Priority-based Initial Access for URLLC Traffic in Massive IoT Networks: Schemes and Performance Analysis

Thilina N. Weerasinghe, Indika A. M. Balapuwaduge, Frank Y. Li*

Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway

Abstract

At a density of one million devices per square kilometer, the 10's of billions of devices, objects, and machines that form a massive Internet of things (mIoT) require ubiquitous connectivity. Among a massive number of IoT devices, a portion of them require ultra-reliable low latency communication (URLLC) provided via fifth generation (5G) networks, bringing many new challenges due to the stringent service requirements. Albeit a surge of research efforts on URLLC and mIoT, access mechanisms which include both URLLC and massive machine type communications (mMTC) have not yet been investigated in-depth. In this paper, we propose three novel schemes to facilitate priority-based *initial access* for mIoT/mMTC devices that require URLLC services while also considering the requirements of other mIoT/mMTC devices. Based on a long term evolution-advanced (LTE-A) or 5G new radio frame structure, the proposed schemes enable device grouping based on device vicinity or/and their URLLC requirements and allocate dedicated preambles for grouped devices supported by flexible slot allocation for random access. These schemes are able not only to increase the reliability and minimize the delay of URLLC devices but also to improve the performance of all involved mIoT devices. Furthermore, we evaluate the performance of the proposed schemes through mathematical analysis as well as simulations and compare the results with the performance of both the

*Corresponding author
Email address: frank.li@uia.no (Frank Y. Li)

legacy LTE-A based initial access scheme and a grant-free transmission scheme.

Keywords: mIoT and mMTC, URLLC, LTE-A and 5G NR, initial access.

1. Introduction

While the Internet of things (IoT) is revolutionizing our society at an unprecedented pace, more recent research and development focus on IoT is shifting towards the direction of massive IoT (mIoT). In parallel with this trend, *massive machine type communications (mMTC), which is an enabling technology for mIoT*, has been envisaged as one of the three major use cases for the fifth generation (5G) mobile and wireless networks. Indeed, the popularity of mIoT arises from the ever-increasing data traffic spurred by various applications ranging from smart cities to mission critical communications in cyber-physical systems and Industry 4.0 [1]. Consequently, the ever-growing network size, heterogeneity in applications, and energy constraints pose various new challenges for mIoT related research [2]-[4].

Together with mMTC, enhanced mobile broadband (eMBB) and ultra-reliable and low latency communication (URLLC) are the other two use cases for 5G applications. The current standardization activities led by the 3rd generation partnership project (3GPP) focus mainly on eMBB, which represents an evolutionary path from long term evolution-advanced (LTE-A) in order to provide ultra-high data rates to end users for applications like high resolution video streaming. Meanwhile, there is a surge of research interests in mIoT/mMTC and URLLC from both academia and industry [5]-[9]. For mIoT/mMTC applications including automated energy distribution in a large smart grid, control of large-scale industrial processes, and surveillance of critical infrastructure, how to provide medium access to a huge volume of devices appears as a challenging task. In contrast to eMBB, the URLLC use case focuses on achieving ultra-high levels of reliability and low latency for futuristic scenarios like remote surgery, remote monitoring and control, as well as augmented and virtual reality [9][10]. For many applications, it is expected that the reliability level reaches

99.9999% or higher and the device to network latency becomes less than 1 ms [10]. However, achieving stringent URLLC in 5G is extremely challenging especially when considering that ultra-reliability and low latency represent two contradictory requirements. For instance, achieving high reliability requires parity check, coding or link redundancy, and packet retransmissions which in turn increase latency [9].

Addressing these mIoT and 5G challenges calls for novel approaches for system development and protocol design. Although a lot of work on eMBB has been done, URLLC and mIoT/mMTC are expecting more innovative contributions from the research community. Among others, one of the most paradoxical research questions to be answered is *how to satisfy service requirements when both mIoT/mMTC and URLLC are jointly taken into consideration.* This point is especially important for *initial access of IoT devices which occurs before actual data transmissions.* It is known that existing LTE/LTE-A based random access (RA) procedures are inefficient when there are a large number of device arrivals simultaneously, due to the constraint of a limited number of preambles or/and radio resource blocks for uplink or downlink traffic [11]. Although numerous initial access schemes have been proposed for fourth generation (4G) networks, the problem becomes more complex in 5G new radio (NR) since 5G NR Phase 1 is more advanced but still based on orthogonal frequency division multiple access (OFDMA). In an mIoT network, when traffic volume is high especially under bursty traffic conditions, the number of attempts for initial access could rise substantially, leading to high collision, low access success probability, and correspondingly increased latency. As such, it is imperative to develop customized solutions in 5G for devices that require URLLC access among mIoT devices.

In this paper, we propose three initial access schemes addressing the aforementioned research question. Considering a large number of mIoT/mMTC devices covered by a cell, we focus on providing ultra-reliable and low latency access for a portion of devices that require URLLC services. The proposed novel schemes utilize device grouping and resource grouping for low latency

communications based on the LTE-A or NR frame structure. Furthermore, the performance of these schemes is analyzed mathematically based on an existing comprehensive model which was initially developed for LTE traffic but with our extension to fit the proposed schemes in our envisaged LTE-A and NR scenarios. Extensive simulations are performed to validate the model and compare the performance of our schemes with that of three existing schemes.

In brief, the main contributions of this paper are summarized as follows.

- Three initial access schemes are proposed with the aim of providing services for mMTC[1] devices in two scenarios with location-bounded and location-spread URLLC devices, respectively. These schemes are specifically designed considering bursty traffic arrivals, posing a worst case scenario for devices sharing resources for initial access.
- Based on the advanced features of numerology and the frame structure in NR, a novel RA slot allocation method which enables flexible URLLC grouping is proposed. Accordingly, collisions among URLLC access contentions and latency are minimized.
- The performance of the proposed schemes is evaluated through analysis and simulations by taking into account a massive number of devices contending for network access and compared with the performance of both the existing LTE-A RA scheme which serves as a baseline scheme and with a grant-free (GF) transmission scheme.

The rest of the paper is organized as follows. Sec. 2 summarizes the related work. Then, Sec. 3 provides preliminaries to help readers better comprehend the work presented in the paper. In Sec. 4, the network scenarios and assumptions are presented. In Sec. 5, the proposed schemes are explained in details, followed by performance analysis in Sec. 6. Thereafter, Sec. 7 illustrates the numerical results. Finally, the paper is concluded in Sec. 8.

[1]In the rest of this paper, the terminologies, IoT and MTC, or mIoT and mMTC, are interchangeably used.

2. Related Work

As an enabling technology for mIoT operation in licensed bands, mMTC follows the procedures defined by 3GPP. Since these procedures are highly relevant to the work presented in this paper, we first outline existing solutions for RA channel (RACH) congestion avoidance for initial access that occurs prior to data transmissions in LTE-A and 5G NR and then introduce a few mathematical models for LTE-A RA process performance evaluation.

2.1. RACH Congestion in LTE-A: Initial Access and Solutions

A main constraint of the LTE-A RA process is the limited number of preambles available in a cell, e.g., 64 preambles within one RA slot (to be clarified in Sec. 4). Out of these 64 preambles, a certain amount, typically 10, is reserved for contention-free transmissions while the rest is shared by other devices. RA collision occurs when multiple devices select the same preamble to transmit in the same RA slot (to be clarified in the next section), causing unsuccessful detection of transmitted preambles at the evolved nodeB (eNB) [13]. This in turn results in an increased number of retransmissions, further escalating the problem.

In [11][40], 3GPP recommended several solutions to resolve this problem. Two of the most popular approaches are access class barring (ACB) and extended access baring (EAB) [40] [14]. Initially, ACB provides an effective access control mechanism in order to prevent potential overload of a network. In ACB, devices are classified into multiple classes with different priority levels. An eNB broadcasts the configuration information periodically through the master information block (MIB) and system information block (SIB) messages. Via SIB Type 2 (SIB2), the eNB broadcasts the current ACB configurations including a barring rate and a barring timer to guide various classes of devices to run a random access procedure in case of possible network overload. When a device intends to access the channel, it will pursue a random access procedure if its selected random number is lower than the barring rate. Although ACB provides higher priority devices with higher access probabilities, it does not guarantee

their access privilege [15]. This is because ACB schemes still follow contention based access and collisions could still happen for example when there are too many high priority devices.

The performance of ACB schemes may vary with different parameter configurations. In [32], an ACB scheme for dealing with physical RACH (PRACH) overload was studied and the impact of its configuration parameters on network performance was analyzed. In [33], an optimal ACB control and resource allocation scheme to acquire system capacity under a limited total number of resource blocks was proposed.

Furthermore, in order to prevent overload of the network, EAB introduces another more restrictive method to control access attempts from devices that can tolerate more access restrictions for instance MTC devices which can tolerate longer delays. EAB provides a *deterministic* access control mechanism, preventing devices belonging to certain types access classes from obtaining access [41]. If congestion occurs, the network could restrict the access of these classes of EAB devices while still allowing access from other EAB devices specified through the advertised SIB messages and ACB devices according to the barring rate [11].

On the other hand, in both ACB and EAB, the detection of traffic conditions by an eNB is performed in a reactive manner and devices also behave passively based on the received SIB messages. Although these schemes improve the access success of higher priority devices, such behavior will cause additional delays which are detrimental for achieving low latency communications, especially upon the arrival of a traffic burst.

In addition to ACB and EAB, [11] has also proposed several other schemes. For instance, an MTC specific backoff approach introduces separate backoff times for MTC and human type communication (HTC) traffic by assuming that HTC traffic always has higher priority. However, when it comes to URLLC, we cannot prioritize HTC traffic over MTC traffic as both types will have similar importance levels. Other approaches include slot based and pull based access or eNB initiated access. For uplink URLLC access, however, these approaches may

not be efficient since URLLC devices cannot wait until the eNB has initiated a communication process. In [44], the coexistence of scheduled and non-scheduled URLLC services and the difficulties for achieving stringent latency requirements under such a scenario were discussed. Furthermore, grouping based methods have also been studied for collision avoidance in LTE-A RA. In [16], a grouping based method was proposed to diminish collisions at the eNB. Using this method, all group devices send their data to a group coordinator based on device-to-device (D2D) communications and group coordinators transmit uplink data following the standard 4-step RA procedure. This scheme was further analyzed in [17]. Recently, a compressed sensing based RACH protocol was proposed in [18].

Furthermore, cluster based access schemes were proposed in [34] [35] to mitigate potentially severe collisions of MTC devices that access to an eNB concurrently. In another study performed in [36], spatial group based reusable preamble allocation was proposed. According to clustering-reuse preamble allocation proposed in [35], complementary preamble sets are allocated to clusters with similar distances and the same preamble set is allocated to clusters that are far away. In [37], a cluster based group paging scheme for congestion and overload control was proposed. This method is based on IEEE 802.11ah by collecting the sensed data from MTC devices and upload data to the LTE/LTE-A cellular network. However, 802.11ah limits the number of devices.

In a nutshell, although many schemes have contributed to a large extent RACH congestion avoidance, most of them are targeted at LTE-A networks without considering the stringent low latency requirements for URLLC services. Despite much progress, the performance gap for RA in terms of providing ultra-high reliability and low latency simultaneously in mMTC networks remains largely unresolved and calls for more research efforts.

2.2. Initial Access for 5G NR

For medium access in NR Phase 1, an OFDMA based RA scheme similar to the LTE-A RA scheme was recommended [19] [20]. Its main difference in

comparison with LTE-A is the introduction of beam steering techniques for synchronization in higher frequency operations, as further discussed in Sec. 3 below. Additionally, the NR frame structure with shorter transmission time intervals (TTIs) ensures faster RA process and allows more flexible numerology [21][22]. In general, with proper parameter tunning, the ACB and EAB mechanisms presented above which are initially designed for LTE/LTE-A are also applicable to NR Phase 1 initial access.

Additionally, there have been numerous access schemes proposed for 5G NR. Among them, [23] proposed a contention based access scheme by allowing multiple transmissions of the same packet in consecutive TTIs. By deducing the optimal number of consecutive transmissions, the low latency and high reliability requirement can be satisfied. Another type of popular approaches is grant-free access, also known as configured grant [24][25], in which devices are allowed to transmit their data messages without following the standard grant based (GB) process [26][27]. In [26], a GF radio access scheme was proposed for low complexity IoT devices where highly reliable access with bounded delay was achieved with long battery lifetime. Accordingly, devices directly transmit their data packets in pre-configured grant-free slots defined by the next generation NodeB (gNB). Rather than waiting for an acknowledgment (ACK) or negative ACK (NACK) message which takes additional time, a device may transmit replicas of its message up to k times in randomly selected k GF slots within a subframe for achieving high reliability and low latency. When multiple devices transmit at the same time, different techniques like successive interference cancellation (SIC) can be employed to cancel out interference and detect data associated with a specific user. However, GF transmissions are targeted at small size data packets with sporadic arrival patterns [41]. When a large number of devices transmit at the same time, grant-free access could result in high collision probability and increased delay considering the additional time required for resolving collisions [24]. As such, how to ensure URLLC in 5G NR based mIoT networks remains as an open research question.

