thus jeopardizing the transmission reliability.

A standard technique to improve transmission reliability, which has been adopted in various wireless standards, is Hybrid Automatic Repeat reQuest (HARQ) retransmission [13]. Conventional HARQ allows for retransmissions only upon reception of a Negative ACKnowledgement (NACK). This requires the BS to first receive the packet for detection, then issue the feedback. This is the so called Reactive (Reac) HARQ scheme, where retransmissions are triggered only when there is a failure in the previous transmission. However, the Reactive HARQ scheme introduces additional latency, as the UE needs to wait for the feedback before performing a retransmission, which is determined by the HARQ round-trip-time (RTT), i.e., time duration of the cycle from the beginning of the transmission until processing its feedback [14]. Thus, the Reactive HARQ scheme only allows for a limited number of retransmissions due to the stringent latency requirement in URLLC service [14], this fact motivated more research for advanced HARQ schemes to be integrated with GF transmission to provide reduced latency and enhanced reliability.

One candidate scheme is the $K$-repetition (Krep) scheme proposed in the current 3GPP NR Release-15, where the pre-defined number ($K_{Krep}$) of consecutive replicas of the same packet are transmitted without waiting for the feedback, and then the BS performs soft combining of these
repetitions to improve the reliability [15]. Another candidate scheme is known as the Proactive (Proa) scheme, which has been first discussed in [16]. In a Proactive scheme, the UE still repeats transmissions in consecutive transmission time intervals (TTIs) like K-repetition scheme with maximum $K_{Proa}$ times, but if the UE receives and decodes a positive feedback (ACK) from the BS before reaching maximum $K_{Proa}$ times, the repetition will be terminated. It is noted that this scheme is more computational heavy for the UE, as the UE has to monitor the feedback.

Another standard technique to enhance reliability is the efficient random access control mechanism, including the Access Class Barring (ACB), the Back-Off (BO) and the Power Boosting (PB) schemes [17]. However, both the ACB and the BO scheme make a group of UEs completely barred in specific time slots, which will introduce extra latency for these UEs. As such, we just consider the GF HARQ schemes integrated with the Power Boosting, which can quickly compensate unexpected Signal-to-Interference plus-Noise Ratio (SINR) degradations at the initial transmissions [18]. Specifically, if a transmission fails, the UE uses the full path-loss inversion power control to maintain the average received power at a higher power level in the next retransmission, where the power control is one candidate technology component for UL transmission with the focus on improving the reliability.

Despite that the aforementioned GF HARQ access designs are proposed to govern the URLLC service, their theoretical formulations and comparative insights have never been fully established. Recent works [18] [19] have evaluated the Reactive, K-repetition, and Proactive schemes for URLLC service through system-level simulation without analytical characterization. The authors in [18] claimed that the effects of inter- and intra-cell interference, queuing and time-frequency variant channels, are difficult or even infeasible to evaluate with analytical models. This is because existing wireless systems were designed mainly to maximize the data rates of the long packet transmission, the short packet transmission in URLLC service challenges the existing wireless system in terms of the joint reliability and latency requirement. To cope with it, correctly modeling and analyzing the reliability and latency is fundamentally important, but the interplay between latency and reliability brings extra complexity. In this paper, we address the following fundamental questions: 1) how to quantify the URLLC reliability and latency performance; 2) how to examine whether different GF schemes satisfy the URLLC requirements or not; 3) how to identify the GF scheme performs better in a certain specific scenario. To do so, we present a novel spatio-temporal mathematical framework to analyze and evaluate both the reliability and latency performances for three different GF access HARQ schemes. The main contributions of
this paper can be summarized in the following:

• We present a novel spatio-temporal mathematical framework for analyzing contention-based GF access HARQ schemes for URLLC service by using stochastic geometry and probability theory. In the spatial domain, stochastic geometry is applied to model and analyze the mutual interference among active UEs (i.e., those with non-empty data buffer). In the time domain, probability theory is applied to model the correlation of the buffer state and the transmission state over different time slots.

• Based on this framework, we propose a tractable approach to characterize and analyze the URLLC performances of a randomly chosen UE by defining the latent access failure probability. We then derive the exact closed-form expressions for the latent access failure probabilities of the UE under three different contention-based GF HARQ schemes, including Reactive, K-repetition, and Proactive schemes, respectively.

• We develop a realistic simulation framework to capture the randomness locations, pilot and data transmission as well as the real packets of each UE in each TTI to verify our derived latent access failure probability. We compare the effectiveness of the three different GF HARQ schemes. Our results show that the Reactive scheme and Proactive scheme provided the lowest latent access failure probability under shorter latency constraints, while the K-repetition scheme has the lowest latent access failure probability as well as the most improvement with Power Boosting under longer latency constraints, except for the lower SINR threshold and high UE density scenario.

The rest of the paper is organized as follows. Section II provides the problem formulation and system model. Section III analyzes the URLLC performance by deriving the expressions of the latent access failure probability of a randomly chosen UE with three different GF HARQ schemes. Section IV provides numerical results. Finally, Section V concludes the work.

II. PROBLEM FORMULATION AND SYSTEM MODEL

A. Network Model

We consider a single layer cellular network, where the BSs and the UEs are spatially distributed following two independent Poisson Point Processes (PPPs) \( \Phi_B \) and \( \Phi_D \) with intensities \( \lambda_B \) and \( \lambda_D \), respectively. We assume that each UE associates to its geographically nearest BS, where a Voronoi tessellation is formed. The UEs are connected and synchronized to the serving cell.
Moreover, we consider additive noise with average power $\sigma^2$ and a flat Rayleigh fading channel, i.e. the channel response is constant over the selected Resource Blocks (RBs), however, it can vary at every transmission or retransmission. The channel power gain $h$ is assumed to be exponentially distributed with unit mean, i.e., $h \sim \text{Exp}(1)$. All channel gains are assumed to be independent and identically distributed (i.i.d.) in space and time. We consider the path loss model with the path-loss attenuation $x^{-\alpha}$, where $x$ is the propagation distance and $\alpha$ is the path-loss exponent. We apply a full path-loss inversion power control at all UEs to solve the “near-far” problem, where each UE compensates for its own path-loss to keep the average received signal power equal to a same threshold $\rho$. We also assume the density of BSs is high enough and no IoT device suffers from truncation outage [17].

B. Contention-Based Grant-Free Access

In this paper, we consider the uplink contention-based GF access for UEs with sporadic small packets with URLLC requirements, where UEs transmit data in an arrive-and-go manner without sending a scheduling request and receiving resource grant from the network. According to [20] [21], each UE chooses a pilot sequence from a predetermined orthogonal pilot sequence set ($S$) and transmits its selected pilot and data simultaneously. Each UE has a data buffer that stores packets received from higher layers. An i.i.d. Bernoulli traffic generation model with probability of $p_a \in [0, 1]$, is assumed at each buffer. Note that we only consider a single packet sequence arrival. This packet sequence will be removed from the buffer and the buffer becomes empty without new packet, once it has been successfully transmitted, otherwise, this UE will wait and reattempt in the next HARQ retransmission.

GF uplink transmissions occur in a slotted-Aloha system based on OFDM (Orthogonal Frequency Division Multiplexing) within short-TTI [22]. The UEs are configured by radio resource control (RRC) signaling prior to the GF access (as Type 1 UL [23]), with time and frequency resource, modulation and coding scheme (MCS), power control settings, and HARQ related parameters. The configured UEs are connected and synchronized, thus being always ready for

25G NR introduces the concept of ‘mini-slots’ and supports a scalable numerology allowing the sub-carrier spacing (SCS) to be expanded up to 240 kHz. Collectively, this allows transmissions over shorter intervals than that in LTE to meet the stringent latency requirement. In this paper, the TTI refers to a mini-slot, which is shorter than the typical coherence times that are of the order of few milliseconds. But generally, the coherence time could be normalized. In addition, the repetitions could be performed over different RBs in frequency so that the channel gains i.i.d assumption is justified.
a URLLC transmission. We consider $N$ UEs pre-configured with $S$ pilots, i.e., $S$ carriers over one TTI, for their uplink GF transmissions in the frequency domain [24]. In each TTI, UEs randomly move to a new position, and the active ones randomly select one of the available $S$ pilots for their transmissions. A collision occurs when two or more UEs choose the same pilot sequence. According to the thinning process [25], the density of active UEs choosing the same pilot can be derived as

$$\lambda_a = p_a \lambda_D / S.$$  \hspace{1cm} (1)

