Performance Analysis of Random Access NOMA for Critical mIoT with Timer-Power Back-Off Strategy

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

***Abstract*—Massive machine type communication (mMTC) and Internet-of-Things (IoT) networks provide global connectivity for massive number of end devices anytime-anywhere. The most challenging part in implementing such networks is the development of spectrum access strategies to provide ultra-reliable low-latency (URLL) transmissions for a large number of nodes with sporadic traffic behavior. Such networks have to deploy spectrum-efficient transmission schemes, and thus, non-orthogonal multiple-access (NOMA) is considered as a viable solution, that can be used to provide high number of URLL transmissions. We propose a random access NOMA transmission protocol (RA-NOMA) for IoT networks with large number of clustered IoT devices is proposed. The nodes in the proposed scheme adopt timer and power back-off strategies to transmit their short packets in a collision-free NOMA-based manner to achieve the URLL requirements. Closed-form expressions for network metrics, namely, delay violation probability, average packet latency, reliability, and effective sum rate (ESR) are analytically derived. Furthermore, the effect of blocklength, back-off timer (countdown value) and the number of active nodes on network metrics is explored. Additionally, the effect of the estimation error on the number of active nodes is analyzed and the impact of preamble transmit power on reliability in the presence of estimation error is investigated. Moreover, mathematical expressions for network metrics are also derived for NOMA-ALOHA with transmission diversity (NOMA-ALOHA-TD) in the underlying scenario, and the obtained results from NOMA-ALOHA-TD and the proposed RA-NOMA are compared.**

***Index Terms*—Internet-of-Things, Latency, NOMA, Random Access, Reliability.**

I. Introduction

ULTRA-reliable and low-latency communications (URLLC) and (massive) machine type communication (m)MTC are two key features of the fifth-generation (5G) networks [1]. Critical mMTC simultaneously needs to deal with emerging use cases such as disaster monitoring in wide areas, wireless factory automation in delay-sensitive industry 4.0 scenarios, industrial internet of things (IIoT), and autonomous vehicles [2]. However, existing standardized wireless communication protocols are inefficient to support URLL requirements in mMTC [3,4]. This is because reliability, latency, and massive connectivity are conflicting requirements that impose challenges and trade-offs on the protocol design [5]–[7]. Therefore, such applications call for novel spectrum-efficient transmission techniques to meet their demands [8]–[10].

Mohammad Reza Amini is with Department of Systems and Computer Engineering, Carleton University, (email: mohammadreza-amini@cunet.carleton.ca)

Ala'a Al-Habashna is with Department of Systems and Computer Engineering, Carleton University, (email: AlaaAlHabashna@cmail.carleton.ca)

Gabriel Wainer is with Department of Systems and Computer Engineering, Carleton University, (email: gwainer@sce.carleton.ca)

Gary Boudreau , Senior Member Ericsson Canada, Ottawa, ON, L4W 5K4, Canada, (email:gary.boudreau@ericsson.com)

A. Related Work

In recent years, various strategies have been proposed to fulfill the URLL requirement in critical MTC [11,12]. One approach to meet the latency requirements is to use signature-based non-orthogonal multiple access (NOMA) [13,14]. The key idea of NOMA is to serve multiple users within the same time-frequency resource blocks (RB), with the aid of superposition coding (SC) techniques at the transmitter and successive interference cancellation (SIC) at the receiver. This is fundamentally different from the classic orthogonal multiple access (OMA) techniques wherein orthogonal resources are assigned to different users [15]. NOMA has also shown to be superior to the conventional OMA in terms of network throughput [16]. Another way to achieve low latency communications is to exploit short packet transmissions via finite blocklength (FBL) regime [17,18]. On the other hand, ultra reliability can be achieved by diversity transmission techniques [19]–[21].