2.3. Modelling LTE-A RA Process

Modeling precisely an LTE-A RA procedure is not an easy task. As mentioned in Sec. I of [28], the performance evaluation of RA schemes is oftentimes conducted by means of simulations due to the fact that the RA procedure of LTE-A is difficult to model analytically. Among the research efforts reported in the literature, [29] provided a model with a focus on the first preamble transmission. Although few other analytical models that consider the complete RA process exist, the accuracy of these models needs to be improved when comparing with simulation results. In [42], a general model to analyze the performance of the RACH procedure was proposed and validated via simulations, focusing on the case of highly synchronized MTC traffic. Furthermore, an in-depth review on the accuracy of existing models was presented in [28]. However, most of these models have ignored access delay which is a key performance indicator. This aspect is especially important in the case of URLLC since the latency performance needs to be properly analyzed. [30] presented a comprehensive analytical model for performance evaluation of the LTE based RA process which also serves as the basis for our performance analysis presented later in Sec. 6. Therein, the authors adopted Stirling numbers of the second kind to derive an exact expression for the probability distribution of the number of successful preamble transmission attempts over multiple RACH slots. Moreover, the drift approximation was used to model a complete and detailed LTE RA procedure based on a 3GPP standard [12].

Furthermore, it is worth mentioning that the schemes proposed in this paper differ from existing work in several ways. Firstly, a salient feature of this work is the consideration of both mMTC and URLLC requirements that is largely overlooked in most other studies. Secondly, the proposed schemes are built on top of the LTE-A or NR based RA procedure and we advance the state-of-the-art techniques by introducing priority based grouping approaches for initial access of URLLC traffic. Thirdly, unlike other existing priority based approaches for instance ACB and EAB, which do not provide guaranteed access with low latency, our schemes ensure access privilege based on device grouping or RA slot

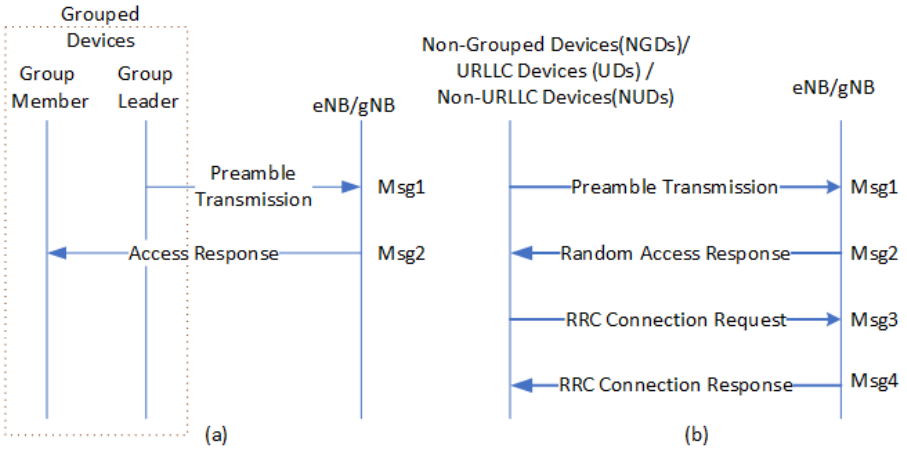

Figure 1: Illustration of (a) the 2-step access procedure for UDs and (b) the 4-step access procedure for LTE-A, NGDs, UDs, and NUDs.

grouping, providing URLLC devices with guaranteed or highly probable access. Lastly, while most other schemes like ACB and EAB follow a *reactive* principle as mentioned above, our schemes behave in a *proactive* manner which is beneficial for achieving low latency and the parameters are reconfigurable. By proactive, it is meant that device grouping is performed in an intended manner and a dedicated preamble is assigned to each group leader. The parameters involved in this procedure, e.g., number of devices in each group, are configurable, however, over a comparatively long period much larger than a MIB or SIB cycle.

3. Preliminaries

This section provides preliminaries that form the bases for the schemes to be presented in the rest of this paper.

3.1. RA Process in LTE/LTE-A and 5G NR

An RA process occurs when devices require initial access, e.g., upon network deployment or update, or transition from an idle mode to a connected mode. Such an RA process needs to be performed for initial access, after a signaled disconnection from the gNB, or a device has just woken up from the power

saving or sleep mode. The LTE/LTE-A RA process recommended by 3GPP consists of the exchange of four handshake messages between a device and its associated eNB, as illustrated in Fig. 1(b).

- *Step 1 (Msg1): Preamble transmission.* Whenever a device needs to communicate with an eNB, it first selects an RA preamble from a set of available preambles and transmits it in the next available RA slot. *An RA slot is a subframe within which devices are allowed to send their selected preambles.* It is defined by eNB and broadcast periodically over paging cycles via the SIB2 messages.

- *Step 2 (Msg2): Random access response (RAR).* When the eNB receives preamble transmissions without collision, it transmits Msg2 in the handshake process. Through RAR, the eNB schedules uplink resources for the transmission of the next message. Additionally, RAR contains also information about the detected RA preamble sequence, for which the response is valid, timing advance details, and a cell radio-network temporary identifier (C-RNTI) for further communication of a particular device.

- *Step 3 (Msg3): Radio resource control (RRC) connection request.* Using the received C-RNTI and uplink resources, the device transmits its RRC request to the eNB based on the uplink radio resources assigned by the RAR message. Msg3 includes the device temporary C-RNTI which is used for contention resolution in the fourth step.

- *Step 4 (Msg4): RRC connection response.* Devices receive the RRC setup message from the eNB. Only the devices which have their transmitted and received identities matched in Msg3 and Msg4 declare their RA procedure to be successful. After this step, the four-step handshake procedure for initial access is complete. Then devices and eNB perform data transmissions based on the C-RNTI of each device.

In case that there is more than one device transmitting the same preamble, a collision occurs and the competing devices may not receive the corresponding

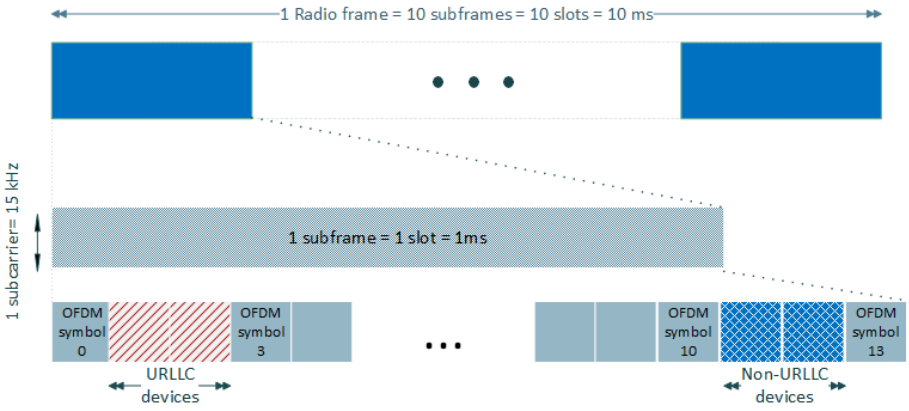

Figure 2: Illustration of the NR frame structure for $\mu = 0$ and OFDM symbol allocation in the second proposed initial access scheme.

RAR message. If any step in one of the four handshake steps fails[2] the involved device will wait for a random backoff period from a window of size w_{BO} and repeat the RA process by retransmitting an RA preamble. The maximum number of transmissions allowed is limited by a given number, n_{PT}.

In 5G NR, the initial access procedure between a device and its associated gNB is similar to the one employed in LTE-A when operating in the sub-6 GHz frequency range, often referred to as frequency range 1 (FR1). For frequency range 2 (FR2), which includes frequency bands from 24.25 GHz to 52.6 GHz, the initial access involves procedures for cell search and synchronization using beam sweeping [20] [22]. However, to study these physical layer details is beyond the scope of this paper.

3.2. 5G NR Frame Structure and Numerologies

NR introduces novel scalable numerology and frame structure with the aim of facilitating the expected capacity and latency requirements in 5G. In contrast to the 15 kHz only option in LTE/LTE-A, NR supports multiple subcarrier spacing. NR defines 15 kHz as a baseline and introduces 5 numerologies based

[2] An unsuccessful message transmission may also occur due to channel impairments for uplink and/or downlink. This effect is partially reflected in the message error probability expression presented later in Sec.6.

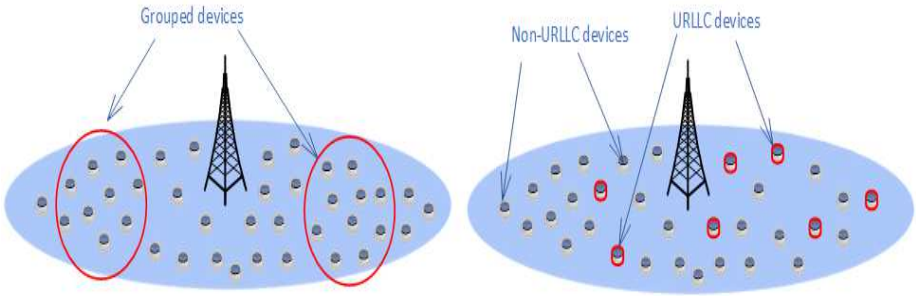

Figure 3: (a) Scenario 1: Location-bounded URLLC devices versus (b) Scenario 2: Location-spread URLLC devices.

on subcarrier spacing $\Delta f = 2^{\mu} * 15$ kHz where $\mu = 0, 1, ..., 4$ is the numerology index [22]. The radio frame duration in NR is the same as in LTE/LTE-A, i.e., 10 ms, and one *frame* consists of 10 *subframes* each with 1 ms duration, as shown in Fig. 2. Moreover, one NR subframe may have one or more *slots* based on the numerology index. For $\mu = 3$ and $\mu = 4$ which are used in our study, the number of slots per subframe would be 8 and 16, respectively. With the increased subcarrier spacing and a larger value of μ, the slot duration reduces according to $1/2^{\mu}$ ms. When $\mu = 3$ and $\mu = 4$, the slot duration would be 125 μs and 62.5 μs respectively. Furthermore, each slot contains 14 (or 12 for extended cyclic prefix (CP)) OFDM *symbols*. However, not all numerologies are applicable to any type of physical channels. Instead, a specific numerology is used only for a given type of physical channels. For more details about NR numerology, refer to [22] [31].

3.3. A 3GPP Model for Bursty Traffic

A bursty traffic arrival process occurs when a large number of IoT devices attempt to access the same network simultaneously during a short period of time. This is especially observable under mMTC scenarios where the number of devices could be huge. In [11], 3GPP recommends applying a Beta distribution based arrival process to model the arrival intensity during bursty traffic arrivals,

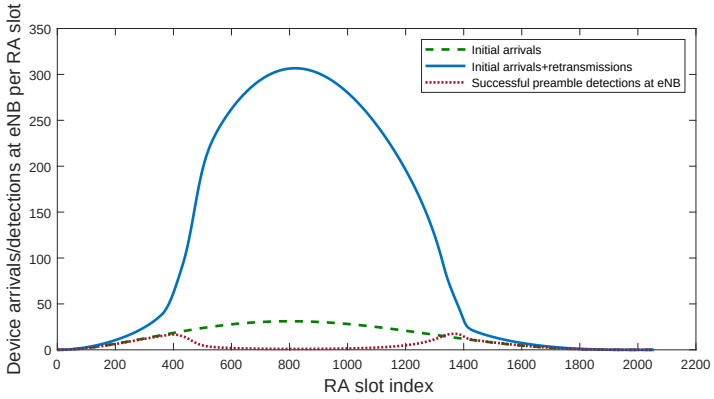

Figure 4: Number of initial arrivals, retransmissions, and detections in LTE-A random access for 30k devices with 54 preambles following a bursty arrival process [43].

shown as follows.

$$A(i) = L \int_{t_i}^{t_{i+1}} p(t)dt, \tag{1}$$

where $A(i)$ represents the access intensity for a total number of L devices contending in an RA slot i between time t_i and t_{i+1}. In (1), $p(t) = (t^{\alpha-1}(T-t)^{\beta-1})/(T^{\alpha+\beta-1}Beta(\alpha,\beta))$ where $Beta(\alpha,\beta)$ is the Beta function with $\alpha = 3$ and $\beta = 4$. T is the total observation time for traffic arrivals [11].

