

Received November 21, 2020, accepted December 26, 2020, date of publication January 1, 2021, date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048824

Analysis of Two-Step Random Access Procedure for Cellular Ultra-Reliable Low Latency Communications

JUN-BAE SEO¹, (Member, IEEE), WAQAS TARIQ TOOR²,
AND HU JIN³, (Senior Member, IEEE)

¹Department of Information and Communication Engineering, Gyeongsang National University, Tongyeong 53064, Republic of Korea

²Department of Electrical Engineering, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan

³Division of Electrical Engineering, Hanyang University, Ansan 15588, Republic of Korea

Corresponding author: Hu Jin (hjin@hanyang.ac.kr)


This work was supported in part by the 5G based IoT Core Technology Development Project Grant funded by the Korean Government (MSIT) (No. 2020-0-00167, Core Technologies for Enhancing Wireless Connectivity of Unlicensed Band Massive IoT in 5G+ Smart City Environment), in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant NRF-2017K1A3A1A19071179, and in part by the Development Fund Foundation, Gyeongsang National University, 2020.

ABSTRACT Due to the emergence of Internet of Things (IoTs), it can be expected that the bandwidth provided by cellular systems might be consumed up soon. Some applications of them are delay-sensitive such that it would be critical to guarantee random access (RA) delay less than a threshold. Since the existing Long-Term Evolution-Advanced (LTE-A) RA procedure is a four-step signaling procedure, it may not be suitable for such delay-sensitive applications due to its time-consuming procedure. This work investigates a two-step RA procedure for 5G New Radio systems, where RA preamble and bandwidth request message are transmitted at the same time. First we show that the operating region of two-step RA procedure can be divided into three regions such as unsaturated stable, bistable, and saturated regions in terms of a packet generation probability, retransmission probability, the number of devices, and the number of RA preambles. To see whether RA delay requirement of delay-sensitive applications can be guaranteed, this work shows that the system should run under unsaturated region and derives RA delay distribution when IoT devices employ geometric probability backoff (GPB) or uniform window backoff (UWB) algorithm. We then examine the probability that the RA delay would be larger than some threshold depending on the operation regions.

INDEX TERMS Random access procedure, URLLC, 5G, long-term evolution advanced.

I. INTRODUCTION

In realizing machine-to-machine (M2M) communications and Internet-of-Things (IoT), remote devices such as sensors, metering, monitoring, or charging devices should be wirelessly connected in order to exchange and report measured data. The collected data in applications of smart city, smart grids, connected vehicles, etc., help us to make better controls, decisions, and planning. As the number of devices grows tremendously large in proportion to upcoming IoT and M2M applications, it might be readily expected that frequent information requests and exchanges from a huge number of devices to servers or vice versa can clog up the

The associate editor coordinating the review of this manuscript and approving it for publication was Zubair Md. Fadlullah .

access channels. In addition, demands on real-time based and/or interactive applications such as autonomous vehicles and healthcare [1] grow rapidly as well, where loss of some information or some delayed ones beyond a threshold could be critically damages for the systems. Therefore, ultra-reliable low latency communications (URLLC) in random access (RA) systems are more demanded than ever. It requires the existing or upcoming RA procedure, e.g., Long-Term Evolution-Advanced (LTE-A) and the fifth generation (5G) New Radio (NR), to guarantee low latency with high reliability.

The LTE-A system would be the basis for 5G NR systems due to backward compatibility. It will also coexist with 5G networks in order to support them. The entire RA procedure of LTE-A system [2] and 5G NR *release 15* is a four-step

handshake procedure: First, IoT devices select randomly one RA preamble and transmit it to the serving eNodeB. When receiving the response message from the serving eNodeB as the second message, the devices send a bandwidth request message as the third message. The fourth message is the response message from the eNodeB as well. It was not until the device received the fourth message that they could know the success of their RA. Note that the devices of sending a non-duplicate RA preamble make a successful RA. Therefore, this procedure might not be suitable for delay-sensitive applications.

In order to improve the performance of LTE-A RA procedure in presence of massive IoT devices, researches thus far in the literature have focused on maximizing system throughput by optimizing backoff algorithm for retransmission control, and/or introducing access barring mechanisms that suppress new packet transmissions depending on access priority in [3]–[5] and reference therein. The efforts made for improving the current LTE-A RA procedure however is fundamentally limited for URLLC, since the four-step handshake procedure of causing long and heavy signalling overheads remains the same. Recently, a two-step RA procedure has been proposed in [6]–[9] and standardized in 5G NR (*release 16*) in [10], [11] in order to overcome such signalling overheads of the four-step signalling procedure. It allows IoT devices that have transmitted an RA preamble to physical random access channel (PRACH) to transmit the bandwidth request message to physical uplink shared channel (PUSCH) *without confirming the outcome of RA preamble transmission*. Therefore, upon receiving a non-duplicate RA preamble transmission in PRACH, the eNodeB can read the bandwidth request message from PUSCH without sending IoT devices RA response message to allocate PUSCH resource for the bandwidth request message. In this case, the resource in PUSCH for the bandwidth request message is mapped to RA preambles in [11]. In other words, when IoT devices select a RA preamble for two-step RA procedure, they know where to send the bandwidth request message in PUSCH following PRACH. Since the signalling procedure is shortened in the two-step procedure, two performance metrics need to be characterized in realizing URLLC for real-time IoT:

- The *distribution* of RA delay is important for the system to dimension the system parameters, e.g., retransmission probability, the number of devices, a packet generation probability, and the number of RA preambles in order to keep the probability that access delay exceeds a delay constraint below a threshold.
- To run a *reliable* RA system, the stable operating region should be identified in terms of the system parameters as well. Moreover, the two-step RA procedure can coexist with the four-step one by allocating some specific RA preambles for it. Thus, dimensioning RA preambles is important to achieve URLLC in the two-step RA procedure.

Note that the performance of two-step RA procedure particularly with respect to RA delay distribution and the

operating region has been not analytically characterized yet in [6]–[9].

While it has long been known that RA systems based on S-ALOHA have inherently bistability [12]–[19], it has been less concerned to characterize bistability latent in the RA procedure built upon multichannel S-ALOHA. It is notable that when a RA system is trapped into a bistable state, the number of backlogged devices (attempting to retransmit) swings from a small to a large number, back and forth over time. Meanwhile, the devices experience a low throughput and excessive access delay. It is thus critical to eliminate or avoid bistability to run reliable RA procedure. It can be triggered by a joint force of some new packet arrival rates and retransmission rates from the backlogged devices. We shall see later that in order to prevent it or get the system out of it quickly, either new packet arrivals or retransmissions should be controlled below some thresholds or more RA preambles are allocated.

As prior work, bistability on S-ALOHA systems has been investigated in [12]–[19]. An analytical framework based on catastrophe theory has been established in [13], where new packet arrival rate and retransmission probabilities (or rate) are identified as two parameters of causing bistability. Based on this framework, the effects of propagation delay and capture, various multipacket reception channels on bistability have been investigated in [14]–[16]. Especially, it has been found that the bistability region is reduced when the number of retransmissions is limited [17]. However, this is achieved by getting rid of the number of backlogged devices upon the maximum number of retransmissions, i.e., packet dropping. Thus far, bistability of multichannel S-ALOHA systems is not examined in [12]–[17]. Especially in [18], bistability of LTE-A RA procedure has been examined with focus on the effect of limiting the number of retransmissions as in [17]. Depending on traffic intensity (new packet arrival rates per RA preamble), sudden throughput collapse and excessive jump of the mean access delay are observed. In addition, it has been examined in [19] when devices have two queues, i.e., one for data packets and the other for access request packets. While bistability of LTE-A RA procedure has been considered in [18], [19], its region has not been explicitly characterized yet. Compared to [12]–[19], this work investigates bistability of two-step RA procedure by applying catastrophe theory and characterizes its operating region with respect to the system parameters of interest.

On the other hand, as research on RA delay distribution, it was examined for S-ALOHA and carrier sense multiple access (CSMA) with geometric probability backoff (GBP), uniform window backoff (UWB), binary exponential backoff (BEB) algorithms in [20] and for S-ALOHA with BEB algorithm in [21]. Compared to [20], [21], where single-channel systems are considered, the two-step RA procedure employs *multichannel* S-ALOHA system; that is, the previous results are no longer applicable. For multichannel systems, the mean access delay for UWB algorithm of LTE-A RA procedure was examined in [22] and an

approximate mean access delay (including queueing delay) was studied when each user has a queue and employs GPB algorithm in [23]. Furthermore, RA delay distribution and a lower bound of the mean access delay have been studied for some bursty traffic model in [24]–[26]. Similar to the two-step procedure presented in this work, Choi also considered the two-step RA procedure with fast retransmission by assuming that each device has a queue to store incoming packets [27]. According to fast retransmission, devices in collision retransmit a RA preamble at the next slot right away in order to lower RA access delay. Instead of examining RA delay distribution, it examined device’s queue length distribution of each device. In [28] Centenaro *et al.* examined two-step and four-step RA procedures. They focused on throughput and the probability that RA attempt eventually fails after a certain number of retransmissions. In comparison, our work characterizes RA delay distribution for GPB and UWB algorithms under the operating regions of two-step RA procedure. Our study shall show that study on RA delay distribution is a byproduct of throughput study, when retransmission intervals are randomly drawn based on independent and identical distribution.

