Analyzing Grant-Free Access for URLLC Service

Yan Liu, *Graduate Student Member, IEEE*, Yansha Deng, *Member, IEEE*, Maged Elkashlan, *Senior 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) (i.e., dynamic grant) random access, which can hardly support the URLLC requirements. Grant-free (GF) (i.e., configured grant) access is proposed as a feasible and promising technology to meet such requirements, especially for uplink transmissions, which effectively saves the time of requesting/waiting for a grant. While some basic GF access features have been proposed and standardized in NR Release-15, there is still much space to improve. Being proposed as 3GPP study items, three GF access schemes with Hybrid Automatic Repeat reQuest (HARQ) retransmissions including Reactive, K-repetition, and Proactive, are analyzed in this article. 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 under shorter latency constraints, the Proactive scheme provides the lowest latent access failure probability, whereas, under longer latency constraints, the K-repetition scheme achieves the lowest latent access failure probability, which depends on K. If K is overestimated, the Proactive scheme provides lower latent access failure probability than the K-repetition scheme.**

***Index Terms*— URLLC, 5G NR, grant free access, HARQ.**

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, the 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]. 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 [8]. 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, grant-free (GF) access has been proposed and extensively discussed in 3GPP RAN WG1 [9]–[11] 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 the dedicated resource, known as Semi-Persistent-Scheduling (SPS) [9], is more suitable for periodic traffic with a fixed pattern, whereas contention-based GF transmission over the shared resource is more suitable for sporadic 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

Manuscript received February 1, 2020; revised June 7, 2020; accepted July 17, 2020. Date of publication August 24, 2020; date of current version February 17, 2021. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), U.K., under Grant EP/R006466/1. *(Corresponding author: Yansha Deng.)*

Yan Liu, Maged Elkashlan, and Arumugam Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: yan.liu@qmul.ac.uk; maged.elkashlan@qmul.ac.uk; a.nallanathan@qmul.ac.uk).

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

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

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2020.3018822

[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.

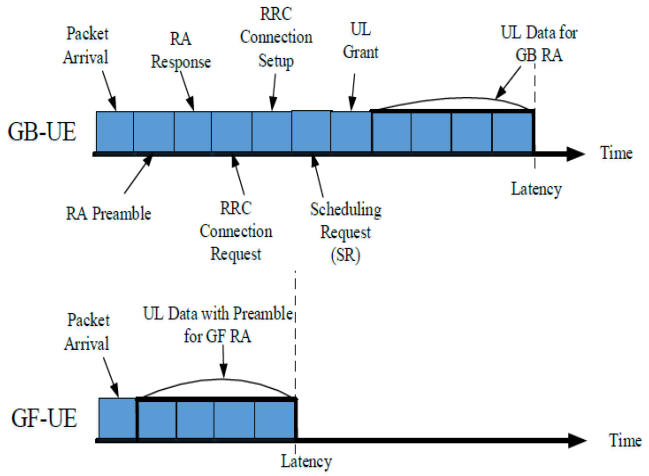

Fig. 1. Uplink transmissions for grant-based and grant-free random access.

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 [12]. 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) scheme, where retransmissions are triggered only when there is a failure in the previous transmission. However, the Reactive 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., the time duration of the cycle from the beginning of the transmission until processing its feedback [13]. Thus, the Reactive scheme only allows for a limited number of retransmissions due to the stringent latency requirement of URLLC service [13], and this fact motivates more research for advanced HARQ schemes to be integrated with GF transmission to provide reduced latency and enhanced reliability.

One scheme is the *K-repetition* (Krep) scheme supported in the 3GPP NR Release-15 [14], 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 discussed in [16], [17]. 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 to reduce latency. 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 [18]. However, both the ACB and BO schemes 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 PB, which can quickly compensate unexpected Signal-to-Interference plus-Noise Ratio (SINR) degradations at the initial transmissions [19]. 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 uplink transmission with the focus on improving the reliability.

Despite that the aforementioned GF access designs are proposed to govern the URLLC service, their theoretical formulations and comparative insights have never been fully established. Recent works [19], [20] have evaluated the Reactive, K-repetition, and Proactive schemes for URLLC service through system-level simulation without analytical characterization. The authors in [19] 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 requirements. 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 article, we address the following fundamental questions: 1) how to quantify the URLLC reliability and latency performances; 2) how to examine whether different GF schemes satisfy the URLLC reliability and latency performances or not; 3) how to evaluate which 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 HARQ schemes. The main contributions of this article can be summarized in the following:

- We present a novel spatio-temporal mathematical framework for analyzing contention-based GF 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 transmissions 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 Proactive scheme provides 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 PB under longer latency constraints.

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 under 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) Φ_{B} and Φ_{D} with intensities λ_{B} and λ_{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 σ^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., $h \sim \mathrm{Exp}(1)$. 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 α 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 ρ. We also assume the density of BSs is high enough and no UE suffers from truncation outage [18].

B. Contention-Based Grant-Free Access

In this article, 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. 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, i.e., the buffer becomes empty without new packets, 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[2] In this article, 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 [21]. The UEs are configured by radio resource control (RRC) signaling prior to the GF access (as Type 1 UL [22]), with time and frequency resource, modulation and coding scheme (MCS), power control settings, and HARQ related parameters [23]. The configured UEs are connected and synchronized, thus being always ready for a GF transmission. According to [24], we consider N UEs pre-configured with S orthogonal pilots, i.e., S sub-carriers over one TTI, for their uplink GF transmissions in the frequency domain [25]. At the beginning of each round trip, UEs randomly move to new positions, and the active ones randomly select one of the available S pilots to transmit with the data simultaneously [26]. A collision occurs when the same pilot is transmitted by two or more UEs using the same sub-carrier, and received successfully by the same BS. According to the thinning process [27], the density of active UEs choosing the same pilot can be derived as

$$\lambda_a = p_a \lambda_D / S. \tag{1}$$

C. Grant-Free Access Schemes

This section provides a general description of the three GF HARQ schemes considered in this article. 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 is represented by T_{fd}. After having received and processed 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 [28]. 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 $T_{\mathrm{fa}} = 1$ TTI and $T_{\mathrm{tx}} = 1$ TTI in this work same as [29]. 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

[2]5G 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. In contrast with the LTE slot duration of 14 OFDM symbols per TTI, mini-slots in 5G NR can compose of 1-13 symbols. Collectively, this allows shorter transmission slots to meet the stringent latency requirement.

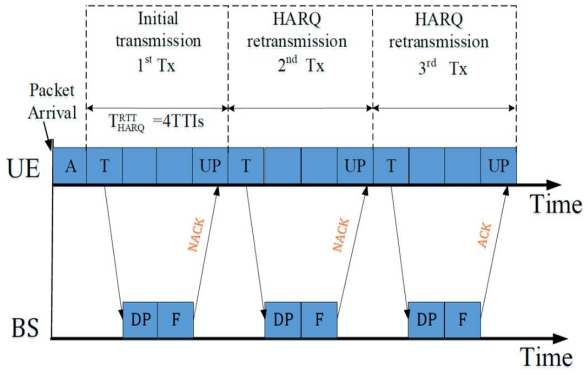

Fig. 2. Reactive GF transmission.

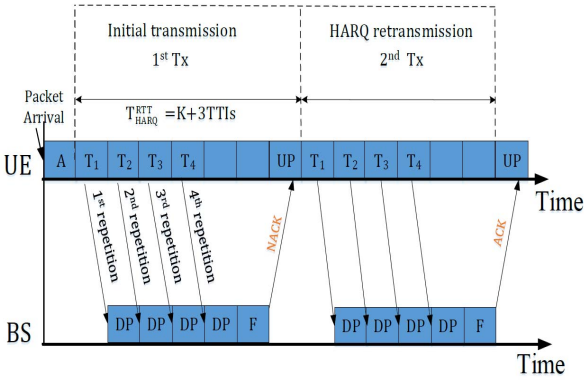

Fig. 3. K-repetition GF transmission with $K_{\mathrm{Krep}} = 4$ repetitions.

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

$$T^{\mathrm{RTT}}_{\mathrm{Reac}} = 4 \text{ TTIs}. \tag{2}$$

Then the latency after m HARQ round trips is obtained as

$$\begin{aligned} T_{\mathrm{Reac}}[m] &= T_{\mathrm{fa}} + mT^{\mathrm{RTT}}_{\mathrm{Reac}} \\ &= T_{\mathrm{fa}} + m(T_{\mathrm{tx}} + T_{\mathrm{dp}} + T_{\mathrm{fb}} + T_{\mathrm{up}}) \\ &= 1 + 4m \text{ TTIs}. \end{aligned} \tag{3}$$

*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_{Krep} repetitions in consecutive TTIs. At the end of K_{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

$$T^{\mathrm{RTT}}_{\mathrm{Krep}} = (K_{\mathrm{Krep}} + 3) \text{ TTIs}. \tag{4}$$

Then the latency after m HARQ round trips is defined as

$$\begin{aligned} T_{\mathrm{Krep}}[m] &= T_{\mathrm{fa}} + mT^{\mathrm{RTT}}_{\mathrm{Krep}} \\ &= T_{\mathrm{fa}} + m(K_{\mathrm{Krep}}T_{\mathrm{tx}} + T_{\mathrm{dp}} + T_{\mathrm{fb}} + T_{\mathrm{up}}) \\ &= 1 + m(K_{\mathrm{Krep}} + 3) \text{ TTIs}. \end{aligned} \tag{5}$$

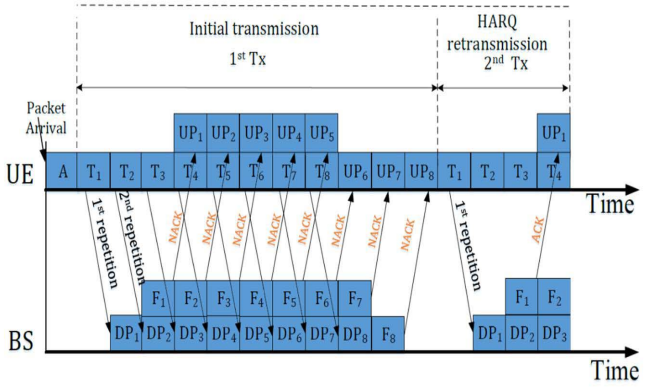

Fig. 4. Proactive GF transmission with maximum $K_{\mathrm{Proa}} = 8$ repetitions.