As an example, we illustrate in Fig. 4 the numbers of initial arrivals, initial arrivals plus retransmissions, and successful detections within an RA slot under a traffic burst of 10 sec based on 30k devices and 54 preambles [43]. It is clear that the actual number of arrivals consisting of both initial arrivals and retransmissions is much higher than the initial arrivals itself. With such bursty traffic arrivals, the number of devices competing for access in an RA slot is unusually high and providing URLLC services in such a scenario is a challenging task since GF based access schemes which were discussed in Subsec. 2.2 above would result in high collisions. For this reason, the proposed schemes in this paper focus on grant-based initial access instead of GF transmissions and radio resource allocation. Later on in Subsec.7.6, we provide a brief comparison of our schemes versus a GF scheme.

4. Network Scenarios and Assumptions

The envisaged network scenarios in this work are inspired by the futuristic cyber-physical mIoT applications recently presented by 3GPP in [10]. In many such applications, devices are battery powered with power saving mode enabled. Upon the occurrence of a mission critical event, for instance, it is likely that many devices will require initial access at almost the same time leading to a traffic burst as presented above.

In this study, we consider that all devices are covered by one cell although some of them may lie comparatively far away from the eNB and that proper preamble formats are allocated to all RA slots [39]. For each of the two scenarios shown in Fig. 3, there are L number of IoT devices within the coverage area of an eNB or a gNB[3] and ϕ number of preambles that can be allocated to this cell in a given RA slot. The number of orthogonal preambles that can be allocated in a given RA slot depends on the cell coverage [38]. According to [12][39], there are 64 preambles that can be allocated in a cell with a coverage radius of 7.4 km and a delay spread of 6 μs and these preambles are designed to be orthogonal to each other.

Scenario 1: Location-bounded URLLC Devices. Although a large number of mMTC devices are deployed across a cell, a set of devices in the immediate vicinity of a point of interest are monitoring the same natural or physical phenomenon, e.g., for process automation within a service area of 100 m $\times$ 100 m as given in [10].

In this scenario, we categorize the total population of L IoT devices into γL grouped devices (GDs) and $(1-\gamma)L$ non-grouped devices (NGDs) where γ is a scalar with $0 < \gamma < 1$. The traffic generated by IoT devices could be deterministic periodic, deterministic aperiodic, or non-deterministic [10]. In this work, we focus on a case where devices abruptly require uplink access after sensing an event triggered in a *non-deterministic* and bursty manner, thus represent-

[3]For the rest of this paper, the abbreviations eNB and gNB are interchangeably used as both LTE/LTE-A and 5G NR Phase 1 follow the same procedure for initial access.

ing a worst case scenario among the aforementioned traffic types. Accordingly, the GDs require URLLC access while NGDs still generate traffic but without demanding URLLC services. Although semi-persistent scheduling for URLLC access is another option, it may not guarantee the required performance due to the stringent delay requirements especially when the number of URLLC devices is huge. Furthermore, maintaining semi-persistent scheduling for a massive number of devices is rather difficult and costly as mMTC traffic is often sporadic. For this reason, we propose to reserve merely a small amount of resources (preambles) for grouped devices and obtain the necessary amount of uplink resources for all other group devices through group leader's communications with the gNB (to be clarified in the next section).

Furthermore, we assume that device grouping including group leader selection is performed beforehand based on a specific criterion, e.g., the functionality or geographic proximity of the IoT devices. Device grouping is reconfigurable, however, over a comparatively long period much longer than a SIB2 cycle. A triggering event would be detected by all IoT devices in the same group including the group leader. All the GDs that sensed the triggering event need their measurements to be transmitted to the gNB as each device may report a different facet of the same event. Once a preamble is received, the gNB is assumed to have enough radio resources to allocate to all these grouped devices.

The rationale of the above assumption is as follows. Although the amount of available physical downlink (PDCCH) resources is always limited in reality, the flexibility provided by NR enables the use of more PDCCH resources compared with that of LTE-A. Based on the NR numerology and frame structure presented above and the flexibility provided for PDCCH scheduling [22], more downlink control information (in terms of both information volume and broadcast interval) can be transmitted via PDCCH within a given 5G NR subframe compared with what is possible in LTE-A. Moreover when considering the privilege of URLLC traffic, it is common in the literature to employ techniques such as preemptive scheduling which provides immediate downlink resources to URLLC traffic by overriding parts of already assigned resources for eMBB or

another type of lower priority traffic. Such a mechanism is justifiable considering the stringent latency requirements of URLLC devices. Accordingly, we may introduce a potential solution which combines preemptive scheduling with the NR frame structure to accommodate extra PDCCH resources to URLLC traffic. In this way, resource constraint which might appear as a bottleneck to complete the initial access procedure could be abbreviated.

Scenario 2: Location-spread URLLC Devices. Consider another scenario where the IoT devices that require URLLC services are not confined to certain areas within the coverage but could be spread anywhere across the cell. The devices in this scenario could be process monitoring devices which are *static* or mobile robots which are *non-static* [10]. Among these L devices, a certain portion, i.e., ηL where η is a scalar with $0 < \eta < 1$, of devices are considered to require URLLC services whereas the remaining $(1-\eta)L$ devices do not have such a requirement. Hereafter, these two categories of IoT devices are denoted as URLLC device (UD) and non-URLLC device (NUD), respectively.

Further Clarification: Different from GDs in Scenario 1 which are restricted to certain small areas, UDs in Scenario 2 could be distributed geographically throughout the cell. During the bursty traffic arrival duration, all these devices are considered to be active, i.e., having at least one packet to transmit. The portions of devices which belong to GDs or UDs, i.e., γ and η, are determined by the eNB as a compromise of performance (collision probability, delay, etc.) and configurable parameters. Since these values are configured periodically and the gNB needs to inform all devices about any update, extra signaling overhead is expected. However, to study such extra overhead is beyond the scope of this paper.

Furthermore, in both scenarios, a single frequency band is considered. For NR frame structure based initial access scheme design, the parameter configurations and assumptions including numerologies, PRACH selection, and slot scheduling will be explained in the next section.

5. Proposed Initial Access Schemes

Based on the scenarios presented above, we propose three schemes for initial access of mMTC devices. While the first two schemes are tailored to the two scenarios (device grouping with dedicated preambles (DGDP) for scenario 1 and RA-slot based URLLC grouping (RAUG) for scenario 2), respectively, the third one combines the merits of the first two schemes and applies to both scenarios.

5.1. Device Grouping with Dedicated Preambles

The main feature of the DGDP scheme is that GDs obtain access privilege to the network through *a contention-free 2-step scheme* [8], as illustrated in Fig. 1(a) and explained below. Meanwhile, NGDs follow the legacy LTE-A 4-step contention based RA procedure, as shown in Fig. 1(b). It is expected that a 2-step RACH scheme will bring benefits to channel access in terms of both reduced latency and lower overhead. Although 2-step RACH approaches are presently under discussion within 3GPP, the current draft [45] does not state which type(s) of traffic should apply the 2-step scheme.

5.1.1. Access Scheme for Grouped Devices

Consider a single group as an example. At the initial network deployment phase, devices communicate and register themselves with their associated eNB. During the registration process, the eNB collects information about all IoT devices inside the group and their location information to infer the required timing advance details. A *unique and permanent* address, which is different from the C-RNTI mentioned in Sec. 3, is allocated to each device and the group also receives *a dedicated preamble for uplink communication to be used by the group leader*. The eNB stores these details in a database for further references.

Furthermore, a group leader is selected by the eNB based on a given criterion, e.g., device battery level, device location, or uplink channel quality among group members. All group members will periodically communicate with the eNB and the updated information will be used for group leader selection in the

next period of time. In other words, the group leader could be dynamically changed based on the adopted criterion by the eNB and newly collected information from group members. To tackle a rare case where the group leader's preamble transmission fails, e.g., due to uplink channel impairment, the eNB also assigns a backup group leader. A backup leader may also initiate a preamble transmission if necessary. The coordination between a serving group leader and the backup group leader can be performed by various methods with or without the involvement of the gNB. For instance, we can set a timer which expires after a pre-defined period from an event and triggers the backup leader to act as the serving group leader. Alternatively, we can assume an out-of-band D2D communication protocol between the serving leader and the backup leader. However, to design a protocol or procedure for group leader and backup group leader selection is beyond the scope of this paper. In what follows, we explain the 2-step scheme illustrated in Fig. 1(a).

Step 1 (Msg1): Event triggered dedicated preamble transmission. Once the deployment phase is finished, IoT devices enter into the operational stage. In an event where the observed measurements of IoT devices exceed a pre-defined threshold, a triggering event will be initiated. We assume that the group leader can sense this triggering event and correspondingly it immediately transmits its allocated preamble in the next available RA slot. Other GDs in the same group will not transmit any preamble but they overhear this transmission and wait for the access response from the eNB. In a rare case if the group leader does not sense the triggering event, or the group leader's uplink channel quality is below the required level, the backup group leader will transmit the preamble *after the timeout duration of the access response has elapsed.*

Step 2 (Msg2): Access response from the eNB: When the eNB receives a preamble that is reserved for a specific group, it identifies the group from the preamble. Since each group leader in different groups has its own dedicated preamble, this access process is collision-free. Once the eNB identifies the corresponding group which the received preamble belongs to, it retrieves the information about the registered group members. The eNB is aware of the immediate

access requirement of these GDs. *It then allocates resource blocks to individual group members* based on the addresses assigned during the registration process. The eNB transmits the relevant timing advance information for each group member based on the calculations from the registration process so that each member can adjust their transmission time accordingly for radio frame synchronization. Since devices are static, the timing advance values would remain the same unless an update is performed.

5.1.2. Access for Non-grouped Devices

The NGDs inside the same cell follow the legacy LTE-A RA scheme [12] with a 4-step procedure for initial access as explained in Sec. 3.1. Since n_G preambles are reserved for n_G group leaders, the number of available preambles for NGDs is reduced by n_G (where $n_G < \phi$), i.e., it becomes $\phi - n_G$. Concurrently, the number of NGDs competing for the $\phi - n_G$ preambles also shrinks to $(1 - \gamma)L$. If a collision happens, the collided devices will retransmit their preambles after waiting for a backoff interval based on a random number selected from a uniformly distributed range $[0 \sim w_{BO} - 1]$. For successfully transmitted preambles, Msg3 and Msg4 will be transmitted subsequently to complete the RA process as shown in Fig. 1(b). In this paper, we do not consider explicitly how a message transmission could be affected by channel impairment for any specific type of channels between the gNB and devices. However, the transmissions of Msg3 and Msg4 are subject to failures as presented in the next section.

As mentioned earlier, the group formation of IoT devices in the DGDP scheme is pre-defined and the parameters are reconfigurable. While having a higher n_G would enable access for a larger number of grouped devices, the selection of n_G and γ needs to be performed carefully to avoid performance degradation of NGDs. Generally, the number of devices per preamble gives an indication about the possibility of different devices selecting the same preamble and thereby causing collisions. In LTE-A without grouping, this ratio is L/ϕ. In DGDP with n_G number of groups and γL grouped devices, this ratio is given by $(1 - \gamma)L/(\phi - n_G)$ for NGDs. In order to improve the performance level that

will be achieved by NGDs without grouping, the following condition must hold

$$\frac{(1-\gamma)L}{(\phi - n_G)} < \frac{L}{\phi}. \tag{2}$$

Reformulating the above inequality into $(1-\gamma)L\phi < L(\phi - n_G)$, (2) can be expressed in a simplified form, as $n_G < \gamma\phi$. This relationship can be utilized when deciding n_G and γ so that the performance of NGDs is not compromised.

5.2. RA-slot based URLLC Grouping

Consider now an mIoT cell as presented earlier in Scenario 2 where the number of IoT devices that require URLLC services could be potentially large and their locations may spread across the cell. In this case, it is prohibitive to assign many dedicated preambles to these UDs as we did in DGDP since the total number of preambles in cell, i.e., ϕ, is very small. In what follows, we propose another scheme, RAUG, which grants access privilege to certain devices *without assigning dedicated preambles*. This scheme is designed largely based on the NR frame structure and numerology outlined in Subsec. 3.2.

5.2.1. The Principle of RAUG

In RAUG, all devices follow the 4-step RA initial access procedure but separate RA slot resources are assigned to URLLC and non-URLLC preamble transmissions respectively. As depicted in Fig. 2, each subframe provides RA opportunities and dedicated RA slots are reserved for UDs in order to provide them with URLLC access. As mentioned in Sec. 4, only a portion of IoT devices, i.e., ηL of them, will have URLLC requirements during a given period of time. Note that although it is possible to form groups with very small URLLC device population, very little benefit would be observed if the group size is too small considering the scarcity of the number of preambles. Accordingly, each particular device will transmit its preamble only in the assigned RA slot for UDs that is broadcast by the gNB beforehand and periodically, e.g., via the SIB2 message.