The main contributions of this work can be summarized as follows:

- Three operating regions of two-step RA procedure such as unsaturated stable, bistable, and saturated regions are characterized in terms of the number of devices, the number of RA preambles, retransmission probability, and a packet generation probability. This would be useful to set those system parameters so as to run the system reliably.
- RA delay distribution is derived for two backoff algorithms, i.e., GPB and UWB algorithms in two-step RA procedure. This enables us to examine an access delay violation probability subject to a delay constraint, which is critical to delay-sensitive IoT applications. Moreover, we present how throughput and access delay distribution behave according to the operating regions of the system.

This paper is organized as follows: Section II introduces the two-step RA procedure. Analysis on bistability and RA delay distribution is presented in Section III. Numerical studies are discussed in Section IV and concluding remarks are given in Section V.

II. SYSTEM MODEL

A RA system for two-step RA procedure is introduced with focus on LTE-A systems. As far as medium access control (MAC) layer is concerned, LTE-A and 5G NR might not have significant differences. In what follows, the difference regarding two-step RA procedure shall be mentioned if necessary.

In uplink channel of LTE-A systems, time is organized as a unit of subframe, which takes 1 msec. One frame consists of 10 subframes numbered from 0 to 9 and each frame is also numbered. Three physical uplink channels are defined as follows: Physical uplink control channel (PUCCH) is used for

hybrid automatic repeat request (H-ARQ) acknowledgements and channel-state reports. The other two channels are PRACH and PUSCH, which have been mentioned before. Both in the two-step and four-step LTE-A RA procedures, a RA preamble called Msg 1 is transmitted to PRACH, whereas a bandwidth request message called Msg 3 is sent to PUSCH, where other radio resource control messages are also carried.

The PRACH can appear periodically in a frame. Its periodicity is determined by PRACH configuration index. Among a total of 64 configuration indices, one is selected depending on traffic load and coverage area of the eNodeB. The PRACH appears in subframe 1 every frame for PRACH configuration index 3 and every odd subframe for configuration index 13. In Appendix B, the information regarding 64 PRACH configuration indices are given.

Without loss of generality, we define one *RA slot period* as a time period of PRACH¹; that is, it starts with the beginning of PRACH and includes the following PUSCH. For example, PRACH appears every even-numbered subframe in Fig. 1. Then, one RA slot is composed of two subframes; that is, one with PRACH and the other without PRACH. In the two-step RA procedure, PUSCH resource for device to transmit Msg 3 is predefined for each RA preamble as shown in Fig. 1. We assume that each device can hold one packet to transmit which is a general consequence of IoT applications. When a device has a packet, we call it backlogged. At each RA slot period, a non-backlogged device can generate a packet with probability p , which is called packet generation probability. While different packet generation probability for each device may reflect a more practical network scenario, the analysis for the the same packet generation probability introduces the worst case scenario [30]. The two-step RA procedure is described as follows:

¹Although one subframe in 5G NR is 1 msec long, it consists of multiple slots so that the length of a slot is much shorter than 1 msec. The slot duration of PRACH in a subframe, its periodicity, and location can be determined by one of 256 PRACH configuration indices in 5G NR [29]. Notice that since two-step RA procedure works in MAC layer, it can be effortlessly integrated with PRACH configuration in physical layer due to independence between layers. The results of MAC layer presented here can be scaled down in time-domain, when a much smaller time-scale in physical layer is applied.

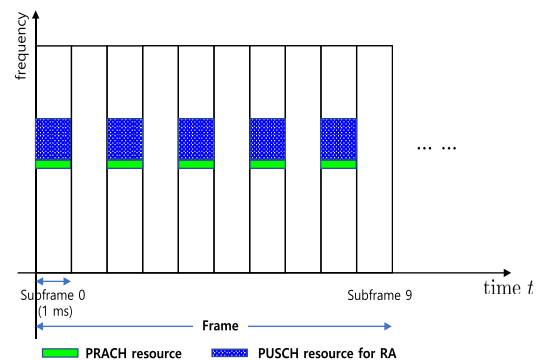


FIGURE 1. The location of RACH and PUSCH for two-step RA with PRACH configuration index 12.

- *Step 1:* Each backlogged IoT device generates a random number in the unit interval $[0, 1]$. If it is less than r , the device chooses one of L RA preambles and transmits it to PRACH and at the same time transmits Msg 3 to a resource in PUSCH designated for the RA preamble selected, as shown in Fig. 1. This Msg 3 can be a connection request message including device ID.
- *Step 2:* The eNodeB sends RA response (RAR) message over the downlink, which specifies a bandwidth for the devices to send a data.

Notice that in Step 1, when a device transmits a non-duplicate RA preamble, its Msg 3 can be successfully decoded in PUSCH. Otherwise, its RA attempt fails. It has been assumed in Step 1 that retransmissions are made with probability r at each RA slot, i.e., GPB algorithm. In practice, devices use UWB algorithm with retry limit. More specifically, the eNodeB broadcasts a uniform window size U called a backoff indicator. In Step 1, the devices pick up a random integer in the interval $[1, U]$ and count it down every RA slot period. If the counter hits zero, the device (re)transmits its packet. While we make use of GPB for Markovian analysis due to its memoryless property in the next section, we also carry out the analysis on RA delay distribution for UWB algorithm as well.

Let us make some notes on the differences between the two-step and four-step RA procedures: *First*, in the four-step RA procedure, when the eNodeB sends RAR message for the RA preambles transmitted, it specifies uplink resource in PUSCH for Msg 3 transmission. Accordingly, PUSCH resource for Msg 3 is not allocated for the RA preambles not transmitted. When a RA preamble is transmitted in subframe 0, the PUSCH for this RA preamble shall be allocated somewhere in a subframe following subframe 0 in 5G NR. In addition, however many devices transmit a specific RA preamble, the eNodeB allocates them one unit of PUSCH resource as if only one device transmits the RA preamble.

In contrast, in two-step RA procedure, the resources in PUSCH for Msg 3 is predefined for each RA preamble, i.e., whether a RA preamble is transmitted or not, as shown in Fig. 1. In this work, we assume that PUSCH resource is located in the same subframe, where PRACH is. When a RA preamble transmission is successful, the RA procedure can be finished in the same subframe. It could be said that the two-step RA procedure shortens access delay in expense of more PUSCH resources (predefined) for Msg 3. *Second*, since the two-step RA procedure can be finished before the next PRACH begins, upon unsuccessful RA attempt a RA preamble retransmission can be made at the next PRACH. This may not hold for the four-step procedure, because the devices come to realize their unsuccessful RA after receiving the (fourth message) response to Msg 3 transmission. *Third*, according to the four-step procedure, the devices have to transmit Msg 3 with H-ARQ up to the maximum of retransmissions, say R times. More than one devices get the same PUSCH resource for Msg 3 when they retransmits the same RA preamble. This PUSCH resource will be wasted over

R RA slot periods in the four-step RA procedure, whereas it takes long times for those devices to resume RA. In the two-step RA procedure, it can happen that either RA preamble, or PUSCH can be successfully transmitted. For analytical simplicity, we assume that both can be successfully transmitted due to heavy coding and a high transmit power for the message transmitted to PUSCH, as long as RA preamble is successfully transmitted.² *Finally*, two-step and four-step RA procedures can coexist by allocating two different sets of RA preambles to them. Therefore, depending on RA preambles received, the eNodeB figures out whether it belongs to two-step or four-step RA procedure.

III. ANALYSIS

In Section III-A we carry out the performance analysis of two-step RA procedure. Time index t means the beginning of RA slot t . The operating regions of the system are analyzed in Section III-B. Section III-C presents an approximate analysis to deal with a large population size as an alternative to the analysis in Section III-A.

A. MARKOVIAN ANALYSIS

Suppose that the system has a total of N devices and one serving eNodeB. Let X_t denote the number of backlogged devices at RA slot t , which is called system state. Then, the number of non-backlogged devices at time t is $N - X_t$. Let \mathcal{S}_t and \mathcal{A}_t be the number of RA preambles successfully (re)transmitted and the devices to have a new packet to send at RA slot t , respectively. In the course of time, X_t evolves as

$$X_{t+1} = X_t - \mathcal{S}_t + \mathcal{A}_t \quad \text{for } \mathcal{S}_t \leq \min(L, X_t). \quad (1)$$

Let us define π_i for $i \in \{0, 1, \dots, N\}$ and $\boldsymbol{\pi} = [\pi_i]$ as the steady-state probability that i backlogged devices are in the system, i.e., $\pi_i = \lim_{t \rightarrow \infty} \Pr[X_t = i]$ and its row probability vector, respectively. Let $q_{n,m}$ for $n, m \in \{0, 1, \dots, N\}$ denote a state transition probability, i.e., $q_{n,m} = \Pr[X_{t+1} = m | X_t = n]$. We can obtain $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = \boldsymbol{\pi} \boldsymbol{Q} \quad \text{and} \quad \sum_{i=0}^N \pi_i = 1, \quad (2)$$

where $\boldsymbol{Q} = [q_{n,m}]$ is the state transition probability matrix whose n -th row and m -th column element is $q_{n,m}$. The following lemma helps to obtain $q_{n,m}$.

