DQ-Based Random Access NOMA for Massive Critical IoT Scenarios in 5G Networks

Mohammad Reza Amini, *Senior Member, IEEE,* Ala'a Al-Habashna, *Member, IEEE,* Gabriel Wainer, *Senior Member, IEEE*, and Gary Boudreau, *Senior Member, IEEE*

Abstract—

Internet-of-Things (IoT) networks provide massive connectivity for many application scenarios. Recently, much work has been dedicated to develop spectrum access strategies for IoT networks with a massive number of nodes and sporadic data traffic behavior. The case becomes more challenging in critical applications when Ultra-Reliable Low-Latency (URLL) transmissions are required. Such networks entail spectrum-efficient transmission schemes in which Non-Orthogonal Multiple-Access (NOMA) is considered a key enabler. We proposed a Distributed Queuing (DQ) approach in NOMA for critical massive IoT (mIoT) applications. More specifically, we introduce a frame structure to support DQ-based NOMA so that dynamic NOMA clustering (at the nodes) and dynamic Successive Interference Cancellation (SIC) ordering at the Base Station (BS) are supported. We also use adaptive power back-off strategy to reduce power collisions by utilizing both nodes' and clusters' activation index. We investigate network performance metrics, such as reliability, delay violation probability, and effective sum rate. These metrics are derived analytically, and the effect of different network parameters such as blocklength, active node arrival rate, and the number of contention subslots on the network metrics are investigated and compared with the S-ALOHA-TD benchmark.

Index Terms—Distributed Queuing, Internet-of-Things, Packet Latency, NOMA, Random Access, Reliability.

1 Introduction

Internet of Things (IoT) networks are on the way to provide massive connectivity in extensive emerging mission-critical applications and use cases such as tactile Internet (involving remote motion control, telesurgery, etc.), factory automation, Industrial IoT (IIoT), and those under the Industry 4.0 paradigm [1], [2], [3], [4]. By the end of 2023, Cisco estimates 13.1 billion mobile users. The number of Internet-enabled devices is projected to increase from 18.4 billion in 2018 to 29.3 billion in 2023 [5]. Such an explosive growth in the number of wireless devices and diverse wireless services comes with challenges in designing network structure, protocols, and access strategies [6]. Therefore, providing massive and spectrum-efficient access techniques is one of the most important targets for the next generation of wireless networks, namely the sixth-generation. Non-Orthogonal Multiple Access (NOMA) techniques promise to be a key enabler to improve spectrum efficiency [7].

On the other hand, nodes in massive IoT (mIoT) applications (also called End Devices (EDs) or User Equipment (UEs)) have sporadic data traffic behavior which makes the use of the Random Access (RA) strategies inevitable since devoting time-frequency resources to each node is inefficient. However, conventional RA techniques (e.g. those defined in the current RA-LTE standard) cannot be used for critical low-latency transmissions due to grant acquisition delay and the excessive signaling overhead [8].

Furthermore, sporadic data traffic pattern creates new challenges in NOMA-based transmission schemes since dynamic NOMA clustering entails active user detection, dynamic Successive Interference Cancellation (SIC) ordering, and dynamic transmit power adjustment [6].

Our research aims to propose a transmission scheme that exploits a collision resolution strategy into NOMA in order to achieve a random access protocol suitable for massive critical IoT. Due to the sporadic data traffic behavior by the users, the proposed framework provides active user detection, dynamic NOMA clustering and SIC ordering, as well as adaptive power control based on the active UE index and number of active UEs in the NOMA cluster.

In the following subsections we discuss the background of our research, give further detail on our motivation and goals, and briefly discuss the contributions of this work.

1.1 Background

NOMA has been identified as one of the key enabling technologies improving spectrum efficiency and throughput [9], [10] which are crucial for Next-Generation mIoT networks. NOMA also allows to fulfill the latency targets for low-latency communications [11], [12], [13]. Furthermore, next generation mIoT networks need to exploit random access techniques to support the sporadic transmissions of each node [14], [15].

To avoid collisions due to overload in RA-LTE, Access Class Barring (ACB) schemes have been proposed in the lit-

- *Mohammad Reza Amini is with the Department of Systems and Computer Engineering, Carleton University.*
  E-mail: Mohammadrezaamini@cunet.carleton.ca
- *Ala'a Al-Habashna is with the Department of Systems and Computer Engineering, Carleton University.*
  E-mail: AlaaAlHabashna@cmail.carleton.ca
- *Gabriel Wainer is with the Department of Systems and Computer Engineering, Carleton University.*
  E-mail: gwainer@sce.carleton.ca
- *Gary Boudreau , Senior Member Ericsson Canada, Ottawa, ON, L4W 5K4, Canada.*
  E-mail: gary.boudreau@ericsson.com

Manuscript received ....; revised ...

erature. The idea is that each node initiates its RA procedure based on the *barring factor* configured by the Base station (BS) [16].

Dynamic ACB with an adaptive barring factor based on load estimation and the number of backlogged nodes has been proposed to improve the performance of ACB techniques [17], [18], [19]. RA-NOMA techniques have also been proposed to exploit NOMA advantages in RA networks. Non-orthogonal random access (NORA) in cellular-based Machine Type Communication (MTC) has been proposed in [6]. The scheme has five steps, namely, cluster establishment by the BS, preamble transmission by cluster heads, Random Access Response (RAR) by BS, power adjustment and data transmission by the nodes, and performing SIC and sending acknowledge (ACK) by BS. The optimum power for each node is then formulated to maximize the Energy Efficiency (EE). Exploiting the difference of Time of Arrival (ToA), the study in [20] has identified several UEs with the same preamble in a NOMA scenario. The analyses show that the proposed structure can serve 30% more users compared to conventional RA schemes. In a 5G scenario, a NORA protocol has been proposed in [21]. By exploiting channel inversion technique, the UEs adjust their transmit power so that their received power at the BS lies in one of two predefined levels. The throughput of the proposed scheme has been shown to be about twice the conventional S-ALOHA scenario. In [22], NOMA-based S-ALOHA with a p-persistent strategy was proposed. The UEs set their transmit power randomly at one of the two prescribed levels with specified probabilities. Extending the channel inversion strategy proposed in [21], [23] has attempted to use multichannel selection diversity, enabling the users to select the best available channel. This helps the users avoid transmitting data with unacceptable transmit power levels which improves the network EE. In [24], the authors have generalized the channel inversion technique for L target levels of transmit power. More specifically, each UE that has data to transmit, adjusts its transmit power randomly based on one of L specific target values. In [25], a NOMA-based random access scheme with two levels of priority has been proposed in an MTC scenario. Each type of device has been allocated a specific preamble set. Delay sensitive devices (high priority) select their preambles from an orthogonal set. Due to their orthogonality, preamble collision is reduced, and preamble re-transmissions is avoided, decreasing the net access delay. To lower the computational complexity, delay-tolerant devices (low priority) are allocated non-orthogonal preamble set yet with larger size. The detection performance of transmitted preambles by the two device types has been explored and the low-complexity preamble detection method has been further proposed. In [26], the performance of NOMA-based RA scheme consisting of two users including near and far users in Finite Block Length (FBL) regime has been analyzed over Nakagami-m fading channels. The authors have shown that the packet error rate for both near and far users is improved by employing ARQ strategy. The optimal blocklength has then been achieved to maximize the throughput.

It is worth noting that the data traffic behavior in almost all the mentioned studies has been considered saturated and non-stochastic. However, it is of prominent importance in mMTC and mIoT to consider the sporadic data traffic nature of the nodes. Furthermore, as mentioned, sporadic traffic behavior of the nodes brings major challenges in RA-NOMA-based scenarios including dynamic NOMA clustering, active node detection, and adaptive transmit power control. Note that a practical RA-NOMA that can handle massive communications in critical applications with low-latency and ultra-reliability is still lacking in the literature.

Recently, a class of more viable collision resolution protocols known as Distributed Queuing (DQ) has been proposed to manage massive connectivity without relying on ACB. Such protocols provide a near-optimum contention resolution access scheme [27]. An initial access collision is resolved using a splitting tree in DQ-based protocols, eliminating the instability issues in RA protocols under heavy load. Among the pioneering studies, [28] proposed a Contention Resolution Queue (CRQ) protocol to tackle simultaneous access of a few thousand devices. Authors in [29] proposed DQ-based protocol in RA-LTE in which the BS roughly probes the number of colliding devices. Based on the probing result, the BS randomly divides these devices into several groups and places these groups at the end of logical queue. An analytic model to estimate the average access delay in the proposed protocol was derived. Authors in [30] then extended their previous work in LTE [29] to incorporate priority in the proposed scheme by exploiting information related to congestion level from the DQ process. Using simulation study, they showed the superiority of the proposed algorithm over B-ACB in terms of both access delay and energy consumption. An extension of the DQ algorithm in LoRa networks, namely DQ-LoRa was proposed in [6]. The results were compared with S-ALOHA showing a significant performance improvement in terms of throughput and average delay. At present, [31] is the only study that incorporates DQ in NOMA-based transmissions. The research uses DQ-NOMA in LoRa networks. The random access procedure was discussed in detail and performance was compared using simulations that study throughput and number of users served. Since DQ-based transmission mechanisms have high performance in massive networks, we propose a method to incorporate DQ in RA-NOMA scenarios. This work is mainly different from [31] in that it provides an analytical framework to evaluate the network metrics such as reliability, delay violation probability, and effective sum rate. Furthermore, a frame structure is devised so that dynamic NOMA clustering, active user detection, dynamic SIC ordering, and adaptive power control are supported. Moreover, by exploiting short packet transmissions in finite blocklength regime in the proposed DQ-NOMA, Ultra-Reliable Low-Latency (URLL) applications in B5G and 6G are intended by considering the decoding block error rate.