### C. Grant-Free Access Schemes

This section provides a general description of the three GF HARQ schemes considered in this paper. For ease of description, we first present definitions for general variables. As illustrated in Fig. 2, the frame alignment (A) delay is denoted as $T_{fa}$, the packet transmission (T) time is denoted as $T_{tx}$, and the processing (DP) time at the BS is denoted as $T_{dp}$. If the packet is successfully decoded, the BS sends an ACK feedback, otherwise it sends a NACK, where the ACK/NACK feedback (F) time are represented by $T_{fd}$. After having received and decoded the feedback, the UE can decide whether to perform a retransmission. The processing time at the UE is denoted by $T_{up}$. The frame alignment delay $T_{fa}$ is a random variable uniformly distributed between zero and one TTI [26]. Depending on the packet size, channel quality and scheduling strategy, the transmission time $T_{tx}$ can vary from one to multiple TTIs. Considering the small packets of URLLC traffic, we assume $T_{fa} = 1$ TTI and $T_{tx} = 1$ TTI in this work same as [27]. The BS feedback time $T_{fb}$ and the BS (UE) processing time $T_{dp}$ ($T_{up}$) are also assumed to be one TTI. Then, the latency framework of the three GF HARQ schemes are described as follows.

1) **Reactive scheme:** The Reactive scheme is illustrated in Fig. 2. After the UE finalizes its initial uplink transmissions (T), its signal will be processed at the BS (DP) for a HARQ feedback (F) (ACK/NACK). After processing the HARQ feedback (UP), the UE retransmits the same packet upon reception of a NACK. In this scheme, we note that the HARQ round trip time

$$T_{RTT}^{Reac} = 4\text{TTIs.}$$  \hspace{1cm} (2)

Then the latency after $m$ HARQ round trips is obtained as

$$T_{Reac}[m] = T_{fa} + mT_{RTT}^{Reac} = T_{fa} + m(T_{tx} + T_{dp} + T_{fb} + T_{up}).$$  \hspace{1cm} (3)
2) **K-repetition scheme**: The K-repetition scheme is illustrated in Fig. 3, where the UE is configured to autonomously transmit the same packet for \( K_{\text{Krep}} \) repetitions in consecutive TTIs. At the end of \( K_{\text{Krep}} \) repetitions, the BS needs to combine the received repetitions, process the received packet, and feedback to the UE. In this scheme, the HARQ round trip time

\[
T_{\text{RTT}}^{\text{Krep}} = (K_{\text{Krep}} + 3) \text{TTIs.}
\]

Then the latency after \( m \) HARQ round trips is defined as

\[
T_{\text{Krep}}[m] = T_{fa} + mT_{\text{RTT}}^{\text{Krep}} = T_{fa} + m(K_{\text{Krep}}T_{tx} + T_{dp} + T_{fb} + T_{up}) = 1 + m(K_{\text{Krep}} + 3) \text{ TTIs.}
\]

3) **Proactive scheme**: The Proactive scheme is illustrated in Fig. 4. Similarly to the K-repetition scheme, the UE is configured to repeat the transmission for a maximum number of \( K_{\text{Proa}} \) repetitions, but can receive the feedback after each repetition. This allows the UE to terminate repetitions earlier once receiving the positive feedback (ACK). We note that the UE could receive the 1st feedback 3TTIs after the 1st repetition. That is to say, the minimum HARQ round trip time is 4. For \( K_{\text{Proa}} \leq 4 \), the UE continues repetitions until maximum \( K_{\text{Proa}} \). For \( K_{\text{Proa}} \geq 5 \), the UE needs to continue repetitions until either the UE receives ACK from the BS, or the number of repetitions reaches maximum \( K_{\text{Proa}} \) times [28]. Let us denote that the 1st access success of the typical UE occurs in the \( l \)th repetition during one HARQ round trip. Thus,
we have the single HARQ round trip time for the Proactive scheme as:

\[ T_{\text{RTT}, \text{Proa}, K, l} = \begin{cases} 
K_{\text{Proa}} + 3, & l = 0, \\
l + 3, & 1 \leq l \leq K_{\text{Proa}}. 
\end{cases} \]  

(6)

Note that if \( l = 0 \), i.e., all the \( K_{\text{Proa}} \) repetitions in one HARQ round trip are not successful with \( T_{\text{Proa}, K, 0} = K_{\text{Proa}} + 3 \), the UE will perform HARQ retransmission in next round trip. Thus, the latency after \( m \) HARQ round trips for the Proactive scheme with a maximum \( K_{\text{Proa}} \) repetitions can be derived as

\[ T_{\text{Proa}}[m] = T_{\text{fa}} + (m - 1)T_{\text{RTT}, \text{Proa}, K, 0} + T_{\text{RTT}, \text{Proa}, K, l} \]

\[ = T_{\text{fa}} + (m - 1)(K_{\text{Proa}} + 3) + T_{\text{Proa}, K, l} = l + 4 + (m - 1)(K_{\text{Proa}} + 3) \text{TTIs} \quad (1 \leq l \leq K_{\text{Proa}}), \]  

(7)

where I denotes that the transmissions in all the former \( (m - 1) \) HARQ round trips are not successful; and II implies the possible case in the final \( m \)th HARQ round trip given in (6).

D. Signal to Noise plus Interference Ratio (SINR)

Note that the GF access outage occurs due to the following two reasons: 1) a pilot cannot be recognized by the received BS, due to its lower received SINR than the SINR threshold \( \gamma_{th} \); 2) the BS successfully received two or more same pilots simultaneously, such that the collision
occurs, and the BS cannot decode any collided pilots. Our model follows the assumption of collision model in [20] [29], where all these collision UEs would not be decoded at the BS. Different from the data transmission with no intra-cell interference due to orthogonal resource allocation, the GF access analysis in this work needs to take into account both the inter- and intra-cell interference. We formulate the SINR of a typical BS located at the origin as

$$\text{SINR}_m = \frac{g_m \rho h_0}{I_{\text{intra}} + I_{\text{inter}} + \sigma^2},$$

where $\rho$ is the full path-loss inversion power control threshold, $h_0$ is the channel power gain from the typical UE to its associated BS, $\sigma^2$ is the noise power, $I_{\text{intra}}$ and $I_{\text{inter}}$ are the aggregate intra-cell and inter-cell interference, which will be discussed in detail in Section III, and $g_m$ denotes the power level unit in the $m$th retransmission by adjusting the target received power at the BS equal to $g_m \rho$ [20] [31] (i.e., $g_1 < g_2 < \ldots < g_m < \ldots < g_J$). Note that $g_J$ is the maximum allowable power level unit.

### E. Problem Formulation and Objectives

The URLLC requirement of the UL GF transmission is that the UEs can successfully complete their payload delivery within a limited time, i.e., $T_{\text{latency}} \leq T$, with a failure probability lower than a certain target, i.e., $P_F \leq \varepsilon$. For its performance characterization, we define the latent access failure probability in the following.

**Definition 1.** The latent access failure probability is defined as

$$P_F(T_{\text{latency}} \leq T) \leq \varepsilon$$

In 5G specification, $\varepsilon = 10^{-5}$ and $T = 1$ ms for general URLLC requirement [1].

## III. Performance Analysis and Evaluation

This section presents a general analytical model for three different GF HARQ schemes. We perform the analysis on a randomly chosen active UE in terms of the latent access failure probability under different latency constraints for three different GF schemes, respectively, in the following.

---

3We consider intra-cell interference because the UEs in the same cell associated with the same BS may choose the same pilot. We consider the inter-cell interference due to that the UEs in different cells share the preamble sequence pool among BSs. Similar with [30], we focus on providing a general analytical framework of cellular network, considering both the inter- and intra-interference.
A. Reactive scheme

In the Reactive scheme as illustrated in Fig. 2, the latent access failure probability remains unchanged at the beginning of each HARQ round trip and only changes at the end of each HARQ round trip (after processing the feedback at the UE, i.e., in the 4th TTI of this round trip), as the UE needs time to transmit packet and receive feedback. For example, in one HARQ round trip (e.g., the \( m \)th round trip), for \( \mathcal{T} = (m - 1)T_{\text{Reac}}^{\text{RTT}} + 2, (m - 1)T_{\text{Reac}}^{\text{RTT}} + 3, (m - 1)T_{\text{Reac}}^{\text{RTT}} + 4 \) TTIs, the latent access failure probabilities are the same as \( \mathcal{P}_F[T_{\text{latency}} \leq \mathcal{T} - 1] \), since the UE cannot receive feedback on time; for \( \mathcal{T} = (m - 1)T_{\text{Reac}}^{\text{RTT}} + 5 = mT_{\text{Reac}}^{\text{RTT}} + 1 \) TTIs, the latent access failure probability \( \mathcal{P}_F[T_{\text{latency}} \leq \mathcal{T}] \) changes determined by the UE’s retransmission or not after receiving NACK or ACK, respectively.