*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_{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 $K_{\mathrm{Proa}} \leqslant 4$, the UE continues repetitions until maximum K_{Proa}. For $K_{\mathrm{Proa}} \geqslant 5$, the UE continues repetitions until either the UE receives ACK from the BS, or the number of repetitions reaches maximum K_{Proa} times [30]. Let us denote the 1st access success of the typical UE occurs in the lth repetition during one HARQ round trip. Thus, we have the single HARQ round trip time for the Proactive scheme as:

$$T^{\mathrm{RTT}}_{\mathrm{Proa},K,l} = \begin{cases} K_{\mathrm{Proa}} + 3, & l = 0, \\ l + 3, & 1 \leqslant l \leqslant K_{\mathrm{Proa}}. \end{cases} \tag{6}$$

Note that if $l = 0$, i.e., all the K_{Proa} repetitions in one HARQ round trip are not successful with $T^{\mathrm{RTT}}_{\mathrm{Proa},K,0} = K_{\mathrm{Proa}} + 3$, 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_{Proa} repetitions can be derived as

$$\begin{aligned} T_{\mathrm{Proa}}[m] &= T_{\mathrm{fa}} + \underbrace{(m-1)T^{\mathrm{RTT}}_{\mathrm{Proa},K,0}}_{\mathrm{I}} + \underbrace{T^{\mathrm{RTT}}_{\mathrm{Proa},K,l}}_{\mathrm{II}} \\ &= T_{\mathrm{fa}} + (m-1)(K_{\mathrm{Proa}} + 3) + T^{\mathrm{RTT}}_{\mathrm{Proa},K,l} \\ &= l + 4 + (m-1)(K_{\mathrm{Proa}} + 3) \text{ TTIs} \\ &\quad \times (1 \leqslant l \leqslant K_{\mathrm{Proa}}), \end{aligned} \tag{7}$$

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 mth HARQ round trip given in (6).

D. Signal to Noise Plus Interference Ratio (SINR)

Note that the GF access failure 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 $\gamma_{\rm th}$; 2) the BS successfully receives 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 [26], [31], 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.[3] We formulate the SINR of a typical BS located at the origin as

$$\mathrm{SINR}_m = \frac{g_m \rho h_0}{\mathcal{I}_{\rm intra} + \mathcal{I}_{\rm inter} + \sigma^2}, \tag{8}$$

where ρ is the full path-loss inversion power control threshold, h_0 is the channel power gain from the typical UE to its associated BS, σ_2 is the noise power, $\mathcal{I}_{\rm intra}$ and $\mathcal{I}_{\rm inter}$ 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 mth retransmission by adjusting the target received power at the BS equal to $g_m\rho$ [26], [33] (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., $T_{\rm latency} \leqslant \mathcal{T}$, with a failure probability lower than a certain target, i.e., $\mathcal{P}_F \leqslant \varepsilon$. For its performance characterization, we define the *latent access failure probability* as $\mathcal{P}_F(T_{\rm latency} \leqslant \mathcal{T})$. To address this inherent limitation of URLLC requirements, it is meaningful to consider a probabilistic Quality of Service (QoS) in the following form:

Definition 1. *(URLLC QoS): We say the URLLC QoS of the UE is satisfied in a given frame if:*

$$\mathcal{P}_F(T_{\rm latency} \leqslant \mathcal{T}) \leqslant \varepsilon. \tag{9}$$

In 5G specification, $\varepsilon = 10^{-5}$ and $\mathcal{T} = 1$ 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.

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 (i.e., after processing the feedback at the UE 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 mth round trip), for $\mathcal{T} = (m-1)T_{\rm Reac}^{\rm RTT}+2, (m-1)T_{\rm Reac}^{\rm RTT}+3, (m-1)T_{\rm Reac}^{\rm RTT}+4$ TTIs, the latent access failure probabilities are the same as $\mathcal{P}_F[T_{\rm latency} \leqslant \mathcal{T}-1]$, since the UE can not receive feedback on time; for $\mathcal{T} = (m-1)T_{\rm Reac}^{\rm RTT} + 5 = mT_{\rm Reac}^{\rm RTT} + 1$ TTIs, the latent access failure probability $\mathcal{P}_F[T_{\rm latency} \leqslant \mathcal{T}]$ 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 maximum number M of HARQ round trips allowed under the latency constraint $\mathcal{T}$ TTIs. For ease of presentation, we define

$$M = \lfloor (\mathcal{T}-1)/T_{\rm Reac}^{\rm RTT} \rfloor, \tag{10}$$

with $T_{\rm Reac}^{\rm RTT} = 4$ 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 mth ($1 \leqslant m \leqslant M$) 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 mth round trip, is obtained as

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

with $T_{\rm Reac}[m-1]$ 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**.

Theorem 1. *The latent access failure probability of a randomly chosen UE with the Reactive HARQ scheme under the latency constraint $\mathcal{T}$* TTIs *is derived as*

$$\mathcal{P}_F^{\rm Reac}[T_{\rm latency} \leqslant \mathcal{T}] = \begin{cases} 1, & M = 0, \\ 1 - \sum_{m=1}^{M} \mathcal{A}_m^{\rm Reac} \mathcal{P}_m^{\rm Reac}, & M \geqslant 1, \end{cases} \tag{12}$$

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

$$\mathcal{A}_m^{\rm Reac} = \begin{cases} 1, & m = 1, \\ 1 - \sum_{i=1}^{m-1} \mathcal{A}_i^{\rm Reac} \mathcal{P}_i^{\rm Reac}, & m \geqslant 2, \end{cases} \tag{13}$$

*and $\mathcal{P}_m^{\rm Reac}$ is the GF access success probability of the typical UE in the mth 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 mth round trip with the Reactive scheme is given by*

$$\mathcal{P}_m^{\rm Reac} = \sum_{n=0}^{\infty} \Big\{ \underbrace{\mathrm{O}[n,m]}_{\rm I} \underbrace{\Theta^{\rm Reac}[n,m]}_{\rm II} \underbrace{\Big(1 - \Theta^{\rm Reac}[n,m]\Big)^n}_{\rm III} \Big\}, \tag{14}$$

[3] We 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 [32], we focus on providing a general analytical framework of cellular network, considering both the inter- and intra-interference.

where

$$\mathrm{O}[n,m]=\frac{c^{(c+1)}\Gamma(n+c+1)\big(\mathcal{A}_m^{\mathrm{Reac}}\lambda_a/\lambda_B\big)^n}{\Gamma(c+1)\Gamma(n+1)\big(\mathcal{A}_m^{\mathrm{Reac}}\lambda_a/\lambda_B+c\big)^{n+c+1}}, \tag{15}$$

and

$$\begin{aligned}\Theta^{\mathrm{Reac}}[n,m]&\\ &=\exp\big(-\frac{\gamma_{\mathrm{th}}\sigma^2}{g_m\rho}\big)(1+\gamma_{\mathrm{th}})^{-n}\\ &\quad\times\exp\Big(-(\gamma_{\mathrm{th}})^{\frac{1}{2}}\mathcal{A}_m^{\mathrm{Reac}}\lambda_a/\lambda_B\arctan((\gamma_{\mathrm{th}})^{\frac{1}{2}})\Big),\\ &\quad(g_m\leqslant g_J).\end{aligned} \tag{16}$$

Part I is the probability of the number of intra-cell interfering UEs for a typical BS $N=n^4$ derived following [34, 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 of the UE.

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 γ_{th} and the density ratio λ_a/λ_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 γ_{th} and the density ratio λ_a/λ_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 $T_{\mathrm{Krep}}^{\mathrm{RTT}}$ TTIs given in (4). More specifically, in one HARQ round trip (e.g., the mth round trip) of the K-repetition scheme, for $\mathcal{T}=(m-1)T_{\mathrm{Krep}}^{\mathrm{RTT}}+2,(m-1)T_{\mathrm{Krep}}^{\mathrm{RTT}}+3,\ldots,(m-1)T_{\mathrm{Krep}}^{\mathrm{RTT}}+K_{\mathrm{Krep}}+3$ TTIs, the latent access failure probabilities are the same as $\mathcal{P}_F[T_{\mathrm{latency}}\leqslant\mathcal{T}-1]$; for $\mathcal{T}=(m-1)T_{\mathrm{Krep}}^{\mathrm{RTT}}+K_{\mathrm{Krep}}+4=mT_{\mathrm{Krep}}^{\mathrm{RTT}}+1$ TTIs, the latent access failure probabilities $\mathcal{P}_F[T_{\mathrm{latency}}\leqslant\mathcal{T}]$ changes determined by the UE's retransmission or not after receiving NACK or ACK, respectively. Let us define

$$M=\lfloor(\mathcal{T}-1)/T_{\mathrm{Krep}}^{\mathrm{RTT}}\rfloor, \tag{17}$$

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

[4]Note that $N=n$ means there are n number of intra-cell interfering UEs (without the typicals UE), i.e., $n+1$ number of active UEs in one cell.