Format	L_{RA}	N_u	N_{CP}^{RA}
$A1$	139	$2 \times 2048 \times \kappa \times 2^{-\mu}$	$288 \times \kappa \times 2^{-\mu}$

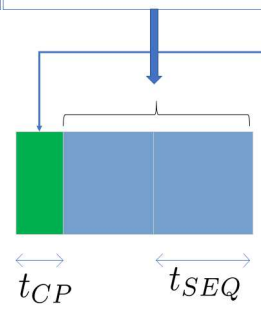

Figure 5: Illustration of the format of preamble type A1.

Different from ACB [11], RAUG does not assign any probabilities for any type of devices to transmit their preambles. In other words, *both UDs and NUDs have equal opportunity when competing for network access, however, through dedicated RA slots assigned inside a 5G NR subframe.* Although having a dedicated RA slot for URLLC devices significantly increases access probability, the time interval between two consecutive RA slots for UDs needs to be minimized in order to reduce latency. Distinct from the slotted access schemes presented in [11] where low latency is not a priority concern, the RAUG scheme utilizes the 5G NR frame structure and numerology concept for the purpose of latency reduction.

5.2.2. Frame Format in RAUG

To demonstrate the concept of RAUG, we use numerology $\mu = 0$ as an example. It corresponds to the 15 kHz subcarrier spacing and each subframe in the radio frame structure consists of a single slot. Among the 13 preamble formats available in NR, a short sequence can be used for numerology $\mu = 0$ [22]. Fig. 5 illustrates the preamble format adopted in this study, known as A1, and the values mentioned therein will be used to calculate the preamble duration. The value of L_{RA}, which is the preamble sequence length, is related to the short sequence while N_u and N_{CP}^{RA} provide the total sequence length and the CP

length of the preamble in samples respectively. To convert them into seconds, we need to multiply the given values by $T_c = 0.509 \times 10^{-6}$ ms where T_c denotes the basic time unit in NR.

Denote by term κ the ratio between the basic time unit of LTE/LTE-A (T_s) and T_c. According to 3GPP [22], it ends up with $\kappa = 64$. Based on Fig. 5, the total duration of preamble format A1 is equal to $t_{cp} + 2 \times t_{seq}$ where $t_{cp} = 288\kappa \times 2^{-\mu} = (288 \times 64 \times 0.509 \times 10^{-6}) = 0.0094$ ms and $t_{seq} = 2048\kappa \times 2^{-\mu} = 0.0667$ ms. Hence, the total duration can be calculated as $t_{cp} + 2 \times t_{seq} = 0.0094 + 2 \times 0.0667 = 142.8$ μs. Note that this duration is similar to the time duration of two OFDM slots in $\mu = 0$ and hence, the preamble can be transmitted using two OFDM slots including CP. Similarly we adopt index 106 mentioned in Table 6.3.3.2-2 in [22] for the PRACH configuration in this study. As such, it is possible to transmit a PRACH preamble in every subframe.

In order to provide priority to devices with low latency requirements, we introduce an option to allocate two RA slots inside a given subframe for initial access. This is possible for specific types of preamble formats available under a given numerology that satisfies the preamble length and OFDM symbol duration requirements mentioned above. Further details regarding these formats can be found in Table 6.3.3.1-2 of [22]. Accordingly, their duration can be calculated similar to the aforementioned calculation. Hence, considering the above configuration by having two RA slots inside a slot (one slot equals to one subframe for $\mu = 0$), both UDs and NUDs obtain an opportunity for an initial access attempt in every slot. Table 11.1.1-1 in [19] defines which symbols could be allocated for uplink and downlink transmissions. However, different from the legacy initial access procedure, the RA slot OFDM symbols in RAUG are different for UDs and NUDs. Correspondingly, *both types of IoT devices can share the same set of preambles in the same subframe, however, in different RA slots.* Furthermore, once the gNB receives a set of preambles in the URLLC RA slot, it treats these requests with higher priority. Hence, the required timing for the transmission of remaining messages is reduced.

Further discussions on the distinctions between RAUG and DGDP: Firstly,

Table 1: Main features of the three proposed schemes.

	DGDP	RAUG	HS
Type of devices	GDs, NGDs	UDs, NUDs	GDs, UDs, NUDs
Pre-grouping of devices	Yes	No	Yes
RA slot based grouping	No	Yes	Yes
URLLC enabled for	GDs only	UD only	GDs, UDs
Guaranteed reliability	for GDs	No	for GDs

no prior grouping based on service types or device location is involved when deciding UDs in RAUG. UDs could be deployed in any location inside a cell and do not have to share any common application with their neighboring devices. Furthermore, each UD could perform its individual task supporting a specific application. Secondly, unlike GDs, UDs need to transmit the preambles themselves and compete with other UDs for initial access. Thirdly, UDs do not necessarily need to be static in deployment whereas GDs are considered to be static for timing advance synchronization purposes needed in the 2-step initial access procedure. However, different from the legacy RA scheme, UDs do not need to compete with NUDs since they have their separate RA slots to transmit the selected preambles. This would ensure better access opportunities for UDs in comparison with devices in the legacy scheme. Furthermore, unlike NGDs in DGDP which compete for $\phi - n_G$ preambles, NUDs in RAUG have all ϕ available preambles for access competition in their allocated RA slot.

5.3. Hybrid Scheme (HS)

While DGDP is designed for providing URLLC services for a specific set of GDs, it cannot be applied to a large number of IoT devices with such requirements. RAUG releases this constraint by providing *high reliability and low latency access* for a potentially much larger number of UDs inside a cell regardless of their locations. However, since RAUG follows a 4-step contention based RA procedure, the achieved reliability and latency could be lower than what is obtained in DGDP. In this subsection, we propose a hybrid scheme which combines the merits of the other two schemes proposed above.

More specifically, HS is a combined access scheme in which both device

based grouping and slot based allocation apply. In this scheme, we still have GDs and NGDs but NGDs are further categorized into UDs and NUDs. UDs will use the first RA slot to transmit its preambles but still follow a contention based procedure. GDs and NUDs will use the second RA slot inside the same subframe, however, GDs still have dedicated preambles. In this way, *GDs and UDs can share the same preambles but in different slots*. Hence, a larger number of IoT devices with URLLC requirements can be accommodated via GDs and UDs while utilizing the benefits of having multiple RA slots inside a subframe.

Accordingly, there will be γL GDs. Among the remaining $(1-\gamma)L$ NGDs, $\eta(1-\gamma)L$ will be UDs and $(1-\eta)(1-\gamma)L$ devices will be NUDs. As a result, $\eta(1-\gamma)L$ UDs will compete for ϕ preambles inside the first RA slot in a subframe whereas $(1-\eta)(1-\gamma)L$ NUDs will compete for $\phi - n_G$ preambles in the second RA slot inside the same subframe.

Moreover, it is worth reiterating that the proposed schemes for IoT device initial access in this paper are targeted at both 4G and 5G NR Phase 1, i.e., OFDMA based networks, and the operation of RAUG and HS relies on the support of NR numerologies. Enabled by the flexibility supported through different numerologies in 5G NR, allocating two RA slots inside one subframe becomes configurable. Meanwhile, reservation of radio resources is also feasible in both 4G and 5G NR. Therefore, to apply the proposed scheme(s) to a specific type of IoT technology, e.g., narrowband IoT (NB-IoT), proper parameter tuning based on the corresponding physical layer specifications is required. In Table 1, we summarize the main features of the three proposed initial access schemes.

6. Performance Analysis

In this section, the performance of the proposed schemes is analyzed. Recall that a contention-free 2-step procedure applies to GDs whereas the other types of IoT devices, i.e., NGDs, UDs, and NUDs follow a contention based 4-step procedure however *with different number of preambles and different number of device arrivals for each type of devices*. Therefore, the same analytical model

Table 2: $\hat{L}$ and $\hat{\phi}$ values for different type of devices in the three proposed schemes.

	Initial Access Scheme						
	DGDP		RAUG		HS		
	GDs	NGDs	UDs	NUDs	GDs	UDs	NUDs
$\hat{L}$	γL	$(1-\gamma)L$	ηL	$(1-\eta)L$	γL	$\eta(1-\gamma)L$	$(1-\eta)(1-\gamma)L$
$\hat{\phi}$	n_G	$\phi - n_G$	ϕ	ϕ	n_G	ϕ	$\phi - n_G$

applies to these three types of devices. In Table 2, we summarize the number of IoT devices and the number of available preambles per RA slot in each type, denoted as $\hat{L}$ and $\hat{\phi}$ respectively, for our performance evaluation. The main notations, their meanings, and the respective numerical values[4] used in this study are listed in Table 3.

In the rest of this section, the performance evaluation of GDs is presented first. Then, an analytical model used to evaluate the performance of NGDs, UDs, and NUDs is developed. For performance evaluation, three metrics which are recommended by 3GPP [11], i.e., preamble collision probability, access success probability, and average delay for successful transmissions, are selected as our performance metrics.

6.1. Performance of GDs

Since each group has its dedicated preamble reserved for GDs, the access process for GDs is contention-free. Hence, the probability of occurring a preamble collision at the eNB is 0. However, although there is no preamble collision, there is no guarantee that the preamble will be successfully received considering the effect of channel impairments. This is represented by the preamble detection probability P_j at the eNB for the j^{th} preamble transmission of the group leader. The value of P_j is calculated based on $P_j = (1 - e^{-j})$, as recommended by 3GPP [11], and it monotonically increases as more transmission attempts are conducted. Although the detection probability is not high enough after the first few attempts, it reaches the value of $P_j > 0.9999$ when $j = n_{PT} = 10$.

[4]In Table 3, the numbers inside () corresponded to values used by UDs.

Table 3: Notations, explanations, and values [11][30].

Notation	Explanation	Value
t_{AP}	Duration of an arrival period (in terms of subframes).	10000
L	Total number of devices in a cell which request service during t_{AP}	10000-300000
w_{BO}	Backoff window size (in terms of subframes)	21, (1)
t_{RAS}	Interval between two successive RA slots (in terms of subframes). The t_{RAS} value in RAUG is 8 OFDM symbols (Refer to Fig. 2)	5, 1
ϕ	Total number of preambles in an RA slot available for access competition	54
n_{PT}	Maximum number of preamble transmissions	10
w_{RAR}	Length of the RA response window (in terms of subframes)	5, (2)
p_j	Preamble detection probability of the j^{th} preamble transmission	$p_j = 1 - \frac{1}{e^j}$
p_f	HARQ retransmission probability for Msg3 and Msg4	0.1
n_{HARQ}	Maximum number of HARQ transmissions for Msg3 and Msg4	5
t_{HARQ}	Time interval required for receiving HARQ ACK (in terms of subframes)	4, (1)
t_{RQ}	Gap of Msg 3 retransmission	4, (1)
t_{RAR}	Processing time required by the eNB to detect transmitted preambles (in terms of subframes)	2, (1)
n_G	Number of groups	5, 10, 15
γ	Portion of devices from L that are grouped	0.1, 0.2, 0.3
η	Portion of devices from L that require URLLC services	0.1, 0.2, 0.3, 0.5
n_{UL}	Maximum number of devices acknowledged within an RA response window	15
t_D	Delay from a preamble transmission to the reception of the RAR response	$w_{RAR} + t_{RAR}$
μ	5G NR subcarrier spacing configuration numerology	0 - 4

Accordingly, we claim that the access success probability for GDs will be 1 even in the worst case given that up to $n_{PT} - 1$ retransmissions can be performed.

For detecting a preamble successfully, at least one transmission attempt is required from the group leader. Whether a retransmission is needed or not depends on the detection status of the previous transmission, up to $n_{PT} - 1$ times. Let $s(j)$ be the probability of success after the j^{th} preamble transmission and it is given by $s(j) = (1 - P_1)(1 - P_2) \cdots (1 - P_{j-1})P_j$. This expression is equivalent to the probability mass function of success at the j^{th} preamble transmission. Therefore, the expected value of the number of preamble transmissions required for a successful detection can be obtained by $\sum_{j=1}^{n_{PT}} js(j)$. After a t_D duration from a successful preamble transmission, the group members receive Msg2 from the eNB with the granted access and allocated radio resources, as shown in Fig. 1(a). Correspondingly, the group leader will wait for a duration

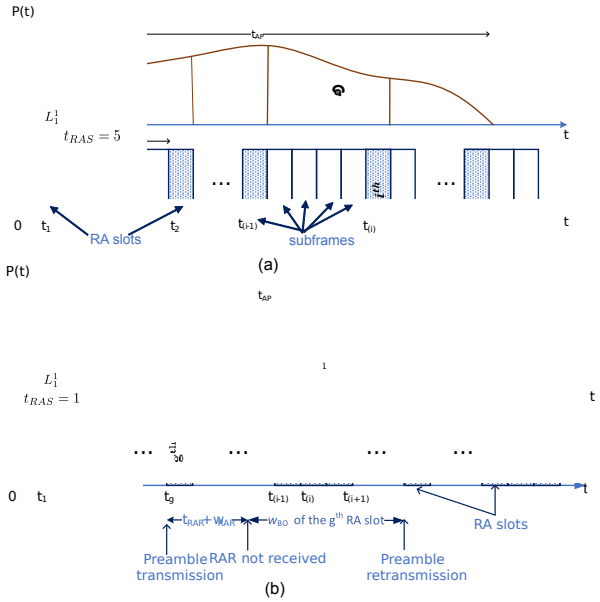

Figure 6: Timing diagram denoting RA slots, initial bursty arrivals per slot and the related timing parameters (a) $t_{RAS} = 5$ (b) $t_{RAS} = 1$.

of t_D before initiating a retransmission attempt. Therefore, considering the number of required retransmissions, the average delay for successfully transmitting a preamble and receiving the corresponding Msg2, denoted as D_a, can be calculated as follows

$$D_a = t_D \sum_{j=1}^{n_{PT}} js(j) = t_D \sum_{j=1}^{n_{PT}} \left(jP_j \prod_{k=1}^{j-1} (1 - P_k) \right). \tag{3}$$

To be more precise, the access delay for grouped devices would be slightly different from the access delay of their group leader if other factors such as the location of devices and extra cost for intra-group communications are included in this calculation. For analysis simplicity, we do not consider additional delay occurred for intra-group communications. Instead, the delay obtained in (3) is considered as an representative value since grouped devices are normally deployed in relatively close proximity to their group leader.