1.2 Motivation and Contributions

Motivated by the features of both NOMA and contention resolution protocols, our research introduces the DQ mechanism in NOMA-based transmissions, namely DQ-NOMA to support critical IoT with massive number of nodes. Employing DQ in NOMA requires to further devise the basic DQ MAC protocol since the DQ-NOMA protocol needs to cover all the problems in RA-NOMA strategies as well as new challenges emanated from massive number of nodes tha has sporadic data traffic behavior. Particularly, NOMA clustering should be dynamically formed at the beginning of each frame at the transmitter's side. Hence, the BS should be able to perform active node detection and to order SIC dynamically. Furthermore, power control, which profoundly affects the performance of the NOMA-based transmissions, should be done adaptively in the resultant dynamic cluster. To this aim, nodes are divided into geographical clusters to which a unique preamble is assigned from a set of low-dimension orthogonal space to avoid implementation complexity and high processing delay at the BS. Furthermore, the proposed frame structure provides dynamic NOMA clustering (at the nodes side) and dynamic SIC ordering (at the BS side) to fully utilized NOMA capacity in sporadic data traffic environment. Moreover, network metrics such as reliability, delay violation probability, average packet latency, and effective sum rate are derived mathematically to analyze the performance of underlying network in the FBL regime. The impact of different network parameters (i.e., blocklength, active node arrival rate, and the number of contention subslots) on the derived metrics is investigated, and the results are compared to the NOMA-based S-ALOHA with Transmission Diversity (TD). Furthermore, to avoid power collision between the nodes sporadically transmitting data packets, adaptive back-off power strategy is adopted. In such a power control strategy, each node adjusts its transmit power at the beginning of each frame according to its own activation index and the total number of active nodes in its NC.

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

- Introduced a RA-NOMA scheme based on DQ, namely, DQ-NOMA for critical massive IoT networks. Particularly, to incorporate DQ notion into NOMA for mIoT with sporadic traffic behavior, nodes are divided into limited number of geographical clusters to which a preamble (signature) is assigned from a set of orthogonal space. This reduces the signature space and helps lower processing delay in active node detection stage at the BS. Moreover, the frame structure provides dynamic NOMA clustering (at the nodes side) and dynamic SIC ordering (at the BS side) to fully utilize NOMA capacity in sporadic data traffic environment. Furthermore, to avoid power collision between sporadically transmitted data by each node, an adaptive power control strategy is adopted which utilizes both nodes' and clusters' activation index.
- Derived network metrics and analyzed the performance of the proposed DQ-NOMA for URLL-mIoT networks. Specifically, network metrics such as reliability, delay violation probability, average packet latency, and effective sum rate are theoretically derived and the performance of the proposed scheme is analyzed and investigated for different network parameters such as, blocklength, active node's arrival rate, and number of contention sub-slots. Additionally, network metrics under the same assumptions yet for S-ALOHA-TD are derived to make a proper comparison with the proposed structure.

The rest of this paper is organized as follows. Section 2 describes system model. Analytical derivations of the proposed scheme and S-ALOHA with transmission diversity are presented in Section 3 and 4 respectively. Numerical results are provided in Section 6. Finally, some conclusions are given in Section 8.

2 System Model

2.1 IoT network model and Link Specifications

In this work, we consider an IoT network with a BS and a number of UEs with prescribed reliability and latency requirements. The BS is located at the center of a cell with a radius of d_{max} which consists of some spatial sectors. Without loss of generality, we consider a typical sector consisting of a massive number of UEs/nodes transmitting their data packets to the BS in a frame-based structure through a DQ-based random access procedure described in the following[1]. The sector coverage area is partitioned into N_c^g annuli called *Geographical Cluster (GC)*. For the sake of simplicity, it is assumed that all the UEs are evenly distributed within the GCs. Furthermore, the number of newly arrived active UEs at each GC in a typical frame, M^a, is assumed to follow a Poisson distribution with rate λ_u^g i.e., the probability that there are a active UEs in a typical GC at each frame equals $\Pr(M^a = a) = \frac{(\lambda_u^g T_f)^a e^{-\lambda_u^g T_f}}{a!}$, in which T_f is the frame length[2]. Fig. 1 depicts the network architecture in the discussed scenario.

The received power from a typical UE in the i^{th} GC, U_i $(i \in \{1, \cdots, N_c^g\})$ at the BS is $P_i = P_i^t |h_i|^2$ where P_i^t and h_i^2 are the U_i's transmit power and the channel gain, respectively. It is also assumed that all the channels between the UEs and the BS experience independent but not necessarily identically distributed Rayleigh block fading. The channel gain is assumed constant during each transmission frame. Therefore, $|h_i|^2$ follows an exponential distribution with mean $d_i^{-\nu}$ where d_i is the distance between U_i and the BS, while ν is the path-loss exponent. Furthermore, the background noise in all communication links is assumed to be independent and identically distributed zero-mean additive white Gaussian noise with variance $\sigma^2 = BN_0$, where N_0 and B are the noise spectral density and bandwidth, respectively. There is a reference channel through which the BS broadcasts a reference signal enabling each UE to estimate the channel and the distance to the BS. This further enables the UEs to perform power control so that their received power at the BS is at a prescribed level for each user.

[1] Each sector in a cell is served via spatial diversity through a specific antenna, assuming no inter-sector interference. Therefore, all the analyses in this study hold true for any sector.

[2] Different distributions and traffic models have been proposed for the massive MTC such as Beta, Uniform, CMMPP. Each model is for a specific use case. However, for simplicity of the analyses and to derive network metrics analytically, we used Poisson process in this paper.

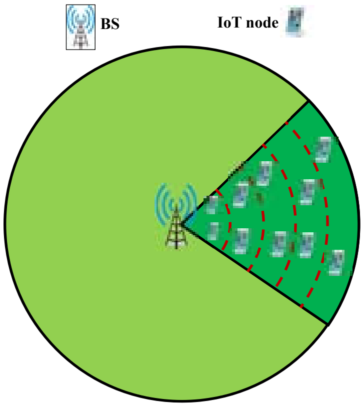


Fig. 1: An illustration of the considered IoT network model highlighting a sector with five GCs (5 rings) and two UEs at each GC.

2.2 Frame Structure For DQ-Based NOMA

Since there is a massive number of UEs in the network, it is not possible to devote orthogonal resources to each user in advance. Therefore, RA-NOMA based on DQ is proposed in this study. DQ is a MAC protocol exploiting tree-splitting algorithm. Since devices are required to contend in a tree-splitting fashion before data transmission
their reliability is independent of the number
that share the same communication channel, mal
able for massive communication scenarios. Ther
logical distributed queues for each Resource Bl
the proposed DQ-NOMA, namely *Transmission* (
and *Collision Queue* (CQ). Each UE follows a s
explained in Section 2.3 to control its transmissi
In order for the BS to detect active UEs and s
ing order at each frame dynamically, orthogona
transmission is included in the frame structure. A
Fig. 2, each frame consists of three slots, namely
Slot (CS), Feedback Slot (FS), and Data Slot (D
the length of T_c seconds, CS contains $\mathcal{S}$ subslo
Contention SubSlots (CSS). In CS, all active UEs
randomly select a Resource Block (RB) from an
set $RB = \{1, \dots, R_b\}$ to send a specific preambl
assigned to each GC on one CSS which is al
randomly[3]. Note that the number of preamble
equals to the number of GCs, N_c^g, avoiding a lar
of preambles that should be detected at the BS
ing implementation and computational complexi ,
more, all active UEs transmit their preambles such that all preambles are received at the BS with the same power level[4]. It is noteworthy that collision between preambles at the same GC in the same CSS (intra-GC collision) may happen when the same RB is selected by more than one active UE. Accordingly, each CSS in a specific RB can experience three different states; empty (selected by no active UE), success (selected by only one active UE), collision (selected by more than one active UE). To avoid intra-GC collision on the data slot, DQ mechanism is employed herein which is explained later. After receiving the preambles, the BS can distinguish active UEs in different GCs even on the same RB and in the same CSS due to their orthogonality. Additionally, it can also detect any intra-GC preamble collision[5].

In the feedback slot, the second slot of a frame, which lasts for T_{fb} seconds, the BS broadcasts feedback information on its observations in CS[6]. Generally, FS has two parts with the duration of T_{fb_1} and T_{fb_2} as shown in Fig. 3. The first part of FS contains $\mathcal{S}$ Feedback SubSlots (FSSs), each contains N_c^g minislots called Feedback MiniSlots (FMS). At each FMS, the BS broadcasts the CSS state relating to that GC. For example, in the i^{th} FMS of s^{th} FSS, the trilateral (empty, success, collision) state of the s^{th} CSS in the i^{th} GC is reported. Therefore, all the UEs know the status of their selected RB both in their own GC and in the other GCs. In the second part of FS, the BS broadcasts the length of CQ and TQ to all UEs which is explained later.

As its name implies, the third slot of a frame, DS, is allocated to data transmission. All the UEs from different GCs that are permitted to transmit data schedule their NOMA-based transmissions in DS[7]. The details of transmission rules are explained in Section 2.3. The BS then applies SIC to decode the signal of IoT UEs in the NC on all RBs[8].

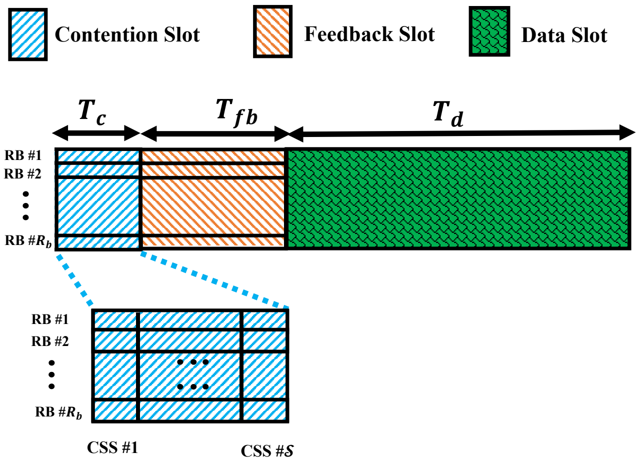


Fig. 2: Frame Structure of IoT UEs

Definition 1 (NOMA Cluster)**.** *All the UEs in different GCs that select the same RB and are allowed to send data in a DS of specific frame forms the NOMA Cluster (NC) on that RB. Therefore, there exist R_b NCs in the network.*

Definition 2 (Active GC on j^{th} RB Cluster)**.** *The i^{th} GC is said to be active on j^{th} RB if a UE from that GC wins the contention for transmitting data on j^{th} RB.*

[3] Orthogonal preamble transmissions are already used in RA-LTE, and defined in 3GPP RA [32], [33].