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

$$\mathcal{P}_F^{\mathrm{Krep}}[T_{\mathrm{latency}}\leqslant\mathcal{T}]=\begin{cases}1, & M=0\\ 1-\sum_{m=1}^{M}\mathcal{A}_m^{\mathrm{Krep}}\mathcal{P}_m^{\mathrm{Krep}}, & M\geqslant 1\end{cases}, \tag{18}$$

where M is given in (17), $\mathcal{A}_m^{\mathrm{Krep}}$ is obtained according to (11) as

$$\mathcal{A}_m^{\mathrm{Krep}}=\begin{cases}1, & m=1,\\ 1-\sum_{i=1}^{m-1}\mathcal{A}_i^{\mathrm{Krep}}\mathcal{P}_i^{\mathrm{Krep}}, & m\geqslant 2,\end{cases} \tag{19}$$

*and $\mathcal{P}_m^{\mathrm{Krep}}$ is the GF access success probability of the typical UE in the mth 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 mth HARQ round trip with the K-repetition scheme is derived as*

$$\begin{aligned}\mathcal{P}_m^{\mathrm{Krep}}=\sum_{n=0}^{\infty}\Big\{&\underbrace{\mathrm{O}[n,m]}_{\mathrm{I}}\underbrace{\Theta^{\mathrm{Krep}}[n,m,K_{\mathrm{Krep}}]}_{\mathrm{II}}\\ &\times\underbrace{\Big(1-\Theta^{\mathrm{Krep}}[n,m,K_{\mathrm{Krep}}]\Big)^n}_{\mathrm{III}}\Big\},\end{aligned} \tag{20}$$

where

$$\mathrm{O}[n,m]=\frac{c^{(c+1)}\Gamma(n+c+1)\big(\mathcal{A}_m^{\mathrm{Krep}}\lambda_a/\lambda_B\big)^n}{\Gamma(c+1)\Gamma(n+1)\big(\mathcal{A}_m^{\mathrm{Krep}}\lambda_a/\lambda_B+c\big)^{n+c+1}}, \tag{21}$$

and

$$\begin{aligned}&\Theta^{\mathrm{Krep}}[n,m,K_{\mathrm{Krep}}]\\ &=\sum_{k=1}^{K_{\mathrm{Krep}}}(-1)^{k+1}\binom{K_{\mathrm{Krep}}}{k}\exp\big(-\frac{k\gamma_{\mathrm{th}}\sigma^2}{g_m\rho}\big)(1+\gamma_{\mathrm{th}})^{-kn}\\ &\quad\times\exp\Big(-\mathcal{A}_m^{\mathrm{Krep}}\lambda_a/\lambda_B\Big({}_2F_1\Big(-\frac{2}{\alpha},k;\frac{\alpha-2}{\alpha};-\gamma_{\mathrm{th}}\Big)-1\Big)\Big).\end{aligned} \tag{22}$$

Similar to **Lemma 1.**, *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 of the UE.*

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_{Krep}. Therefore, there exists a tradeoff between transmission success probability and non-collision probability. For illustration, the relationship among GF access success probability ($\mathcal{P}_1^{\mathrm{Krep}}$), the transmission success probability ($\mathcal{P}_1^{\mathrm{Krep}}$ with III = 1), and the non-collision probability ($\mathcal{P}_1^{\mathrm{Krep}}$ with II = 1) versus repetition values are shown in Fig. 5. We can see that in certain scenario*

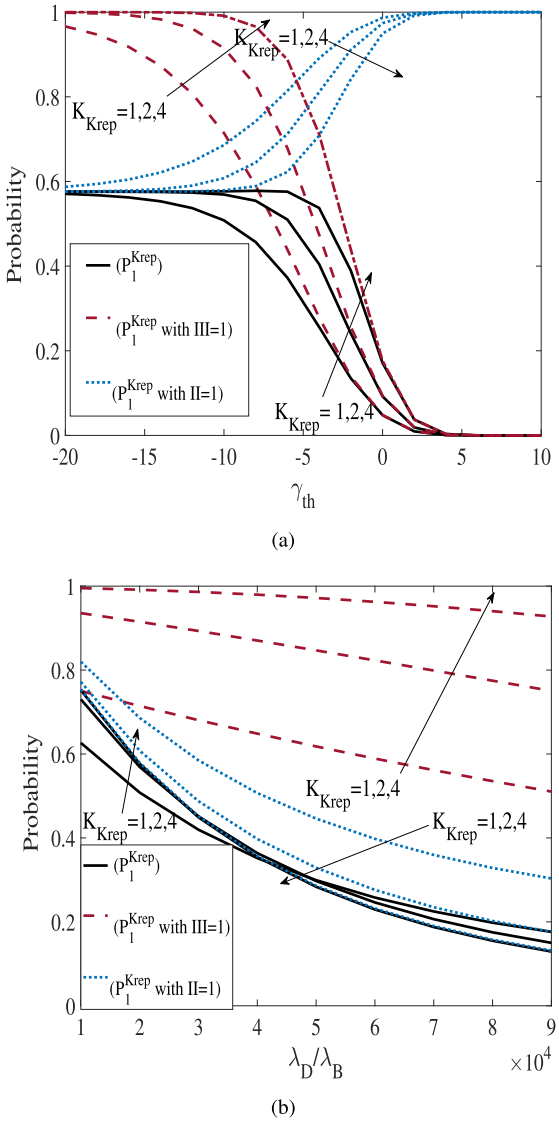

Fig. 5. Comparing GF access success probability ($\mathcal{P}_1^{\mathrm{Krep}}$), transmission success probability ($\mathcal{P}_1^{\mathrm{Krep}}$ with III $=$ 1), and non-collision probability ($\mathcal{P}_1^{\mathrm{Krep}}$ with II $=$ 1). The parameters are $\lambda_B = 1$ BS/km^2, $\lambda_D =$ 20000 UEs/km^2, $p_a = 0.0011$, $\rho =$ 130 dBm, $\gamma_{\mathrm{th}} = -10$ dB and $\sigma^2 =$ 126.2 dBm.

in (b) (i.e., $\gamma_{\mathrm{th}} = -10$ dB and $\lambda_D/\lambda_B > 4\times10^4$), the increase of repetition value K_{Krep} 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 and Reactive schemes 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 $K_{\mathrm{Krep}} = 1$. We assume m is a variable that denotes the HARQ round trip 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:

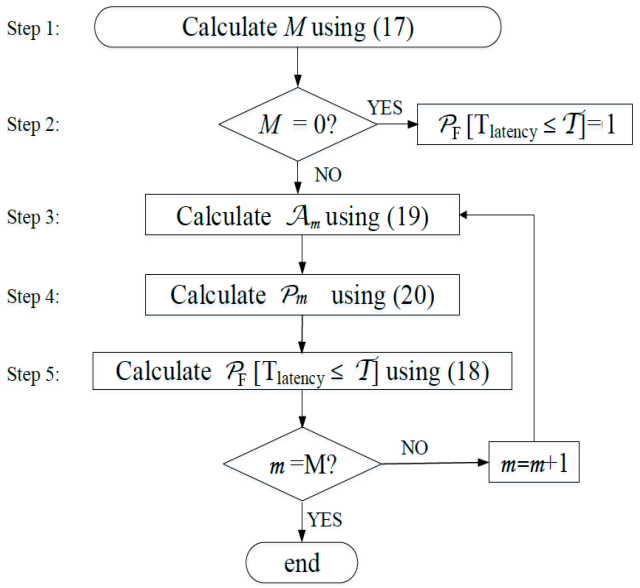

Fig. 6. Flowchart for deriving the latent access failure probability of the K-repetition scheme and the Reactive scheme.

- Step 1: Calculate the maximum number of HARQ round trips M under the given latency constraint $\mathcal{T}$ TTIs using (17).
- Step 2: If $M = 0$, $\mathcal{P}_F^{\mathrm{Reac}}[T_{\mathrm{latency}} \leqslant \mathcal{T}] = 1$, otherwise go to Step 3;
- Step 3: Calculate the active probability $\mathcal{A}_m$ in the mth HARQ round trip using (19);
- Step 4: Calculate the GF access success probability in the mth round trip using (20);
- Step 5: Calculate the latent access failure probability in the mth round trip using (18).

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

C. Proactive Scheme

The analytical model for the Proactive scheme is more complicated compared with the Reactive and 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 $\mathcal{T} \leqslant K_{\mathrm{Proa}} + 4$ TTIs without HARQ retransmissions.