6.2. Performance of NGDs, UDs, and NUDs

6.2.1. Modeling the Initial Access Procedure

Consider a burst of initial traffic arrivals for the duration of t_{AP}. Fig. 6 illustrates the timing diagram with RA slots and arrivals. As explained earlier, the initial access procedure for NGDs, UDs, and NUDs follows the legacy RA process. Hence, a common analytical model is adopted as the baseline for analyzing these three types of devices. Based on a comprehensive analytical model proposed in [30] which provides sufficiently high accuracy for LTE-A RA processes, we present below our analysis tailored for performance evaluation of mMTC networks consisting of four types of devices according to the envisaged scenarios and the proposed schemes.

Initial arrivals: The average number of device arrivals at the i^{th} RA slot is calculated by the following equation

$$L_i^1 = \hat{L} \int_{t_{i-1}+1}^{t_i+1} p(t)\, dt, \tag{4}$$

where $p(t)$ is based on Beta distribution and t_i is the starting time of the i^{th} RA slot as explained in Sec. 3. The superscript of L_i^1 represents the initial arrival, i.e., $j = 1$. Term $\hat{L}$ in (4) denotes the total number of IoT devices based on each device type and the adopted access scheme, as illustrated in Table 2. Accordingly, the initial access device intensity at a given RA slot, L_i^1, is the integration of the number of new device arrivals between the end points of the previous and current RA slots.

Retransmissions: For a given RA slot i, in addition to the initial arrivals, there would be IoT devices attempting their j^{th} preamble transmissions ($1 < j \leq n_{PT}$) due to previously failed $(j-1)^{th}$ preamble transmissions at the g^{th} RA slot. The positions of the g^{th} and i^{th} RA slots are demonstrated in Fig. 6. The number of IoT devices performing their j^{th} preamble transmission on the i^{th} RA slot, denoted by L_i^j, is calculated as follows

$$L_i^j = \sum\nolimits_{g=G_{\min}}^{G_{\max}} \alpha_{g,i} L_{g,F}^{j-1}, \tag{5}$$

where G_{min} and G_{max} denote respectively the lower and upper limit of the window of the RA slot values that g could take. That is, in order to transmit the j^{th} transmission on the i^{th} RA slot, the $(j-1)^{th}$ transmission failure should occur between G_{min} and G_{max} time before t_i. $\alpha_{g,i}$ denotes the percentage of the backoff interval of the g^{th} RA slot that overlaps with the transmission interval of the i^{th} RA slot. The G_{min}, G_{max}, and $\alpha_{g,i}$ values are calculated as follows [30],

$G_{\min} = (i-1) - \frac{t_D + w_{BO} - 1}{t_{RAS}}$, $G_{\max} = i - \frac{t_D + 1}{t_{RAS}}$.

$$\alpha_{g,i} = \begin{cases} \frac{t_g + t_D + w_{BO} - t_{i-1}}{w_{BO}}, & \text{if} \quad G_{min} \leq g \leq i - \frac{t_D + w_{BO}}{t_{RAS}}; \\ \frac{t_{RAS}}{w_{BO}}, & \text{if } i - \frac{t_D + w_{BO}}{t_{RAS}} < g < (i-1) - \frac{t_D}{t_{RAS}}; \\ \frac{t_i - (t_g + t_D)}{w_{BO}}, & \text{if } (i-1) - \frac{t_D}{t_{RAS}} \leq g \leq G_{max}; \\ 0, \text{ otherwise.} \end{cases}$$

Furthermore, the number of IoT devices that failed their j^{th} preamble transmission at the i^{th} RA slot, $L_{i,F}^j$, can be calculated from the relationship $L_i^j = L_{i,S}^j + L_{i,F}^j$, where

$$L_{i,S}^j = \begin{cases} L_i^j e^{-\frac{L_i}{\phi - n_G} p_n}, & \text{if } \sum_{j=1}^{n_{PT}} L_i^j e^{-\frac{L_i}{\phi - n_G} p_n} \leq n_{UL}; \\ \frac{L_i^j e^{-\frac{L_i}{\phi - n_G} p_n n_{UL}}}{\sum_{j=1}^{n_{PT}} L_i^j e^{-\frac{L_i}{\phi - n_G} p_j}}, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $L_i = \sum_{j=1}^{n_{PT}} L_i^j$. Note that, even if the preamble transmission is performed without collision, there is no guarantee on the successful reception of the RA response due to channel impairments as discussed above and the constraint on the maximum number of IoT devices that would be acknowledged within an RA response window, denoted by n_{UL}. Hereafter, term $L_{g,F}^{j-1}$ in (5) can be calculated accordingly.

As mentioned earlier, the transmissions of Msg3 and Msg4 may not always be successful due to channel impairments. A message transmission is considered to be failed if the transmission of Msg3 or MSg4 exceeds n_{HARQ} times. Accordingly, we calculate the error probability of message transmission, $P_{e,MSG}$,

including the hybrid automatic repeat request (HARQ) process as follows,

$$P_{e,MSG} = p_f^{n_{HARQ}} + \sum_{k=0}^{n_{HARQ}-1} p_f^k (1-p_f) p_f^{n_{HARQ}}. \tag{7}$$

6.2.2. Performance Metrics

Using the outcome from the above modeling, we are able to obtain the number of initial arrivals and retransmissions at each RA slot as well as the number of successful and failed devices at each RA slot. Based on this information, closed-form expressions for the three performance parameters of interest are obtained as follows.

Collision probability, denoted as P_c, is the ratio between the number of collided preambles and the total number of preambles transmitted. As the number of collided preambles equals to the total number of preambles minus the number of successful and idle preambles, P_c is obtained as follows

$$P_c = \frac{\sum_{i=1}^{I_R} \left(\hat{\phi} - L_i e^{-\frac{L_i}{\hat{\phi}}} - \hat{\phi} e^{-\frac{L_i}{\hat{\phi}}} \right)}{I_R \hat{\phi}} = \frac{\sum_{i=1}^{I_R} \left(\hat{\phi} - e^{-\frac{L_i}{\hat{\phi}}} \left(L_i + \hat{\phi} \right) \right)}{I_R \hat{\phi}}. \tag{8}$$

In (8), term I_R denotes the number of RA slots inside the observation time duration. Term $\hat{\phi}$ denotes the total number of preambles available for each type of IoT devices under a specific asccess scheme, as explained in Table 2.

Access success probability, denoted by P_s, is the probability that an IoT device successfully completes the RA procedure within n_{PT} transmission attempts. That is, P_s = (total number of successfully accessed devices) / (total number of devices arrived in t_{AP}), as given in (9). Note that an access success means not only a successful preamble transmission but also the completion of all four steps in the RA procedure. Therefore, the number of successfully accessed devices that transmit the j^{th} preamble within the i^{th} RA slot is equal to $L_{i,S}^j (1 - P_{e,MSG})$. Considering that the values for $P_{e,MSG}$ are negligibly low in reality, P_s can be expressed and estimated as follows,

$$P_s = \frac{\sum_{i=1}^{I_R} \sum_{j=1}^{n_{PT}} L_{i,S}^j (1 - P_{e,MSG})}{\hat{L}} \approx \frac{\sum_{i=1}^{I_R} \sum_{j=1}^{n_{PT}} L_{i,S}^j}{\hat{L}}. \tag{9}$$

Average delay for successful devices, denoted by D'_a, equals to the accumulated access delay experienced by those devices which experience successful access divided by the total number of successfully accessed devices. It is given by

$$D'_a = \frac{\sum_{i=1}^{I_R} \sum_{j=1}^{n_{PT}} L_{i,S}^{j} T_n}{\sum_{i=1}^{I_R} \sum_{j=1}^{n_{PT}} L_{i,S}^{j}}, \tag{10}$$

where T_n is the average access delay of a successfully accessed device that performs exactly n preamble transmissions.

Moreover, it is well understood that backoff mechanisms may lead to long delays and induce heavy-tailed delay distributions, especially when the number of competing devices is large. In our schemes, however, the number of preamble transmissions is strictly bounded by a parameter, n_{PT}. Therefore, the time an RA request can wait for access is also bounded by this constraint.

7. Numerical Results and Discussions

This section presents the numerical results obtained from both the analytical model and simulations for an mMTC cell with its parameters configured as listed in Table 3. The analytical results are obtained following the model presented in Sec. 6. For simulations, we develop a program in MATLAB which mimics the behavior of the proposed schemes as well as the baseline scheme for LTE-A based initial access and the GF transmission scheme. The results reported in this section are the average values obtained from a large number of simulation runs for all considered schemes. For traffic arrivals, the Beta distribution based arrival intensity function expressed in (1) is adopted. The performance of the studied schemes is evaluated by configuring ϕ, γ, η, and n_G to certain values according to Table 3 while varying the number of IoT devices, i.e., L, in each case. More specific configuration details will be elaborated when presenting the performance under each scenario. Consequently, each configuration will in turn affect the $\hat{L}$ and $\hat{\phi}$ values in each scheme, as explained in Table 2. In order to

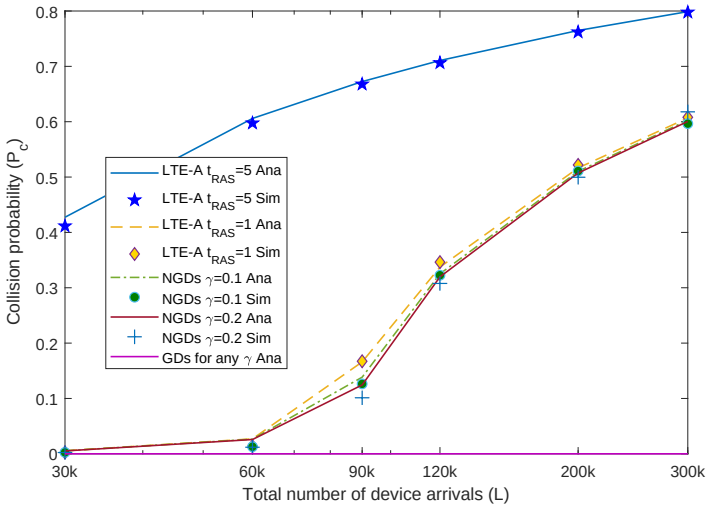

Figure 7: Collision probability in DGDP: GDs versus NGDs.

reflect bursty traffic in *a massive MTC network*, we let L vary from 30k up to 300k which is 10 times as large as what was typically considered in early studies, e.g., in [30] which considered merely an MTC network with a moderate size.

The performance of the proposed schemes is first evaluated and compared with that of the legacy LTE-A RA scheme. Then the access success probability is compared with that of GF transmission. To perform the comparison, we enable two PRACH configurations by selecting the t_{RAS} value alternatively between 5 and 1. When $t_{RAS} = 5$, the access schemes behave as what is commonly used in LTE-A PRACH [11][8], i.e, an IoT device gets an initial access opportunity in every fifth subframe. By configuring $t_{RAS} = 1$, which is a feature supported by multiple PRACH configurations in NR and also supported in LTE-A, IoT devices are entitled to transmit their preambles in every subframe. These two initial access options are illustrated in Fig. 6(a) and Fig. 6(b), respectively.