Due to resource scarcity, if a massive number of devices (nodes or users) with sporadic data traffic behavior needs to be connected in a network, the network resources cannot be allocated to each device. In such case, random access (RA) strategies can be used [11,12] in which each device contends with the others for accessing the network resources. However, conventional RA techniques (e.g. those available in the current RA-LTE standard) cannot be used for low-latency transmissions due to grant acquisition delay and excessive signaling overhead [22]. Hence, proposing and analyzing RA-NOMA for IoT networks with URLL requirements is of the essence. In [4], the non-orthogonal random access (NORA) scheme in cellular-based MTC was presented. The scheme has five steps, namely, cluster establishment by base station (BS), preamble transmission by cluster centers, random access response by BS, power adjustment and data transmission by the nodes, and performing SIC and sending acknowledge (ACK) by BS. The authors then found optimum power allocation by formulating an energy efficiency maximization problem for their scheme.

The authors claim that the proposed protocol reduces the signaling overhead as well as the complexity of the resource allocation problem.

In [23], the difference in the time of arrival is exploited to identify several UEs with the same preamble. The authors have leveraged power domain multiplexing of collided UEs with the BS performing SIC based on the information achieved in the preamble detection phase. Results show that the proposed structure can provide 30% increase in the number of supported UEs when compared to conventional orthogonal RA (ORA) scheme.

In [24], the authors proposed a NORA scheme in 5G in which the UEs exploit channel inversion so that their received power at BS sets at one of two predefined levels. The throughput performance of this scheme is about twice the conventional S-ALOHA. In [24], the authors introduce NOMA-based S-ALOHA with a p-persistent strategy in which the UEs adjust their transmit powers at random, with certain probabilities for the two predefined transmit power values. To extend the channel inversion strategy used in [24], the authors in [25] have attempted to incorporate multichannel selection diversity to give the users the capability of selecting the best available channel. This helps users to avoid high power transmission, improving network energy efficiency (EE).

In [26], the authors 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 the L target values. Furthermore, the UEs are assumed to be uniformly located in a cell of radius R. The performance of the proposed scheme in terms of throughput, energy efficiency, and bistability has been investigated.

In [27], a NOMA-based random access scheme in MTC scenario with two levels of priority was proposed. Two sets of preambles are allocated for two types of devices. High priority devices (delay-sensitive devices) select their preambles from an orthogonal set. Since they are orthogonal, preamble collision is reduced, and preamble re-transmissions is avoided. Therefore, access delay is decreased. To lower the computational complexity, low priority devices (delay-tolerant devices) are allocated non-orthogonal preamble sets, with higher number of preambles. The detection performance of transmitted preambles by high and low priority devices has been investigated and a low-complexity preamble detection method has been also proposed.

In [28], the performance of NOMA-based RA scheme with two users (near and far users) transmitting data through FBL regime has been analyzed over Nakagami-m fading channels. The authors have shown that exploiting ARQ strategy improves the packet error rate for both near and far users. Furthermore, the optimal block length to maximize the effective throughput has been identified.

It is worth noting that the data traffic behavior in most of the mentioned studies has been considered saturated and non-stochastic which is not the case in mIoT scenario, where small packets are transmitted sporadically by a large number of nodes.

B. Motivation and Contributions

As mentioned, the lack of communication protocols for critical mMTC and critical mIoT scenarios has motivated us to focus on proposing and designing NOMA-based transmission schemes in this paper. The main contributions of this paper are summarized as follows:

- introduced a contention-based collision-free RA-NOMA scheme for massive IoT networks with URLL requirements. Particularly, the proposed scheme is considered in a URLL-IoT network with geographically clustered nodes exploiting RA-NOMA transmission. We use a Timer-Power back-off strategy in which each node controls its transmit power in a distributed fashion to minimize power interference and also contends with other nodes to avoid inter-cluster collision in FBL transmissions which helps achieve URLL requirements.
- Analyzed the performance of the proposed scheme. More specifically, mathematical expressions for network metrics such as, effective sum rate, reliability, delay violation probability, and average packet latency are derived. Moreover, analytical derivations are numerically validated and the effect of back-off timer, blocklength, number of active nodes, back-off timer (countdown value) and their trade-offs on network metrics are explored. Furthermore, theoretical expression for reliability in the presence of estimation error on the number of active nodes is obtained and the impact of preamble transmit power on reliability along with other network parameters is investigated.
- Proposed frame structure that enables BS to perform dynamic SIC ordering at each frame where dynamic NOMA clustering is inherently done due to the random access nature of the nodes with sporadic data traffic behavior. Such a structure can be easily modified into heterogeneous networks where nodes with different QoS requirements and priority levels co-exist.
- Compared the derived network metrics with network metrics derived specifically for NOMA-ALOHA with transmission diversity in the underlying scenario.