In order to calculate the latent access failure probabilities under various latency constraints, we need to know the number of HARQ round trips allowed under the latency constraint \( \mathcal{T} \). For ease of presentation, we define

\[
M = \left\lfloor (\mathcal{T} - 1)/T_{\text{Reac}}^{\text{RTT}} \right\rfloor, \tag{10}
\]

to imply the maximum number of HARQ round trips allowed under the latency constraint \( \mathcal{T} \) with \( T_{\text{Reac}}^{\text{RTT}} = 4 \) TTIs given in (2).

Note that inactive UEs (with empty data buffer) do not transmit, such that they do not generate interference. A UE is still active in the \( m \)th (\( 1 \leq m \leq M \)) round trip if none of its GF access in the last \( (m - 1) \) round trips are successful. Mathematically, the active probability \( \mathcal{A}_m \) of the UE in the \( m \)th round trip (\( 1 \leq m \leq M \)), is obtained as

\[
\mathcal{A}_m = 1 - \mathcal{P}_F[T_{\text{latency}} \leq T_{\text{Reac}}[m - 1]], \tag{11}
\]

with \( T_{\text{Reac}}[m - 1] \) obtained from (3).

Based on (11), the latent access failure probability of a randomly chosen UE with the Reactive scheme is derived in the following Theorem 1.

---

\(^4\)Note that we focus on the latent access failure probability of a randomly chosen UE in the UL transmission, where the UE finishes one transmission and knows if it is successful or not until it receives the feedback from the BS (e.g., in the 4th TTI in a round trip in the Reactive scheme). This is different from the BS perspective, where one transmission is finished until the BS processes the signal from the UE and knows if it is successful (e.g., in the 2th TTI in a round trip in the Reactive scheme).
Theorem 1. The latent access failure probability of a randomly chosen UE with the Reactive HARQ scheme under the latency constraint \( T \) is derived as

\[
P^\text{Reac}_T[T_{\text{latency}} \leq T] = \begin{cases} 
1, & M = 0, \\
1 - \sum_{m=1}^{M} A^\text{Reac}_m P^\text{Reac}_m, & M \geq 1,
\end{cases}
\]

where \( M \) is given in (10), \( A^\text{Reac}_m \) is given according to (11) as

\[
A^\text{Reac}_m = \begin{cases} 
1, & m = 1, \\
1 - \sum_{i=1}^{m-1} A^\text{Reac}_i P^\text{Reac}_i, & m \geq 2,
\end{cases}
\]

and \( P^\text{Reac}_m \) is the GF access success probability of the typical UE in the \( m \)th round trip with the Reactive scheme that derived in (14) of the following Lemma 1.

Proof. See Appendix A.

Lemma 1. The GF access success probability of the typical UE in the \( m \)th round trip with the Reactive scheme is given by

\[
P^\text{Reac}_m = \sum_{n=0}^{\infty} \left\{ O[n,m] \Theta^\text{Reac}[n,m] \left( 1 - \Theta^\text{Reac}[n,m] \right)^n \right\},
\]

where

\[
O[n,m] = \frac{c^{(c+1)} \Gamma(n + c + 1)(A^\text{Reac}_m \lambda_a/\lambda_B)^n}{\Gamma(c + 1) \Gamma(n + 1)(A^\text{Reac}_m \lambda_a/\lambda_B + c)^{n+c+1}},
\]

and

\[
\Theta^\text{Reac}[n,m] = \frac{\exp \left(-\gamma_\text{th}\sigma^2/(g_m \rho) - (\gamma_\text{th})^{\frac{1}{2}} A^\text{Reac}_m \lambda_a/\lambda_B \arctan((\gamma_\text{th})^{\frac{1}{2}})\right)}{(1 + \gamma_\text{th})^n}, (g_m \leq g_J).
\]

Part I is the probability of the number of intra-cell interfering UEs for a typical BS \( N = n \) derived following [32, Eq.(3)], where \( c = 3.575 \) is a constant related to the approximate PMF of the PPP Voronoi cell and \( \Gamma(\cdot) \) is the gamma function. Part II is the transmission success probability of the UE conditioning on \( N = n \). Part III is the transmission failure probability that the transmissions from other \( n \) intra-cell interfering UEs are not successfully received by the BS, i.e., the non-collision probability.

Proof. See Appendix B.
Remark 1. In (16), it can be shown that the transmission success probability (II in (14)) of the typical UE is inversely proportional to the received SINR threshold $\gamma_{th}$ and the density ratio $\lambda_a/\lambda_B$. The transmission failure probabilities of other interfering UEs (III in (14)) (i.e., the non-collision probability of the typical UE) are directly proportional to the received SINR threshold $\gamma_{th}$ and the density ratio $\lambda_a/\lambda_B$. Therefore, a tradeoff between transmission success probability and non-collision probability is observed.

B. K-repetition scheme

In the K-repetition scheme as illustrated in Fig. 3, the latent access failure probability also changes at the end of each HARQ round trip similar to the Reactive scheme, but with longer round trip time $T_{Krept} = (K_{Krept} + 3)$ TTIs given in (4). More specifically, in one HARQ round trip (e.g., the $m$th round trip) of the K-repetition scheme, for $T = (m - 1)T_{Krept}^\text{RTT} + 2, (m - 1)T_{Krept}^\text{RTT} + 3, ..., (m - 1)T_{Krept}^\text{RTT} + K_{Krept} + 3$ TTIs, the latent access failure probabilities are the same as $P_{F}[T_{\text{latency}} \leq T - 1]$; for $T = (m - 1)T_{Krept}^\text{RTT} + K_{Krept} + 4 = mT_{Krept}^\text{RTT} + 1$ TTIs, the latent access failure probabilities $P_{F}[T_{\text{latency}} \leq T]$ changes determined by the UEs retransmission or not after receiving NACK or ACK, respectively. Let us define

$$M = \lfloor (T - 1)/T_{Krept}^\text{RTT} \rfloor,$$

(17)

to imply the maximum number of HARQ round trips allowed under the latency constraint $T$ with $T_{Krept}^\text{RTT} = K_{Krept} + 3$ TTIs, we can derive the latent access failure probability of a randomly chosen UE with the K-repetition scheme in the following Theorem 2.

**Theorem 2.** The latent access failure probability of a randomly chosen UE with the K-repetition scheme under latency constraint $T$ is derived as

$$P_{F}^{\text{Krept}}[T_{\text{latency}} \leq T] = \begin{cases} 1, & M = 0 \\ 1 - \sum_{m=1}^{M} A_{m}^{\text{Krept}} P_{m}^{\text{Krept}}, & M \geq 1, \end{cases}$$

(18)

where $M$ is given in (17), $A_{m}^{\text{Krept}}$ is obtained according to (11) as

$$A_{m}^{\text{Krept}} = \begin{cases} 1, & m = 1, \\ 1 - \sum_{i=1}^{m-1} A_{i}^{\text{Krept}} P_{i}^{\text{Krept}}, & m \geq 2, \end{cases}$$

(19)

and $P_{m}^{\text{Krept}}$ is the GF access success probability of the typical UE in the $m$th round trip with the K-repetition scheme that derived in (20) of the following Lemma 2..
Lemma 2. The GF access success probability of the typical UE in the $m$th HARQ round trip with the K-repetition scheme is derived as

$$P_{m}^{K_{\text{rep}}} = \sum_{n=0}^{\infty} \left\{ O[n, m] \Theta_{K_{\text{rep}}}[n, m, K_{\text{rep}}] \left( 1 - \Theta_{K_{\text{rep}}}[n, m, K_{\text{rep}}] \right)^n \right\}, \quad (20)$$

where

$$O[n, m] = \frac{e^{(c+1)} \Gamma(n + c + 1) \left( A_{m}^{K_{\text{rep}}} \lambda_a / \lambda_B \right)^n}{\Gamma(c + 1) \Gamma(n + 1) \left( A_{m}^{K_{\text{rep}}} \lambda_a / \lambda_B + c \right)^{n+c+1}}, \quad (21)$$

and

$$\Theta_{K_{\text{rep}}}[n, m, K_{\text{rep}}] = \sum_{k=1}^{K_{\text{rep}}} \left( -1 \right)^{k+1} \left( \begin{array}{c} K_{\text{rep}} \\ k \end{array} \right) \exp \left( -k \gamma_{th} \sigma^2 / (g_m \rho) - A_{m}^{K_{\text{rep}}} \lambda_a / \lambda_B \left( 2 \Gamma_{1} \left( -2, k; \frac{\alpha - 2}{\alpha}; -\gamma_{th} \right) - 1 \right) \right) \left( 1 + \gamma_{th} \right)^{kn}. \quad (22)$$

Part I is the probability of the number of intra-cell interfering UEs $N = n$. Part II is the transmission success probability of the UE conditioning on $N = n$. Part III is the non-collision probability.