1) Proactive Scheme Without HARQ Retransmissions, $\mathcal{T} \leqslant K_{Proa} + 4$ TTIs: Compared with the K-repetition scheme, in which the UE is enforced to perform K_{Krep} repetitions no matter if its transmission is successful or not within K_{Krep} times, the UE in the Proactive scheme is allowed to terminate the repetition once the UE receives ACK. Take one example,

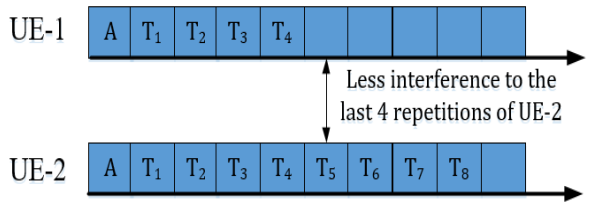

Fig. 7. Early termination reduces UE interference.

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 $K_{\mathrm{Proa}} \leqslant 4$, the UE can not receive feedback before completing K_{Proa} repetitions, thus the UE needs to complete all the K_{Proa} repetitions without terminating earlier and the number of interfering users will not change in each repetition of one round trip.

For $K_{\mathrm{Proa}} \geqslant 5$, 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 lth 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 $(K_{\mathrm{Proa}}-l-3)$ repetitions for this packet any more.

We define the *feedback factor* for the lth $(1 \leq l \leq K_{\mathrm{Proa}})$ 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

$$\eta_{1,l} = \begin{cases} 1, & 1 \leqslant l \leqslant 4, \\ 1-\mathcal{P}^{\mathrm{Proa}}_{1,l-4}, & l \geqslant 5\ , \end{cases} \tag{23}$$

where $\mathcal{P}^{\mathrm{Proa}}_{1,l}$ is derived in the following **Lemma 3**.

Lemma 3. *We define the transmission success probability in the lth repetition as $\mathbb{P}_{1,l}$, the transmission success probability in all l repetitions as $\Theta^{\mathrm{Proa}}[n,1,l]$ (i.e., any one of the l repetitions succeeds ($\mathbb{P}_{1,l}$)), and the access success probability in l repetitions as $\mathcal{P}^{\mathrm{Proa}}_{1,l}$ (considering collision). Then, the GF access success probability of a randomly chosen UE with the Proactive scheme under the latency constraint $\mathcal{T} \leqslant K_{\mathrm{Proa}}+4$* TTIs *is driven as*

$$\mathcal{P}^{\mathrm{Proa}}_{1,l} = \sum_{n=0}^{\infty} \Big\{ \underbrace{\mathrm{O}[n,1,l]}_{\mathrm{I}} \underbrace{\Theta^{\mathrm{Proa}}[n,1,l]}_{\mathrm{II}} \underbrace{\left(1-\Theta^{\mathrm{Proa}}[n,1,l]\right)^{n}}_{\mathrm{III}} \Big\}, \tag{24}$$

where

$$\mathrm{O}[n,1,l] = \frac{c^{(c+1)}\Gamma(n+c+1)\left(\eta_{1,l}\lambda_a/\lambda_B\right)^n}{\Gamma(c+1)\Gamma(n+1)\left(\eta_{1,l}\lambda_a/\lambda_B+c\right)^{n+c+1}}, \tag{25}$$

and for $l \leq 4$,

$$\begin{aligned} &\Theta^{\mathrm{Proa}}[n,1,l] \\ &= 1-\prod_{r=1}^{l}(1-\mathbb{P}_{1,r}) \qquad (26) \\ &= \sum_{r=1}^{l}(-1)^{r+1}\binom{l}{r}\exp\left(-\frac{r\gamma_{\mathrm{th}}\sigma^2}{g_m\rho}\right)(1+\gamma_{\mathrm{th}})^{-kn} \\ &\quad \times \exp\left(-\eta_{1,r}\lambda_a/\lambda_B\left({}_2F_1\left(-\frac{2}{\alpha},k;\frac{\alpha-2}{\alpha};-\gamma_{\mathrm{th}}\right)-1\right)\right), \qquad (27) \end{aligned}$$

and for $l \geqslant 5$,

$$\Theta^{\mathrm{Proa}}[n,1,l] = 1-(1-\Theta^{\mathrm{Proa}}[n,1,4])\prod_{r=5}^{l}\left(1-\mathbb{P}_{1,r}\right), \tag{28}$$

with

$$\mathbb{P}_{1,r} = \eta_{1,r}\mathrm{O}[n,1,r]\Theta^{\mathrm{Proa}}[n,1,1], \tag{29}$$

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

Proof: See Appendix D. □

In order to calculate the latent access failure probabilities under arbitrary latency constraints $\mathcal{T} \leqslant K_{\mathrm{Proa}}+4$ TTIs, we define two indexes for $\mathcal{T}$ as

$$\begin{cases} \mu = \lfloor(\mathcal{T}-2)/T^{\mathrm{RTT}}_{\mathrm{Proa},K,0}\rfloor, \\ \nu = \mathrm{mod}(\mathcal{T}-2, \mathrm{T}^{\mathrm{RTT}}_{\mathrm{Proa,K,0}}), \end{cases} \tag{30}$$

where $T^{\mathrm{RTT}}_{\mathrm{Proa},K,0}$ is given in (6),[6] μ implies the maximum number of the HARQ round trips under the latency constraint (for the Proactive scheme under the latency constraint $\mathcal{T} \leqslant K_{\mathrm{Proa}}+4$, $\mu=0$), ν 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 $\mathcal{T} \leqslant K_{\mathrm{Proa}}+4$ is derived in **Theorem 3**.

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

$$\mathcal{P}^{\mathrm{Proa}}_{F}[T_{\mathrm{latency}} \leqslant \mathcal{T}] = \begin{cases} 1, & \nu \leqslant 2, \text{ and } \mu=0, \\ 1-\mathcal{P}^{\mathrm{Proa}}_{1,\nu-2}, & \nu \geqslant 3, \text{ and } \mu=0\ . \end{cases} \tag{31}$$

*where $\mathcal{P}^{\mathrm{Proa}}_{1,\nu-2}$ 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 $\mathcal{T}$ allowing the maximum M number of HARQ round trips.

[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 lth repetition. Only the failure UEs in the former $(l-4)$ repetitions will transmit in the lth repetition.

[6]In the Proactive scheme with m HARQ round trips, a UE is still active in the mth $(1 \leqslant m \leqslant M)$ HARQ round trip if none of its GF access in the former $(m1)$ HARQ round trips is successful. That is to say, all the maximum K_{Proa} repetitions in the Proactive scheme in the former $(m1)$ HARQ round trips are not successful, i.e., $l=0$.

2) Proactive Scheme With HARQ Retransmissions: In the Proactive scheme with HARQ retransmissions, a UE is still active in the mth ($1 \leqslant m \leqslant M$) HARQ round trip if none of its GF access in the former ($m1$) HARQ round trips are successful. That is to say, all the maximum K_{Proa} repetitions in the Proactive scheme in the former ($m1$) HARQ round trips are not successful. Similar to the other two schemes, we give the active probability $\mathcal{A}_m^{Proa}$ in the mth HARQ round trip in (33). For an arbitrary latency constraint $\mathcal{T}$ TTIs, we first obtain the two indexes μ and ν using (30), 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 $\mathcal{T}$* TTIs *is derived as*

$$\begin{aligned}&\mathcal{P}_F^{\mathrm{Proa}}[T_{\mathrm{latency}} \leqslant \mathcal{T}]\\ &= \begin{cases} 1, & \nu \leqslant 2 \ \& \ \mu = 0,\\ 1-\mathcal{P}_{1,\nu-2}^{\mathrm{Proa}}, & \nu \geqslant 3 \ \& \ \mu = 0,\\ 1-\sum\limits_{m=1}^{M}\mathcal{A}_m^{\mathrm{Proa}}\mathcal{P}_{m,K}^{\mathrm{Proa}}, & \nu \leqslant 2 \ \& \ \mu \geq 1,\\ 1-\sum\limits_{m=1}^{M}\mathcal{A}_m^{\mathrm{Proa}}\mathcal{P}_{m,K}^{\mathrm{Proa}}+\mathcal{A}_{M+1}^{\mathrm{Proa}}\mathcal{P}_{M+1,\nu-2}^{\mathrm{Proa}}, & \\ \qquad\qquad \nu \geqslant 3 \ \& \ \mu \geq 1, \end{cases}\end{aligned} \tag{32}$$

where $\mathcal{A}_m^{\mathrm{Proa}}$ is obtained according to (11) as

$$\mathcal{A}_m^{\mathrm{Proa}} = \begin{cases} 1, & m = 1,\\ 1-\sum\limits_{i=1}^{m-1}\mathcal{A}_i^{\mathrm{Proa}}\mathcal{P}_i^{\mathrm{Proa}}, & m \geqslant 2, \end{cases} \tag{33}$$

*and $\mathcal{P}_{m,l}^{\mathrm{Proa}}$ is the GF access probability of a typical UE in the mth 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 mth HARQ round trip is driven as*

$$\begin{aligned}&\mathcal{P}_{m,l}^{\mathrm{Proa}}\\ &= \sum_{n=0}^{\infty}\Big\{\underbrace{\mathrm{O}[n,m,l]}_{\mathrm{I}}\underbrace{\Theta^{\mathrm{Proa}}[n,m,l]}_{\mathrm{II}}\times\underbrace{\Big(1-\Theta^{\mathrm{Proa}}[n,m,l]\Big)^{n}}_{\mathrm{III}}\Big\},\end{aligned} \tag{34}$$

where

$$\mathrm{O}[n,m,l] = \frac{c^{(c+1)}\Gamma(n+c+1)\big(\eta_{m,l}\mathcal{A}_m^{Proa}\lambda_a/\lambda_B\big)^{n}}{\Gamma(c+1)\Gamma(n+1)\big(\eta_{m,l}\mathcal{A}_m^{Proa}\lambda_a/\lambda_B+c\big)^{n+c+1}}, \tag{35}$$

with

$$\eta_{m,l} = \begin{cases} 1, & \text{if } 1 \leqslant l \leqslant 4,\\ 1-\mathcal{P}_{m,l-4}^{\mathrm{Proa}}, & \text{if } l \geqslant 5\ , \end{cases} \tag{36}$$