The rest of this paper is organized as follows. Section II describes system model. Analytical derivation of the network performance metrics is presented in Section III. Section IV provides the performance metrics for NOMA-ALOHA with transmission diversity in the underlying scenario. Numerical results for the proposed RA-NOMA as well as their comparison with NOMA-ALOHA is presented in Section V. The effect of the estimation error on the node's reliability is presented in Section VI. Section VII outlines some research directions can be pursued by the researchers in the future. Finally, some conclusions are given in Section VIII.

II. System Model

A. IoT network model

An IoT network containing a BS and a number of nodes with a prescribed reliability and latency requirements is considered. The BS is located at the center of a cell with radius d_{max} which consists of some spatial sectors[1]. Without loss of

[1]The multi-cell scenario is out of scope of this paper.

generality, a typical sector is considered in which there is a large number of active nodes transmitting their data packets to the BS through a contention-based random access strategy described in Section II-B[2]. The nodes are randomly deployed in the underlying sector. The sector coverage area is partitioned into N_g annuli called *Geographical Cluster (GC)*. Without loss of generality and to make the analyses simpler, the number of nodes at each GC is assumed to be M_g. Fig. 1 depicts the network architecture.

All the links in the network experience independent but not necessarily identically distributed Rayleigh block fading and that is assumed constant during each transmission time. The channel coefficient between a typical node in the i^{th} GC U_i and the BS is denoted by h_i. Therefore, the corresponding channel gain $|h_i|^2$ 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. Hence, the received power P_i from U_i at the BS is $P_i = P_{i,t}|h_i|^2$, where $P_{i,t}$ is the U_i's transmit power. Furthermore, the background noise in all communication links is assumed to be independent and identically distributed (i.i.d.) 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. Furthermore, the packet arrival process for each IoT node is assumed Poisson with arrival rate of λ_p. All the packets are stored in the nodes' buffer before transmission.

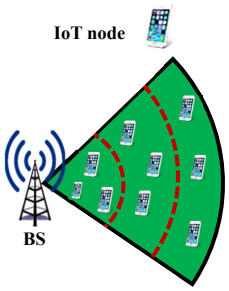

Fig. 1. IoT Network model with three GCs and three nodes at each GC

B. Frame Structure and Channel Access

Since there are massive number of nodes in the network, it is not possible to devote orthogonal resources to each user in advance. Therefore, contention-based RA-NOMA with the frame structure shown in Fig. 2 is adopted in this study. As can be seen, each frame consists of three phases, namely preamble phase, feedback phase, and transmission phase. In the first phase, all active nodes at each GC randomly select a resource block (RB) from R_b total orthogonal RBs to send a specific preamble sequence assigned to each GC[3]. Note that the number of preamble sequences equals to the number of GCs, N_g. Furthermore, all active nodes transmit their preambles so that the preambles are received at the same power level at the BS[4]. Note that preamble collision between nodes in the same GC (intra-GC collision) may happen when the same RB is selected by more than one node. To avoid intra-GC collision, distributed contention mechanism is employed. This strategy exploits back-off timer to determine the winner node, i.e., the node at each GC that can transmit the preamble on each RB. In such a mechanism, after selecting their RB, all active nodes at GCs randomly choose an integer value k from the set $\mathcal{K} = \{1, \ldots, k_d\}$. Based on the selected value of k, each node initializes its back-off timer with that value and then starts counting down. Each node starts transmitting its GC preamble on the selected RB only when its timer goes off and no other node initiates transmitting preambles during countdown. To this aim, all nodes must listen to the selected RB by sensing their GC preamble[5]. Such contention strategy helps reduce intra-GC preamble collisions since the collision occurs only when at least two nodes select the same RB and the same countdown value. Therefore, the preamble phase duration equals $T_p = k_d \times T_{pre}$, where T_{pre} is the preamble sequence duration.