Lemma 1: Let S be the number of RA preambles successfully transmitted among a total of L RA preambles; that is, each of them is chosen by only one device. Additionally, X denotes the number of backlogged devices. Given that there are n backlogged devices and each retransmits a RA preamble with probability r , the probability that k RA preambles are

²In 5G NR, a reference signal receive power (RSRP) from the eNodeB can be used for the devices to determine whether to use the two-step RA procedure. We can add an assumption that RSRP is so good that the message transmitted to PUSCH may not have an error.

chosen and transmitted by k individual devices is obtained as

$$\Pr(S = k | X = n) = \sum_{l=0}^{L-k} \frac{(-1)^l \left(\frac{r}{L}\right)^{k+l} n! L!}{l! k! (n-k-l)! (L-k-l)!} \times \left[1 - \frac{r}{L}(k+l)\right]^{n-k-l}. \quad (3)$$

Proof: See Appendix A. ■

It is notable that computational complexity in (3) lies in calculating $n!$. Now we can get $q_{n,m}$ as follows: For $n = 0$, we have

$$q_{0,m} = \binom{N}{m} p^m (1-p)^{N-m}, \quad (4)$$

which is the probability that m devices have a packet to send, i.e., backlogged, when no devices are backlogged. Using Lemma 1, we can write $q_{n,m}$ for $n \geq 1$ as

$$q_{n,m} = \sum_{k=0}^{\min(L,n)} \Pr(S = k | X = n) \times \binom{N-n}{m-n+k} p^{m-n+k} (1-p)^{N-m-k}. \quad (5)$$

This (conditional) probability shows that k devices make a successful RA among n backlogged devices such that the system might have $n - k$ backlogged devices. At the same time, $m - (n - k)$ devices join the backlog newly from $N - n$ nonbacklogged devices. As a result, at the next RA slot, the system shall have m backlogged devices.

The system throughput denoted by τ is the number of RA preambles successfully transmitted per RA slot period, which is equivalent to the number of Msg 3 successfully sent in two-step RA procedure. We can get τ as

$$\tau = \sum_{n=1}^N \sum_{k=0}^{\min(L,n)} k \Pr(S = k | X = n) \pi_n. \quad (6)$$

Let p_s and D be the RA success probability and the random variable of RA delay for a device to experience in terms of RA slots until it makes a successful RA, respectively. We can obtain them as follows.

Proposition 1: The probability mass function of RA delay for GPB algorithm is expressed as

$$\Pr[D = k] = (1 - rp_s)^{k-1} (rp_s), \quad (7)$$

where p_s is obtained as

$$p_s = \frac{\tau}{r \sum_{k=0}^N k \pi_k}. \quad (8)$$

Proof: Since GPB algorithm is employed, the RA delay distribution follows a geometric distribution in (7). To find p_s , let us consider the mean access delay of (7), i.e.,

$$\bar{D} = E[D] = \frac{1}{rp_s}. \quad (9)$$

According to Little's result [32], the mean access delay is also expressed as a ratio of the average number of backlogged devices to the system throughput; that is,

$$\bar{D} = \frac{\sum_{k=0}^N k \pi_k}{\tau}. \quad (10)$$

Equating (9) to (10), one can get (8). Note that the denominator in (8) indicates the average offered load to the system per RA slot. ■

Now, the probability that the access delay would be larger than threshold d , i.e., access delay violation probability subject to a delay constraint d , is obtained as

$$\Pr[D > d] = (1 - rp_s)^d. \quad (11)$$

Thus far we have considered retransmissions based on GPB with probability r . Let us turn to UWB with window size U . For fair comparison we match the mean retransmission interval of geometric distribution to that of the UWB algorithm, i.e., $1/r = U/2$. Therefore, we get $U = \frac{2}{r}$; that is, the window size is twice larger than the mean interval of geometrically distributed retransmissions. The following proposition gets us $\Pr[D > d]$ for UWB algorithm.

Proposition 2: Let ψ_n denote the probability that a backlogged device makes a (re)transmission at the n -th RA slots. We then obtain $\Pr[D > d]$ as

$$\Pr[D > d] = 1 - p_s \sum_{k=1}^d \psi_k, \quad (12)$$

where a product $p_s \psi_k$ indicates the probability that a user makes a successful RA at the k -th RA slots, where p_s in (8) is used with $r = 2/U$. In (12), ψ_n for $n \geq 2$ is recursively obtained as

$$\psi_n = \sum_{i=1}^{n-1} (1 - p_s) \psi_i q_{n-i} + q_n, \quad (13)$$

where q_k is expressed as

$$q_k = \begin{cases} \frac{1}{U}, & \text{if } k \in \{1, 2, \dots, U\}, \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

For $n = 1$, i.e., retransmission just after one RA slot, we have

$$\psi_1 = \frac{1}{U}. \quad (15)$$

Proof: The derivation on (13) is by induction. For example, the first RA attempt takes one RA slot, if the window size randomly picked up at first is one. This takes place with probability $1/U$. We thus have (15). For $n = 2$, the first RA attempt with window size one should fail and the second RA is tried with window size one, which occurs with probability $1/U$. Or, the first RA attempt should have window size two with probability $1/U$. This is recursively applied. ■

Simulation in Section IV shall validate Proposition 2.

One of the striking differences between two-step and four-step RA procedures is the amount of PUSCH resources

used for Msg 3 (re)transmissions. If we assume that one unit of PUSCH resource is used for one RA preamble, the four-step RA procedure saves PUSCH resources proportional to the number of RA preambles not transmitted, i.e., idle. Let us examine the average number of RA preambles not transmitted per RA slot.

Proposition 3: Let $\Pr(\mathbb{I} = i|m)$ denote the probability that i RA preambles are idle given that m devices (re)transmit a RA preamble among L RA preambles. The average number of idle RA preambles per RA slot period is expressed as

$$\bar{\mathbb{I}} = \sum_{n=0}^N \sum_{m=0}^n \sum_{i=0}^L i \Pr(\mathbb{I} = i|m) \binom{n}{m} r^m (1-r)^{n-m} \pi_n, \quad (16)$$

where π_n is given in (2).

Proof: In [31], $\Pr(\mathbb{I} = i|m)$ is obtained as

$$\Pr(\mathbb{I} = i|m) = \binom{L}{i} \sum_{k=0}^{L-i} (-1)^k \binom{L-i}{k} \left(1 - \frac{i+k}{L}\right)^m. \quad (17)$$

Since it takes place with probability π_n that the system has n backlogged devices, we have (16), which completes the proof. ■

Note that the average number of PUSCH resources used in the four-step procedure is expressed as $L - \bar{\mathbb{I}}$. However, if Msg 3 is retransmitted over R times by H-ARQ in the four-step RA procedure, $R(L - \bar{\mathbb{I}} - \tau)$ PUSCH resources are wasted over R RA slot periods on average.

B. BISTABILITY

Bistability of X_t in the two-step RA procedure is characterized as follows: According to the catastrophe theory [13],

the potential function of the system can be defined as $\mathcal{F} : \mathbb{R}^k \times \mathbb{R}^\ell \rightarrow \mathbb{R}$, where k and ℓ are the number of control variables and system states, respectively. If $k = 2$, we have the fold and the cusp catastrophes and $\ell = 1$, i.e., the system state x . To carry out this, let x denote a normalized backlog size for backlog size $n \in [0, N]$; that is, $x \triangleq \frac{n}{N}$ for $x \in [0, 1]$. Let us define a flow balance $\mathcal{F}(x)$ as

$$\mathcal{F}(x) = E[\mathcal{S}_t] - E[\mathcal{A}_t], \quad (18)$$

where $E[\mathcal{S}_t]$ and $E[\mathcal{A}_t]$ denote the average output (or throughput) of the system and the average input rate of new arrivals to the system, respectively (See (1)). Although we can find them from the Markovian analysis, we now need the expressions of $E[\mathcal{S}_t]$ and $E[\mathcal{A}_t]$ in terms of two control variables p and r and one state variable x for the stability analysis below.