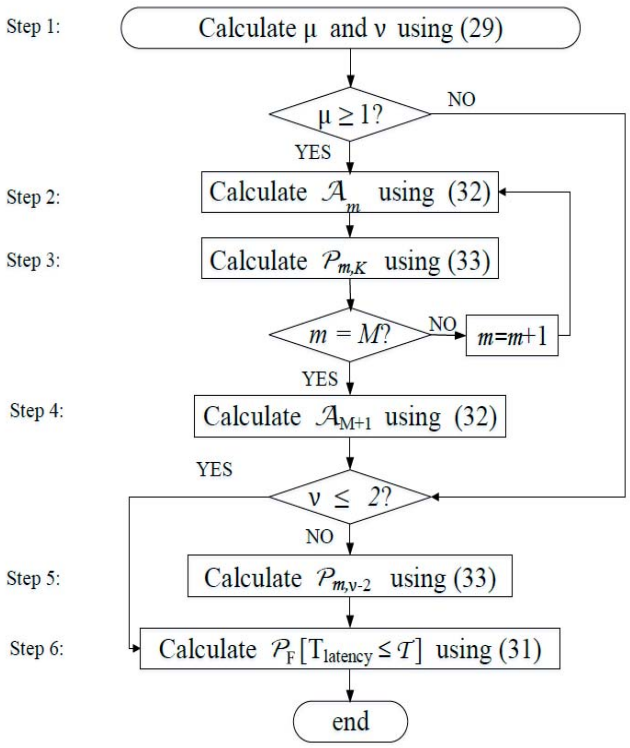

Fig. 8. Flowchart for deriving the latent access failure probability of the Proactive scheme.

and for $l \leq 4$,

$$\begin{aligned}&\Theta^{\mathrm{Proa}}[n,m,l]\\ &= 1-\prod_{r=1}^{l}(1-\mathbb{P}_{m,r})\end{aligned} \tag{37}$$

$$\begin{aligned}&= \sum_{r=1}^{l}(-1)^{r+1}\binom{l}{r}\exp\Big(-\frac{r\gamma_{\mathrm{th}}\sigma^2}{g_m\rho}\Big)(1+\gamma_{\mathrm{th}})^{-kn}\\ &\quad\times\exp\Big(-\eta_{m,r}\mathcal{A}_m^{Proa}\lambda_a/\lambda_B\Big({}_2F_1\Big(-\frac{2}{\alpha},k;\frac{\alpha-2}{\alpha};-\gamma_{\mathrm{th}}\Big)-1\Big)\Big),\end{aligned} \tag{38}$$

and for $l \geq 5$,

$$\Theta^{\mathrm{Proa}}[n,m,l] = 1-(1-\Theta^{\mathrm{Proa}}[n,m,4])\prod_{r=5}^{l}\big(1-\mathbb{P}_{m,r}\big), \tag{39}$$

with

$$\mathbb{P}_{m,r} = \eta_{m,r}\mathrm{O}[n,m,r]\Theta^{\mathrm{Proa}}[n,m,1], \tag{40}$$

where $\Theta^{\mathrm{Proa}}[n,m,1]$ is obtained from (37) and $\mathrm{O}[n,m,r]$ is obtained from (35).

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.

- Step 1: Calculate the indexes μ and ν under the given latency constraint $\mathcal{T}$ TTIs usinig (30). 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 $\mathcal{A}_m^{Proa}$ using (33);
- Step 3: Calculate the GF access success probability in the mth round trip, $\mathcal{P}_{m,K}^{\mathrm{Proa}}$ using (34);
  Repeating Step 2 to 3 until $m = M$;
- Step 4: Calculate non-empty probability $\mathcal{A}_{M+1}^{Proa}$ using (33);
- Step 5: If $\nu \geq 3$, calculate the GF access success probability $\mathcal{P}_{m,\nu-2}^{\mathrm{Proa}}$ using (34);
- Step 6: Calculate the latent access failure probability $\mathcal{P}_{\mathrm{out}}^{\mathrm{Proa}}[T_{\mathrm{latency}} \leqslant \mathcal{T}]$ using (32).

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. At the beginning of each round trip, UEs randomly move to new positions and the active ones randomly choose a pilot from $S = 48$ pilots 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 [30] with mini-slots of 7 OFDM symbols for transmissions in short TTI (0.125ms) using 60 kHz SCS.[7] To focus on the GF access in UL, we assume feedback in DL is error-free.[8] The simulation time is configured to collect at least 5×10^6 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 $\lambda_{\mathrm{B}} = 1$ BSs/km^2, $\lambda_D = 20000$ UEs/km^2, $\gamma_{\mathrm{th}} = -2$ 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 schemes versus SINR thresholds $\gamma_{\mathrm{th}} = -10$dB and $\gamma_{\mathrm{th}} = -2$dB, respectively. The analytical curves of the Reactive and K-repetition schemes are plotted following the flowchart in Fig. 6, and the analytical curves of the Proactive scheme are plotted following the flowchart 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 probabilities stay unchanged for a period of time) is caused by the waiting time between each retransmission.

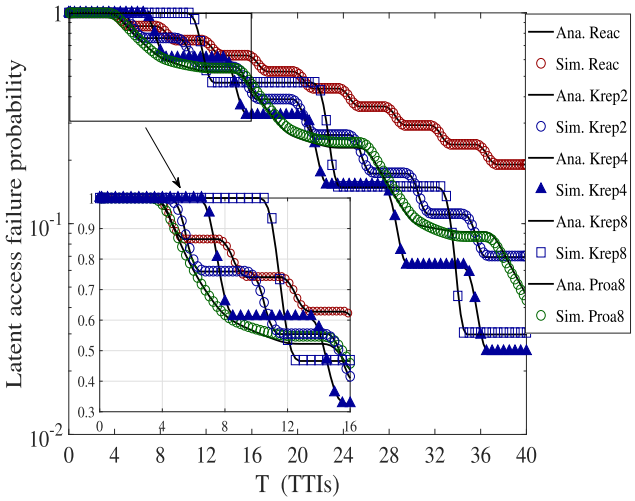

Fig. 9. Latent access failure probability when $\gamma_{\mathrm{th}} = -2$dB.

In Fig. 9, we first observe that the latent access failure probabilities follow Proa=Reac≤Krep under latency constraints $\mathcal{T} \leq 0.625$ms (5TTIs). We also observe that the latent access failure probabilities follow Proa≤Krep≤Reac under latency constraints 0.625 ms $\leq \mathcal{T} \leq$ 1.5ms (12TTIs). In this case, under shorter latency constraints $\mathcal{T} \leq$ 12TTIs, the Proactive scheme should be chosen. This is due to that the Proactive scheme could terminate earlier to reduce latency without waiting for K repetitions, which satisfies the shorter latency constraints. But when the latency constraints $\mathcal{T}$ get longer, the advantage of the Proactive scheme than the K-repetition scheme is not obvious but the advantage of the Proactive and K-repetition schemes than the Reactive scheme is obvious, i.e., Proa&Krep<Reac, due to that the UE has enough time to finish the repetitions and get feedback. We note that increasing repetition value increases the GF access success probability, as it offers more opportunities to retransmit. However, when the repetition value is too large (e.g., $K_{\mathrm{Krep}} = 8$), the latent access failure probabilities are not lower than those of the 4-repetition scheme in most of the time (except $1.5\mathrm{ms} \leq \mathcal{T} \leq 1.8\mathrm{ms}$, $4.2\mathrm{ms} \leq \mathcal{T} \leq 4.5\mathrm{ms}$). This is due to that transmitting 8 repetitions will cost too much waiting time and introduce a much longer delay. 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.

In Fig. 10, we first observe that latent access failure probabilities follow Krep≤Reac≤Proa, under longer latency constraints $\mathcal{T} \geq$ 1.5ms (12 TTIs) for small repetition value $K_{\mathrm{Krep}} = 2$. In this case, under longer latency constraints $\mathcal{T} \geq$ 12TTIs, the K-repetition scheme should be chosen. We also observe that the 8-repetition scheme has the highest latent access failure probabilities. This is due to that there is a trade-off between transmission success probability and non-collision probability when increasing the repetition value, which is in line with Fig. 5 (b). Thus, in Fig. 10, increasing the repetition value to 8 does not decrease but increases the latent access failure probabilities because it introduces longer waiting

[7]Mini-slot durations will depend on the SCS and on the number of OFDM symbols for a given SCS, adopted according to the type of deployment and carrier frequency.

[8]According to Section II.C, a HARQ round-trip includes: 1) UL (UE to BS): the UE transmits the signal to the BS and the BS decodes the received signal; 2) DL (BS to UE): the BS sends an ACK/NACK feedback and the UE processes the feedback to decide whether to perform a retransmission in the next HARQ. This is to say, from the UE perspective, one HARQ round-trip is finished until the UE processes the feedback to know whether it is successful or not, which should consider both the transmission probability in UL and DL. In this article, to focus on the GF access in UL, we assume feedback in DL is error-free. The analysis of the feedback with error probability can be extended following this work.