[4] Note that due to employing reference channel, the UEs can estimate the channel and adjust their transmit powers to target the preambles for the specific power level at the BS

[5] intra-GC preamble collision can be detected by measuring the received power level of preambles. Statistical approaches such as hypothesis testing can also be employed in collision detection [34], [35].

[6] In order to correctly decode the received preambles by the BS, the set of preambles must have fine auto and cross correlation properties. Zadoff-chu [36], Golden codes [37], and Reed-Muller [38] are the examples of appropriate sequences.

[7] From a practical perspective, Inter-frame and Inter-Slot Spacing (ISS) can be added between slots and frames.

[8] Selecting the SIC decoding order is considered based on the users' channel state information.

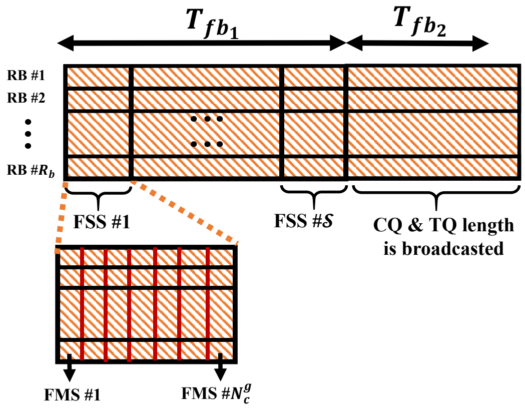


Fig. 3: Feedback slot of each Frame

Definition 3 (GC Activation Index). *$\mathbb{I}_i^j$ is defined as the activation index for the i^{th} GC on j^{th} RB. If the i^{th} GC is active on j^{th} RB, then $\mathbb{I}_i^j = 1$, otherwise $\mathbb{I}_i^j = 0$.*

Without loss of generality, it is assumed that U_i lies in the j^{th} NC ($j \in \{1, \cdots, R_b\}$).

2.3 DQ-Based NOMA MAC Procedure

As mentioned, the DQ MAC protocol uses two logical queues for each RB and a set of rules that should be followed by UEs. Let TQ and CQ be the labels of the transmission queue and collision queue for j^{th} RB, respectively. Furthermore, there is an internal counter assigned to each logical queue where each UE distributively updates its value. Consider a typical UE in the i^{th} GC U_i as the UE of interest. As stated, it holds and updates two internal counters for each of TQ and CQ. The first counter is transmission queue position counter ($TQPC$) showing U_i's position in TQ on its selected RB. The second counter, collision queue position counter ($CQPC$), shows U_i's position in CQ on its selected RB.

At the second part of FS, the BS sends the length of CQ and TQ for all RBs allowing idle UEs to know the queues' length after becoming active, helping them performing DQ process. Suppose U_i has a data packet to transmit at the beginning of the current frame. Therefore, it selects an RB and a CSS randomly and sends its random access preamble assigned to its GC. Upon receiving the FS, U_i checks the state of selected CSS in the selected RB, say s^{th} CSS in j^{th} RB, by observing i^{th} FMS in the s^{th} FSS. If it is 'success', U_i enters the TQ, otherwise, it enters the CQ. At the same time, it updates its internal counters according to the following rules.

- **TQPC Set:** After transmitting its preamble, if U_i receives 'success' on i^{th} FMS in the s^{th} FSS, it joins the TQ and sets its $TQPC$. Assuming that the i^{th} FMS in the s^{th} FSS is the c^{th} 'success' report, $TQPC$ is set to the sum of c and TQ length, i.e., $TQPC = TQ_{length} + c$.
- **TQPC Update:** If U_i has already joined the TQ, it decreases its $TQPC$ by one at the end of each frame provided that $TQPC > 0$. If at the end of frame and after decreasing $TQPC$ by one, it equals zero, U_i becomes the first UE in TQ and it starts data transmission in the next frame.
- **CQPC Set:** After transmitting its preamble, if U_i receives 'collision' on i^{th} FMS in the s^{th} FSS, it joins the CQ and sets $CQPC$. Assuming that the i^{th} FMS in the s^{th} FSS is the d^{th} 'collision' report, $CQPC$ is set to the sum of d and CQ length i.e., $CQPC = CQ_{length} + d$.
- **CQPC Update:** If U_i has already joined the CQ, it decreases the $CQPC$ by one at the end of each frame provided that $CQPC > 0$. If at the end of frame and after decreasing $CQPC$ by one, it equals zero, U_i becomes the first UE in CQ. Hence, it leaves the CQ and attends contention by sending preamble in CS in the next frame.

Note that setting the value of internal counters is performed by UEs at the end of FS at each frame while updating them is done at the end of each frame (or equivalently, the end of DS at each frame). Fig. 4 illustrates the proposed RA process with $N_c^G = 3$, $\mathcal{S} = 2$ and 9 UEs contending to transmit data over only one RB. To make the scenario clearer, UEs are numbered from 1 to 9. UEs with numbers {1, 2, 3} are located in GC #1 marked in red, UEs with numbers {4, 5, 6} are located in GC #2 marked in green, UEs with numbers {7, 8, 9} are located in GC #3 marked in blue. Each frame starts with a preamble phase containing two contention subslots. Subslots in the preamble phases in Fig. 4 are colored in light blue. Since there are three GCs, each feedback subslot contains three FMSs which are colored in light orange. As can be seen, UEs 1, 3 and 6 have selected CSS#1 randomly to transmit their preambles while UEs 2, 5, 7 and 8 have selected CSS#2. Since two UEs, i.e., 1 and 3 from GC#1 have selected the same CSS (CSS#1), FMS related to GC#1 and CSS#1 is reported as collided (C in the figure). However, FMS#2 is reported as a success (S in the figure) since only one UE from GC#2 (UE 6) has transmitted its preamble in CSS#1. Furthermore, no UE from GC#3 has selected CSS#1 for preamble transmission. Following this rule, the feedback phase for CSS#2 results in a success, success, and collision for FMS#1, 2 and 3, respectively. At the end of feedback phase, UEs 6 (in CSS#1) and 2 and 5 (in CSS#2) have been reported as successful UEs. At this point, every UE knows that UEs 6, 2 and 5 should be moved into TQ with 6 having the first place and {2, 5} having both the second place. Moreover, collided UEs are moved to CQ for the next round of contention. The order that the UEs enter the queues is according to their selected CSS. For instance, since UEs {1, 3} have selected CSS#1 and have been reported collided in that CSS, they go first in CQ. However, CSS#2 has been selected by UEs {7, 8}, then they are placed in the second position in the CQ after 1 and 3. Now, it is time for transmission of the first UE(s) in the TQ. Note that each place in the TQ may contain up to three transmitting UEs related to the three GCs. These UEs form the dynamic NC for the underlying frame. For the next frame, UEs in the first position of the TQ (2 and 5) have already been scheduled for transmission, forming NC for that frame. Furthermore, UEs in the first position of the CQ (1 and 3) along with the incoming active UEs in the second frame (4 and 9 in the scenario) try to contend by selecting

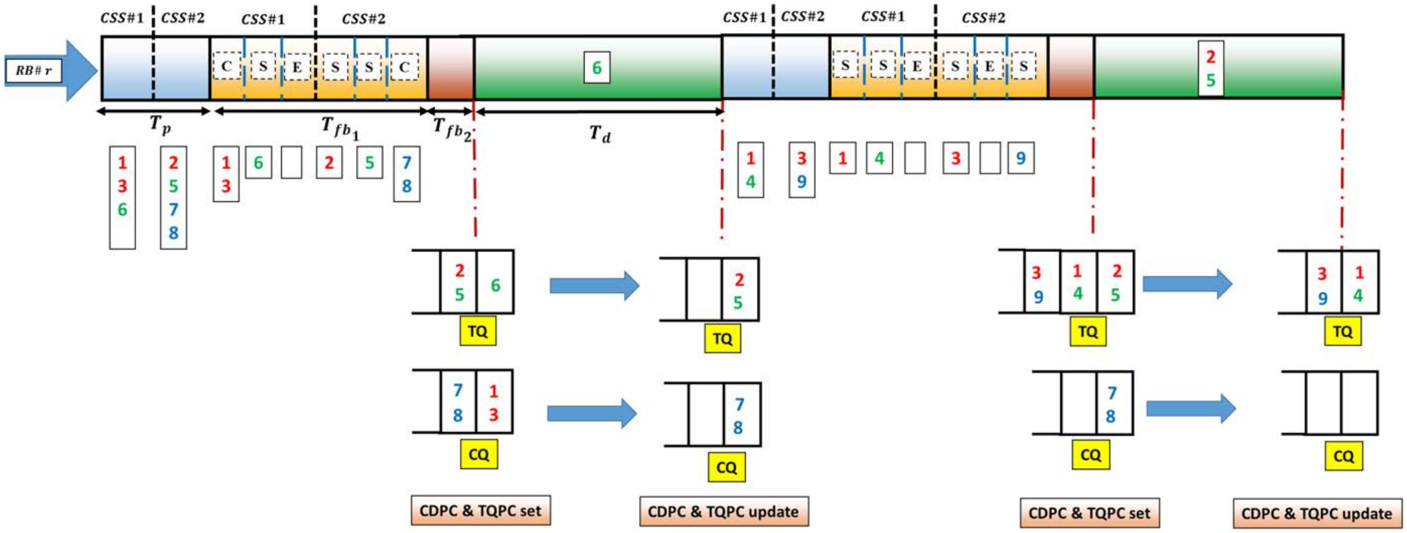


Fig. 4: An illustrative example for DQ-NOMA RA with $N_c^g = 3$ and $\mathcal{S} = 2$

the CSS and transmitting preamble over them. The process will then continue.