In [13], [17] the cusp catastrophe may exist if we can solve $\mathcal{F}(x) = \frac{\partial \mathcal{F}(x)}{\partial x} = \frac{\partial^2 \mathcal{F}(x)}{\partial x^2} = 0$ and $\frac{\partial^k \mathcal{F}(x)}{\partial x^k} \neq 0$ for $k \geq 3$. Under the assumption of its existence, the catastrophe manifold Θ is a surface in three dimensions defined by $\Theta = \{(x, p, r) | \mathcal{F}(x) = 0\}$. Let Θ_B be the fold line consisting of the points of Θ , where the manifold surface folds over. It is defined as

$$\Theta_B = \left\{ (x, p, r) \mid \mathcal{F}(x) = \frac{d\mathcal{F}(x)}{dx} = 0 \right\}. \quad (20)$$

Using $\frac{d^2 \mathcal{F}(x)}{dx^2}$, we can characterize Θ_B as three parts:

$$B^+ = \left\{ (x, p, r) \mid \mathcal{F}(x) = \frac{d\mathcal{F}(x)}{dx} = 0, \frac{d^2 \mathcal{F}(x)}{dx^2} > 0 \right\}, \quad (21)$$

and

$$B^- = \left\{ (x, p, r) \mid \mathcal{F}(x) = \frac{d\mathcal{F}(x)}{dx} = 0, \frac{d^2 \mathcal{F}(x)}{dx^2} < 0 \right\}, \quad (22)$$

$$\begin{aligned} \Pr(S = k) &= \sum_{n=0}^{\infty} \Pr(S = k | X = n) \Pr(X = n) \\ &= \sum_{l=0}^{L-k} \sum_{n=0}^{\infty} (-1)^l \frac{L!}{l!(L-k-l)!} \frac{n!}{k!(n-k-l)!} \left(\frac{r}{L}\right)^{k+l} \left[1 - \frac{r}{L}(k+l)\right]^{n-k-l} \frac{v^n e^{-v}}{n!} \\ &= \sum_{l=0}^{L-k} (-1)^l \frac{L!}{l!k!(L-k-l)!} \left(\frac{vr}{L}\right)^{k+l} e^{-v} \sum_{n=0}^{\infty} \frac{[v(1 - \frac{r}{L}(k+l))]^{n-k-l}}{(n-k-l)!} \\ &= \sum_{l=0}^{L-k} (-1)^l \frac{L!}{l!k!(L-k-l)!} \left(\frac{vr}{L}\right)^{k+l} e^{-v} e^{v(1 - \frac{r}{L}(k+l))} \\ &= \frac{L!}{k!} \left(\frac{vr}{L}\right)^k e^{-\frac{vr}{L}k} \sum_{l=0}^{L-k} \frac{(-1)^l}{l!(L-k-l)!} \left(\frac{vr}{L}\right)^l e^{-\frac{vr}{L}l} \\ &= \frac{L!}{k!(L-k)!} \left(\frac{vr}{L}\right)^k e^{-\frac{vr}{L}k} \sum_{l=0}^{L-k} \frac{(L-k)!}{l!(L-k-l)!} \left(\frac{-vr}{L} e^{-\frac{vr}{L}}\right)^l \\ &= \binom{L}{k} \left(\frac{vr}{L} e^{-\frac{vr}{L}}\right)^k \sum_{l=0}^{L-k} \binom{L-k}{l} \left(\frac{-vr}{L} e^{-\frac{vr}{L}}\right)^l = \binom{L}{k} \left(\frac{vr}{L} e^{-\frac{vr}{L}}\right)^k \left(1 - \frac{vr}{L} e^{-\frac{vr}{L}}\right)^{L-k} \end{aligned} \quad (19)$$

which are called the bifurcation sets. We obtain B^+ and B^- as the projection of the three-dimensional catastrophe manifold (fold line) B^+ and B^- into the control space (p, r) . As a third part, the cusp point, i.e., the locus of the bifurcation sets, is expressed as

$$B_0 = \left\{ (x, p, r) \mid \mathcal{F}(x) = \frac{d\mathcal{F}(x)}{dx} = \frac{d^2\mathcal{F}(x)}{dx^2} = 0 \right\}. \quad (23)$$

To find (21) and (22), we need to get (18) in terms of x . The following lemma helps us find $E[\mathcal{S}_t]$.

Lemma 2: Let us assume that the number of backlogged devices follows a Poisson distribution with mean $\nu = \sum_{k=0}^N k\pi_k$ (devices/RA slot period). The probability that k RA preambles are successfully transmitted is obtained as

$$\Pr(S = k) = \binom{L}{k} \left(\frac{\nu r}{L} e^{-\frac{\nu r}{L}} \right)^k \left(1 - \frac{\nu r}{L} e^{-\frac{\nu r}{L}} \right)^{L-k}. \quad (24)$$

Proof: Let us assume that the number of backlogged devices X follows a Poisson process, i.e.,

$$\Pr(X = n) = \frac{\nu^n e^{-\nu}}{n!}. \quad (25)$$

We shall see in Section IV that this Poisson assumption with mean ν for the number of backlogged devices is valid when the system is in stable region, which is our utmost interest in practice. We can get (24) as in (19), as shown at the bottom of the previous page, where we have used $\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$ and $(a+b)^n = \sum_{i=0}^n \binom{n}{i} a^i b^{n-i}$. ■

As an alternative proof for Lemma 2, we can make use of splitting and merging properties of Poisson process [32]: Since each backlogged device chooses a RA preamble independently upon retransmission with probability $1/L$, we can find a Poisson process with mean $\nu r/L$ for each RA preamble. In (24), $\frac{\nu r}{L} e^{-\frac{\nu r}{L}}$ is the probability that only one device chooses and transmits a specific RA preamble. This also completes the proof of Lemma 2.

Going back to (24), letting $G = r\nu$, we can find $E[\mathcal{S}_t]$ as

$$E[\mathcal{S}_t] = \sum_{k=0}^L k \Pr(S = k) = Ge^{-\frac{G}{L}}. \quad (26)$$

Additionally, $E[\mathcal{A}_t]$ is the expectation of binomial process with parameter p . We can express $E[\mathcal{A}_t]$ as

$$E[\mathcal{A}_t] = (N - \nu)p = \left(N - \frac{G}{r} \right) p \quad (27)$$

with $x = \frac{\nu}{N}$ and $G = r\nu$. Then, $\mathcal{F}(x)$ is obtained as

$$\mathcal{F}(x) = E[\mathcal{S}_t] - E[\mathcal{A}_t] = Ge^{-\frac{G}{L}} - \left(N - \frac{G}{r} \right) p, \quad (28)$$

where $G = rNx$.

Based on (20)-(22), i.e., $\mathcal{F}(x) = \frac{\partial \mathcal{F}(x)}{\partial x} = \frac{\partial^2 \mathcal{F}(x)}{\partial x^2} = 0$, the following lemmas characterize p and r for B^+ and B^- .

Lemma 3: In terms of x and r , the control variable p subject to (20) is expressed as

$$p = -re^{-\frac{G}{L}} \left(1 - \frac{G}{L} \right). \quad (29)$$

Proof: We get $\frac{\partial \mathcal{F}(x)}{\partial x}$ as

$$\frac{\partial \mathcal{F}(x)}{\partial x} = \frac{\partial G}{\partial x} \frac{\partial \mathcal{F}}{\partial G} = rN \left[\left(1 - \frac{G}{L} \right) e^{-\frac{G}{L}} + \frac{p}{r} \right], \quad (30)$$

where $\frac{\partial G}{\partial x} = rN$. Then, setting $\frac{\partial \mathcal{F}(x)}{\partial x} = 0$ and solving it with respect to p , we get (29). ■

Lemma 4: Let x_{\pm} denote x corresponding to B^+ and B^- , respectively. They are found as

$$x_{\pm} = \frac{rN \pm \sqrt{(rN)^2 - 4rNL}}{2rN}. \quad (31)$$

Proof: Based on Lemma 3, we plug (29) into (28) so that $\mathcal{F}(x)$ can be expressed as

$$\begin{aligned} \mathcal{F}(x) &= \frac{1}{L} e^{-\frac{G}{L}} \left(G^2 - rNG + rNL \right) \\ &= \frac{rN}{L} e^{-\frac{rN}{L}x} \left(rNx^2 - rNx + L \right). \end{aligned} \quad (32)$$

We find G as the roots of $\mathcal{F}(x) = 0$; that is, the roots of the quadratic equation $G^2 - rNG + rNL$. If G_{\pm} denotes the roots, we find that

$$G_{\pm} = \frac{rN \pm \sqrt{(rN)^2 - 4rNL}}{2}. \quad (33)$$

Since $G_{\pm} = rNx_{\pm}$, we get (31). This completes the proof. ■

Lemma 5: Let x^* , p^* , and r^* be the cusp point in (23), which can be obtained as

$$x^* = \frac{2L}{Nr^*} = \frac{1}{2}, \quad p^* = \frac{4L}{N} e^{-2}, \quad r^* = \frac{4L}{N}. \quad (34)$$

Proof: From (23), we have

$$\frac{\partial^2 \mathcal{F}}{\partial x^2} = -N \left(\frac{r}{L} \right)^2 e^{-\frac{rNx}{L}} (2L - rNx) = 0. \quad (35)$$

From $\frac{\partial^2 \mathcal{F}}{\partial x^2} = 0$ we get $x^* = \frac{2L}{Nr^*}$. In fact, this x^* is the point, at which x_+ meets x_- . Setting $x_+ = x_-$ in (31), we have $r^* = 4L/N$. Using (29), we find that $p^* = \frac{4L}{N} e^{-2}$. ■

It is notable that we should have that $(rN)^2 - 4rNL \geq 0$ for x (or G) to be nonnegative real in (31), which means that $r \geq 4L/N$. Since x_{\pm} (or G_{\pm}) is a function of r , we can get x_{\pm} for $r \in [4L/N, 1]$ corresponding to B^+ and B^- . Thus, bistability does not occur when $r < 4L/N$. Finally, higher-order derivatives of $\mathcal{F}(x)$ for $k \geq 3$ we have

$$\frac{\partial^k \mathcal{F}}{\partial x^k} = (-1)^{k-1} \left(\frac{rN}{L} \right)^k e^{-\frac{rNx}{L}} (kL - rNx). \quad (36)$$