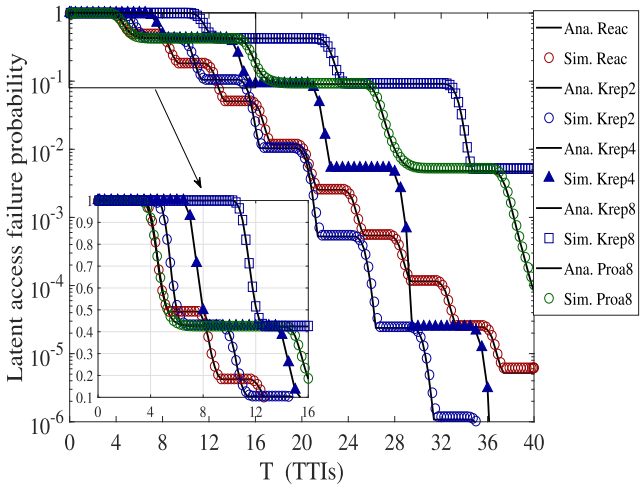

Fig. 10. Latent access failure probability when $\gamma_{\rm th} = -10$ dB.

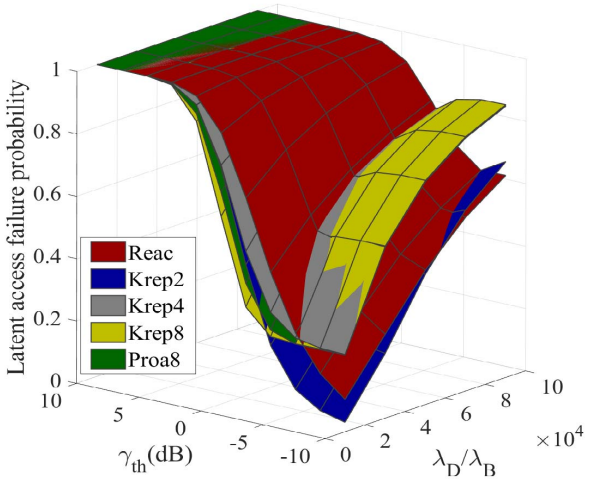

Fig. 12. Latent access failure probability for different density ratios and SINR thresholds when $\mathcal{T} = 12$ TTIs (1.5ms).

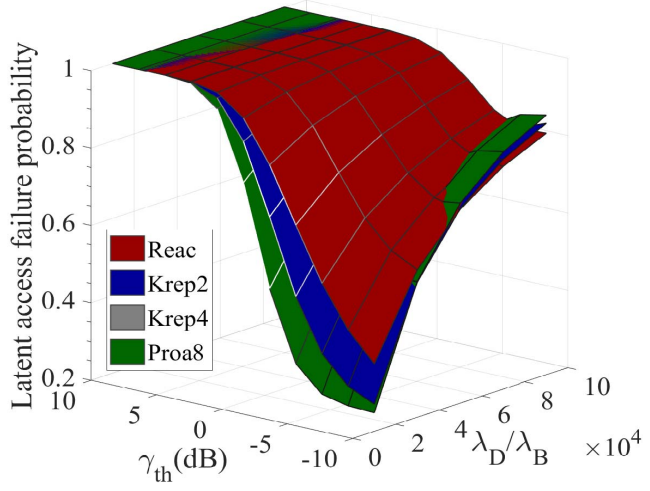

Fig. 11. Latent access failure probability for different density ratios and SINR thresholds when $\mathcal{T} = 8$ TTIs (1ms).

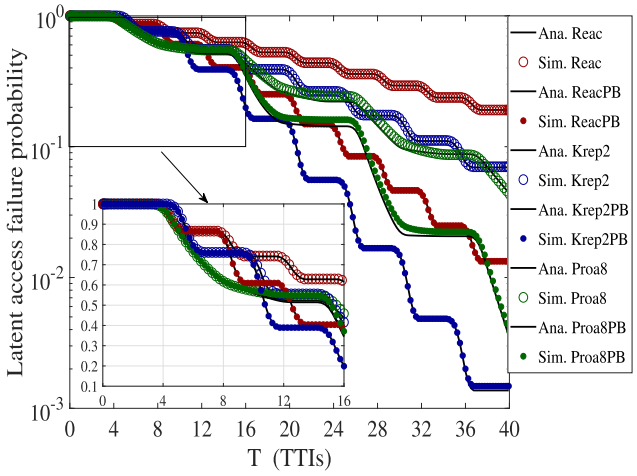

Fig. 13. Latent access failure probability when $\gamma_{\rm th} = -2$dB.

time without increasing the access success probabilities. In this case, when the repetition value is overestimated, the Proactive scheme should be chosen.

Fig. 11-Fig. 12 plot the GF latent access failure probabilities under the latency constraint $\mathcal{T} = 1$ms (8 TTIs) and $\mathcal{T} = 1.5$ms (12 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 also note that, as the latency constraint $\mathcal{T} = 1$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 thresholds scenarios (e.g., $\gamma_{\rm th} \geq -5$dB), while increases in lower SINR thresholds scenarios (e.g., $\gamma_{\rm th} \leq -5$dB) with increasing the repetition value $K_{\rm Krep} > 2$. Thus, despite 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 scenarios.

Fig. 13 plots the GF latent access failure probabilities of the UE under the Reactive, K-repetition, and Proactive

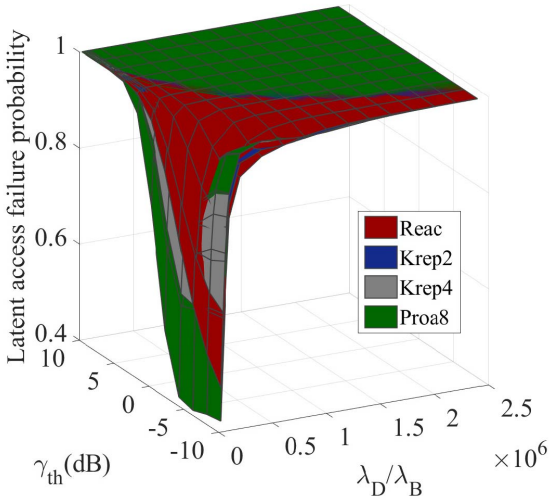

Fig. 14. Latent access failure probability when $\mathcal{T} = 8$ TTIs (1ms) for large densities.

schemes with PB. Interestingly, we observe that PB has greater improvement in the Reactive and K-repetition schemes than the Proactive scheme. For example, without PB, GF latent access failure probabilities of the UE under the K-repetition scheme are similar to those of the Proactive scheme, while with PB, the GF latent access failure probabilities of the UE under the K-repetition scheme is much lower than those of the Proactive scheme.

Fig. 14 plots the GF latent access failure probabilities under the latency constraint $\mathcal{T} = 1$ms (8 TTIs) versus density ratios and SINR thresholds for larger UE densities. We observe that when the density ratios (UE densities) are particularly large ($\lambda_D/\lambda_B \geq 1.5 \times 10^6$), no matter what schemes are taken, the GF access cannot be successful, that is, the network is very crowded. Thus, the number of active UEs that access to the network should be limited to some thresholds.

V. Conclusion

In this article, 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 GF latent access failure probabilities of the UE under the Reactive, the K-repetition, and the Proactive schemes, respectively. Our results have shown that 1) either K-repetition scheme or Proactive scheme provides lower latent access failure probability than the Reactive scheme, except higher density and lower SINR threshold scenarios; 2) under shorter latency constraints ($\mathcal{T} \leq 8$TTIs), the Proactive scheme provides the lowest latent access failure probability; 3) under longer latency constraints, the K-repetition scheme provides the lowest latent access failure probability, which depends on K, i.e., K need to be optimized; 4) if K is overestimated; the Proactive scheme provides lower latent access failure probability than the K-repetition scheme; 5) the Power Boosting can improve the latent access failure probability, especially for the K-repetition scheme (including $K = 1$). The analytical model presented in this article 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 $\mathcal{T}$ TTIs, we have $M = \lfloor (\mathcal{T}-1)/T_{\text{Reac}}^{\text{RTT}} \rfloor$. For $M = 1$, the latent access failure probability is the probability that the UE fails to access in the 1st HARQ round trip, where we can derive

$$\mathcal{P}_F[T_{\text{latency}} \leqslant \mathcal{T}] = 1 - \mathcal{P}_1. \tag{A.1}$$

For $M = 2$, the latent access failure probability is the probability that the UE fails to access in neither two HARQ round trips, where we can derive

$$\mathcal{P}_F[T_{\text{latency}} \leqslant \mathcal{T}] = 1 - \mathcal{P}_1 - (1-\mathcal{P}_1)\mathcal{P}_2. \tag{A.2}$$

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

$$\mathcal{P}_F[T_{\text{latency}} \leqslant \mathcal{T}] = 1 - \sum_{m=1}^{M=2} \mathcal{A}_m \mathcal{P}_m. \tag{A.3}$$