2.4 Frame Duration

- **CS Duration:** The preamble duration is adopted from RA-LTE as PRACH preamble format #4 (short preamble format) for UL-TX [39] which equals to 2 symbols. In this way, the random access overhead is significantly reduced[9]. Hence, the duration of CS is $2\mathcal{S}$ symbols.
- **FS Duration:** FS has two main parts. The first part contains $N_c^g \times \mathcal{S}$ FMS. Since each FMS indicates three states of 'empty', 'success', 'collision', it is enough to consider 2 bits for each FMS. Therefore, the duration of the first part of FS equals $T_{f_1} = 2N_c^g\mathcal{S}$ bits. The duration of the second part, T_{f_2}, is adjusted according to the maximum number of UEs in the network that can be in CQ and TQ. By setting it as two Bytes, 2^{16} UEs can line up for transmissions which is enough to accommodate the UEs in a cell in a realistic scenario. Therefore, $T_f = T_{f_1} + T_{f_2} = 2N_c^g\mathcal{S} + 16$ bits.
- **DS Duration:** To support low latency requirement, short packet transmissions are adopted via FBL regime. Hence, considering the blocklength of n_b the duration of DS equals $T_d = \frac{n_b}{B}$ [40].

2.5 Power Back-off Strategy and Decoding Error

In NOMA-based transmissions, the UEs' transmit power plays an important role in the performance of the whole system, since it significantly affects the decoding error. Therefore, the power control strategy must be employed to reduce what is called power collision as much as possible. The most common power control method used in the UL NOMA is the power back-off strategy [34], [41], [42], [43] in which the transmit power of the a^{th} active UE in its NC at the i^{th} GC is given as,

$$P_{i,t}^a = \min\{P_{max}, P_u - (a-1)\varrho + PL_i\}, \quad (1)$$

where P_{max} and P_u are the maximum transmit power and target received power at the BS, respectively. Furthermore, ϱ is the power back-off step of a target received power and PL_i is the path loss. Hence, if U_i is the a^{th} active UE in its NC, its received power is as,

$$P_i^a = \min\{P_{max} - PL_i, P_u - (a-1)\varrho\}. \quad (2)$$

It is worth noting that (2) necessitates that a^{th} active UE at each NC always targets its received power at $P_u-(a-1)\varrho$, regardless of its location as long as its transmit power does not exceed P_{max}[10].

Moreover, it is assumed that the IoT UEs transmit their data packets in FBL regime to lower the packet latency. However, in such a case, Shannon's capacity is no longer applicable since the decoding block error cannot be ignored. Thus, given a blocklength of $n_b > 100$ with n_d data bits per data packet, the instantaneous block error rate of decoding U_i's signal at the BS– provided that U_i is the a^{th} active UE from the total m active UEs in j^{th} NC cluster– is approximated as [40] ,

$$\epsilon_{i|a,m} = Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{i|a,m}\right)}}\left(\mathcal{C}\left(\gamma_{i|a,m}\right) - \frac{n_d}{n_b}\right)\right), \quad (3)$$

where $\mathcal{C}\left(\gamma_{i|a,m}\right) = \log_2(1 + \gamma_{i|a,m})$ is the Shannon capacity of U_i, while $\chi\left(\gamma_{i|a,m}\right) = \left(1 - \frac{1}{1+\gamma_{i|a,m}^2}\right)(\log_2 e)^2$ represents the channel dispersion. Furthermore, $\gamma_{i|a,m}$ is the received Signal-to-Interference plus Noise ratio (SINR) of U_i's signal at the BS which is given as,

[9] From a practical perspective, it is possible to adopt LTE short packet format for the underlying scenario since the number of orthogonal preambles is low. This is because it is equal to the number of GCs which should be low in NOMA-based transmission.

[10] Through provided information on the feedback slot and by exploiting learning algorithms, a typical active UE can adjust transmit power parameters ϱ and P_u so that the sum rate in its NC at each frame is maximized. However, this is beyond the scope of this study and is left for future studies.

$$\gamma_{i|a,m} = \frac{P_i^a}{\sum_{h=a+1}^{m} P_i^h + \sigma^2}. \tag{4}$$

Table 1 summarizes the main symbols used in this study and their descriptions.

TABLE 1: Notations

Parameter	Description
$\gamma_{i,j\|b,A_j}$	SINR for $U_{i,j}$'s signal at the BS provided that $U_{i,j}$ is the b^{th} active UE of all A_j UEs in j^{th} NC
B	Bandwidth of each RB
$P_{i,j}^b$	Transmit power of $U_{i,j}$ provided that it is the b^{th} active UE in its NC cluster
n_b	Blocklength
n_d	Number of data bits in a packet
N_c^g	Number of Geographical clusters
N_u^t	Total number of active UEs in the network
N_0	Noise spectral density
N_u^g	Number of active IoT UEs in a GC of interest
A_j	Number of active IoT UEs in the j^{th} NC
T_c	Contention slot duration
T_{fb}	Feedback slot duration
T_f	Frame duration
T_d	Data slot duration
ν	Path-loss exponent
R_b	Number of RBs
λ_u^g	Active UE arrival rate at a typical GC
$\mathcal{S}$	Number of contention/feedback subslots

3 Derivation of Performance Metrics

3.1 Distribution of Packet Latency

The Average packet latency $\bar{\mathcal{D}}^{DQ}$ is defined as the average delay of delivering a typical packet and all replicas of that packet to the BS, which includes the transmission delay and the queuing waiting time. To draw $\bar{\mathcal{D}}^{DQ}$ and delay violation probability, the distribution of packet latency is intended. Such a distribution incorporates both transmission delay and waiting times in CQ and TQ. To derive the packet latency in the proposed DQ-NOMA, queuing network analyses are employed in which the whole system is viewed as a network with two queue subsystems; CQ and TQ subsystem. The delay statistics are then calculated considering the statistics of the mentioned subsystem.

3.1.1 Delay Analyses of CQ Subsytem

Consider U_i as the UE of interest. While experiencing the 'collision' state in the selected CSS, U_i must join the CQ subsystem, update the $CQPC$ at each frame and wait until $CQPC = 1$ to retry preamble transmission. Lemma 1 gives the CQ subsystem delay distribution.

Lemma 1. *The Probability Density Function (PDF) of U_i's packet delay in CQ subsystem $f_{\mathcal{D}_{CQ}}(t)$ is obtained as,*

$$f_{\mathcal{D}_{CQ}}(t) = \xi e^{-\xi t}, \tag{5}$$

where ξ is as,

$$\xi = \ln(\frac{1}{1-p_s})T_f^{-1} - \frac{\lambda_u^g(1-p_s)}{R_b}, \tag{6}$$

and p_s is the probability that a typical UE experiences the 'success' state on the selected CSS which is given as,

$$p_s = e^{-\frac{\lambda_u^g}{R_b \mathcal{S}} T_f}. \tag{7}$$

Furthermore, the average packet latency in the CQ subsystem is given as,

$$\bar{\mathcal{D}}_{CQ} = \left[\ln(\frac{1}{1-p_s})T_f^{-1} - \frac{\lambda_u^g(1-p_s)}{R_b}\right]^{-1}, \tag{8}$$

where $T_f = T_c + T_{fb} + T_d$ is set according to the explanations provided in Section 2.4.

Proof: See Appendix A. □

3.1.2 Delay Analyses of TQ Subsytem

As explained in Section 2.3, U_i joins the TQ if it experiences 'success' state in the selected CSS. Then it remains in TQ until its $TQPC$ equals one. It is noteworthy that TQ has two arrival sources. The first one is the UEs that enter the TQ after their first successful attempt on sending a preamble in CS. The second source of arrival is from those UEs leaving the CQ. Considering the two arrival sources, Lemma 2 gives the CQ subsystem delay distribution.

Lemma 2. *The PDF of U_i's packet delay in TQ subsystem $F_{\mathcal{D}_{TQ}}(t)$ is obtained as,*

$$f_{\mathcal{D}_{TQ}}(t) = \sum_{n=1}^{\infty} \pi_{n-1}\delta(t - nT_f), \tag{9}$$

where $\delta(.)$ is the Dirac delta function and π_{n-1} $(n \in \{1, 2, \cdots\})$ can be derived recursively by writing system of equations at equilibrium yielding as,

$$\pi_0 = 1 - N_c^g \lambda_u^g T_f / R_b, \tag{10}$$

and

$$\pi_n = \pi_0 a_n + \sum_{k=1}^{n+1} \pi_k a_{n+1-k}, \qquad n \geq 1 \tag{11}$$

where $a_n = \frac{1}{n!}\left(\frac{N_c^g \lambda_u^g T_f}{R_b}\right)^n e^{-\frac{N_c^g \lambda_u^g T_f}{R_b}}$. Furthermore, the average packet latency in the TQ subsystem is given as,

$$\bar{\mathcal{D}}_{TQ} = T_f\left(\frac{\rho}{2(1-\rho)} + 1\right), \tag{12}$$

where ρ, the transmission queue utilization, is given as,

$$\rho = N_c^g \lambda_u^g T_f / R_b. \tag{13}$$

Proof: See Appendix B. □

Since a typical UE can join either TQ alone or both CQ and TQ, the total packet latency in DQ-NOMA MAC protocol is the weighted sum of delay in TQ and CQ subsystems. Lemma 3 provides the expression for the delay violation probability of UE's packet in the underlying MAC protocol.