Proof. See Appendix C.

Remark 2. It is evident from (22) that the transmission success probability (II in (20)) of the typical UE increases, whereas the non-collision probability (III in (20)) decreases with increasing the repetition value $K_{\text{rep}}$. Therefore, there exists a trade off between transmission success probability and non-collision probability. For illustration, the relationship among GF access success probability, the transmission success probability, and the non-collision probability are shown in Fig. 5. We can see that in certain scenario in (b) (i.e., $\gamma_{th} = -10$ dB and $\lambda_D / \lambda_B > 4 \times 10^4$), the increase of repetition value $K_{\text{rep}}$ could not further improve, and even degrades the GF access success probability. This is due to the fact that increasing the repetition increases the collisions in overloaded traffic scenario, and wastes extra time and frequency resource. Further details will be described later in Section V.

Finally, the latent access failure probabilities under arbitrary latency constraints of a randomly chosen UE with the K-repetition scheme and the Reactive scheme can be derived based on the iteration process (Note that the Reactive scheme is a special case of K-repetition scheme when the repetition value $K_{\text{rep}} = 1$). We assume $m$ is a variable that denotes the HARQ round trip
Fig. 5. Comparing GF access success probability ($P_{\text{Krep}}^1$), transmission success probability ($P_{\text{Krep with III=1}}^1$), and non-collision probability ($P_{\text{Krep with II=1}}^1$). The parameters are $\lambda_B = 1 \text{ BS/km}^2$, $\lambda_D = 20000 \text{ UEs/km}^2$, $p_a = 0.0011$, $\rho = 130 \text{ dBm}$, $\gamma_{th} = -10 \text{ dB}$ and $\sigma^2 = 126.2 \text{ dBm}$.

Fig. 6. Flowchart for deriving the latent access failure probability of the K-repetition scheme and the Reactive scheme.
from 1 to $M$. The iteration process for calculating the latent access failure probability is shown in Fig. 6. Details of this process are described by the following:

- Step 1: Calculate the maximum number of HARQ round trips $M$ under the given latency constraint $T$ using (17).
- Step 2: If $M = 0$, $P_{\text{Reac}}[T_{\text{latency}} \leq T] = 1$, otherwise go to Step 3;
- Step 3: Calculate the active probability $A_m$ in the $m$th HARQ round trip using (19);
- Step 4: Calculate the GF access success probability in the $m$th round trip using (20);
- Step 5: Calculate the latent access failure probability in the $m$th round trip using (18).

Repeating Step 3 to 5 until $m = M$, the latent access failure probability under latency constraint $T$ can be obtained.

C. Proactive scheme

The analytical model for the Proactive scheme is more complicated compared with the Reactive and the K-repetition schemes. In the former two schemes, the latent access failure probabilities only change at the end of each HARQ round trip, as the BS processes the received signal and sends the feedback to the UE once in each round trip. However, in the Proactive scheme, the latent access failure probabilities change at several TTIs in one round trip, as the BS processes each repetition and sends the feedback to the UE at several TTIs. Due to the complexity of the Proactive scheme, we first analyze the latent access failure probability of a randomly chosen UE with the latency constraint $T \leq K_{\text{Proa}} + 4$ TTIs without HARQ retransmissions.

1) Proactive scheme without HARQ retransmissions, $T \leq K_{\text{Proa}} + 4$: Compared with the K-repetition scheme, in which the UE is enforced to perform $K_{\text{Krep}}$ repetitions no matter if its transmission is successful or not within $K_{\text{Krep}}$ times, the UE in the Proactive scheme is allowed to terminate the repetition once the UE receives ACK. Take one example, as shown in Fig. 7, the UE-1 successfully transmits the packet in the 1st repetition, the UE-1 knows the success of its 1st repetition in the 4th repetition, and the UE-1 terminates its 5th repetition. That is to say, if a UE does not have a second packet to be transmitted, the Proactive scheme could help to reduce its interference to other UE(s) that share the same resource and happen to be active at the same time.

Due to the fact that the ACK/NACK feedback can only be received after 3TTIs, for the maximum repetition value $K_{\text{Proa}} \leq 4$, the UE can not receive feedback before completing $K_{\text{Proa}}$
Fig. 7. Early termination reduces UE interference

repetitions, thus the UE needs to complete all the \( K_{\text{Proa}} \) repetitions without terminating earlier and the number of interfering users will not change in each repetition of one round trip.

For \( K_{\text{Proa}} \geq 5 \), the UE can receive feedback from the BS to determine retransmission or not. For instance, the ACK feedback decreases the number of interfering users in the later repetitions as shown in Fig. 7. Let us denote that the 1st successful transmission occurs in the \( l \)th repetition, thus the feedback of this repetition will be received in the \( (l + 3) \)th repetition, which affects the latent access failure probability of the UE in the \( (l + 3) \)th repetition, and from the \( (l + 4) \)th repetition, these successful UEs will not repeat the rest \( (K_{\text{Proa}} - l - 3) \) repetitions for this packet any more.

We define the feedback factor for the \( l \)th \( (1 \leq l \leq K_{\text{Proa}}) \) repetition as \( \eta_{1,l} \), which means the GF access failure probability in the former \( (l - 4) \) repetitions\(^5\). It is obvious that \( \eta_{1,l} = 1 \) when \( 1 \leq l \leq 4 \). Then we derive the feedback factor as

\[
\eta_{1,l} = \begin{cases} 
1, & 1 \leq l \leq 4, \\
1 - P_{1,l-4}^{\text{Proa}}, & l \geq 5 ,
\end{cases}
\]  

(23)

where \( P_{1,l-4}^{\text{Proa}} \) is derived in the following Lemma 3.

**Lemma 3.** We define the transmission success probability in the \( l \)th repetition as \( P_{1,l} \), the transmission success probability in all \( l \) repetitions as \( \Theta_{\text{Proa}}^{\text{Proa}}[n, 1, l] \) (i.e., any one of the \( l \) repetitions succeeds \( (P_{1,l}) \)), and the access success probability in \( l \) repetitions as \( P_{1,l}^{\text{Proa}} \) (considering

\(^5\)Note that the ACK/NACK feedback can only be received after 3TTIs, thus the feedback from the former \( (l - 4) \) repetitions will affect the \( l \)th repetition. Only the failure UEs in the former \( (l - 4) \) repetitions will transmit in the \( l \)th repetition.
collision). Then, the GF access success probability of a randomly chosen UE with the Proactive scheme under the latency constraint \( T \leq K_{Proa} + 4 \) is driven as

\[
P_{1,l}^{Proa} = \sum_{n=0}^{\infty} \left\{ O[n, 1, l] \right. \Theta_{Proa}^{[n, 1, l]} \left( 1 - \Theta_{Proa}^{[n, 1, l]} \right)^n \left\}, \right.
\]

where

\[
O[n, 1, l] = \frac{c^{(c+1)} \Gamma(n+c+1) (\eta_{1,l} \lambda_a/\lambda_B)^n}{\Gamma(c+1) \Gamma(n+1) (\eta_{1,l} \lambda_a/\lambda_B + c)^{n+c+1}},
\]

and for \( l \leq 4 \),

\[
\Theta_{Proa}^{[n, 1, l]} = 1 - \prod_{r=1}^{l} (1 - P_{1,r})
\]

\[
= \sum_{r=1}^{l} (-1)^{r+1} \binom{l}{r} \exp(-r \gamma_{th} \sigma^2/(g_m \rho) - \eta_{1,r} \lambda_a/\lambda_B \left( 2 F_1 \left( -\frac{2}{\alpha}, k; \frac{\alpha-2}{\alpha}; -\gamma_{th} \right) - 1 \right) \right)}{(1 + \gamma_{th})^\alpha}^{kn},
\]

and for \( l \geq 5 \),

\[
\Theta_{Proa}^{[n, 1, l]} = 1 - (1 - \Theta_{Proa}^{[n, 1, 4]} \prod_{r=5}^{l} (1 - P_{1,r}) \right),
\]

with

\[
P_{1,r} = \eta_{1,r} O[n, 1, r] \Theta_{Proa}^{[n, 1, 1]}.
\]

where \( \Theta_{Proa}^{[n, 1, 1]} \) is obtained from (26), and \( O[n, 1, r] \) is obtained from (25).