C. APPROXIMATION FOR LARGE POPULATION SIZE

As N grows large, Markovian analysis in Section III-A suffers from computational burden in calculating $n!$, e.g., $N \geq 170$. In such cases we can use the following approximation: As $t \rightarrow \infty$, the system is in equilibrium such that in (1) we can see $\mathbb{E}[X_t] = c$ for $c \in (0, N)$. Furthermore, from $\mathbb{E}[X_{t+1}] = \mathbb{E}[X_t - \mathcal{S}_t + \mathcal{A}_t]$ and $E[X_{t+1}] = E[X_t]$ for $t \rightarrow \infty$, we should have $\mathbb{E}[\mathcal{S}_t] = \mathbb{E}[\mathcal{A}_t]$. This implies that $\mathcal{F}(x) = 0$ in (18). Let x^* be a single root of $\mathcal{F}(x) = 0$ for an unsaturated stable system; that is, we have

$$\mathcal{F}(x) = 0 \Rightarrow rxe^{-\frac{rNx}{L}} - p(1 - x) = 0. \quad (37)$$

It is not difficult to see that the root of (37) shows $x^* \approx \frac{\nu}{N} = \frac{1}{N} \sum_{k=0}^N k\pi_k$. This approximation is accurate especially when π_k follows a Poisson distribution with mean $\nu = \sum_{k=0}^N k\pi_k$. Furthermore, based on the argument in the alternative proof of Lemma 2, we can find p_s in (7) for a large N . Suppose that a *tagged* device transmits a RA preamble. For it to make a successful RA, no device shall transmit the RA preamble. Thus, we obtain

$$p_s = e^{-\frac{\nu r}{L}} = e^{-r \frac{Nx^*}{L}}. \quad (38)$$

From (8) and (38), we also have

$$e^{-\frac{\nu r}{L}} = \frac{\tau}{r \sum_{k=0}^N k\pi_k} \Rightarrow r \nu e^{-\frac{\nu r}{L}} = \tau. \quad (39)$$

For a large N , we can approximate RA delay distribution using (38) together with (11), or (12). Note that $\mathcal{F}(0) = -p < 0$ and $\mathcal{F}(1) = r e^{-\frac{rN}{L}} > 0$. Therefore, if $\mathcal{F}(x)$ is an increasing function of x , i.e.,

$$\frac{\partial \mathcal{F}(x)}{\partial x} = p + r \left(1 - \frac{rNx}{L}\right) e^{-\frac{rNx}{L}} > 0, \quad (40)$$

then, (37) has a unique solution $x^* \in (0, 1)$. Additionally, one can get x^* as a fixed point iteration:

$$x = f(x) = \frac{p}{r} (1 - x) e^{\frac{rNx}{L}}. \quad (41)$$

Convergence of this fixed point iteration can be guaranteed if $|f'(x)| < 1$. This condition is identical to (40). It is notable that from $|f'(0)| = \frac{p}{r} (1 - \frac{rN}{L})$ and $|f'(1)| = \frac{p}{r} e^{\frac{rN}{L}}$, we have a necessary condition for (41) to converge as $p \ll r$.

D. STABILIZING THE SYSTEM

Let us consider how to maximize $E[S_i]$ and its impact on the system bistability. When maximizing $E[S_i]$ in (26) with respect to r using $\frac{dE[S_i]}{dr} = 0$, we can find $r = \frac{L}{\nu}$. This implies that the system throughput can be maximized when a retransmission probability is controlled in the course of time as $r = \frac{L}{\nu}$. To do this, it is necessary to estimate ν , i.e., the (mean) number of backlogged devices. It can be also found by using $\frac{\partial \mathcal{F}(x)}{\partial r} = 0$, which yields $r = \frac{L}{xN}$. Notice that $\nu = xN$. Plugging this r into $\mathcal{F}(x)$ we have

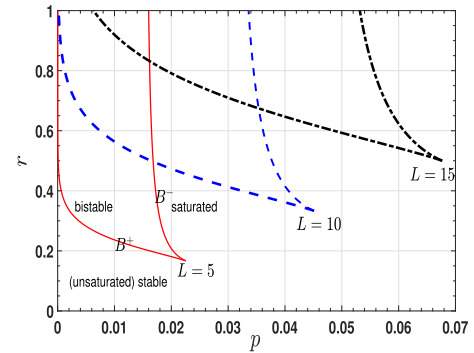
$$\mathcal{F}(x) = L e^{-\frac{1}{L}} - N(1-x)p, \quad (42)$$

which is a linear function of x . In steady-state, the system has the average normalized backlog size $x^* = 1 - \frac{L}{Np} e^{-\frac{1}{L}}$. This implies that the system keeps its average backlogged device to Nx^* over time.

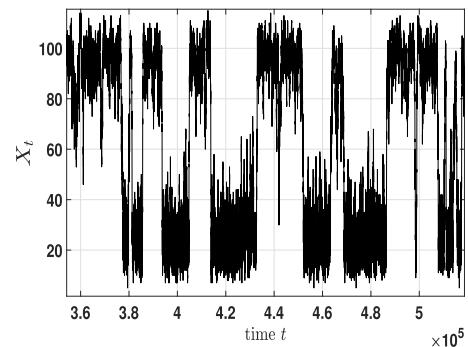
IV. NUMERICAL STUDIES

In the figures presented in this section, symbols show simulation results, while lines depict analytical results. We set simulation run length to 10^6 RA slot periods and get the time-averaged results.

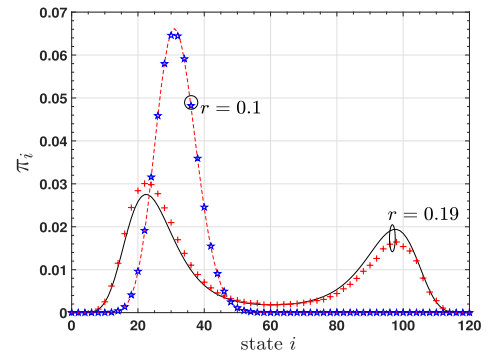
In Figs. 2(a)–2(d), we first examine bistability of the system for $N = 120$. In Fig. 2(a), (unsaturated) stable, bistable and saturated regions are specified for $L = 5$. As the number



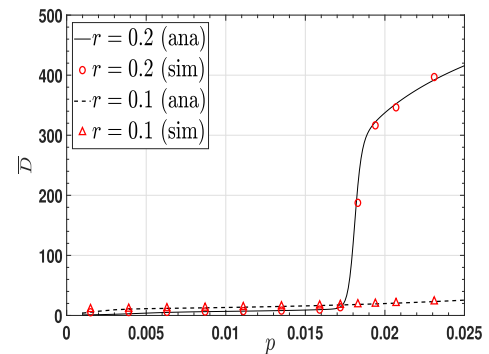
(a) Bifurcation sets



(b) Sample path of backlog size in bistable system



(c) State probability distribution with $L = 5$



(d) Mean access delay with $L = 5$

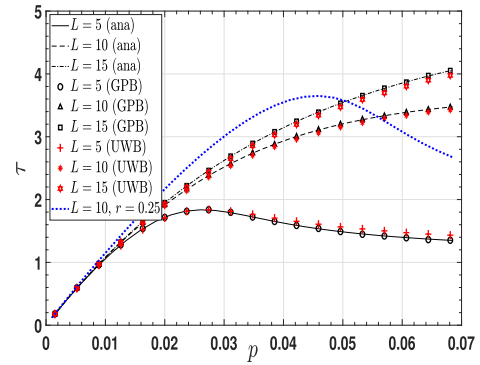
FIGURE 2. Bistability of two-step LTE-A RA procedure: $N = 120$.

of RA preambles L allocated for two-step RA procedure increases, the stable region becomes larger. Thus, larger p (higher packet generation) and r (aggressive retransmission)

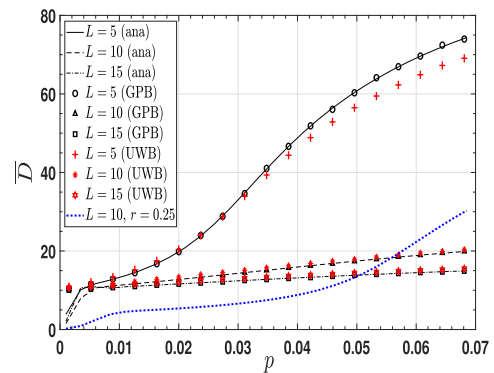
can be used for the system. In Fig. 2(b), a sample path of backlog size X_t is depicted for $r = 0.19$, and $p = 0.019$. This pair of p and r is inside the bistable region in Fig. 2(a). It can be seen that X_t switches back and forth from 100 to 20 over time. In Fig. 2(c), π_k 's are depicted for stable ($r = 0.1$) and bistable systems ($r = 0.19$), and $p = 0.019$. In the bistable system with $r = 0.19$, it can be observed that the state probability distribution takes a bimodal shape; that is, two peaks are found around state 20 and 100. This corresponds to the alternating behavior of X_t from 100 to 20 in Fig. 2(b).