7.1. DGDP Performance

The performance of the DGDP scheme is evaluated based on the n_G and γ values configured as $\gamma = 0.1, 0.2$ with corresponding $n_G = 5, 10$, respectively. In order to further reduce latency in the 2-step handshake procedure, GDs need

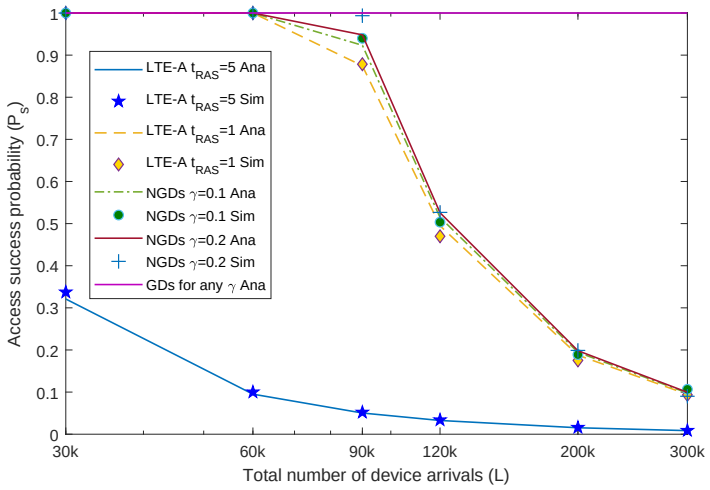

Figure 8: Access success probability in DGDP: GDs versus NGDs.

faster responses from eNB. Accordingly, $w_{RAR} = 2$ and $t_{RAR} = 1$ are configured for the initial access of UDs.

7.1.1. Collision Probability and Access Success Probability

As discussed in Sec. 6, $P_c = 0$ for GDs since the initial access of GDs is contention-free. Furthermore, by allowing up to $n_{PT} - 1$ retransmissions, GDs have guaranteed access even when channel impairment is taken into account, leading to $P_s = 1$. In Fig. 7 and Fig. 8, we depict respectively the collision probability and access success probability achieved by DGDP, for both GDs and NGDs, and compare them with the performance of the legacy LTE-A scheme. It is evident that, in addition to the guaranteed performance of GDs, NGDs have also achieved better or much better performance over the legacy scheme for both γ values. The same observation applies to the other figures illustrated later in this section, even though not explicitly highlighted in result discussions.

For NGDs, P_c monotonically increases as the number of IoT devices, L, grows. With a large device population, a higher number of devices will select the same preamble and transmit it in the same RA slot, resulting in collisions. The collided transmissions prompt more retransmissions, leading to further collisions

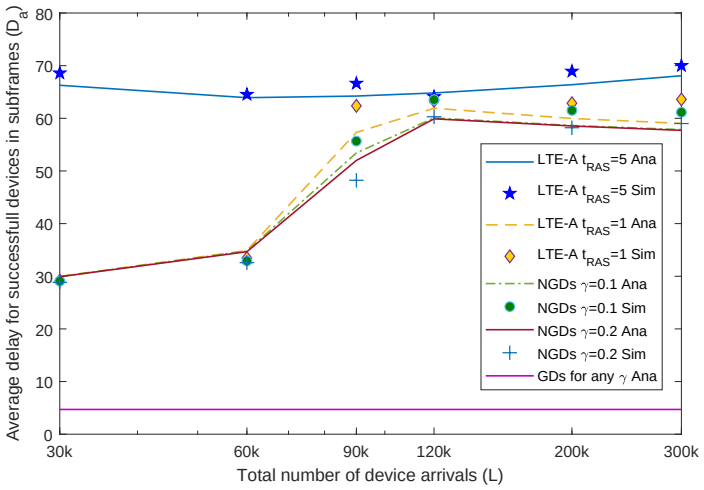

Figure 9: Average delay of the successfully accessed devices in DGDP.

per RA slot. As a result, P_s for NGDs decreases with a larger L. With $\gamma = 0.2$, which means that more IoT devices are grouped in comparison with $\gamma = 0.1$, the performance of both metrics is marginally better. This is due to the fact that, although the number of competing NGDs is reduced with a larger γ, the number of available preamble, $\hat{\phi}$, has also shrunk, leading to limited performance gain. In Subsection 7.4 below, we will further elaborate this relationship.

7.1.2. Average Delay for Successfully Accessed Devices

The average delay for the successfully accessed GDs obtained based on (3) equals approximately to 5 subframes according to our parameter configuration. This is significantly lower in comparison with the delay that a successful IoT device would experience without grouping, i.e., via LTE-A based access, as presented in Fig. 9.

Note that the delay behavior of the GDs is governed by (3) and it is independent of the number of IoT devices in the group. For NGDs, in all configurations except when $t_{RAS} = 5$ for LTE-A, the average delay of the successfully accessed devices increases up to when there are $L = 120$k devices. Beyond this point, the average delay exhibits a slightly descending trend. This behavior can be

explained by referring back to Fig. 8 which shows that the P_s values obtained at 200k is approximately 1/3 of the value at 120k. That is, the total number of successful devices is much lower at 200k in comparison with when there are 120k IoT devices. Among these successful ones, transmission successes occur at the initial or final phase of an arrival burst since heavy losses happened during the peak of the burst. In other words, the successful devices have experienced relatively low access delays, leading to a slightly lower average delay.

7.2. RAUG Performance

The performance of RAUG needs to be evaluated with respect to UDs and NUDs. As the number of UDs and NUDs depends on the value of η, we evaluate the impact of η on the performance of each type of IoT devices.

As introduced in Sec. 5, UDs and NUDs transmit their preambles in separate RA slots of the same subframe. This enables eNB to recognize UDs from the arriving RA slot in a subframe and to perform the remaining handshake steps faster. For this purpose, we adopt two different timing values for UDs and NUDs in our network configuration. This is a reasonable approach since UDs require minimum latency. The flexible frame structure in NR with shorter TTI values also enables such a privilege for UDs. Accordingly, the backoff window size w_{BO} is reduced to 1 in order to speed up the retransmission process in case of a transmission failure due to collisions or channel impairment. Furthermore, we configure the w_{RAR} value as $w_{RAR} = 2$ [19]. In addition to LTE-A with $t_{RAS} = 1$, we have considered another scheme that follows the legacy LTE-A access procedure but allows two RA slots within a subframe for the purpose of further comparison. Hereafter this scheme is referred to as *legacy 5G* as this configuration is possible considering the flexibility provided by the 5G NR frame structure. Note however that although RAUG also provides two RA slots per subframe, each type of devices (UD or NUD) has only one RA slot available within one subframe.

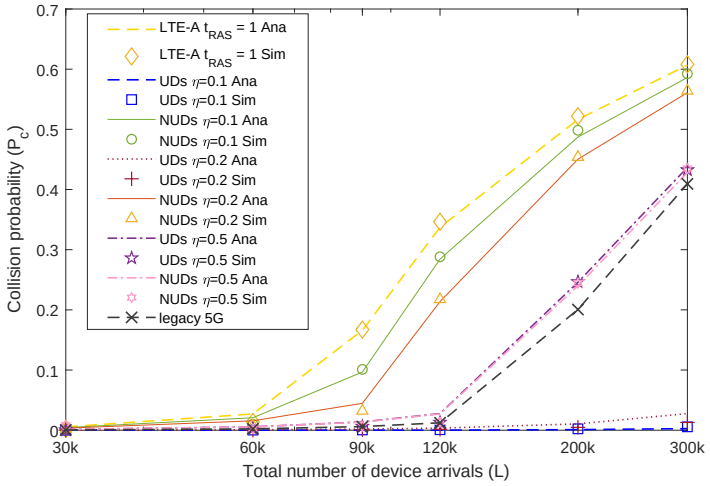

Figure 10: Collision probability in RAUG: GDs, UDs, versus NUDs.

7.2.1. Collision Probability and Access Success Probability

As expected, UDs achieve lower P_c and higher P_s for all ranges of L when $\eta = 0.1, 0.2$, as illustrated in Figs. 10 and 11. Having all ϕ preambles available for access competition of a small fraction of L enables such significant improvements. On the other hand, the performance of NUDs deteriorates with larger L values. However, NUDs still exhibit better performance when compared with the baseline scheme and also with NGDs when γ is configured with the same value as η. This comparison will be further discussed in Subsec. 7.5. For an extreme case with $\eta = 0.5$, the performance of UDs also degrades when $L > 120$k. However, this configuration will significantly improve the performance of NUDs as the number of NUDs would reduce substantially. In Subsec. 7.4, the performance tradeoff between UDs and NUDs with respect to the value of η will be further elaborated.

As shown in Figs. 10 - 11, the performance of the legacy 5G scheme is similar to that of the UDs and NUDs given that $\eta = 0.5$. Since legacy 5G does not employ device grouping, the number of devices competing for RA slots is twice as many as for UDs and NUDs with $\eta = 0.5$. At the same time, the total amount

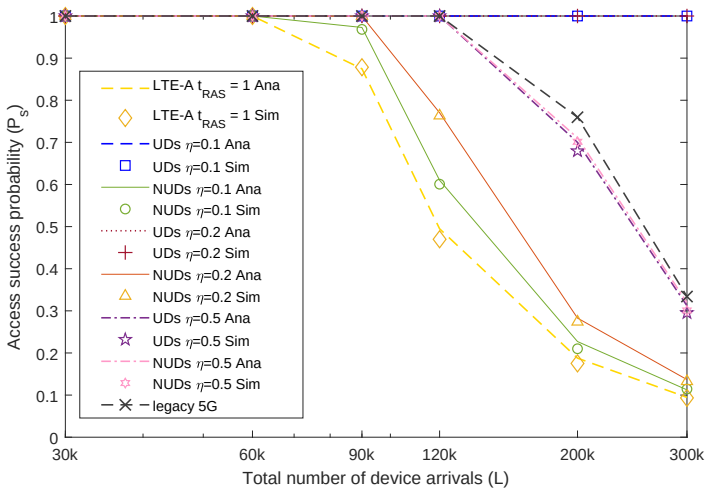

Figure 11: Access success probability in RAUG: GDs, UDs, vs. NUDs.

of available resources for legacy 5G is also doubled for UDs and NUDs with the same η value due to the fact that there are two RA slots within each subframe. Accordingly, the amount of resources used by each device type is half of what is available for legacy 5G. Therefore, the performance of these three schemes is similar based on the given configuration. From these figures, it is clear that the UDs still exhibit better performance when the $\eta = 0.1$ or 0.2 thanks to the concept of having separate resources for URLLC traffic.

7.2.2. Average Delay for Successfully Accessed Devices

As shown in Fig. 12, when $\eta = 0.1, 0.2$, the achieved average delay for UDs is approximately 10 subframes and this value keeps comparatively stable regardless of the IoT device population. With a low collision probability as presented above, devices can transmit their preambles successfully with a low number of transmission attempts, resulting in reduced overall delay. Additionally, the shortened response time configured for UDs further contributes to latency reduction. Compared with UDs, and legacy 5G, NUDs have a significantly higher delay and the corresponding value generally increases with a higher L. However, in comparison with the baseline scheme, NUDs still attain lower latency. When

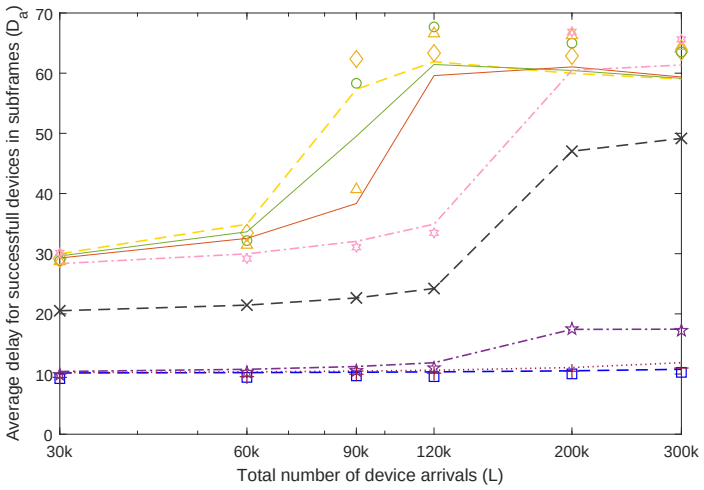

Figure 12: Average delay of the successfully accessed devices in RAUG (The legend is identical to the ones shown in Fig.11).

$\eta = 0.5$, which indicates lesser competition among NUDs, shorter latency for NUDs is achieved at a cost of slightly increased latency for UDs.

7.3. HS Performance

The performance of the HS scheme is illustrated in Figs. 13 - 15. It is clear that the performance of GDs in HS is similar to what is observed in DGDP. Furthermore, UDs, which are a subset of NGDs, exhibit also similar performance as what is observed in the RAUG scheme. Recall, however, that the number of competing IoT devices in each device type will be different when NGDs are categorized into UDs and NUDs.