**Proof.** See Appendix D.

In order to calculate the latent access failure probabilities under arbitrary latency constraints \( T \leq K_{Proa} + 4 \), we define two indexes for \( T \) as

\[
\left\{ \begin{array}{l}
\mu = \lfloor (T - 2)/T_{Proa,K,0}^{RTT} \rfloor, \\
\nu = \text{mod}(T - 2, T_{Proa,K,0}^{RTT}).
\end{array} \right.
\]

where \( T_{Proa,K,0}^{RTT} \) is given in (6). \( \mu \) implies the maximum number of the HARQ round trips under the latency constraint (for the Proactive scheme under the latency constraint \( T \leq K_{Proa} + 4 \).

6In the Proactive scheme with \( m \) HARQ round trips, a UE is still active in the \( m \)th \( (1 \leq m \leq M) \) HARQ round trip if none of its GF access in the former \( (m-1) \) HARQ round trips is successful. That is to say, all the maximum \( K_{Proa} \) repetitions in the Proactive scheme in the former \( (m-1) \) HARQ round trips are not successful, i.e., \( l = 0 \).
\( \mu = 0 \), \( \nu \) implies the updated TTI index for the latent access failure probability in each HARQ round trip.

Then, the latent access failure probability of a randomly chosen UE with the Proactive scheme under the latency constraint \( T \leq K_{Proa} + 4 \) is derived in the following **Theorem 3**.

**Theorem 3.** The latent access failure probability of a randomly chosen UE with the Proactive scheme under the latency constraint \( T \leq K_{Proa} + 4 \) is derived as

\[
P_{F}^{Proa}[T_{\text{latency}} \leq T] = \begin{cases} 
1, & \nu \leq 2, \text{ and } \mu = 0, \\
1 - P_{1,\nu-2}^{Proa}, & \nu \geq 3, \text{ and } \mu = 0.
\end{cases} \tag{30}
\]

where \( P_{1,\nu-2}^{Proa} \) is obtained from (24) of Lemma 3.

Next, we extend the analysis of the latent access failure probabilities of the typical UE with the Proactive scheme to an arbitrary latency constraint \( T \) allowing the maximum \( M \) number of HARQ round trips.

2) **Proactive scheme with HARQ retransmissions:** In the Proactive scheme with \( M \) HARQ round trips, a UE is still active in the \( m \)th (\( 1 \leq m \leq M \)) HARQ round trip if none of its GF access in the former (\( m-1 \)) HARQ round trips are successful. That is to say, all the maximum \( K_{Proa} \) repetitions in the Proactive scheme in the former (\( m-1 \)) HARQ round trips are not successful. Similar to the other two schemes, we give the active probability \( A_{m}^{Proa} \) in the \( m \)th HARQ round trip in (32). For an arbitrary latency constraint \( T \), we first obtain the two indexes \( \mu \) and \( \nu \) using (29), i.e., the maximum number of the HARQ round trips under the latency constraint is \( M = \mu \). Then, the latent access failure probability can be obtained in the following **Theorem 4**.

**Theorem 4.** The latent access failure probability of a randomly chosen UE with the Proactive HARQ scheme under arbitrary latency constraint \( T \) is derived as

\[
P_{F}^{Proa}[T_{\text{latency}} \leq T] = \begin{cases} 
1, & \nu \leq 2, \text{ and } \mu = 0, \\
1 - P_{1,\nu-2}^{Proa}, & \nu \geq 3, \text{ and } \mu = 0, \\
1 - \sum_{m=1}^{M} A_{m}^{Proa} P_{m,K}^{Proa}, & \nu \leq 2, \text{ and } \mu \geq 1, \\
1 - \sum_{m=1}^{M} A_{m}^{Proa} P_{m,K}^{Proa} + A_{M+1}^{Proa} P_{M+1,\nu-2}^{Proa}, & \nu \geq 3, \text{ and } \mu \geq 1.
\end{cases} \tag{31}
\]
where $A_{m}^{Proa}$ is obtained according to (11) as
\[
A_{m}^{Proa} = \begin{cases} 
1, & m = 1, \\
1 - \sum_{i=1}^{m-1} A_{i}^{Proa} P_{i}^{Proa}, & m \geq 2,
\end{cases}
\]
(32)

and $P_{m,l}^{Proa}$ is the GF access probability of a typical UE in the $m$th HARQ round trip, given in the following Lemma 4.

**Lemma 4.** The GF access success probability of a randomly chosen UE with the Proactive HARQ scheme in the $m$th HARQ round trip is driven as
\[
P_{m,l}^{Proa} = \sum_{n=0}^{\infty} \left\{ O[n, m, l] \Theta_{Proa}[n, m, l] \left( 1 - \Theta_{Proa}[n, m, l] \right)^n \right\},
\]
(33)

where
\[
O[n, m, l] = \frac{c^{(c+1)} \Gamma(n + c + 1) \left( \eta_{m,l} A_{m}^{Proa} \lambda_{a} / \lambda_{B} \right)^n}{\Gamma(c + 1) \Gamma(n + 1) \left( \eta_{m,l} A_{m}^{Proa} \lambda_{a} / \lambda_{B} + c \right)^{n+c+1}},
\]
(34)

with
\[
\eta_{m,l} = \begin{cases} 
1, & \text{if } 1 \leq l \leq 4, \\
1 - P_{m,l}^{Proa}, & \text{if } l \geq 5,
\end{cases}
\]
(35)

and for $l \leq 4$,
\[
\Theta_{Proa}[n, m, l] = 1 - \prod_{r=1}^{l} (1 - P_{m,r})
\]
(36)

\[
= \sum_{r=1}^{l} (-1)^{r+1} \binom{l}{r} \exp \left( -r \gamma_{th} \sigma^2 / (g_{m} \rho) - \eta_{m,r} A_{m}^{Proa} \lambda_{a} / \lambda_{B} \left( 2 F_{1} \left( -\frac{2}{\alpha}; \frac{\alpha - 2}{\alpha}; -\gamma_{th} \right) - 1 \right) \right) (1 + \gamma_{th})^{kn},
\]

and for $l \geq 5$,
\[
\Theta_{Proa}[n, m, l] = 1 - (1 - \Theta_{Proa}[n, m, 4]) \prod_{r=5}^{l} (1 - P_{m,r}),
\]
(37)

with
\[
P_{m,r} = \eta_{m,r} O[n, m, r] \Theta_{Proa}[n, m, 1],
\]
(38)

where $\Theta_{Proa}[n, m, 1]$ is obtained from (36) and $O[n, m, r]$ is obtained from (34).

Finally, the latent access failure probabilities for the Proactive scheme under an arbitrary latency constraint can be obtained using the iteration process shown in Fig. 8 with the details described in the following.
Fig. 8. Flowchart for deriving the latent access failure probability of the Proactive scheme.