When r is reduced from 0.19 to 0.1, (farther away from the bistable region in Fig. 2(a)), π_k becomes a unimodal distribution. It is notable that the unimodal distribution with $r = 0.1$ is very close to a Poisson distribution with mean $\sum_{k=0}^N k\pi_k = 31.6909$. In general, it holds true that π_k takes a shape very close to a Poisson distribution, if p and r are chosen inside the unsaturated stable region. What is interesting to note is that as r is raised further from 0.19, the first peak found around state 20 becomes lower, while the second peak around state 100 gets higher and moves to a higher state. Eventually, even in the bistable region, the system behaves like the one in saturated region with a very low throughput and large access delay. It should be noted that the switching behavior of X_t from a low to a high state can be found when we pick up the parameters p and r around the boundary line between unsaturated stable and bistable regions. It can be said that for a given p , if too large r (ill-designed backoff algorithm) is chosen inside the bistable region, the system becomes bistable. On the other hand, if p increases (excessive traffic load) for a given $r = 0.2$ as in Fig. 2(a), the system gets into the bistable region around $p = 0.0162$. When the system does, the mean access delay suddenly jumps up as depicted in Fig. 2(d). Although not presented here, the throughput also drops abruptly as sudden jump in \bar{D} . When a small r is chosen such that increasing p does not let the system pass the bistable region in Fig. 2(a), e.g., $r = 0.1$, \bar{D} does not have such sudden jumps in Fig. 2(d). Thus, as a way of avoiding the bistability region, a traffic-adaptive backoff algorithm should be employed so that a retransmission probability reduces progressively to prevent the bistability.

In Figs. 3(a) and 3(b), the throughput τ and mean access delay of unsaturated stable systems are examined for $N = 120$. We use GPB with $r = 0.1$ and compare the simulation results of UWB with window size $U = 2/r = 20$. Even though the distributions of backoff intervals are different in GPB and UWB, it can be seen that the analysis agrees well with simulation results and simulation results of two backoff algorithms are very close to each other. It can be concluded that when independent and identically distributed backoff intervals are used upon retransmissions, the performance may depend only on the mean of backoff intervals, not much on the distribution itself. It is notable that for $L = 5$, the retransmission probability $r = 0.1$ can make the system saturated around $p^* = (4L/N)e^{-2} = 0.0226$. It occurs for the system with $L = 10$, referring to Fig. 2(a) around a much higher r , e.g., $r = 0.25$, such that the system becomes



(a) Throughput



(b) Average access delay

FIGURE 3. Two-step LTE-A RA procedure with $N = 120$.

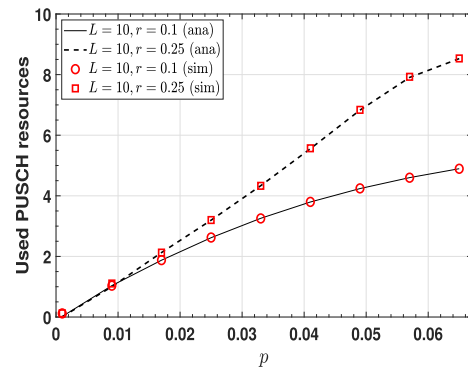
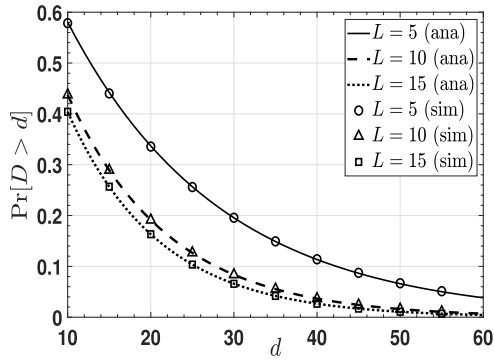


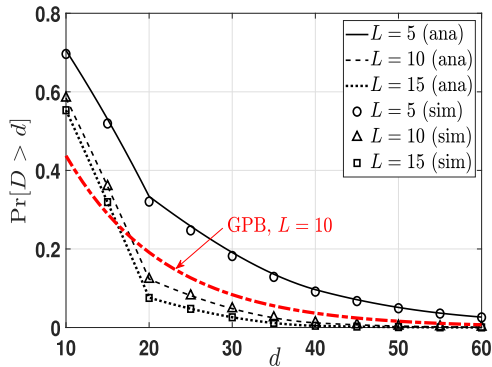
FIGURE 4. Used PUSCH resources per RA slot with $N = 120$ and $L = 10$.

saturated around $p^* = 0.0452$. If the system gets saturated without passing through the bistable region, the performance degradation does not show a sudden collapse.

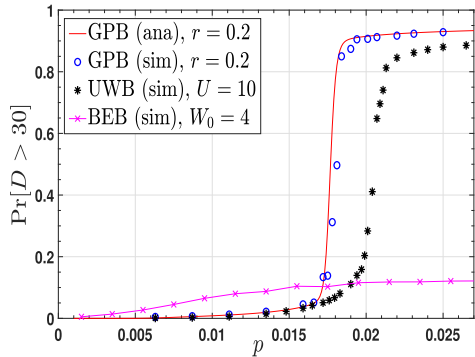
Fig. 4 depicts the average of PUSCH resources used over one RA slot in the four-step RA procedure without H-ARQ retransmission. We assume that one unit of PUSCH resource is allocated for Msg 3 transmission. In the two-step RA procedure, the PUSCH resource used is always L regardless of p . However, in the four-step procedure, PUSCH resources are less used for low p . If a higher r is used, more PUSCH resources are used. In other words, \bar{D} drops rapidly in the four-step procedure.



(a) Delay violation probability with GPB



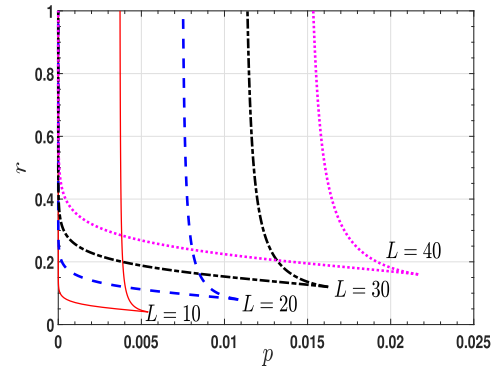
(b) Delay violation probability with UWB



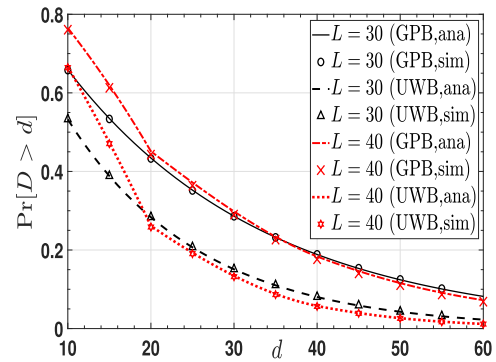
(c) Delay violation probability of bistable system

FIGURE 5. Access delay violation with $N = 120$.

Figs. 5(a) and 5(b) depict the access delay violation (probability) for GPB $r = 0.1$ and UWB $U = 20$, respectively. We set $p = 0.019$ and $L = 5$. The system is in unsaturated stable region with these parameters. Good agreements can be observed between analysis and simulation results. As expected, the more RA preambles are used, the lower the delay violation probability given d in both figures. Compared to Fig. 5(a), the delay violation probability with UWB drops remarkably at $d = 20$ in Fig. 5(b). Moreover, the violation probability of UWB is higher than that of GPB when d is less than some d^* , whereas it becomes lower for $d \geq d^*$. Additionally, Fig. 5(c) depicts $\Pr[D > d]$ for the system with GPB $r = 0.2$ and UWB $U = 10$. As p increases, the system is driven from unsaturated stable to bistable region



(a) Bifurcation sets



(b) Delay violation probability with GPB

FIGURE 6. Two-step LTE-A RA procedure with $N = 1000$.

as in Fig. 2(d). As we have seen in the mean access delay, sudden performance collapse in $\Pr[D > 30]$ is also observed; that is, as a larger p drives the system bistable, the violation probability jumps from around 0.05 to 0.9. Until then, the analysis and simulation for GPB agree with each other well and even with simulation for UWB. The access delay violation of UWB takes a high jump around $p = 0.0175$. Note that the bistability region in Fig. 2(a) is obtained from the system with GPB. Accordingly, we do not have the information when the system with UWB goes bistable in terms of p . However, from Fig. 5(c) it can be conjectured that the bistable region of UWB can be found at a higher p than that of GPB. Even though the mean RA access delay is quite insensitive to GPB and UWB, it seems that the operating region is sensitive to the distribution of backoff interval. Since BEB algorithm progressively reduces its retransmission probability by increasing its window size, we compare it with GPB and UWB in Fig. 5(c) with simulation. Upon the k -th retransmission, devices pick up a random integer $B \in [1, W_0 2^k]$ and $W_0 = 4$. Even with low p , the delay violation probability of BEB is higher than GPB and UWB, since it can use a larger window size with small collisions. However, it can be conjectured that BEB can alleviate the onset of bistability that can take place for GPB and UWB.