As a result, NUDs in HS achieve much better performance compared with NUDs in RAUG and NGDs in DGDP even though their available number of preamble, $\hat{\phi}$, is lower than in RAUG or DGDP. Furthermore, since both GDs and UDs coexist in HS, a much larger number of devices with URLLC requirements can be accommodated when HS is employed. Observe Fig. 14 and take $L =$ 200k, $\gamma = \eta = 0.3$, and $n_G = 15$ as an example. The total number of IoT devices that achieve $P_s = 1$ would be as many as 102k including $\gamma L = 60$k GDs grouped in 15 groups plus $\eta(1-\gamma)L = 42$k UDs.

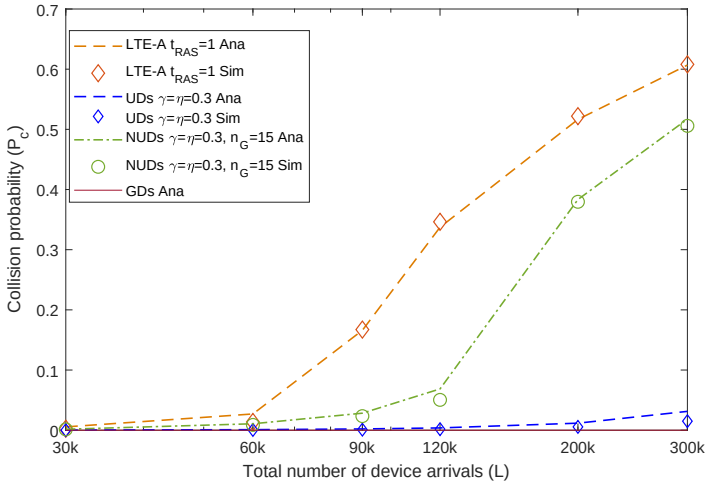

Figure 13: Collision probability in HS: GDs, UDs, versus NUDs.

7.4. The Impact of γ, η, and n_G

As mentioned earlier, the values of γ, η, and n_G are reconfigurable. In a cell with other parameters like L and ϕ fixed, the adopted values of these three variables play a significant role in determining the performance of the proposed schemes. A higher γ value means a larger number of GDs and accordingly n_G also needs to be enlarged. The performance of NGDs in DGDP depends on the *joint configuration* of γ and n_G values. Similarly, increasing η would lead to a higher number of UDs in RAUG indicating more competition among UDs and better access opportunities for NUDs, respectively.

To achieve optimal performance from the proposed initial access schemes, it is vital to configure network parameters appropriately so that, while GDs and UDs enjoy URLLC service, NGDs and NUDs could also achieve better or at least similar performance in comparison with the baseline scheme. Observing the presented numerical results for DGDP, it is evident that the selected γ values satisfy the criterion given in (2). Any violation of this criterion would deteriorate the performance of NGDs as further discussed in [8]. Furthermore, the impact of η values on the performance of NUDs has a simpler proportional relationship. Whenever η is increased, NUDs will obtain better performance owing to reduced

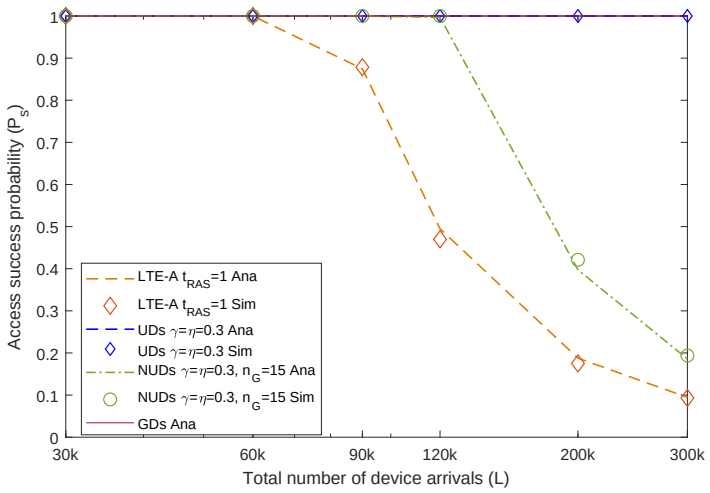

Figure 14: Access success probability in HS: GDs, UDs, versus NUDs.

competition, as observed in the numerical results for RAUG. However, η *should only be enlarged to a level up to which the required performance for UDs is still guaranteed.*

7.5. Performance Comparison among Our Schemes and versus LTE-A

As demonstrated above, the proposed schemes outperform the baseline scheme under all studied configurations. To elaborate the performance distinctions, we further differentiate the results obtained from the baseline scheme with two configuration options, i.e., when the interval between two successive RA slots is configured as $t_{RAS} = 5$ and $t_{RAS} = 1$, respectively.

The baseline scheme with $t_{RAS} = 5$ performs worst among all the studied schemes. Although this configuration is commonly adopted in LTE-A, our results reveal that this is not an effective option when the number of IoT devices could increase promptly, e.g., under mMTC bursty traffic scenarios. When $t_{RAS} = 1$, the performance of the baseline scheme improves significantly, thanks to a much higher number of RA slots (10000 for $t_{RAS} = 1$ versus 2000 for $t_{RAS} = 5$) available for preamble transmissions of arriving devices. However, when the number of IoT devices is very large, i.e., $L >$ 90k, the performance

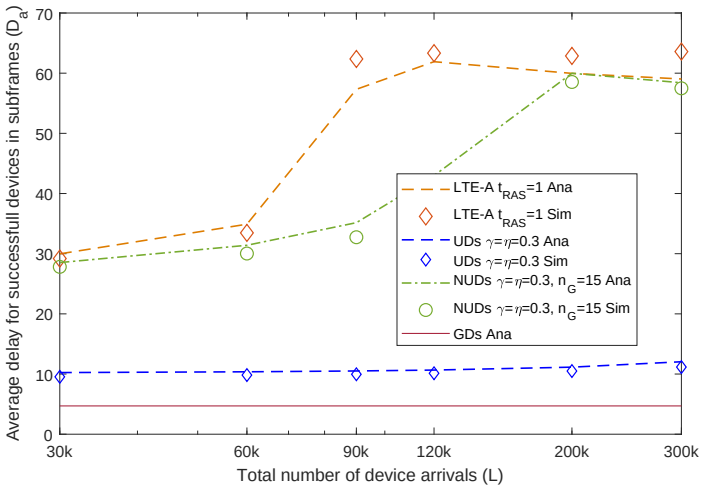

Figure 15: Average delay of the successfully accessed devices for different types of devices in HS.

of this configuration also degrades more seriously than what is achieved in our proposed schemes.

Among the proposed schemes, DGDP provides the best URLLC performance for GDs since GDs always enjoy guaranteed access privilege with null collision based on their contention-free access. The proposed 2-step handshake procedure combined with lower response times further reduces the latency for GDs. The performance of UDs in the HS and RAUG schemes is better than any NUDs or NGDs in all cases. UDs benefit from the proposed dedicated RA slots with reduced latency obtained by allowing multiple slots inside one subframe for preamble transmission and also from the shortened response times. However, the performance of UDs in RAUG is not as superb as GDs in DGDP since UDs in RAUG still need to follow the 4-step RA procedure and to compete with other UDs. Nevertheless, unlike GDs, UDs have more flexibility in terms of device implementation and the support of various IoT applications (since no pre-grouping is required and no requirement on static deployments). Moreover, with the same γ and η configuration, NUDs in RAUG achieve generally better performance in comparison with NGDs. Since two dedicated RACH slots are

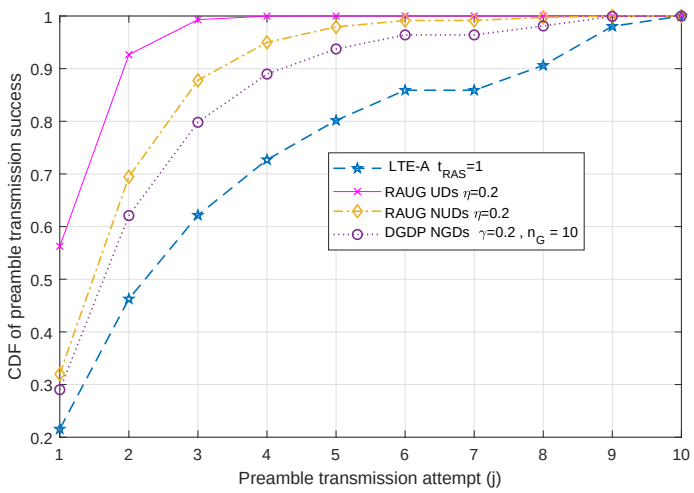

Figure 16: CDF of successful preamble transmissions for different types of devices under LTE-A, DGDP, and RAUG respectively.

enabled inside a subframe and no preambles are pre-allocated to GDs, the access opportunities for NUDs are based on all ϕ preambles. In contrast, NGDs in DGDP have only $(\phi - n_G)$ preambles, leading to slightly degraded performance in comparison with NUDs in RAUG.

Furthermore, for the purpose of performance comparison under a medium size device population, we reconfigure the network as $L = 90$k, $\gamma = \eta = 0.2$, and $n_G = 10$. In Fig. 16, we illustrate the cumulative distributed function (CDF) of successful preamble transmissions for different types of IoT devices under LTE-A, DGDP, and RAUG, respectively. As can be observed, almost all UDs in RAUG obtain network access within three preamble transmissions. Moreover, NGDs and NUDs have also achieved higher CDF values compared with the baseline scheme. With a cross-reference of the respective P_s values in Fig. 16, we ascertain that DGDP and RAUG provide faster access to the network than the baseline scheme does. For instance, to achieve $P_s = 95\%$, NUDs in RAUG need on average merely 4 preamble transmissions whereas about 6 and $7 \sim 8$ transmissions are required for NGDs in DGDP and devices in LTE-A respectively. In Fig. 17, we further illustrate the CDF of the access latency

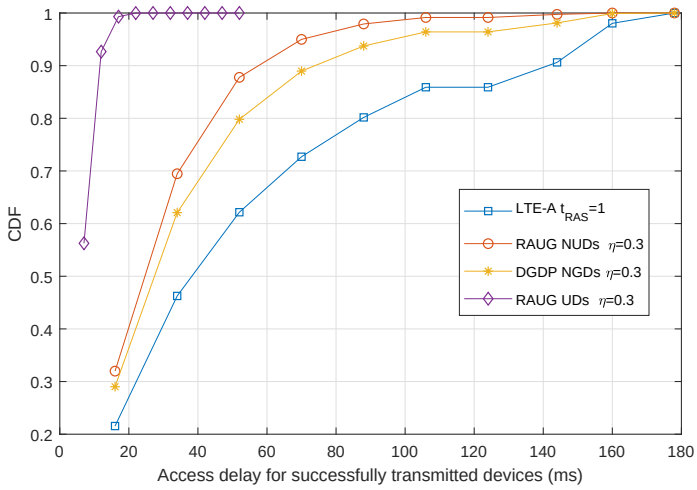

Figure 17: CDF of access delay for successful UDs, NGDs, and NUDs: Comparison of RAUG, DGDP, and LTE-A.

experienced by successfully transmitted devices in milliseconds based on the four studied schemes. It is evident that all devices in our schemes including UD, NUDs, and NGDs have achieved better performance in comparison with that of LTE-A and among them UDs obtain the best performance.

Moreover, HS offers best opportunities to all types of IoT devices owing to its hybrid nature. When HS is employed, both GDs and UDs could coexist without compromising each other's performance, thus supporting a higher number of IoT devices with URLLC requirements. Although NUDs in HS possess a smaller set of preambles, i.e., $(\phi - n_G)$, the same as NGDs in DGDP, the number of NUDs is meanwhile significantly reduced to $(1-\eta)(1-\gamma)L$ which is lower than that of NGDs in DGDP, i.e., $(1-\gamma)L$. In this way, the performance of NUDs in HS is also improved.

7.6. Access Success Probability Comparison with Grant-free Transmission

As mentioned in Subsec. 2.2, GF transmission appears as an attractive mechanism for data reporting in various mMTC and URLLC scenarios, especially for small data and sporadic traffic. In this subsection, we compare through simulations the performance of the proposed schemes with GF in terms of access

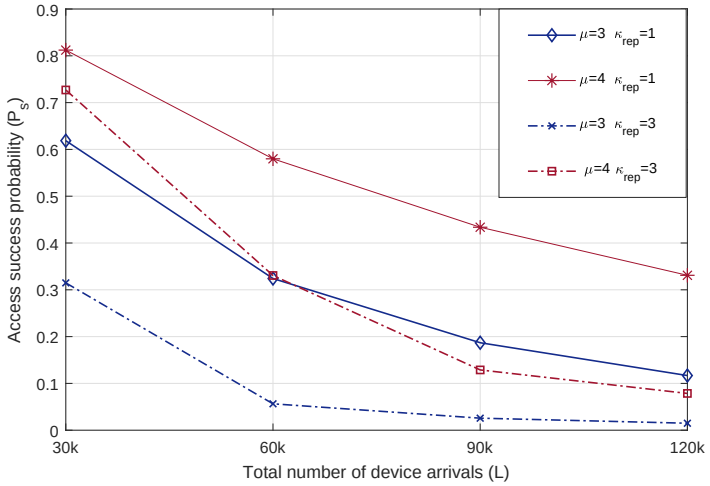

Figure 18: Access success probability for GF transmissions under bursty traffic.

success probability by considering two numerologies $\mu = 3$ and $\mu = 4$ which have 8 and 16 slots respectively. To maximize resource allocation for GF transmissions, we assume that all these available slots can be utilized by GF traffic. For GF transmissions, we adopt a popular transmission scheme known as *k repetitions* [46]. Accordingly, a number of k_{rep} replicas of the same packet will be transmitted within a subframe. A packet transmission is regarded as successful if at least one of these k_{rep} transmissions is successful.