- Step 1: Calculate the indexes $\mu$ and $\nu$ under the given latency constraint $T$ using (29). If $\mu \geq 1$, go to Step 2; If $\mu = 0$, $\nu \leq 2$, go to Step 6; If $\mu = 0$, $\nu \geq 3$, go to Step 5;
- Step 2: Calculate non-empty probability $A^{P_{\text{Proa}}}_m$ using (32);
- Step 3: Calculate the GF access success probability in the $m$th round trip, $P^{P_{\text{Proa}}}_{m,K}$ using (33); Repeating Step 2 to 3 until $m = M$;
- Step 4: Calculate non-empty probability $A^{P_{\text{Proa}}}_{M+1}$ using (32);
- Step 5: If $\nu \geq 3$, calculate the GF access success probability $P^{P_{\text{Proa}}}_{m,\nu-2}$ using (33);
- Step 6: Calculate the latent access failure probability $P^{P_{\text{Proa}}}_{\text{out}}[T_{\text{latency}} \leq T]$ using (31).
IV. SIMULATION AND DISCUSSION

In this section, we verify our analytical results by comparing the theoretical GF latent access failure probabilities with the results from Monte-Carlo simulations, where the simulations are performed using the system model described in Section II in MATLAB. The BSs and UEs are deployed via independent HPPPs in a 1600 km$^2$ circle area with each UE associated with its nearest BS. In each TTI, IoT devices randomly move to a new position and the active ones randomly choose a pilot to transmit. The channel fading gains between the UEs and BSs are modeled by exponentially distributed random variables. The simulation parameters used for this study are in line with the main guidelines for 3GPP NR performance evaluations presented in [28] with mini-slots of 7 OFDM symbols for transmissions in short TTI (0.125ms) using 60 kHz SCS. Grant-free transmissions use all available $S = 48$ pilots, to transmit the small packet. We assume that the receiver can ideally decode the signal if there is no collision and that the signal processing as well as the feedback are error-free. The simulation time is configured to collect at least $5 \times 10^6$ samples to ensure a sufficient confidence level on the $10^5$ quantile. In all figures of this section, Analytical and Simulation are abbreviated as Ana. and Sim., respectively. Unless otherwise stated, we consider $\lambda_B = 1$ BSs/km$^2$, $\lambda_D = 20000$ UEs/km$^2$, $\gamma_{th} = -2$ dB, $\alpha = 4$, $\rho = 130$ dBm, $p_a = 0.0011$, $g_J = g_1 = 1$, the noise $\sigma^2 = 174 + 10\log_{10}(60000) = 126.2$ dBm.

Fig. 9-Fig. 10 plot the GF latent access failure probabilities of the UE with the Reactive, K-repetition, and Proactive GF HARQ schemes versus SINR thresholds $\gamma_{th} = -10$ dB and $\gamma_{th} = -2$ dB, respectively. The analytical curves of the Reactive scheme and the K-repetition scheme are plotted following the flowchart in Fig. 6, and the analytical curves of the Proactive scheme are plotted following the flowcharts in Fig. 8. The close match between the analytical curves and simulation points validates the accuracy of the developed spatio-temporal mathematical framework. The stair behaviour (i.e., the latent access failure probability stay unchanged for a period of time) is caused by the waiting time between each retransmission.

In Fig. 9, we first observe that the latent access failure probabilities of the K-repetition scheme with different repetition values are lower than those of the Reactive scheme under longer latency constraints $\mathcal{T} \geq 1.5$ ms (12 TTIs). This is due to that increasing repetition value increases the GF access success probability, as it offers more opportunities to retransmit. However, we note that when the repetition value is too large (e.g., $K_{\text{Krep}} = 8$), the latent access failure probabilities are not lower than those of the 4-repetition scheme in most of the time (except $1.5$ ms $\leq \mathcal{T} \leq 1.8$ ms,
4.2ms \leq T \leq 4.5ms). This is due to that transmitting 8 repetitions will cost too much waiting time and introduce a much longer delay. In this case, it is obvious that if the repetition value is overestimated, the K-repetition scheme will waste the potential resource and lead to lower resource efficiency. Interestingly, the Proactive scheme could provide the lowest latent access failure probability under the shorter latency constraints $T \leq 1.5$ms (12 TTI), as the Proactive scheme could terminate earlier to reduce latency without waiting for 8 repetitions. But when the latency constraint $T$ gets longer, the UE has enough time to finish the $K_{\text{K-rep}}$ repetitions and get feedback, the advantage of the Proactive scheme than the K-repetition scheme is not obvious.

In Fig. 10, we observe that the 2-repetition scheme outperforms the Reactive scheme after 2 HARQ round trips, that the 4-repetition scheme outperforms the Reactive scheme after 4 HARQ round trips, and that the 8-repetition scheme has the highest latent access failure probabilities. Usually, increasing repetition value increases the GF access success probability, as it offers more opportunities to retransmit with time and frequency diversity. But there is a trade-off between transmission success probability and non-collision probability when increasing the repetition value in this scenario, which is in line with Fig. 5 (b). Thus, in Fig. 10, increasing the repetition value does not decrease the latent access failure probabilities because it introduces longer waiting time without increasing the access success probabilities.

Fig. 11-Fig. 12 plot the GF latent access failure probabilities under the latency constraint
\( T = 1 \text{ms} (8 \text{TTIs}) \) and \( T = 1.5 \text{ms} (12 \text{TTIs}) \) for different density ratios and SINR thresholds. We observe that the GF latent access failure probability increases with increasing density ratio which is due to the following two reasons: 1) increasing the number of UEs generating interference leads to lower received SINR at the BS; 2) increasing the number of UEs leads to higher probability of collision. We also observe that the GF latent access failure probabilities decreases with decreasing SINR threshold. This is due to the lower SINR threshold leading to higher access success probability.

In Fig. 11, we observe that the GF latent access failure probabilities decrease in light load scenario (e.g., \( \lambda_D/\lambda_B \leq 40000 \)), while increases in high load scenario (e.g., \( \lambda_D/\lambda_B \geq 40000 \)) with increasing the repetition value, which is in line with Fig. 5 (b). This is due to the fact that increasing the repetition increases the collisions in overloaded traffic scenario, and wastes extra time and frequency resource. We should also note that, as the latency constraint \( T = 1 \text{ms} \) (8 TTIs), so the 8-repetition scheme can not be adopted because its waiting time for the 1st transmission is more than 1ms. But the Proactive scheme with a maximum of 8 repetitions could have as good performance as the 4-repetition scheme.

In Fig. 12, we observe that the GF latent access failure probabilities decrease in higher SINR threshold scenario (e.g., \( \gamma_{th} \geq -5 \text{dB} \)), while increases in lower SINR threshold scenario (e.g.,
\( \gamma_{th} \leq -5 \text{dB} \) with increasing the repetition value. Thus, despite that the usage of the K-repetition scheme can cope with tight time constraints by allowing a number of consecutive repetitions in a short time, the interference due to the multiple repetitions is the major impacting factor and surpasses the benefits of the combining gain in lower SINR threshold and high density scenario.

![Graph](image)

Fig. 13. Latent access failure probability when \( g_1 = 1 \), and \( g_J = g_2 = 2 \).

Fig. 13 plots the GF latent access failure probabilities of the UE under the Reactive, K-repetition, and Proactive GF HARQ schemes with Power Boosting. We observe that the GF latent access failure probabilities of the GF HARQ schemes with Power Boosting outperform those of the schemes without Power Boosting. This is due to that the failed UEs are favored by stepping up the transmit power, which significantly increases the transmission success probability. Interestingly, we also observe that Power Boosting has a greater improvement on the K-repetition scheme than the other two schemes.

V. CONCLUSION

In this paper, we developed a spatio-temporal mathematical model to analyze and compare the grant-free access latent access failure probabilities of a randomly chosen UE with three different GF HARQ schemes for URLLC requirements. We defined the latent access failure probability to characterize the URLLC performance. We proposed a tractable approach to derive and analyze the contention-based grant-free latent access failure probabilities of the UE with the Reactive, the K-repetition, and the Proactive GF HARQ schemes, respectively. Our results shown that 1) the Reactive scheme and the Proactive scheme can provide the lowest latent access failure
probability under the shorter latency constraints $T \leq 0.625\text{ms};$ 2) the Proactive schemes can provide the lowest latent access failure probability under shorter latency constraints $T \leq 1.5\text{ms}$ in the higher SINR threshold scenario; 3) the K-repetition scheme can provide the lowest latent access failure probability under longer latency constraints $T \geq 2.5\text{ms}$ except for lower SINR threshold and higher UE density scenario, but the repetition value needs to be optimized; 4) the Power Boosting can improve the latent access failure probability, especially for the K-repetition scheme. The analytical model presented in this paper can also be applied for the reliability and latency performance evaluation of other types of GF HARQ schemes in the cellular-based networks.