Remark 1. *The BS can distinguish active nodes in different GCs transmitting preambles within the same RB because the nodes in different GCs use orthogonal preambles. On the other hand, it can also figure out any intra-GC preamble collision by measuring the received power level of a specific preamble. To this aim, statistical approaches such as hypothesis testing can also be employed [31,32].*

Remark 2. *Practically, in order to correctly decode the received preambles by both BS and the nodes in the GC, the set of preambles must have fine auto and cross correlation properties. Zadoff-chu [33], Golden codes [34], and Reed Muller [35] are the examples of appropriate sequences.*

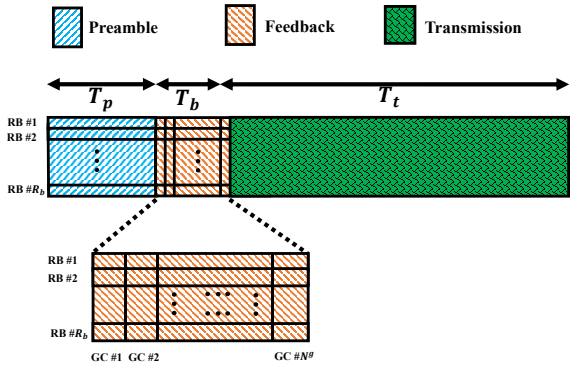

Fig. 2. Frame Structure of IoT nodes

Remark 3. *Transmitting preambles by each node enables the BS not only to detect active nodes on each RB, but also to*

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

[3]Orthogonal preamble transmissions are used in RA-LTE, and defined in 3GPP RA [29,30].

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

[5]Note that since all the nodes in the GC are at the proximity to each other, preamble sensing is applicable and hidden node problem is resolved.

figure out the SIC decoding order dynamically at each frame[6].

The second phase of a frame consists of N_g mini slots on each RB. This phase lasts for T_b seconds within which the BS puts all the received information of the underlying sector into a table-like entity called ***RB-GC Map*** and then broadcasts it to all the nodes of the sector as shown in Fig. 2. The RB-GC Map has R_b rows and N_g columns. Each cell in the RB-GC map informs about the status of its RB in the corresponding GC and can get three states: 1. empty (no device contends), 2. success (only one device has won the contention), and 3. collision (two or more devices have been collided). In this way, the nodes are informed about the status of their selected RB both in their GC and in other GCs. Hence, a typical node starts transmitting data in the third phase if it is reported success by the BS on its selected RB at its GCs.

Definition 1 (NOMA Cluster)**.** *All the Nodes in different GCs that select the same RB and are the winner of the contention constitutes the NOMA Cluster (NC) on that RB.*

Note that if an IoT node from the i^{th} GC starts transmitting data on the j^{th} RB, that GC is said to be active on that RB. Furthermore, $\mathbb{I}_i^j$ is defined as the activation index for the i^{th} GC on j^{th} RB where $\mathbb{I}_i^j = 1$ indicates that the i^{th} GC is active on j^{th} RB and $\mathbb{I}_i^j = 0$ otherwise.

Note that the total number of NCs is R_b. Without loss of generality, it is assumed that U_i, which is the nodes of interest, selects the j^{th} RB which means U_i lies in the j^{th} NC. Note that, through RB-GC Map, each node knows all the necessary information on its own NC, such as its index, number of nodes in NC, and the GC index of each active node in its NC.

In the third phase of the frame which lasts for T_t seconds, IoT nodes in different GCs that win the contention start transmitting their data packets on the selected RBs by employing PD-NOMA. The BS then applies SIC to decode the node's received signal on all RBs.