For GF transmissions, all devices that arrived during a given subframe will compete for transmission in the next subframe. Each device will randomly select one or more (if $k_{rep} > 1$) slots based on the configuration and transmit k_{rep} replicas of its packet in the selected slot(s) *within the same subframe*. A collision happens if two or more devices have selected the same slot for transmitting any replica of their packets.

Fig. 18 illustrates the obtained access success probability of GF devices according to a bursty traffic arrival pattern which was presented in Subsec. 3.3. As expected, the success probability monotonically decreases with a higher number of device arrivals. When comparing the results for $\mu = 3$ and $\mu = 4$, it is clear that providing a higher number of slots for GF data transmission would result

in a higher access success probability. On the other hand, it is counter-intuitive that having a higher number of repetitions does not help to increase access success. This is because with $k_{rep} = 3$, the number of competing transmissions per slot increases, leading to an even lower success probability.

Finally, let us compare the access success probability achieved by GF devices with what is achieved in the proposed DGDP and RAUG schemes which belong to GB schemes (with the results shown in Fig. 8 and Fig. 11 respectively). By comparing the curves in these figures, it is evident that the GB schemes perform better. This is because during a traffic burst, a very higher number of arrivals within a short interval have occurred, causing a higher number of collisions for both GB access and GF transmissions. Initially, the number of arrivals for each subframe is the same for both GB and GF. Although a GB scheme has to deal with retransmissions, it has the advantage of transmitting up to n_{PT} transmissions across multiple subframes. On the other hand, a *k repetitions* GF scheme has to finish all k_{rep} transmissions within one subframe without the possibility of retransmissions. As a consequence, GF transmissions experience higher collisions than GB transmissions, resulting in a lower access success probability. Based on this observation, we ascertain that, although GF communication reduces extra overhead by skipping the initial access phase and it provides lower latency when traffic load is light, it is not better suited for providing URLLC services *in the presence of bursty traffic with a high number of arrivals.*

7.7. Further Discussions

The proposed schemes are developed based on multiple assumptions as presented above. For instance, the procedures for intra-group communications between group members and their group leader are not included in our scheme design. Nor is the coordination between a serving group leader and its backup group leader considered. In spite of having a very lower probability, it is not impossible that neither of the group leaders sensed an event or the transmissions

of both leaders failed. If such an extreme case occurs, extra intra-group communication is needed. Although intra-group communications could be performed with or without the involvement of downlink message coordinations through a gNB, extra protocol overhead and longer delay are unavoidable. As such, the reported results in this section represent an upper bound of the performance of our schemes.

8. Conclusions and Future Work

Targeting at two massive IoT traffic scenarios, we have proposed in this paper three LTE-A or 5G NR based initial access schemes which provide URLLC access to a selected portion of mMTC devices. The schemes were developed by considering various mission critical and cyber-physical IoT applications envisaged by 3GPP. The first scheme, DGDP, provides contention-free access with low latency to grouped IoT devices based on dedicated preamble reservation. The second scheme, RAUG, is still contention based but facilitates reserved random access slots allowing multiple occurrences inside each subframe and hence produces lower latency and very high access success probabilities to those IoT devices with URLLC requirements. The third scheme, HS, combines the merits of these two schemes and provides more flexibility to a larger number of URLLC devices as well as non-grouped and non-URLLC devices. Furthermore, the performance of all four types of IoT devices under these three schemes has been evaluated based on both analysis and simulations, in comparison with the legacy LTE-A initial access as well as grant-free transmission. Through performance comparison, we demonstrate that, by fine-tuning a few configurable network parameters, the proposed schemes are able to provide ultra-high reliability and low latency to grouped devices and URLLC devices while still improving the performance of non-grouped and non-URLLC devices. As future work, we will further study both inter- and intra-group communications in a two-tier architecture for mMTC networks, intra-group communications among devices and group leaders, and initial access for beyond 5G networks together with data transmission

and radio resource allocation after the initial access phase. For protocol design, we will also consider more realistic channel conditions, the constraint of radio resource blocks, as well as minimized extra protocol overhead.

References

[1] H. Habibzadeha, T. Soyataa, B. Kantarci, A. Boukerche, C. Kaptan, Sensing, communication and security planes: A new challenge for a smart city system design, Comput. Netw. 144 (2018) 163–200.

[2] O. Galinina, S. Andreev, M. Komarov, S. Maltseva, Leveraging heterogeneous device connectivity in a converged 5G-IoT ecosystem, Comput. Netw. 128, (2017) 123-132.

[3] 3GPP TS 22.368, Service requirements for machine-type communications (MTC); Stage 1, R15, v15.0.0, 2019.

[4] M.S. Ali, E. Hossain, D.I. Kim, LTE/LTE-A random access for massive machine-type communications in smart cities, IEEE Commun. Mag. 55 (1) (2017) 76–83.

[5] G. Hampel, C. Li, J. Li, 5G ultra-reliable low-latency communications in factory automation leveraging licensed and unlicensed bands, IEEE Commun. Mag. 57 (5) (2019), 117–123.

[6] S. Zhang, Y. Wang, W. Zhou, Towards secure 5G networks: A Survey, Comput. Netw. 162 (2019) 1–22.

[7] G.J. Sutton, J. Zeng, R.P. Liu, W. Ni, D.N. Nguyen, B.A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, T. Lv, Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives, IEEE Commun. Surv. Tut. 21 (3) (2019) 2488–2524

[8] T.N. Weerasinghe, I.A.M. Balapuwaduge, F.Y. Li, Preamble reservation based access for grouped mMMTC devices with URLLC requirements, in: Proc. IEEE ICC, 2019, pp.1-6.

[9] M. Bennis, M. Debbah, H.V. Poor, Ultra reliable and low-latency wireless communication: Tail, risk, and scale, Proc. IEEE, 106 (10) (2018) 1834–1853.

[10] 3GPP TS 22.104, Service requirements for cyber-physical control applications in vertical domains, R17, v17.0.0, 2019.

[11] 3GPP TR 37.868, Study on RAN improvements for machine type communications, R11, v11.0.0, 2011.

[12] 3GPP TS 36.321, Evolved universal terrestrial radio access (e-UTRA), R15, v15.6.0, 2019.

[13] P. Castagno, V. Mancuso, M. Sereno, M.A. Marsan, Limitations and sidelink-based extensions of 3GPP cellular access protocols for very crowded environments, Comput. Netw. 168, (2020) 1–15.

[14] M. Tavana, A. Rahmati, V. Shah-Mansouri, Congestion control with adaptive access class barring for LTE M2M overload using Kalman filters, Comput. Netw. 141 (2018) 222–233.

[15] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks, in: Proc. IEEE ICC, 2016, pp. 1–6.

[16] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, N. Alonistioti, C. Zhou, J. Eichinger, Ö. Bulakci, On the way to massive access in 5G: Challenges and solutions for massive machine communications, in: Proc. International Conference on Cognitive Radio Oriented Wireless Networks (CROWNCOM), 2015, pp. 708–717.

[17] B. Han, H.D. Schotten, Grouping-based random access collision control for massive machine-type communication, in: Proc. IEEE GLOBECOM, 2017, pp. 1–7.

[18] H. Seo, J. Hong, W. Choi, Low latency random access for sporadic MTC devices in Internet of things, IEEE Internet Things J. 6 (3) (2019) 5108-5118.

[19] 3GPP TS 38.213, NR; Physical layer procedures for control, R15, v15.6.0, 2019.

[20] G. Sanfilippo, O. Galinina, S. Andreev, S. Pizzi, G. Araniti, A concise review of 5G new radio capabilities for directional access at mmWave frequencies, in: Proc. International Conference on Next Generation Wired/Wireless Advanced Networks and Systems (NEW2AN), 2018, pp. 340–354.

[21] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, H. Wei, 5G new radio: Waveform, frame structure, multiple access, and initial access, IEEE Commun. Mag. 55 (6) (2017) 64–71.

[22] 3GPP TS 38.211, NR; Physical channels and modulation, R16, v16.1.0, 2020.

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

[24] 3GPP TR 38.824, Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC), R16, v16.0.0, 2019.

[25] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H.V. Poor, B. Vucetic, High-reliability and low-latency wireless communication for Internet of things: Challenges, fundamentals and enabling technologies, IEEE Internet Things J. 6 (5) (2019) 7946–7970.

[26] A. Azari, P. Popovski, G. Miao, C. Stefanovic, Grant-free radio access for short-packet communications over 5G networks, in: Proc. IEEE GLOBECOM, 2017, pp. 1-7.

[27] A.T. Abebe, C.G. Kang, Comprehensive grant-free random access for massive and low latency communication, in: Proc. IEEE ICC, 2017, pp. 1-6.

[28] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme, IEEE Trans. Wireless Commun. 16 (12) (2017) 7785–7799.

[29] P. Zhou, H. Hu, H. Wang, H.H. Chen, An efficient random access scheme for OFDMA systems with implicit message transmission, IEEE Trans. Wireless Commun. 7 (7) (2008) 2790-2797.

[30] C. Wei, G. Bianchi, R. Cheng, Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks, IEEE Trans. Wireless Commun. 14 (4) (2015) 1940–1953.

[31] 3GPP TS 38.331, NR; Radio resource control (RRC) protocol specification, R15, v15.6.0, 2019.

[32] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. Vidal, V. Casares-Giner, L. Guijarro, Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic, IEEE Trans. Veh. Technol. 67(4) (2018) 3505-3520.

[33] J. Li, Q. Du, L. Sun, P. Ren, Queue-aware joint ACB control and resource allocation for mMTC networks, in Proc. IEEE GLOBECOM Workshops, 2018, pp. 1-6.

[34] L. Liang, L. X. B. Cao, Y. Jia, A cluster-based congestion-mitigating access scheme for massive M2M communications in Internet of things, IEEE Internet Things J. 5(3) (2018) 2200–2211.

[35] F. Wu, B. Zhang, W. Fan, X. Tian, S. Huang, C. Yu, Y. Liu, An enhanced random access algorithm based on the clustering-reuse preamble allocation in NB-IoT system, IEEE Access 7 (2019) 183847–183859.

[36] T. Kim, H. S. Jang, D. K. Sung, An enhanced random access scheme with spatial group based reusable preamble allocation in cellular M2M networks, IEEE Commun. Lett., 19(10) (2015) 1714–1717.

[37] Q. Pan, X. Wen, Z. Lu, W. Jing, L. Li, Cluster-based group paging for massive machine type communications under 5G networks, IEEE Access 6 (2018) 64981–64904.

[38] S. Sesia, I. Toufik, M. Baker, *LTE - the UMTS long term evolution: From theory to practice*, 2nd Edition, John Wiley & Sons Ltd., 2011.

[39] M. Rahnema, M. Dryjanski, *From LTE to LTE-Advanced Pro and 5G*, Artech House, 2017.

[40] 3GPP TS 36.331, Radio resource control (RRC); Protocol specification, R15, v15.8.0, Dec. 2019.

[41] I. Leyva-Mayorga, C. Stefanovic, P. Popovski, V. Pla, J. Martinez-Bauset, Random access for machine-type communications, Wiley 5G Ref: The Essential 5G reference Online, 2019.

[42] O. Arouk, A. Ksentini, General model for RACH procedure performance analysis, IEEE Commun. Lett. 20 (2) (2016) 37275.

[43] T. Weerasinghe, I. A. M. Balapuwaduge, F. Y. Li, Supervised learning based arrival prediction and dynamic preamble allocation for bursty traffic, in Proc. IEEE INFOCOM Workshops, 2019, pp.1-6.

[44] A. Azari, M. Ozger, C. Cavdar, Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning, IEEE Commun. Mag. 57(3) (2019) 4248.

[45] 3GPP RP-191677, Revised work item proposal: 2-step RACH for NR, 3GPP TSG RAN Meeting #85, 2019.

[46] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, T. H. Jacobsen, Uplink grant-free random access solutions for URLLC services in 5G new radio, in Proc. IEEE ISWCS, 2019, pp. 607-612.