Lemma 3. *The delay violation probability of UE's packet in the proposed DQ-NOMA MAC protocol is obtained as,*

$$\begin{aligned} F_{\mathcal{D}}^C(t) = Pr(\mathcal{D} > t) = p_s\Big(1 - \sum_{n=1}^{\infty} \pi_{n-1}\,\mathcal{U}(t - nT_f)\Big) \\ +\ (1-p_s)\sum_{n=1}^{\infty} \pi_{n-1}\, e^{-\xi(t-nT_f)}, \end{aligned} \tag{14}$$

where the $\mathcal{D}$ is the random variable indicating node's packet delay. Moreover, the average packet latency can be derived as,

$$\mathcal{R}_i^{DQ} = \sum_{a=1}^{i} \sum_{m=a}^{a+N_c^g-i} \binom{i-1}{a-1} \binom{N_c^g-i}{m-a} \left(\lambda_u^g T_f/R_b\right)^{m-1} \left(1-\lambda_u^g T_f/R_b\right)^{N_c^g-m} \prod_{l=1}^{a} \left(1-\epsilon_{i|l,m}\right). \tag{16}$$

$$\begin{aligned}\bar{\mathcal{D}}^{DQ} = {} & T_f\left(\frac{\rho}{2(1-\rho)}+1\right) \\ & + \left[\ln(\frac{1}{1-p_s})T_f^{-1} - \lambda_u^g(1-p_s)/R_b\right]^{-1}(1-p_s),\end{aligned} \tag{15}$$

where p_s and ρ are given in (7) and (13), respectively.

Proof: See Appendix C. □

Note that the first part in (15) is related to the TQ waiting time and the second part is related CQ waiting time. As can be inferred, the CQ waiting time is dependent on p_s. However, TQ does not depend on p_s. To reduce the CQ waiting time, one can conclude on increasing p_s by an increase in $\mathcal{S}$ as (7) implies. However, increasing $\mathcal{S}$ will increase T_f and it in turn increases ρ which has a detrimental effect on the first part, TQ waiting time. The negative effect is more essential when ρ approaches 1, threatening the queue stability. Hence, some sensible values for the number of CSS should be taken. Increasing blocklength for the sake of increasing reliability also makes the waiting time in both subsytems longer as an increase in n_b will increase T_f and ρ and decrease success probability, p_s. Hence, a reasonable value for blocklength to yield a target reliability and acceptable latency is the objective.

3.2 Reliability

The transmission reliability metric for U_i, $\mathcal{R}_i^{DQ}$, is defined as the probability that a typical packet transmitted by U_i is received successfully at the BS (i.e. without any decoding errors [11]). Such a metric is obtained in Lemma 4.

Lemma 4. *The U_i's reliability $\mathcal{R}_i^{DQ}$ is obtained as per (16) where $\epsilon_{i|l,m}$ is given in (3).*

Proof: See Appendix D. □

We can come to some important conclusions from (16). Firstly, we note that the term that has the most significant influence on $\mathcal{R}_i^{DQ}$ is $\prod_{l=1}^{a}\left(1-\epsilon_{i|l,m}\right)$. One can infer that in general, the UE with a higher GC index, i, should have lower reliability. That is, because for higher values of i, the product term has higher number of terms as a increases with an increase in i (this is due to the first summation). Therefore, reliability for UEs with high number of GC index is lower than those with lower GC index. Conceptually, this is due to the fact that in uplink NOMA, decoding of a typical UE packet depends on the decoding accuracy of the previous UEs in its NC. Since UEs with lower GC index have lower active UEs in their NC cluster, they have higher reliability. Another point is that as n_b increases, $\epsilon_{i|l,m}$ will decrease due to (3) and FBL concept. Hence, increasing blocklength up to some values will help increase reliability.

3.3 Effective Sum Rate

The Effective Sum Rate (ESR) is defined as the sum of all UE's effective rate, i.e., error-free "non-redundant" data bits per time unit successfully delivered at the BS. Lemma 5 provides a mathematical expression for ESR.

Lemma 5. *The effective sum rate of the proposed DQ-NOMA MAC protocol is given as (17) where*

$$\epsilon_{l|\mathbf{I}_k} = Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{l|\mathbf{I}_k}\right)}}\left(\mathcal{C}(\gamma_{l|\mathbf{I}_k}) - \frac{n_d}{n_b}\right)\right), \tag{18}$$

and

$$\gamma_{l|\mathbf{I}_k} = \frac{P_l^{J(l)}}{\sum_{h=l+1}^{N_c^g} P_h^{J(h)} + \sigma^2}, \tag{19}$$

where $J(\beta) = \sum_{q=1}^{\beta} \mathbb{I}_{\beta}^{j}$.

Proof: See Appendix E. □

From (17), one can conclude that ESR increases with increasing n_b since the term $\epsilon_{l|\mathbf{I}_k}$ decreases according to (3) and FBL concept. However, the term $\frac{n_d}{T_f}$ shows its negative effects on ESR when n_b increases excessively. This is due to the fact that $\epsilon_{l|\mathbf{I}_k}$ is proved to decrease at lower rate when n_b is high enough where the negative effect of the former term outweighs the positive effect of blocklength. Such effects are discussed in Section 6.

4 S-ALOHA WITH TRANSMISSION DIVERSITY

In this section, network metrics for S-ALOHA with transmission diversity are derived which helps evaluating the proposed DQ-NOMA MAC protocol. In S-ALOHA-TD, all the active UEs randomly select one RB and then start transmitting their data on the selected RBs. Since there is no NOMA strategy, each UE transmits its data packet with the maximum transmit power, i.e., P_{max}. Therefore, there is no need for power control strategy and hence, U_i's received power at the BS is $P_{max} - PL_i$. To be comparable with the proposed DQ-NOMA MAC protocol, it is also assumed that there are N_c^g GCs in the network each with active UE arrival rate λ_u^g. Furthermore, it is also assumed that nodes transmit each data packet L times in L successive frames. It is worth noting that in the underlying S-ALOHA-TD, the arrival rate of active UEs on each RB equals $N_c^g\lambda_u^g/R_b$.

Note that in S-ALOHA-TD, the packet latency for UEs is a constant value and equals to $\mathcal{D}^{AL} = L \times T_f$.

The reliability for U_i in S-ALOHA-TD is derived in Lemma 6.

Lemma 6. *The U_i's reliability metric in S-ALOHA-TD is given as,*

$$\mathcal{R}_i^{AL} = 1 - \left(1 - (1-\epsilon_i)\, e^{-\frac{N_c^g \lambda_u^g T_f}{R_b}}\right)^{L}, \tag{20}$$

where ϵ_i is as per (3) with γ_i as,

$$\gamma_i = \frac{P_{max} - PL_i}{\sigma^2}. \tag{21}$$

[11] Note that there is no intra-cluster collision between the UEs due to the BS feedback. However, channel distortion may cause decoding error.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3319545

$$ESR^{DQ} = R_b \sum_{k=1}^{N_c^g}$$

Proof: See Appendix F.

Furthermore, the ESR in S-ALOI

Lemma 7.

Lemma 7. *The effective sum rate of S-A*

$$ESR^{AL} = \frac{N_c^g \lambda_u^g n_d}{LT_f}$$

Proof: See Appendix G.

5 Numerical Results and (Benchmark S-ALOHA

This section evaluates the analytical r tion 3 and compares them with the c derived for S-ALOHA-ARQ in 4. It is nodes in the i^{th} GC are located at t Particularly, the distance from U_i to $d_i = \frac{(2i-1)d_{max}}{2N_c^g}$ $(i \in \{1, \cdots, N_c^g\})$.

TABLE 2: Simulation Pa

Parameter	Value	Parameter	Value
B	180KHz	R_b	40
ν	2.5	$\mathcal{S}$	2
T_c	0.1 ms	T_{fb}	100 μs
P_u	10^{-6} W	N_0	-174 dBm/Hz
N_c^g	3	n_d	32 Bytes
P_{max}	0.3 W	ϱ	10 dB
d_{max}	1000 m	T_{sim}	20×10^3 s

The effect of blocklength n_b on the ave tency for both DQ-NOMA and S-ALOHA w is shown in Fig. 5. As can be seen, the latency for both DQ-NOMA and S-ALOH/ increase in n_b. This is because the higher the greater the transmission period, and hence average packet latency. Additionally, increas of replicas in S-ALOHA results in higher ra latency rises up. This is the obvious drawl with multiple re-transmissions. Interestingl latency for DQ-NOMA is greater than S-A blocklength values, it then steadily increases lower than S-ALOHA meeting the S-ALOI $L = 4$, $L = 3$, and $L = 2$ at $n_b = 20$, $n_b = 3$ respectively. This is due to the extra signali contention and feedback slots[12]. This fulfills requirements for massive critical IoT netwo

Fig. 6 illustrates the average packet lat and $\mathcal{S}$. The higher the arrival rate of active U the collisions in the contention slot, and hence, the greater the waiting time in CQ on average. This leads to higher packet latency. Another observation is that increasing the number of CSSs increases the average packet latency as well.

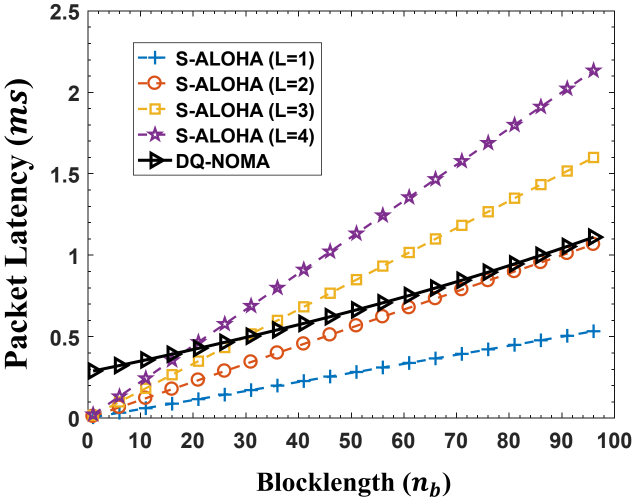


Fig. 5: Average Packet Latency vs. Blocklength - $\lambda_u^g = 10000$.