In Figs. 6(a) and 6(b), we examine the systems with a large population size $N = 1000$, $p = 0.015$, and $r = 0.1$.

Compared to the case with $L = 10$ in Fig. 2(a), as N increases, the unsaturated stable region gets shrunk. Even with much large L , the stable region defined by p and r is small. Consequently, the number of devices allowed for two-step RA procedure should be carefully limited for the number of RA preambles used. Notice that when $L = 15$ in Fig. 2(a), $r = 1$ can be used for small p . This can not be allowed in Fig. 6(a). In Fig. 6(b) the access delay violation probability is presented for unsaturated stable system. Instead of using Markovian analysis, we use the approximation presented in Section III-C. Our approximation seems quite accurate for GPB and UWB. This shows that a proper dimensioning on L is essential in tuning the two-step RA procedure for a large population size.

V. CONCLUSION

This paper has investigated the performance of two-step LTE-A RA procedure for URLLC and discussed its difference from the existing four-step procedure with respect to PUSCH resources used. In particular, we have characterized three operating regions of this two-step RA procedure such as unsaturated stable, bistable, and saturated one in terms of a packet generation probability, retransmission probability, population size, and the number of RA preambles. As a result, it would expect that the desired stable region can be realized in practice with a proper system dimensioning, e.g., the number of RA preambles. In addition, taking into account two backoff algorithms such as GPB and UWB, we have derived the RA access delay distribution, which is essential to study of guaranteeing low latency in the two-step RA procedure. Moreover, in order to overcome computational burden in Markovian analysis for a large population size, we presented an approximation method and discussed its convergence. Under the unsaturated stable region, it can be concluded that our approximation seems quite accurate and the access delay performance depends mainly on the mean of backoff interval, not its distribution.

As future work, we are interested in the following: First, since it is shown that the bistability is eliminated by retransmission probability r that can keep track of the number of backlogged devices over time, i.e., $r = L/\nu$, it may be important to develop a control algorithm in LTE-A as well as 5G NR systems, whose physical layer is based on millimeter wave (mmWave). The control algorithm would be based on Bayesian online learning so that burstiness of traffic can be detected and alleviated to maximize the resource utilization according to channel outcomes. In addition, we assumed that the message transmitted to PUSCH may not have an error. However, it could be relaxed to examine how erroneous transmissions of the message for PUSCH affect the overall performance.

**APPENDIX A
PROOF OF LEMMA 1**

Let ϕ_l denote the probability that l distinct RA preambles are selected by only one device out of m devices. These l devices

can be selected in $\binom{m}{l}$ ways, whereas they seize RA preambles in $l!$ ways. Thus, the remaining $m - l$ devices are distributed among $L - l$ RA preambles in $(L - l)^{m-l}$ ways. Dividing these multiplications by L^m , i.e., a total number of ways for m devices to choose L RA preambles, we can write ϕ_l as

$$\phi_l = \binom{m}{l} \frac{l!}{L^m} (L - l)^{m-l} = \binom{m}{l} \frac{l!}{L^l} \left(1 - \frac{l}{L}\right)^{m-l} \tag{43}$$

Let S_l denote the probability that any l RA preambles are selected by only one device. From (43), S_l is obtained as

$$S_l = \binom{L}{l} \pi_l = \binom{L}{l} \binom{m}{l} \frac{l!}{L^l} \left(1 - \frac{l}{L}\right)^{m-l} \tag{44}$$

Let \mathbb{S}_j be the event that RA preamble j is chosen by only one device. According to the inclusion-exclusion procedure, the probability that at least one RA preamble is selected by only one device is expressed as

$$\Pr\left(\bigcup_{j=1}^L \mathbb{S}_j\right) = \sum_{l=1}^L (-1)^{l+1} S_l = \sum_{l=1}^L (-1)^{l+1} \binom{L}{l} \binom{m}{l} \frac{l!}{L^l} \left(1 - \frac{l}{L}\right)^{m-l} \tag{45}$$

Let $\Theta(m, L)$ denote the probability that each RA preamble is selected by more than only one device or by none. This is the complementary event to the event $\bigcup_{j=1}^L \mathbb{S}_j$ and hence its probability can be written as

$$\begin{aligned} \Theta(m, L) &= 1 - \Pr\left(\bigcup_{j=1}^L \mathbb{S}_j\right) \\ &= 1 - \sum_{l=1}^L (-1)^{l+1} \binom{L}{l} \binom{m}{l} \frac{l!}{L^l} \left(1 - \frac{l}{L}\right)^{m-l} \\ &= \sum_{l=0}^L (-1)^l \binom{L}{l} \binom{m}{l} \frac{l!}{L^l} \left(1 - \frac{l}{L}\right)^{m-l} \end{aligned} \tag{46}$$

The probability that there are k successful devices out of m , denoted by $\Phi(S = k|m)$, can be derived as follows. There are $\binom{L}{k}$ ways of choosing k RA preambles from L and we also have $\binom{m}{k}$ ways of choosing k successful devices out of m . The devices can be placed in k successful RA preambles in $k!$ ways. The remaining $m - k$ devices are distributed among $L - k$ RA preambles and this can be done in $(L - k)^{m-k}$ $\Pr(m - k, L - k)$. Dividing all these multiplications by L^m , we get

$$\Phi(S = k|m) = \binom{L}{k} \binom{m}{k} \frac{k! (L - k)^{m-k}}{L^m} \Theta(m - k, L - k) \tag{47}$$

For notational simplicity, let $\Psi(m, n, x)$ denote a binomial distribution, i.e., $\Psi(m, n, x) = \binom{n}{m} x^m (1 - x)^{n-m}$. Now the probability that there are k successful devices given that

TABLE 1. RACH location and PRACH configuration indice in LTE systems.

PRACH configuration index	Preamble format	System frame number	Subframe number	PRACH configuration index	Preamble format	System frame number	Subframe number
0	0	Even	1	32	2	Even	1
1	0	Even	4	33	2	Even	4
2	0	Even	7	34	2	Even	7
3	0	Any	1	35	2	Any	1
4	0	Any	4	36	2	Any	4
5	0	Any	7	37	2	Any	7
6	0	Any	1,6	38	2	Any	1,6
7	0	Any	2,7	39	2	Any	2,7
8	0	Any	3,8	40	2	Any	3,8
9	0	Any	1,4,7	41	2	Any	1,4,7
10	0	Any	2,5,8	42	2	Any	2,5,8
11	0	Any	3,6,9	43	2	Any	3,6,9
12	0	Any	0,2,4,6,8	44	2	Any	0,2,4,6,8
13	0	Any	1,3,5,7,9	45	2	Any	1,3,5,7,9
14	0	Any	0,1,2,3,4,5,6,7,8,9	46	N/A	N/A	N/A
15	0	Even	9	47	2	Even	9
16	1	Even	1	48	3	Even	1
17	1	Even	4	49	3	Even	4
18	1	Even	7	50	3	Even	7
19	1	Any	1	51	3	Any	1
20	1	Any	4	52	3	Any	4
21	1	Any	7	53	3	Any	7
22	1	Any	1,6	54	3	Any	1,6
23	1	Any	2,7	55	3	Any	2,7
24	1	Any	3,8	56	3	Any	3,8
25	1	Any	1,4,7	57	3	Any	1,4,7
26	1	Any	2,5,8	58	3	Any	2,5,8
27	1	Any	3,6,9	59	3	Any	3,6,9
28	1	Any	0,2,4,6,8	60	N/A	N/A	N/A
29	1	Any	1,3,5,7,9	61	N/A	N/A	N/A
30	N/A	N/A	N/A	62	N/A	N/A	N/A
31	1	Even	9	63	3	Even	9

the system, i.e., $\Pr(S = k|n)$, has n backlogged devices, is expressed as

$$\begin{aligned}
 &\Pr(S = k|X = n) \\
 &= \sum_{m=0}^n \Phi(S = k|m) \Psi(m, n, p) \\
 &= \sum_{m=0}^n \Psi(m, n, p) \binom{L}{k} \binom{m}{k} \frac{k! (L-k)^{m-k}}{L^m} \Theta(m-k, L-k).
 \end{aligned} \tag{48}$$