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

$$\begin{aligned}
&\mathcal{P}_F[T_{\text{latency}} \leqslant \mathcal{T}] \\
&\quad = 1 - \mathcal{P}_1 - (1-\mathcal{P}_1)\mathcal{P}_2 - (1 - \mathcal{P}_1 - (1-\mathcal{P}_1)\mathcal{P}_2)\mathcal{P}_3 \\
&\quad = 1 - \mathcal{P}_1 - \mathcal{A}_2\mathcal{P}_2 - \mathcal{A}_3\mathcal{P}_3 = 1 - \sum_{m=1}^{M=3} \mathcal{A}_m \mathcal{P}_m.
\end{aligned} \tag{A.4}$$

For $M > 3$, the latent access failure probability $\mathcal{P}_F[T_{\text{latency}} \leqslant \mathcal{T}]$ 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

$$\begin{aligned}
&\Theta^{\text{Reac}}[n, m] \\
&\quad = \mathbb{P}[\text{SINR}_m \geqslant \gamma_{\text{th}} | N = n] \\
&\quad = \mathbb{P}\left\{ \frac{g_m \rho h_0}{\mathcal{I}_{\text{inter}}^m + \mathcal{I}_{\text{intra}}^m + \sigma^2} \geqslant \gamma_{\text{th}} \middle| N = n \right\} \\
&\quad = \exp\left(-\frac{\gamma_{\text{th}}}{g_m \rho}\sigma^2\right) \mathcal{L}_{\mathcal{I}_{\text{inter}}^m}\left(\frac{\gamma_{\text{th}}}{g_m\rho}\right) \mathcal{L}_{\mathcal{I}_{\text{intra}}^m}\left(\frac{\gamma_{\text{th}}}{g_m\rho} \middle| N = n\right).
\end{aligned} \tag{B.1}$$

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

$$\begin{aligned}
\mathcal{L}_{\mathcal{I}_{\text{intra}}^m}(s|N=n) &= E\left[\exp\left(-s\sum_{\beta=1}^{n} g_m \rho h_\beta\right)\right] \\
&= \left(\frac{1}{1+s g_m \rho}\right)^n,
\end{aligned} \tag{B.2}$$

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

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

$$\begin{aligned}
&\mathcal{L}_{\mathcal{I}^m_{\text{inter}}}(s)\\
&= E\Big[\exp\Big(-s\sum_{i\in\mathcal{Z}_{\text{inter}}} g_m P_i h_i \|u_i\|^{-\alpha}\Big)\Big]\\
&\overset{\text{(a)}}{=} E\Big[\prod_{i\in\mathcal{Z}_{\text{inter}}}\frac{1}{1+sg_mP_iy_i^{-\alpha}}\Big]\\
&\overset{\text{(b)}}{=} \exp\Big(-2\pi\mathcal{A}^{\text{Reac}}_m\lambda_a\int_{(\frac{P}{g_m\rho})^{\frac{1}{\alpha}}}^{\infty} E_P\Big[1-\frac{1}{1+sg_mPy^{-\alpha}}\Big]ydy\Big)\\
&\overset{\text{(c)}}{=} \exp\Big(-2\pi\mathcal{A}^{\text{Reac}}_m\lambda_a(g_ms)^{\frac{2}{\alpha}}E_P[P^{\frac{2}{\alpha}}]\int_{(sg_m\rho)^{\frac{-1}{\alpha}}}^{\infty}\frac{x}{1+x^{\alpha}}dx\Big)\\
&= \exp\Big(-2\mathcal{A}^{\text{Reac}}_m\lambda_a/\lambda_B(\gamma_{\text{th}})^{\frac{2}{\alpha}}\int_{(\gamma_{\text{th}})^{\frac{-1}{\alpha}}}^{\infty}\frac{x}{1+x^{\alpha}}dx\Big),
\end{aligned} \tag{B.3}$$

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 $E_P[P^{\frac{2}{\alpha}}] = \rho^{\frac{2}{\alpha}}/(\pi\lambda_B)$ 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 mth round trip as

$$\begin{aligned}
&\Theta^{\text{Reac}}[n,m] = \exp\big(-\frac{\gamma_{\text{th}}\sigma^2}{g_m\rho}\big)(1+\gamma_{\text{th}})^{-n}\\
&\times\exp\Big(-\mathcal{A}^{\text{Reac}}_m\lambda_a/\lambda_B\Big({}_2F_1\Big(-\frac{2}{\alpha},1;\frac{\alpha-2}{\alpha};-\gamma_{\text{th}}\Big)-1\Big)\Big).
\end{aligned} \tag{B.4}$$

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

$$\begin{aligned}
&\Theta^{\text{Reac}}[n,m]\\
&= \exp\big(-\frac{\gamma_{\text{th}}\sigma^2}{g_m\rho}\big)(1+\gamma_{\text{th}})^{-n}\\
&\quad\times\exp\Big(-(\gamma_{\text{th}})^{\frac{1}{2}}\mathcal{A}^{\text{Reac}}_m\lambda_a/\lambda_B\arctan((\gamma_{\text{th}})^{\frac{1}{2}})\Big).
\end{aligned} \tag{B.5}$$

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_{Krep} repetitions conditioning on n number of intra-cell interfering UEs based on the SINR outage as

$$\begin{aligned}
&\Theta^{\text{Krep}}[n,m,K_{\text{Krep}}]\\
&\qquad= 1-\prod_{k=1}^{K_{\text{Krep}}}\Big(1-\mathbb{P}[\text{SINR}^m_k\geqslant\gamma_{\text{th}}|N=n]\Big).
\end{aligned} \tag{C.1}$$

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

$$\begin{aligned}
&\Theta^{\text{Krep}}[n,m,K_{\text{Krep}}]\\
&= \sum_{k=1}^{K_{\text{Krep}}}(-1)^{k+1}\binom{K_{\text{Krep}}}{k}\\
&\quad\times\mathbb{P}[\text{SINR}^m_1\geqslant\gamma_{\text{th}},\ldots,\text{SINR}^m_k\geqslant\gamma_{\text{th}}|N=n],
\end{aligned} \tag{C.2}$$

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

$$\begin{aligned}
&\mathbb{P}[\text{SINR}^m_1\geqslant\gamma_{\text{th}},\ldots,\text{SINR}^m_k\geqslant\gamma_{\text{th}}|N=n]\\
&\quad= \exp\Big(-\frac{k\gamma_{\text{th}}}{g_m\rho}\sigma^2\Big)\mathcal{L}_{I^m_{\text{inter}}}\Big(\frac{\gamma_{\text{th}}}{g_m\rho}\Big)\mathcal{L}_{I^m_{\text{intra}}}\Big(\frac{\gamma_{\text{th}}}{g_m\rho}\Big|N=n\Big).
\end{aligned} \tag{C.3}$$

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

$$\begin{aligned}
\mathcal{L}_{I^m_{\text{intra}}}(s|N=n) &= E\Big[\exp\Big(-s\sum_{\beta=1}^{n}g_m\rho\sum_{r=1}^{k}h^r_\beta\Big)\Big]\\
&= \Big(\frac{1}{1+sg_m\rho}\Big)^{kn},
\end{aligned} \tag{C.4}$$

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

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

$$\begin{aligned}
&\mathcal{L}_{I^m_{\text{inter}}}(s)\\
&= E\Big[\exp\Big(-s\sum_{i\in\mathcal{Z}_{\text{inter}}}g_mP_i(\sum_{r=1}^{k}h^r_i)\|u_i\|^{-\alpha}\Big)\Big]\\
&= E\Big[\prod_{i\in\mathcal{Z}_{\text{inter}}}\Big(\frac{1}{1+sg_mP_iy_i^{-\alpha}}\Big)^k\Big]\\
&= \exp\Big(-2\pi\mathcal{A}^{Krep}_m\lambda_a\int_{(\frac{P}{\rho})^{\frac{1}{\alpha}}}^{\infty}E_P\Big[1-(\frac{1}{1+sg_mPy^{-\alpha}})^k\Big]ydy\Big)\\
&= \exp\Big(-2\mathcal{A}^{Krep}_m\frac{\lambda_a}{\lambda_B}(\gamma_{\text{th}})^{\frac{2}{\alpha}}\int_{(\gamma_{\text{th}})^{\frac{-1}{\alpha}}}^{\infty}\Big[1-(\frac{1}{1+x^{-\alpha}})^k\Big]xdx\Big).
\end{aligned} \tag{C.5}$$

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 mth roudn trip with the K-repetition scheme as

$$\begin{aligned}
&\Theta^{\text{Krep}}[n,m,K_{\text{Krep}}]\\
&= \sum_{k=1}^{K_{\text{Krep}}}(-1)^{k+1}\binom{K_{\text{Krep}}}{k}\exp\big(-\frac{k\gamma_{\text{th}}\sigma^2}{g_m\rho}\big)(1+\gamma_{\text{th}})^{-kn}\\
&\quad\times\exp\Big(-\mathcal{A}^{\text{Krep}}_m\lambda_a/\lambda_B\Big({}_2F_1\Big(-\frac{2}{\alpha},k;\frac{\alpha-2}{\alpha};-\gamma_{\text{th}}\Big)-1\Big)\Big).
\end{aligned} \tag{C.6}$$

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

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

where (D.2), as shown at the top of the next page. For $l \geq 5$, the UE can receive feedback from the 4th repetition, thus the number of interfering users changes from the

$$\mathbb{P}_{1,r} = \eta_{1,r}\Theta^{\mathrm{Proa}}[n,1,1] = \frac{\eta_{1,r}\exp\left(\frac{-\gamma_{\mathrm{th}}\sigma^2}{\rho} - \eta_{1,r}\frac{\lambda_a}{\lambda_B}\left({}_2F_1\left(-\frac{2}{\alpha},1;\frac{\alpha-2}{\alpha};-\gamma_{\mathrm{th}}\right)-1\right)\right)}{\left(1+\gamma_{\mathrm{th}}\right)^n} \tag{D.2}$$

5th repetition. So we have

$$\Theta^{\mathrm{Proa}}[n,1,l] = 1-(1-\Theta^{\mathrm{Proa}}[n,1,4])\Big(\prod_{r=5}^{l} 1-\mathbb{P}_{1,r}\Big), \tag{D.3}$$

where

$$\mathbb{P}_{1,r} = \eta_{1,r}\mathrm{O}[n,1,r]\Theta^{\mathrm{Proa}}[n,1,1], \tag{D.4}$$

with

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

REFERENCES

[1] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document 3GPP, TS 38.913 v15.2.0, Jun. 2018.