The packet latency drastically rises up at high active UE arrival rates since high waiting times in CQ coincides with higher signaling overhead (and longer frame duration) due to exploiting a high number of CSSs. Note that Fig. 6 reveals that increasing the number of CSS to have less collisions does not reduce latency in FBL regime since the negative effect of signaling overhead outweighs the productive effect of lowering the waiting time in the CQ. Hence, choosing small values (e.g., $\mathcal{S} = 1, 2$) works for mIoT networks to

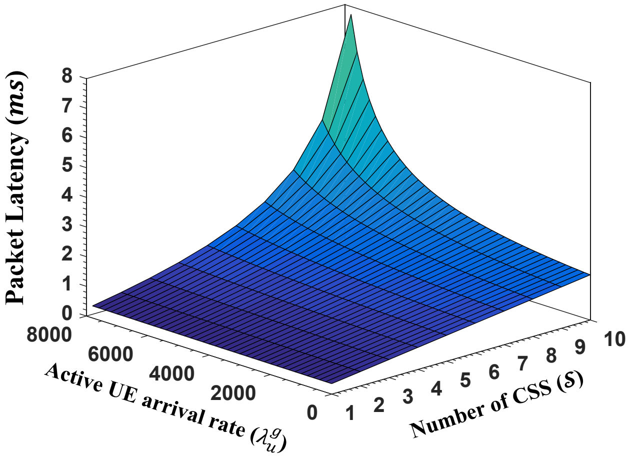


Fig. 6: Average Packet Latency vs. UE arrival rate and number of CSS- $n_b = 45$.

Fig. 7 shows the effect of n_b on U_i's reliability $(i = 1, 2, 3)$ and compares it to that of S-ALOHA, with different number of transmissions $L = 1, 2, 3, 4$. Generally speaking, when n_b is small, the decoding error probability in FBL regime is significant and hence, the reliability has the lowest values for both DQ-NOMA and S-ALOHA. It then rises up sharply with increase in n_b until it reaches its maximum value

[12] Note that for small values of n_b (near zero), the transmission duration will also be near zero. However, the average packet latency is not zero and is about 0.3 ms which is due to the overhead of contention and feedback phases.

of 1 for DQ-NOMA and remains al
excessive increase in blocklength doe
the decoding error probability in FBL
of S-ALOHA experiences a sharp i
and touches it maximum of 0.88 fo
0.95 for $L = 2, 3, 4$. It then starts c
leaving its peak to 0.25 for $L = 1$
for $L = 2, 3, 4$. This decay is due to
increases, the frame length also incr
number of active users are increased as
collisions. In general, despite DQ-NO
fulfills the URLL requirements. Some
seen in the figure. One is that $\mathcal{R}_1^{DQ} \geq$
is due to the fact that reliability of e
GC in DQ-NOMA depends on the re
decoded nodes in its NC. The other ob
experiences a decrease in the interval
first rise up. The reason falls in the
nodes' data traffic at each NC. Partic
only active node in its NC, the first a
two, and the first among the three total active nodes in the NC. This significantly affects its received SINR at the BS, and hence, the block error probability in the decoding process and consequently, affects the reliability. In
observed that in the mentioned interval,
blocklength is, the lower the reliability. Th
number of active users is increased with th
and frame length. Therefore, U_i experience

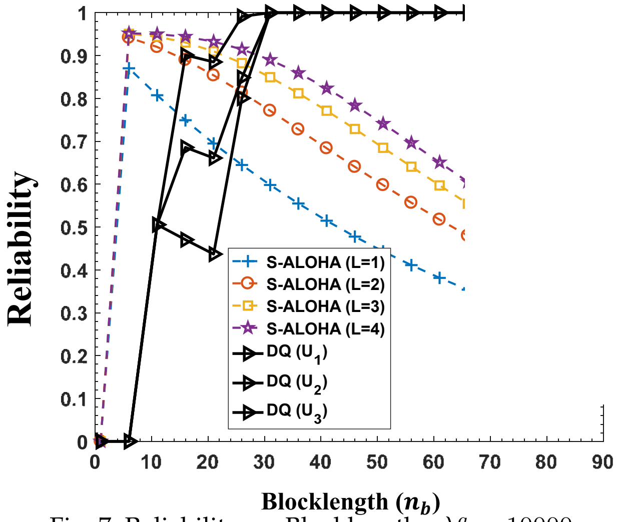


Fig. 7: Reliability vs. Blocklength - $\lambda_u^g = 10000$.

The reliability versus active UE arrival rate is depicted in Fig. 8. As can be seen, increasing λ_u^g has no effect on the reliability in DQ-NOMA. Despite DQ-NOMA, increasing λ_u^g significantly deteriorates the reliability of S-ALOHA which is the result of more collisions due to higher arrival rates. It starts from 0.95 for a very low rate of $\lambda_u^g = 10$ and decreases to 0.5, 0.68, 0.78, and 0.83 for $L = 1$, $L = 2$, $L = 3$, and $L = 4$, respectively.

Fig. 9 illustrates the effect of both n_b and λ_u^g on the reliability of the node U_1[13]. It can be seen that $\mathcal{R}_1^{DQ}$ (as well as $\mathcal{R}_2^{DQ}$ and $\mathcal{R}_3^{DQ}$, as discussed for Fig. 7) can reach to almost 1 by setting the appropriate value of n_b using network traffic predictive algorithms that provide an integral part of self-configured next generation networks.

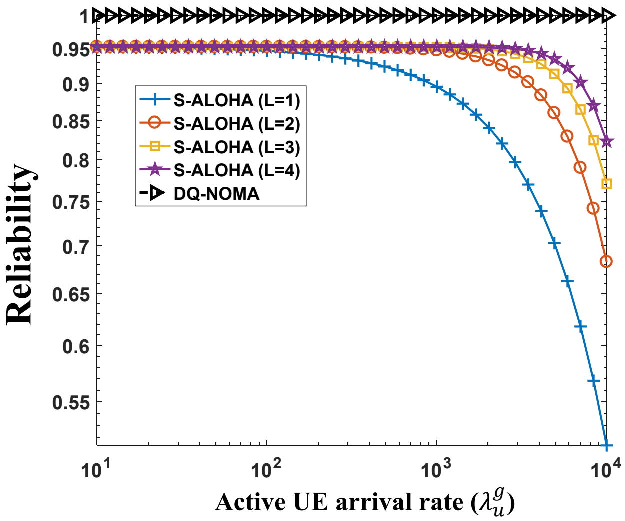


Fig. 8: Reliability vs. UE arrival rate - $n_b = 45$.

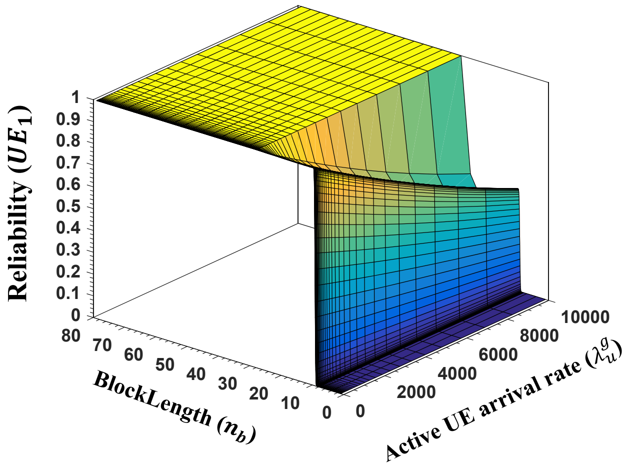


Fig. 9: Reliability of U_1 vs. Blocklength and UE arrival rate.

Fig. 10 shows ESR as a function of blocklength for both DQ-NOMA and S-ALOHA. As can be observed, ESR for S-ALOHA experiences a peak around $n_b = 7$ for $L = 1, 2, 3, 4$, and then decreases sharply. This happens because increasing blocklength from small values significantly helps improving the decoding error probability as discussed in Fig. 7. Therefore, the effective data bits decoded at the BS is increased, which improves the ESR. However, excessive increase in n_b does not help improving the decoding error probability, yet increases the frame duration at the same number of data bits (n_d) transmitted in a frame. As a result, it incurs the redundancy and lowers the ESR. Notably, S-ALOHA with $L = 1$ shows a higher ESR than other S-ALOHA scenarios which indicates that signaling overhead due to transmission of multiple replicas outweighs its positive effect on decoding performance. On the other hand, ESR for DQ-NOMA experiences several sharp rises and falls. Such changes coincide with the change in the reliability curve in Fig. 7. Particularly, the increasing (decreasing) parts in

[13] To avoid redundancy, the reliability for nodes U_2 and U_3 is not plotted

ESR are related to the improvement (degradation) of the decoding error performance as discussed for Fig. 7.

Another observation is that ESR levels-off for $n_b > 32$ where the reliability reaches its maximum value and remains constant. However, for $n_b > 132$, ESR starts decaying. Recall that increasing n_b when the reliability reaches its peak increases both redundant data b
of active nodes in a frame. Such a trer
can be viewed as the interaction betwee
utilization (due to the rise in the numb
a frame) and increasing the redundant
figured out from Fig. 11. In Fig. 11, it is
the arrival rate of active nodes is, the lo
value at which ESR starts decaying. T
higher values of λ_u^g, the communication
at lower values of n_b. This shows the ef
on determining n_b as an important ne

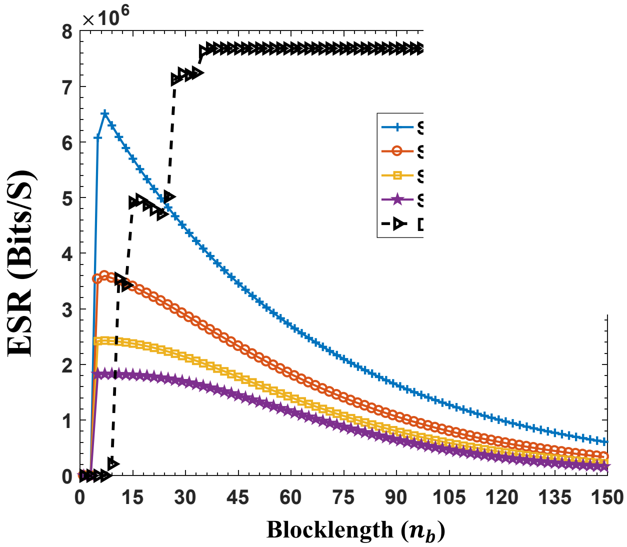


Fig. 10: ESR vs. Blocklength - $\lambda_u^g = 10000$.