APPENDIX A

A PROOF OF THEOREM 1

For a given latency constraint $T,$ we have $M = \lfloor (T - 1)/T_{\text{RTT}} \text{Reac} \rfloor.$ For $M = 1,$ the latent access failure probability is the probability that the UE successfully accesses in the 1st HARQ round trip, where we can derive

$$P_F[T_{\text{latency}} \leq T] = 1 - P_1.$$  \hspace{1cm} (A.1)

For $M = 2,$ the latent access failure probability is the probability that the UE successfully accesses in either 1st or 2nd HARQ round trips, where we can derive

$$P_F[T_{\text{latency}} \leq T] = 1 - P_1 - (1 - P_1)P_2.$$  \hspace{1cm} (A.2)

Substituting (11) into (A.2), we have

$$P_F[T_{\text{latency}} \leq T] = 1 - \sum_{m=1}^{M=2} A_m P_m.$$  \hspace{1cm} (A.3)

For $M = 3,$ the latent access failure probability means the probability that the UE successfully accesses in all the three HARQ round trips. So we can derive

$$P_F[T_{\text{latency}} \leq T] = 1 - P_1 - (1 - P_1)P_2 - (1 - P_1 - (1 - P_1)P_2)P_3$$ \hspace{1cm} (A.4)

$$= 1 - P_1 - A_2 P_2 - A_3 P_3 = 1 - \sum_{m=1}^{M=3} A_m P_m.$$  

For $M > 3,$ the latent access failure probability $P_F[T_{\text{latency}} \leq T]$ can be derived based on the iteration process following $M = 2$ and 3.
APPENDIX B
A PROOF OF LEMMA 1

We derive the GF transmission success probability conditioning on \( n \) number of intra-cell interfering UEs based on the SINR outage as

\[
\Theta_{\text{Reac}}[n, m] = \mathbb{P}[\text{SINR}_m \geq \gamma_{th}] = \mathbb{P}\left\{ \frac{g_m \rho h_0}{\mathcal{T}_{\text{inter}}^m + \mathcal{T}_{\text{intra}}^m + \sigma^2} \geq \gamma_{th} \mid N = n \right\} \quad (B.1)
\]

\[
= \exp\left(-\frac{\gamma_{th}^2}{g_m \rho} \right) \mathcal{L}_{\mathcal{T}_{\text{inter}}^m} \left( \frac{\gamma_{th}}{g_m \rho} \right) \mathcal{L}_{\mathcal{T}_{\text{intra}}^m} \left( \frac{\gamma_{th}}{g_m \rho} \mid N = n \right).
\]

The Laplace Transform of aggregate intra-cell interference received at the typical BS conditioning on \( N = n \) is derived as

\[
\mathcal{L}_{\mathcal{I}_{\text{intra}}^m}(s \mid N = n) = E\left[ \exp\left(-s \sum_{\beta=1}^{n} g_m \rho h_{\beta} \right) \right] = \left( \frac{1}{1 + sg_m \rho} \right)^n, \quad (B.2)
\]

where \( s = \gamma_{th} / (g_m \rho) \).

The Laplace Transform of aggregate inter-cell interference received at the BS is derived as

\[
\mathcal{L}_{\mathcal{I}_{\text{inter}}^m}(s) = E\left[ \exp\left(-s \sum_{i \in \mathcal{Z}_{\text{inter}}} g_m P_i h_i \|u_i\|^{-\alpha} \right) \right] \overset{(a)}{=} E\left[ \prod_{i \in \mathcal{Z}_{\text{inter}}} \frac{1}{1 + sg_m \rho y_i^{-\alpha}} \right] \quad (B.3)
\]

\[
\overset{(b)}{=} \exp\left(-2\pi A^\text{Reac}_m \lambda_a \int_{\frac{P}{g_m \rho}}^{\infty} E_P \left[ 1 - \frac{1}{1 + sg_m \rho y^{-\alpha}} \right] y dy \right)
\]

\[
\overset{(c)}{=} \exp\left(-2\pi A^\text{Reac}_m \lambda_a (g_m s)^{\frac{\alpha}{2}} E_P[P_{\frac{2}{\alpha}}] \int_{(sg_m \rho)^{\frac{1}{\alpha}}}^{\infty} \frac{x}{1 + x^{\alpha}} dx \right)
\]

\[
= \exp\left(-2A^\text{Reac}_m \lambda_a / \lambda_B (\gamma_{th})^{\frac{2}{\alpha}} \int_{(\gamma_{th})^{\frac{1}{\alpha}}}^{\infty} \frac{x}{1 + x^{\alpha}} dx \right),
\]

where (a) is obtained by taking the average with respect to \( h_{i} \), (b) follows from the probability generation functional (PGFL) of the PPP, (c) follows by changing the variables \( x = y / (sP)^{\frac{1}{\alpha}} \) and \( E_P[P_{\frac{2}{\alpha}}] = \rho_{\alpha}^2 / (\pi \lambda_B) \) is the moments of the transmit power. Substituting Eq. (B.2) and Eq. (B.3) into (B.1), we derive the transmission success probability in the \( m \)th round trip as

\[
\Theta_{\text{Reac}}[n, m] = \exp\left(-\gamma_{th}^2 / (g_m \rho) - A^\text{Reac}_m \lambda_a / \lambda_B \left(2F_1\left(-\frac{2}{\alpha}, 1; \frac{\alpha - 2}{\alpha} : -\gamma_{th} \right) - 1\right)\right) / (1 + \gamma_{th})^n. \quad (B.4)
\]

We consider a general fading with the path loss exponent \( \alpha = 4 \) to simplify our results as

\[
\Theta_{\text{Reac}}[n, m] = \exp\left(-\gamma_{th}^2 / (g_m \rho) - (\gamma_{th})^{\frac{2}{4}} A^\text{Reac}_m \lambda_a / \lambda_B \arctan((\gamma_{th})^{\frac{1}{2}})\right) / (1 + \gamma_{th})^n. \quad (B.5)
\]
APPENDIX C

A PROOF OF LEMMA 2

For the K-repetition scheme, the GF transmission in one HARQ round trip is successful if any of the repetition succeeds. We derive the GF transmission success probability under $K_{\text{Krep}}$ repetitions conditioning on $n$ number of intra-cell interfering UEs based on the SINR outage as

$$\Theta^{\text{Krep}}[n, m, K_{\text{Krep}}] = 1 - \prod_{k=1}^{K_{\text{Krep}}} \left(1 - \mathbb{P}[\text{SINR}_k^m \geq \gamma_{th}]\right). \quad (C.1)$$

Based on the Binomial theorem, (C.1) can be rewritten as

$$\Theta^{\text{Krep}}[n, m, K_{\text{Krep}}] = \sum_{k=1}^{K_{\text{Krep}}} (-1)^{k+1} \binom{K_{\text{Krep}}}{k} \mathbb{P}[\text{SINR}_1^m, \text{SINR}_2^m, \ldots, \text{SINR}_k^m | N = n], \quad (C.2)$$

where $\binom{K_{\text{Krep}}}{k} = \frac{K_{\text{Krep}}!}{k!(K_{\text{Krep}}-k)!}$ is the binomial coefficient and

$$\mathbb{P}[\text{SINR}_1^m \geq \gamma_{th}, \ldots, \text{SINR}_k^m \geq \gamma_{th} | N = n] = \exp\left(-\frac{k\gamma_{th}^2}{g_m \rho \sigma^2}\right) \mathcal{L}_{\text{inner}}^{m}(\frac{\gamma_{th}}{g_m \rho}) \mathcal{L}_{\text{inner}}^{m}(\frac{\gamma_{th}}{g_m \rho} | N = n). \quad (C.3)$$

The Laplace Transform of aggregate intra-cell interference conditioning on $N = n$ is derived as

$$\mathcal{L}_{\text{inner}}^{m}(s | N = n) = E\left[\exp\left(-s \sum_{\beta=1}^{n} g_m \rho \sum_{r=1}^{k} h_{\beta}^r\right)\right] = \left(\frac{1}{1 + sg_m \rho}\right)^{kn}, \quad (C.4)$$

where $s = \gamma_{th}/(g_m \rho)$.