[2] *IMT Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document M.2083-0, Recommendation ITU-R, Sep. 2015, p. 2083.

[3] B. Holfeld *et al.*, "Wireless communication for factory automation: An opportunity for LTE and 5G systems," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 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.*, vol. 24, no. 6, pp. 38–45, Dec. 2017.

[5] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42–48, Mar. 2019.

[6] M. Yao, M. Sohul, V. Marojevic, and J. H. Reed, "Artificial intelligence defined 5G radio access networks," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 14–20, Mar. 2019.

[7] *5G; Service Requirements for the 5G System*, document 3GPP, TS 22.261 v17.2.0, Mar. 2020.

[8] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document 3GPP, TS 36.213 V14.2.0, Mar. 2017.

[9] *Semi-Persistent Scheduling for 5G New Radio URLLC*, document R1-167309, 3GPP TSG-RAN WG1 #86, Aug. 2016.

[10] *Dynamic Scheduling Based Transmission for URLLC*, document R1-1705245, 3GPP TSG-RAN WG1 #88, Apr. 2017.

[11] *UL Grant-Free Transmission for URLLC*, document R1-1705654, 3GPP TSG-RAN WG1 #88, Apr. 2017.

[12] L. Vangelista and M. Centenaro, "Performance evaluation of HARQ schemes for the Internet of Things," *Computers*, vol. 7, no. 4, p. 48, Sep. 2018.

[13] S. Sesia, I. I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*. Hoboken, NJ, USA: Wiley, 2011.

[14] *5G; NR; Physical Layer Procedures for Data*, document 3GPP TS 38.214 v15.9.0, Mar. 2020.

[15] *UL Grant-Free Transmission for URLLC*, document R1-1705246, Apr. 2017.

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

[17] *Discussion on Explicit HARQ-ACK Feedback for Configured Grant Transmission*, document R1-1903079, 3GPP TSG RAN WG1 #96, Mar. 2019.

[18] 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.*, vol. 66, no. 11, pp. 5788–5803, Nov. 2018.

[19] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. Kovacs, and P. Mogensen, "Power control optimization for uplink grant-free URLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.

[20] N. H. Mahmood, R. Abreu, R. Bohnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.

[21] 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.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[22] *RAN1 Chairman's Notes*, document 3GPP, TSG RAN WG1 NR Ad-Hoc#2, Jun. 2017.

[23] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, document 3GPP, TR 38.824 v16.0.0, Mar. 2019.

[24] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[25] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 182–185, Apr. 2018.

[26] 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.*, vol. 17, no. 10, pp. 6853–6870, Oct. 2018.

[27] J. F. C. Kingman, *Poisson Processes*. Hoboken, NJ, USA: Wiley, Jan. 1993.

[28] 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 *Proc. IEEE Wireless Commun. Netw. Conf. Workshop (WCNCW)*, Apr. 2019, pp. 1–6.

[29] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovács, and P. Mogensen, "System level analysis of K-repetition for uplink grant-free URLLC in 5G NR," in *Proc. 25th Eur. Wireless Conf. Eur. Wireless*, May 2019, pp. 1–5.

[30] *Study on New Radio Access Technology-Physical Layer Aspects*, document 3GPP, TR 38.802 v14.0.0, Mar. 2017.

[31] *Study on RAN Improvements for Machine-Type Communications*, document 3GPP TR 37.868 v.11.2.0, Sep. 2011.

[32] M. Gharbieh, H. ElSawy, A. Bader, and M.-S. Alouini, "Spatiotemporal stochastic modeling of IoT enabled cellular networks: Scalability and stability analysis," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3585–3600, Aug. 2017.

[33] Y. Yang and T.-S.-P. Yum, "Analysis of power ramping schemes for UTRA-FDD random access channel," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2688–2693, Nov. 2005.

[34] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *Proc. IEEE 11th Int. Symp. Workshops Modeling Optim. Mobile, Ad Hoc Wireless Netw. (WiOpt)*, Tsukuba, Japan, May 2013, pp. 119–124.

Yan Liu (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in electronic engineering with the Queen Mary University of London, London, U.K. Her research interests include the Internet of Things and ultrareliability and low-latency communications.

Yansha Deng (Member, IEEE) received the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2015. From 2015 to 2017, she was a Post-Doctoral Research Fellow with Kings College London, U.K., where she is currently a Lecturer (Assistant Professor) with the Department of Engineering. Her research interests include molecular communications, machine learning, and 5G wireless networks. She has served as a TPC Member of many IEEE conferences, such as the IEEE GLOBECOM and ICC. She was a recipient of the Best Paper Awards from ICC 2016 and GLOBECOM 2017 as the first author. She is an Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATIONS and a Senior Editor of the IEEE COMMUNICATION LETTERS. She received the Exemplary Reviewers of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2016 and 2017 and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2018.

Maged Elkashlan (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from The University of British Columbia in 2006.

From 2007 to 2011, he was with Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia. During this time, he held visiting faculty appointments at the University of New South Wales, The University of Sydney, and the University of Technology Sydney. In 2011, he joined the School of Electronic Engineering and Computer Science, Queen Mary University of London. He also holds a visiting faculty appointment at the Beijing University of Posts and Telecommunications. His research interests include communication theory and statistical signal processing. He was a co-recipient of the Best Paper Awards at the IEEE International Conference on Communications (ICC) in 2016 and 2014, the International Conference on Communications and Networking in China (CHINACOM) in 2014, and the IEEE Vehicular Technology Conference (VTC-Spring) in 2013. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2013 to 2018 and the IEEE COMMUNICATIONS LETTERS from 2012 to 2016. He is an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATIONS.

Arumugam Nallanathan (Fellow, IEEE) has been a Professor of wireless communications and the Head of Communication Systems Research (CSR) Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since September 2017. He was with the Department of Informatics, King's College London, from December 2007 to August 2017, where he was a Professor of wireless communications from April 2013 to August 2017 and has been a Visiting Professor since September 2017. He was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, from August 2000 to December 2007. His research interests include artificial intelligence for wireless systems, beyond 5G wireless networks, the Internet of Things (IoT), and molecular communications. He has published nearly 500 technical papers in scientific journals and international conferences.

Dr. Nallanathan was a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016 (ICC'2016), the IEEE Global Communications Conference 2017 (GLOBECOM'2017), and the IEEE Vehicular Technology Conference 2018 (VTC'2018). He received the IEEE Communications Society SPCE Outstanding Service Award in 2012 and the IEEE Communications Society RCC Outstanding Service Award in 2014. He has been selected as a Web of Science Highly Cited Researcher in 2016 and the AI 2000 Internet of Things Most Influential Scholar in 2020. He has served as the Chair of the Signal Processing and Communication Electronics Technical Committee of the IEEE Communications Society and the Technical Program Chair and a member of Technical Program Committees in numerous IEEE conferences. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2006 to 2011, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2006 to 2017, and the IEEE SIGNAL PROCESSING LETTERS. He is an Editor-at-Large of the IEEE TRANSACTIONS ON COMMUNICATIONS and a Senior Editor of the IEEE WIRELESS COMMUNICATIONS LETTERS. He is an IEEE Distinguished Lecturer.

George K. Karagiannidis (Fellow, IEEE) was born in Pythagoreio, Greece. He received the University Diploma (five years) and Ph.D. degrees in electrical and computer engineering from the University of Patras in 1987 and 1999, respectively. From 2000 to 2004, he was a Senior Researcher with the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In June 2004, he joined the faculty of the Aristotle University of Thessaloniki, Greece, where he is currently a Professor with the Electrical and Computer Engineering Department and the Head of the Wireless Communications Systems Group (WCSG). He is also Honorary Professor with Southwest Jiaotong University, Chengdu, China. His research interests include digital communications systems and signal processing, with emphasis on wireless communications, optical wireless communications, wireless power transfer and applications and communications, and signal processing for biomedical engineering. He is one of the highly cited authors across all areas of electrical engineering, recognized from Clarivate Analytics as Web-of-Science Highly Cited Researcher in the five consecutive years 2015–2019. He has been involved as the General Chair, the Technical Program Chair, and a member of the Technical Program Committees in several IEEE and non-IEEE conferences. He was an Editor in several IEEE journals and the Editor-in Chief of the IEEE COMMUNICATIONS LETTERS from 2012 to 2015. He serves as an Associate Editor-in-Chief of the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.