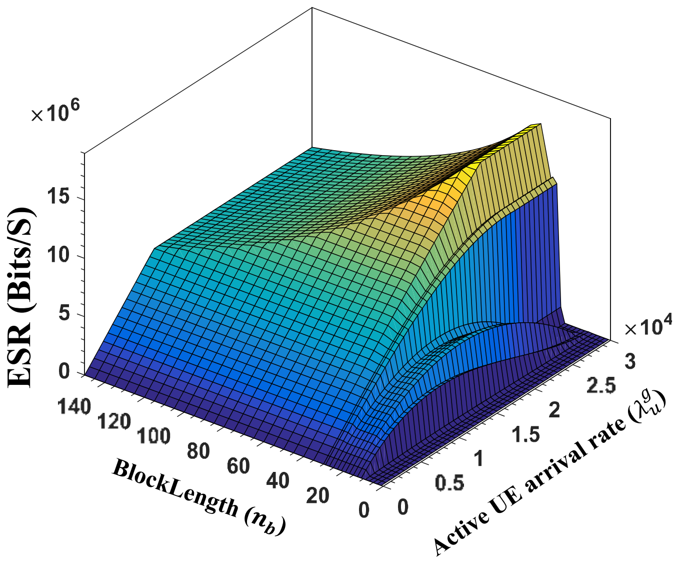


Fig. 11: ESR vs. Blocklength and UE arrival rate - $n_b = 45$.

Finally, ESR as a function of λ_u^g is depicted in Fig. 12. An important point is that unlike S-ALOHA, ESR for DQ-NOMA rises up when λ_u^g increases to 18500 to reach its maximum at 14.4×10^6 and remains constant thereafter. While it experiences a decreasing trend for S-ALOHA after touching its maximum. Not to mention that the maximum value of ESR in DQ-NOMA is far greater than that of

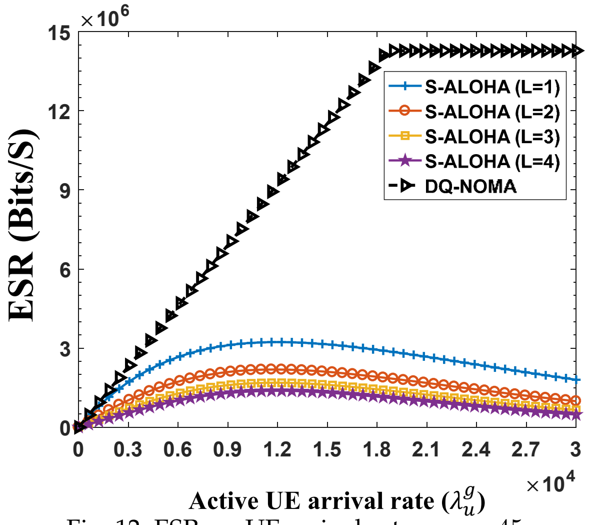
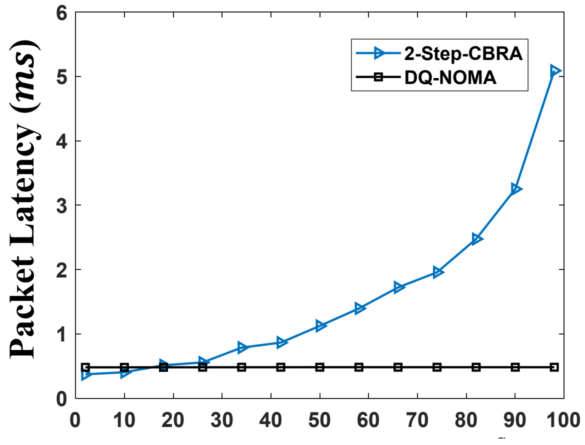


Fig. 12: ESR vs. UE arrival rate - $n_b = 45$.

6 Simulation Results and Comparison with 2-Step CBRA in 5G

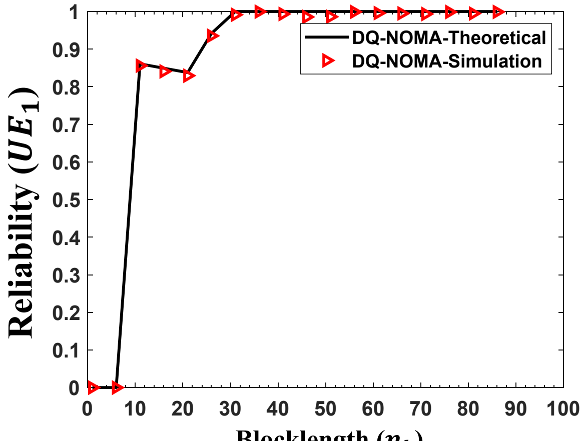
In this section, multiple simulation scenarios are executed. The purpose of the first simulation scenario is to make a comparison between the proposed structure with the 2-step RA techniques as the-state-of-the-art RA scheme. The second simulation scenario is performed to justify our theoretical derivations. These scenarios are explained following along with the descriptions of the corresponding figures.

To evaluate our proposed scheme, the DQ-NOMA is compared to the 2-step contention-based random access (CBRA) in Release 16 5G in Fig 13. In 2-step procedure, Message A contains both a preamble on Physical Random Access Channel (PRACH) and a payload on Physical Uplink Shared Channel (PUSCH). The payload corresponds to Message 3 in conventional 4-step CBRA in LTE [44]. After transmitting a Message A with a preamble, a UE waits for Message B from the gNB on Physical Downlink Shared Channel (PDSCH) to receive the corresponding configuration. The gNB takes different actions based on the status of the received Message A. Note that in both 2-step and 4-step, the UE has to wait for System Information Block (SIB 2) to start RA process. To make a comparison, simulation scenario is set up for 2-step CBRA. Discrete Event Simulation (DES) was adopted to simulate such a scenario. To this aim, all the events in the simulation scenario were scheduled based on their statistics. Starting from the first event, we calculated all the necessary metrics and proceeded the simulation time to the next nearest event. The simulation was performed 50 times, each for sufficient duration (final simulation time) in order to have adequate number of events at the underlying



scenario. Then, the calculated
all runs to obtain the final re
generating active UEs based o
(similar to the proposed study)
transmitting payload upon co
stamped to calculate the temp
time was set to 20,000 second
for each event. Each scenario
then all the results were avera
mentioned that in order to be c
scenario, we have made som
2-step CBRA and our propo
short packets with finite block
in CBRA instead of normal p
latency for low latency applic
have the same number of RE
exist a mapping between pre
Uplink Shared Channel) Reso
CBRA involving 16 RBs, the n
RA has been set to 16. It is
study needs just a few preambles (
underlying work) which eases the
reduces interference floor. However,
in 2-step CBRA since the performa
on the number of preambles. Fina
latency for 2-step CBRA and DQ-N
13. It is observed that the 2-step C
latency than the proposed DQ-NC
arrival rate, i.e., $\lambda_u^g < 20$ for $n_b = 4$
the latter is significantly superior to
number of active UEs grows. No
average packet latency for DQ-NC
Referring to Fig. 6, when the num
$\mathcal{S} = 2, 3, 4$, the average packet lat
for the high λ_u^g, making the propos
for massive URLL IoT. Note that we
the reliability of the proposed DQ-N
since the latter performance suffers
the number of preambles is very lin

To supplement our derivations
MATLAB. To this aim, all the events
scenario were scheduled based on the
events include generating active UEs,
and CSS randomly, updating DQPC and
transmitting packets, and updating TQ a
executing the simulation scenario 50 tin
seconds, the results have been averaged
14 and 15.



7 Discussions

This section discusses the energy consu
posed scheme. Furthermore, some adva
compared to the current ACB techniques

Since energy consumption is also a
massive IoT scenarios, it is worth discus
consumption. The energy consumption
proposed protocol is expected to be more

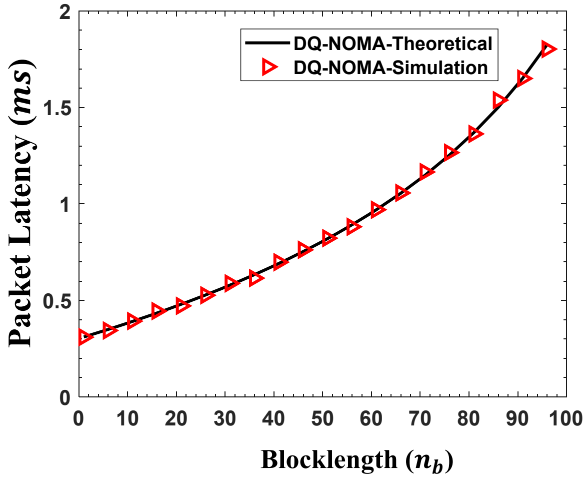


Fig. 15: Simulation and theoretical results for packet latency vs. blocklength - $\mathcal{S} = 2$, $R_b = 40$, $\lambda_u^g = 12000$.

regular RA-NOMA protocols since a modified version of the back-off power control is employed. That is because instead of assigning the power level based on the distance or the

[14] Note that Not all the 64 preambles in the 2-step CBRA 5G and even in 4-step case is used for random access procedure. However, we set the number of preambles to 64 to show the performance of the study

GC index, the modified version in our proposal relates the power level to the network's traffic by incorporating q in (1). To explain this further, consider the UEs in GCs other than the first GC. In the traditional back-off power strategies, q is replaced with the GC's index or other parameters related to UE's distance to the BS and not related to the number of active UEs. However, in our scenario with three GCs, q is equal to one when UE_2 is the first active UE in its NC, which means higher transmit power compared to a regular strategy. Although adopting the proposed algorithm can increase the energy consumption of the UEs, this does not necessary mean decreasing their energy efficiency. Since the energy efficiency is defined as the number of GoodBits per unit of consumed energy, the energy efficiency of the UEs with modified power control would be higher as the reliability and ESR are higher. This topic should be investigated comprehensively in future study as energy consumption and energy efficiency are beyond the scope of this work. Furthermore, the power back-off step, q, can be optimized for maximum energy efficiency under URLL constraint. The effect of network parameters on energy efficiency can also be investigated. As an example, energy consumption grows with an increase in n_b since more data bits are transmitted. However, there might be an optimum value for n_b to achieve maximum energy efficiency since low values of blocklength have low energy consumption but low reliability (low GoodBit) and high values have high energy consumption but high reliability, which means high GoodBit.