C. Power Back-off Strategy and Decoding Error

In NOMA-based transmissions, the nodes' transmit power plays an important role in the performance of the whole system, since it significantly affects the decoding error. Therefore, a power control strategy must be employed to reduce co-channel interference as much as possible. Power back-off strategy is the most common power control method used in the UL NOMA [31,36]–[38] in which the transmit power of the q^{th} active node in the i^{th} GC at a typical NC is expressed as,

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

where P_{max} and P_u are the maximum transmit power and target arrived 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 q^{th} active node in its NC, its received power is given as,

$$P_i^q = \min\{P_{max} - PL_i, P_u - (q-1)\varrho\} \tag{2}$$

Moreover, it is assumed that the IoT nodes 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 the signal of node U_i at the BS is approximated as [39],

$$\epsilon_i = Q\left(\sqrt{\frac{n_b}{\chi(\gamma_i)}\left(\mathcal{C}(\gamma_i) - \frac{n_d}{n_b}\right)}\right), \tag{3}$$

where $\mathcal{C}(\gamma_i) = \log_2(1+\gamma_i)$ is the Shannon capacity of U_i, while $\chi(\gamma_i) = \left(1 - \frac{1}{1+\gamma_i^2}\right)(\log_2 e)^2$ represents the channel dispersion. Furthermore, γ_i is the Signal-to Interference plus Noise Ratio (SINR) for U_i's signal at the BS.

Note that since short packet transmissions are adopted, the transmission phase duration equals $T_t = \frac{n_b}{B}$. Table I summarizes the main symbols used in this study and their descriptions.

TABLE I
NOTATIONS

Parameter	Description
$\gamma_{i\|q,\alpha}$	SINR for U_i's signal at the BS provided that U_i is the q^{th} active node of total α nodes in its NC
B	Bandwidth of each RB
P_i^q	Transmit power of U_i provided that it is the b^{th} active node in its NC cluster
n_b	Blocklength
n_d	Number of data bits in a packet
N_g	Number of Geographical clusters
N_0	Noise spectral density
M_g	Number of IoT nodes in a GC of interest
T_p	Preamble phase duration
T_b	Feedback phase duration
T_t	Transmission phase duration
T_{pre}	Preamble sequence duration
ν	Path-loss exponent
R_b	Number of RBs
T_f	Frame duration
P_{max}	Maximum transmit power
ϱ	power back-off step
L	number of packet replicas in NOMA-ALOHA-TD
N_0	Noise power spectral density

III. Derivation of Performance Metrics

A. Definitions

Definition 2 (Average Packet Latency)**.** *The Average packet latency $\bar{\mathcal{D}}$ is defined as the average delay of delivering a typical packet and all replicas of that packet to the BS*[7]*, which includes the transmission delay and the queuing (buffer) waiting time.*

Definition 3 (Reliability)**.** *The transmission reliability $\mathcal{R}$ is defined as the probability that a typical packet transmitted*

[6]SIC decoding order is performed for the nodes in the same RB and in the same NC. Since the preambles are received with the same power, the famous Near-Far effect is no longer impacts the detection of different orthogonal preambles of the nodes in the same RB and the same NC. Furthermore, by receiving preambles, the BS knows which GCs are active at each RB and hence,it is ready to perform SIC absed on the power control scheme.

[7]Note that transmitting a packet replicas is defined in Section IV for NOMA-ALOHA protocol and it is not considered in the proposed scheme.

$$\mathcal{P}_{T_f} \triangleq \Pr\left(\mathcal{T}=T_f\right)=\sum_{a=0}^{M_g-1}\sum_{t=0}^{k_d-1}\sum_{n_a=a}^{M_g-1}\binom{N_a-1}{a}\binom{M_g}{n_a}\frac{1}{k_d}\left(\frac{k_d-t-1}{k_dR_b}\right)^{a}\left(1-\frac{1}{R_b}\right)^{M_g-a-1}\times\left(1-e^{-\lambda_pT_f}\right)^{n_a}\left(e^{-\lambda_pT_f}\right)^{M_g-n_a-1}. \tag{5}$$

$$\mathcal{R}_i^{NM}=\sum_{q=1}^{i}\sum_{\alpha=q}^{q+N_g-i}\left[\prod_{l=1}^{q}\left(1-\epsilon_{i|l,\alpha}\right)\binom{i-1}{q-1}\binom{N_g-i}{\alpha-q}\left(\mathcal{P}_{\mathbb{I}}\right)^{\alpha-1}\left(1-\mathcal{P}_{\mathbb{I}}\right)^{N_g-\alpha}\right], \tag{7}$$

$$\mathcal{P}_{\mathbb{I}}=\sum_{m=0}^{M_g}\left[\sum_{t=0}^{k_d-1}\frac{m}{k_d}\left(\frac{k_d-t-1}{k_d}\right)^{m-1}\right]\binom{M_g}{m}\left(\frac{1}{R_b}\right)^{m}\left(1-\frac{1}{R_b}\right)^{M_g-m}. \tag{8}$$

from an IoT node is received successfully at the BS (i.e. without any decoding errors[8]).