The Laplace Transform of aggregate inter-cell interference is derived as

$$\mathcal{L}_{\text{inter}}^{m}(s) = E\left[\exp\left(-s \sum_{i \in Z_{\text{inter}}} g_m \rho_{i} \left(\sum_{r=1}^{k} h_{i}^r\right)\|u_i\|^{-\alpha}\right)\right] = E\left[\prod_{i \in Z_{\text{inter}}} \left(\frac{1}{1 + sg_m \rho y_i^{-\alpha}}\right)^{k}\right] \quad (C.5)$$

$$= \exp \left(-2\pi A_m^{\text{Krep}} \lambda_a \int_{\left(\frac{1}{\rho}\right)^{1/2}}^{\infty} E_P \left[1 - \left(\frac{1}{1 + sg_m \rho y^{-\alpha}}\right)^k\right] dy\right)$$

$$= \exp \left(-2A_m^{\text{Krep}} \lambda_a / \lambda B(\gamma_{th})^{-\frac{2}{\alpha}} \int_{(\gamma_{th})^{-\frac{1}{\alpha}}}^{\infty} \left[1 - \left(\frac{1}{1 + x^{-\alpha}}\right)^k\right] dx\right).$$

Substituting (C.4) and (C.5) into (C.3) and then substituting (C.3) into (C.2), we derive the transmission success probability in the $m$th round trip with the K-repetition scheme as

$$\Theta^{\text{Krep}}[n, m, K_{\text{Krep}}] \quad (C.6)$$

$$= \sum_{k=1}^{K_{\text{Krep}}} (-1)^{k+1} \binom{K_{\text{Krep}}}{k} \exp\left(-k\gamma_{th}^2/(g_m \rho) - A_m^{\text{Krep}} \lambda_a / \lambda B \left(\frac{2F_1(-2; k; \frac{\alpha - 2}{\alpha}; -\gamma_{th}) - 1}{(1 + \gamma_{th})^{kn}}\right)\right).$$
APPENDIX D

A PROOF OF LEMMA 3

For \( l \leq 4 \), the UE can not receive feedback, thus the number of interfering users remains unchanged in each repetition. So we have

\[
\Theta_{\text{Proa}}[n, 1, l] = 1 - \prod_{r=1}^{l} (1 - \mathbb{P}_{1,r}), \quad (D.1)
\]

where

\[
\mathbb{P}_{1,r} = \eta_{1,r} \Theta_{\text{Proa}}[n, 1, 1] = \eta_{1,r} \exp \left( -\gamma_{th}^2 \sigma^2 / \rho - \eta_{1,r} \lambda_a / \lambda_B \left( 2 F_1 \left( -\frac{2}{\alpha}; 1; \frac{\alpha - 2}{\alpha}; -\gamma_{th} \right) - 1 \right) \right) / (1 + \gamma_{th}^n) \quad \text{.} \quad (D.2)
\]

For \( l \geq 5 \), the UE can receive feedback from the 4th repetition, thus the number of interfering users changes from the 5th repetition. So we have

\[
\Theta_{\text{Proa}}[n, 1, l] = 1 - (1 - \Theta_{\text{Proa}}[n, 1, 4]) \left( \prod_{r=1}^{l-4} (1 - \mathbb{P}_{1,r}) \right), \quad (D.3)
\]

where

\[
\mathbb{P}_{1,r} = \eta_{1,r} O[n, 1, r] \Theta_{\text{Proa}}[n, 1, 1] \quad (D.4)
\]

with

\[
O[n, 1, r] = \frac{c^{(c+1)} \Gamma(n + c + 1) (\eta_{1,r} \lambda_a / \lambda_B)^n}{\Gamma(c + 1) \Gamma(n + 1) (\eta_{1,r} \lambda_a / \lambda_B + c)^{n+c+1}} \quad . \quad (D.5)
\]

REFERENCES

[1] Study on scenarios and requirements for next generation access technologies (release 15). 3GPP, TS 38.913 v.15.2.0, Jun. 2018.

[2] M Series. IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond. Recommendation ITU, pages 2083–0, Sep. 2015.

[3] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari. Wireless communication for factory automation: an opportunity for LTE and 5G systems. IEEE Commun. Mag., 54(6):36–43, Jun. 2016.

[4] C. Campolo, A. Molinaro, A. Iera, and F. Menichella. 5G network slicing for vehicle-to-everything services. IEEE Wireless Commun., 24(6):38–45, Dec. 2017.

[5] B. Chang, G. Zhao, Z. Chen, L. Li, and M. A. Imran. Packet-drop design in URLLC for real-time wireless control systems. IEEE Access, pages 1–1, Jul. 2019.

[6] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung. Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges. IEEE Commun. Mag., 55(8):138–145, Aug. 2017.

[7] J. H. Kim, K. J. Choi, K. Lee, and K. S. Kim. Grant-free multiple access for ultra-reliable low-latency communications in a large-scale antenna system. In Proc. Int. Conf. Inf. Commun. Technol. Converg.(ICTC), pages 466–470, Oct. 2016.
[8] 5G service requirements for next generation new services and markets (release 15). 3GPP, TS 22.261 v.15.5.0, Release 15, Jul. 2018.

[9] Evolved universal terrestrial radio access (E-UTRA); physical layer procedures. 3GPP, TS 36.213 V14.2.0, Mar. 2017.

[10] Semi-persistent scheduling for 5G new radio URLLC. R1-167309, 3GPP TSG-RAN WG1 #86, Aug. 2016.

[11] Dynamic scheduling based transmission for URLLC. R1-1705245, 3GPP TSG-RAN WG1 #88, Apr. 2017.

[12] UL grant-free transmission for URLLC. R1-1705654, 3GPP TSG-RAN WG1 #88, Apr. 2017.

[13] L. Vangelista and M. Centenaro. Performance evaluation of HARQ schemes for the internet of things. Computers, 7(4):48, Jul. 2018.

[14] S. Sesia, Issam I. Toufik, and M. Baker. LTE-the UMTS long term evolution: from theory to practice. John Wiley and Sons, 2011.

[15] UL grant-free transmission for URLLC. R1-1705246, Apr. 2017.

[16] Discussion on HARQ support for URLLC. R1-1612246, 3GPP TR-RAN1 #87, Nov. 2016.

[17] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan. Random access analysis for massive IoT networks under a new spatio-temporal model: A stochastic geometry approach. IEEE Trans. Commun., pages 1–1, Jul. 2018.

[18] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. Kovcs, and P. Mogensen. Power control optimization for uplink grant-free URLLC. In 2018 IEEE Wireless Commun. Netw. Conf. (WCNC), pages 1–6, Apr. 2018.

[19] N. H. Mahmood, R. Abreu, R. Bhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen. Uplink grant-free access solutions for URLLC services in 5G new radio. In 2019 16th Inter. Symposium Wireless Commun. Systems (ISWCS), pages 607–612, Aug. 2019.

[20] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek. Analyzing random access collisions in massive IoT networks. IEEE Trans. Wireless Commun., 17(10):6853–6870, Oct. 2018.

[21] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen. 5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance. IEEE Access, 6:72253–72262, Nov. 2018.

[22] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska. A flexible 5G frame structure design for frequency-division duplex cases. IEEE Commun. Mag., 54(3):53–59, Mar. 2016.

[23] RAN1 chairmans notes. 3GPP, TSG RAN WG1 NR Ad-Hoc #2, Jun. 2017.

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

[25] J. F. C. Kingman. Poisson processes. Wiley Online Library, Jan. 1993.

[26] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen. 5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance. IEEE Access, 6:72253–72262, Nov. 2018.

[27] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva. On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio. In 2019 IEEE Wireless Commun. Netw. Conf. Workshop (WCNCW), pages 1–6, Apr. 2019.

[28] Study on new radio access technology-physical layer aspects. 3GPP, TR 38.802 v14.0.0, Mar. 2017.

[29] Study on RAN improvements for machine-type communications. 3GPP TR 37,868 v.11.2.0, Sept. 2011.

[30] M. Gharbieh, H. ElSawy, A. Bader, and M. Alouini. Spatiotemporal stochastic modeling of IoT enabled cellular networks: Scalability and stability analysis. IEEE Trans. on Commun., 65(8):3585–3600, Aug. 2017.

[31] Y. Yang and T. P. Yum. Analysis of power ramping schemes for UTRA-FDD random access channel. IEEE Trans. Wireless Commun., 4(6):2688–2693, Nov. 2005.

[32] S. M. Yu and S. L. Kim. Downlink capacity and base station density in cellular networks. In 11th Int. Symp. Model. Optim. Mobile Ad Hoc Wireless Netw. (WiOpt), pages 119–124, May. 2013.