One of the well-known and promising techniques to control the congestion in communication networks is ACB. Such a technique basically controls the number of simultaneous access attempts by adjusting the barring factor, which is the probability of attempting RA for each device. Since the ACB factor is affected by the number of preamble collisions and/or the number of active devices, ACB techniques are able to control the congestion only by the available resources associated with the first step of Random Access Procedure (RAP). Hence, they can not adapt to the dynamic network scenarios with non-periodic and bursty traffic as the whole network resources including the number of preambles, number of backlogged users, number of resource blocks, etc., must be considered for massive communications. Even if a large number of UEs can transmit collision-free preambles in the first step of RAP, numerous bottlenecks may occur during the remaining steps of RAP. However, the proposed DQ-based is not just a MAC congestion control and it affects the whole transmission scheme considering the number of RBs, blocklength, contention subslots, number of preambles (which is related to the number of GCs). The proposed protocol can be adaptive in terms of the number of CSS that can be learned frame by frame based on the network traffic load to run a trade-off between TQ and CQ length and reach the optimum value of packet latency (this will be considered in future work).

8 Conclusions

This paper introduced distributed queuing (DQ) in NOMA-based transmissions to support massive connectivity in next generation networks. Particularly, a random access scheme for massive IoT networks that can support URLL requirements has been proposed based on DQ in uplink RA-NOMA transmissions, where short packet transmission in FBL regime is employed to meet the target reliability and average packet latency. A frame structure is proposed to provide the dynamic NOMA clustering of the nodes' sporadic data transmissions. Furthermore, to avoid power collision between the nodes sporadically transmitting data packets, adaptive back-off power strategy is adopted. In such a power control strategy, each node adjusts its transmit power at the beginning of each frame according to its own activation index and the total number of active nodes in its NC. Network metrics such as reliability, delay violation probability, and effective sum rate have been analytically derived. The effect of different network parameters such as blocklength, active user arrival rate, and number of contention subslots, have been investigated on the derived metrics.

Acknowledgement

This work is funded by Ericsson Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] G. Hampel, C. Li, and J. Li, "5G Ultra-Reliable Low-Latency Communications in Factory Automation Leveraging Licensed and Unlicensed Bands," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 117–123, 2019.

[2] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High-Reliability and Low-Latency Wireless Communication for Internet of Things: Challenges, Fundamentals, and Enabling Technologies," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7946–7970, Oct. 2019.

[3] M. A. Siddiqi, H. Yu, and J. Joung, "5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices," *Electronics*, vol. 8, no. 9, 2019, article No. 981. [Online]. Available: https://www.mdpi.com/2079-9292/8/9/981

[4] M. Fuentes, J. L. Carcel, C. Dietrich, L. Yu, E. Garro, V. Pauli, F. I. Lazarakis, O. Grondalen, O. Bulakci, J. Yu, W. Mohr, and D. Gomez-Barquero, "5G New Radio Evaluation Against IMT-2020 Key Performance Indicators," *IEEE Access*, vol. 8, pp. 110 880–110 896, Jun. 2020.

[5] Cisco, "Cisco Annual Internet Report (2018–2023) white paper," 2020.

[6] W. Wu, Y. Li, Y. Zhang, B. Wang, and W. Wang, "Distributed Queueing-Based Random Access Protocol for LoRa Networks," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 763–772, 2020.

[7] Y. Wang, Z. Tian, and X. Cheng, "Enabling Technologies for Spectrum and Energy Efficient NOMA-Mmwave-MMIMO Systems," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 53–59, 2020.

[8] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive Non-Orthogonal Multiple Access for Cellular IoT: Potentials and Limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

[9] Y. Yuan, S. Wang, Y. Wu, H. V. Poor, Z. Ding, X. You, and L. Hanzo, "NOMA for Next-Generation Massive IoT: Performance Potential and Technology Directions," *IEEE Communications Magazine*, vol. 59, no. 7, pp. 115–121, 2021.

[10] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, Jul. 2017.

[11] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, "How URLLC can Benefit from NOMA-based Retransmissions," 2020.

[12] M. M. Ebrahimi, K. Khamforoosh, M. Amini, A. Sheikhahmadi, and H. Khamfroush, "Adaptive–Persistent Nonorthogonal Random Access Scheme for URLL Massive IoT Networks," *IEEE Systems Journal*, pp. 1–12, 2022.

[13] M. R. Amini and M. W. Baidas, "Performance Analysis of Grant-Free Random-Access NOMA in URLL IoT Networks," *IEEE Access*, vol. 9, pp. 105 974–105 988, 2021.

[14] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-Free Non-Orthogonal Multiple Access for IoT: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1805–1838, May 2020.

[15] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[16] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance Analysis and Optimal Access Class Barring Parameter Configuration in LTE-A Networks With Massive M2M Traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2018.

[17] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9847–9861, 2016.

[18] M. Bouzouita, Y. Hadjadj-Aoul, N. Zangar, G. Rubino, and S. Tabbane, "Applying Nonlinear Optimal Control Strategy for the Access Management of MTC Devices," in *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2016, pp. 901–906.

[19] W. Yu, C. H. Foh, A. U. Quddus, Y. Liu, and R. Tafazolli, "Throughput Analysis and User Barring Design for Uplink NOMA-Enabled Random Access," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6298–6314, 2021.

[20] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-Orthogonal Random Access for 5G Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4817–4831, Jul. 2017.

[21] J. Seo, B. C. Jung, and H. Jin, "Nonorthogonal Random Access for 5G Mobile Communication Systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7867–7871, 2018.

[22] Z. Chen, Y. Liu, S. Khairy, L. X. Cai, Y. Cheng, and R. Zhang, "Optimizing Non-Orthogonal Multiple Access in Random Access Networks," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[23] J.-B. Seo, H. Jin, and B. C. Jung, "Multichannel Uplink NOMA Random Access: Selection Diversity and Bistability," *IEEE Communications Letters*, vol. 23, no. 9, pp. 1515–1519, 2019.

[24] J.-B. Seo, B. C. Jung, and H. Jin, "Performance Analysis of NOMA Random Access," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2242–2245, 2018.

[25] J. Choi, "Random Access With Layered Preambles Based on NOMA for Two Different Types of Devices in MTC," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 871–881, 2021.

[26] F. Cao, Y. Song, and Y. Yang, "ARQ Assisted Short-Packet Communications for NOMA Networks Over Nakagami-m Fading Channels," *IEEE Access*, vol. 8, pp. 158 263–158 272, 2020.

[27] W. Xu and G. Campbell, "A Near Perfect Stable Random Access Protocol for a Broadcast Channel," in *[Conference Record] SUPERCOMM/ICC '92 Discovering a New World of Communications*, 1992, pp. 370–374 vol.1.

[28] A. Laya, L. Alonso, and J. Alonso-Zarate, "Contention resolution queues for massive machine type communications in LTE," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2015, pp. 2314–2318.

[29] A.-T. H. Bui, C. T. Nguyen, T. C. Thang, and A. T. Pham, "Design and Performance Analysis of a Novel Distributed Queue Access Protocol for Cellular-Based Massive M2M Communications," *IEEE Access*, vol. 6, pp. 3008–3019, 2018.

[30] A. T. H. Bui, C. T. Nguyen, T. C. Thang, and A. T. Pham, "A Comprehensive Distributed Queue-Based Random Access Framework for mMTC in LTE/LTE-A Networks With Mixed-Type Traffic," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 107–12 120, 2019.

[31] W. Wu, W. Wang, J. Yang, and B. Wang, "A Random Access Control Scheme for a NOMA-Enabled LoRa Network", booktitle="Communications and Networking," H. Gao, P. Fan, J. Wun, X. Xiaoping, J. Yu, and Y. Wang, Eds. Cham: Springer International Publishing, 2021, pp. 403–420.

[32] 3GPP TS 36.211 V13.2.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," 2016. [Online]. Available: www.portal.3gpp.org

[33] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice, 2nd Ed.* Wiley, 2011.

[34] E. Balevi, F. T. A. Rabee, and R. D. Gitlin, "ALOHA-NOMA for Massive Machine-to-Machine IoT Communication," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–5.

[35] M. Elkourdi, A. Mazin, E. Balevi, and R. D. Gitlin, "Enabling slotted Aloha-NOMA for massive M2M communication in IoT networks," in *2018 IEEE 19th Wireless and Microwave Technology Conference (WAMICON)*, 2018, pp. 1–4.

[36] C.-P. Li and W.-C. Huang, "A Constructive Representation for the Fourier Dual of the Zadoff–Chu Sequences," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4221–4224, 2007.

[37] J.-C. Belfiore, G. Rekaya, and E. Viterbo, "The Golden Code: a 2/spl Times/2 Full-Rate Space-Time code with nonvanishing determinants," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1432–1436, 2005.

[38] T. Yang, L. Yang, Y. J. Guo, and J. Yuan, "A Non-Orthogonal Multiple-Access Scheme Using Reliable Physical-Layer Network Coding and Cascade-Computation Decoding," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1633–1645, 2017.

[39] ETSI TS 136 211 V11.0.0, "3rd Generation Partnership Project; LTE;Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," 2012. [Online]. Available: www.portal.3gpp.org

[40] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[41] Y. Gao, B. Xia, K. Xiao, Z. Chen, X. Li, and S. Zhang, "Theoretical Analysis of the Dynamic Decode Ordering SIC Receiver for Uplink NOMA Systems," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2246–2249, 2017.

[42] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink Nonorthogonal Multiple Access in 5G Systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[43] T.-T. Thi Nguyen, C.-B. Le, and D.-T. Do, *Implementation of a Non-orthogonal Multiple Access Scheme Under Practical Impairments*. Singapore: Springer Singapore, 2021, pp. 107–127.

[44] ETSI TS 138 300 V16.2.0, "5G; NR; NR and NG-RAN Overall description; Stage-2."