We can rewrite (48) as

$$\begin{aligned}
 &\Pr(S = k|X = n) \\
 &= \sum_{m=0}^n \Psi(m, n, r) \binom{L}{k} \binom{m}{k} \frac{k! (L-k)^{m-k}}{L^m} \\
 &\quad \times \sum_{l=0}^{L-k} (-1)^l \binom{L-k}{l} \binom{m-k}{l} \frac{l!}{(L-k)^l} \left(1 - \frac{l}{L-k}\right)^{m-k-l} \\
 &= \sum_{m=0}^n \sum_{l=0}^{L-k} \frac{n! r^m (1-r)^{n-m}}{(n-m)!} \frac{L! (-1)^l (L-k-l)^{m-k-l}}{k! L^m (L-k-l)! (m-k-l)!}
 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{l=0}^{L-k} \frac{(-1)^l L! n!}{k! l! (L-k-l)!} \sum_{m=0}^n \frac{r^m (1-r)^{n-m} (L-k-l)^{m-k-l}}{(n-m)! L^m (m-k-l)!} \\
 &= \sum_{l=0}^{L-k} \frac{(-1)^l L! n!}{k! l! (L-k-l)! (n-k-l)!} \\
 &\quad \times \sum_{m=0}^n \frac{(n-k-l)! (1-r)^{n-m}}{(n-m)! (m-k-l)!} \left(\frac{r}{L}\right)^m (L-k-l)^{m-k-l} \\
 &= \sum_{l=0}^{L-k} \frac{(-1)^l L! n! \left(\frac{r}{L}\right)^{k+l}}{k! l! (L-k-l)! (n-k-l)!} \\
 &\quad \times \sum_{m=0}^n \binom{n-k-l}{m-k-l} \left[\frac{r}{L} (L-k-l)\right]^{m-k-l} (1-r)^{n-m} \\
 &= \sum_{l=0}^{L-k} \frac{(-1)^l L!}{k! l! (L-k-l)!} \frac{n! \left(\frac{r}{L}\right)^{k+l}}{(n-k-l)!} \left[1 - \frac{r}{L} (k+l)\right]^{n-k-l} \\
 &= \sum_{l=0}^{L-k} (-1)^l \left(\frac{r}{L}\right)^{k+l} k! \binom{n}{k} \binom{L}{k} \binom{L-k}{l} \\
 &\quad \times \binom{n-k}{l} l! \left[1 - \frac{r(k+l)}{L}\right]^{n-k-l}.
 \end{aligned} \tag{49}$$

APPENDIX B RACH LOCATION

According to PRACH configuration index in LTE-A, RACH can be found in subframe(s) as summarized in Table I. Note that N/A stands for not available and there are four types of RA preamble formats in LTE, among which two are a short one and the other two a long one. The long one concatenates two copies of a RA preamble.

REFERENCES

- [1] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Pusmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.
- [2] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.
- [3] L. Tello-Quendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3505–3520, Apr. 2018.
- [4] H. Seung Jang, H. Jin, B. Chul Jung, and T. Q. S. Quek, "Versatile access control for massive IoT: Throughput, latency, and energy efficiency," *IEEE Trans. Mobile Comput.*, vol. 19, no. 8, pp. 1984–1997, Aug. 2020.
- [5] C. Di, B. Zhang, Q. Liang, S. Li, and Y. Guo, "Learning automata-based access class barring scheme for massive random access in machine-to-machine communications," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6007–6017, Aug. 2019.
- [6] *Reply Ls on NR RACH Procedure*, Standard 3GPP TSG-RAN2, Reno, USA, NV, Nov. 2016.
- [7] *Physical Channel Design for 2-Step RACH*, Standard 3GPP TSG RAN WG1 Meeting NR, Spokane, WA, USA, Jan. 2017.
- [8] *Discussion on 2 Steps RACH Procedure*, Standard 3GPP TSG RAN WG1 Meeting NR, Spokane, WA, USA, Jan. 2017.
- [9] J. Thota and A. Aijaz, "On performance evaluation of random access enhancements for 5G URLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–7.
- [10] *5G; NR; Medium Access Control Protocol Specification*, Standard 3GPP TS 38.321 Version 16.1.0 Release 16, 2017.
- [11] *5G; NR; Physical Layer Procedures for Control*, Standard 3GPP TS 38.213 Version 16.0.0 Release 16., 2018.
- [12] A. Carleial and M. Hellman, "Bistable Behavior of ALOHA-type systems," *IEEE Trans. Commun.*, vol. COM-23, no. 4, pp. 401–410, Apr. 1975.
- [13] Y. Onozato and S. Noguchi, "On the thrashing cusp in slotted aloha systems," *IEEE Trans. Commun.*, vol. COM-33, no. 11, pp. 1171–1182, Nov. 1985.
- [14] Y. Onozato, J. Liu, S. Shimamoto, and S. Noguchi, "Effect of propagation delays on ALOHA systems," *Comput. Netw. ISDN Syst.*, vol. 12, no. 5, pp. 329–337, Jan. 1986.
- [15] Y. Onozato, J. Liu, and S. Noguchi, "Stability of a slotted ALOHA system with capture effect," *IEEE Trans. Veh. Technol.*, vol. 38, no. 1, pp. 31–36, Feb. 1989.
- [16] K. Sakakibara, M. Hanaoka, and Y. Yuba, "On the stability of five types of slotted ALOHA systems with capture and multiple packet reception," *IEICE Trans. Fundam.*, vol. E81-A, no. 10, Oct, pp. 2092–2100, 1998.
- [17] K. Sakakibara, H. Muta, and Y. Yuba, "The effect of limiting the number of retransmission trials on the stability of slotted ALOHA systems," *IEEE Trans. Veh. Technol.*, vol. 49, no. 4, pp. 1449–1453, Jul. 2000.
- [18] R. R. Tyagi, F. Aurzada, K.-D. Lee, S. G. Kim, and M. Reisslein, "Impact of retransmission limit on preamble contention in LTE-advanced network," *IEEE Syst. J.*, vol. 9, no. 3, pp. 752–765, Sep. 2015.
- [19] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in LTE networks: Modeling and throughput optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771–2785, Apr. 2018.
- [20] Y. Yang and T.-S. Peter Yum, "Delay distributions of slotted ALOHA and CSMA," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1846–1857, Nov. 2003.
- [21] L. Barletta, F. Borgonovo, and I. Filippini, "The throughput and access delay of slotted-aloha with exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 451–464, Feb. 2018.
- [22] J.-B. Seo and V. C. M. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.
- [23] A. Mutairi, S. Roy, and G. Hwang, "Delay analysis of OFDMA-aloha," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 89–99, Jan. 2013.
- [24] M. Koseoglu, "Lower bounds on the LTE—A average random access delay under massive M2M arrivals," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2104–2115, May 2016.
- [25] C.-H. Wei, G. Bianchi, and R.-G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.
- [26] X. Jian, Y. Liu, Y. Wei, X. Zeng, and X. Tan, "Random access delay distribution of multichannel slotted ALOHA with its applications for machine type communications," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 21–28, Feb. 2017.
- [27] J. Choi, "On fast retrieval for two-step random access in MTC," *IEEE Internet Things J.*, early access, Jul. 28, 2020, doi: [10.1109/JIOT.2020.3012449](https://doi.org/10.1109/JIOT.2020.3012449).
- [28] M. Centenaro, L. Vangelista, and S. Saur, "Analysis of 5G radio access protocols for uplink URLLC in a connection-less mode," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3104–3117, May 2020.
- [29] *5G; NR; Physical Channels and Modulation*, Standard 3GPP TS 38.211 Version 16.2.0 Release 16, 2018.
- [30] J.-B. Seo and V. Leung, "Queuing performance of multichannel S-ALOHA systems with correlated arrivals," *IEEE Trans. Veh. Technol.*, vol. 60, no. 9, pp. 4574–4586, Nov. 2011.
- [31] H. Jin, W. Tariq Toor, B. Chul Jung, and J.-B. Seo, "Recursive pseudo-Bayesian access class barring for M2M communications in LTE systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, Sep. 2017.
- [32] H. C. Tijms, *A First Course in Stochastic Models*. Hoboken, NJ, USA: Wiley, Mar. 2003.



JUN-BAE SEO (Member, IEEE) received the B.S. and M.Sc. degrees in electrical engineering from Korea University, South Korea, in 2000 and 2003, respectively, and the Ph.D. degree from The University of British Columbia (UBC), Vancouver, BC, Canada, in 2012. From 2003 to 2006, he was a Member of the Research Staff with the Electronics and Telecommunications Research Institute, Daejeon, South Korea, carrying out research on IEEE 802.16 systems. He was a Postdoctoral

Fellow with UBC until 2014. From 2015 to August 2019, he was an Assistant Professor with IIT Delhi, New Delhi, India. From September 2019 to August 2020, he was a Research Professor with Hanyang University, ERICA. Since September 2020, he has been with the Department of Information and Communication Engineering, Gyeongsang National University, South Korea, where he is currently is an Associate Professor. His research interests include stochastic modeling and optimizing queuing systems with applications to wireless mobile and computer communications networks.



WAQAS TARIQ TOOR received the B.S. degree in electrical engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, in 2007, the M.S. degree in electrical engineering from the University of Management and Technology (UMT), Lahore, in 2011, and the Ph.D. degree in electronics and communication engineering from Hanyang University, Ansan, South Korea, in 2018. From 2009 to 2014, he has worked as a Faculty Member with UMT Lahore. He is currently working as an Assistant Professor with the Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan. His research interests include medium-access control for random access networks, scheduling systems, machine-type communications (MTCs), the Internet of Things (IoT), and non-orthogonal multiple access (NOMA).



HU JIN (Senior Member, IEEE) received the B.E. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2006 and 2011, respectively. From 2011 to 2013, he was a Postdoctoral Fellow with The University of British Columbia, Vancouver, BC, Canada. From 2013 to 2014, he was a Research Professor with Gyeongsang National University, Tongyeong, South Korea. Since 2014, he has been with the Division of Electrical Engineering, Hanyang University, Ansan, South Korea, where he is currently is an Associate Professor. His research interests include medium-access control and radio resource management for random access networks and scheduling systems considering advanced signal processing and queuing performance.

• • •