Definition 4 (Effective Sum Rate)**.** *The number of data bits per time unit which is delivered successfully at the BS is defined as the effective rate for each IoT node. Accordingly, the sum of all node's effective rate is the effective sum rate (ESR) of the network.*

B. Delay distribution and Average Packet Latency

To consider both transmission delay and buffer waiting time in the packet latency, the packet arrival process for each IoT node is assumed Poisson with arrival rate of λ_p[9]. Note that if a node is not able to transmit data in a frame (in the case of losing the contention), it waits for the next frame to try. Therefore, transmission time is a random variable depending on the number of RBs, the number of back-off timer values, and the number of nodes in GCs. Lemma 1 gives the packet delay distribution.

Lemma 1. *The cumulative density function (CDF) of the packet delay for U_i's packet $F_{\mathcal{D}}(t)$ is obtained as,*

$$F_{\mathcal{D}}(t)=Pr(\mathcal{D}<t)=1-e^{-\varpi t}, \tag{4}$$

where $\varpi=\ln(\frac{1}{1-\mathcal{P}_{T_f}})T_f^{-1}-\lambda_p$ and $\mathcal{P}_{T_f}$, which is the probability that U_i succeeds to send its packet by its first attempt, is as per (5). Furthermore, the average packet latency is given as,

$$\overline{\mathcal{D}}^{NM}=\left[\ln(\frac{1}{1-\mathcal{P}_{T_f}})T_f^{-1}-\lambda_p\right]^{-1}, \tag{6}$$

where $T_f=T_p+T_b+\frac{n_b}{B}$.

Proof: See Appendix A. ∎

According to (4), the delay violation probability is concluded as $\Pr(\mathcal{D}>t)=e^{-\varpi t}$, which is used as a statistical constraint when it is set to remain lower than a threshold value.

C. Reliability

The reliability for the node U_i, $\mathcal{R}_i^{NM}$, is obtained in Lemma 2.

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

[9]This model has been widely used in communication networks [40,41].

Lemma 2. *The U_i's reliability, $\mathcal{R}_i^{NM}$, is obtained as per (7), in which $\mathcal{P}_{\mathbb{I}}$, which is the probability that the i^{th} GC is active on the j^{th} RB, is given in (8). Furthermore, $\epsilon_{i|l,\alpha}$ is the conditional block error rate of decoding the signal of node U_i at the BS, provided that U_i is the l^{th} node from the total α active nodes in the j^{th} NC specified as,*

$$\epsilon_{i|l,\alpha}=Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{i|l,\alpha}\right)}}\left(\mathcal{C}\left(\gamma_{i|l,\alpha}\right)-\frac{n_d}{n_b}\right)\right), \tag{9}$$

$\gamma_{i|l,\alpha}$ is the SINR for U_i's signal at the BS, which is given as,

$$\gamma_{i|l,\alpha}=\frac{P_i^l}{\sum_{h=l+1}^{\alpha}P_i^l+\sigma^2}. \tag{10}$$

Proof: See Appendix B. ∎

Remark 4. *Note that in deriving $\mathcal{R}_i^{NM}$ in (7), All the nodes are assumed to have data in buffer in order to investigate the worst case scenario for reliability. This assumption is also considered in deriving ESR in Lemma 3.*

D. Effective Sum Rate

According to the definition provided in subsection III-A, the ESR for IoT node U_i is defined as the sum of all node's effective rate in the network. To derive such a metric, the successfully decoded data bits transmitted within time unit from all nodes to the BS must be considered. To this aim, Define $\mathcal{I}_k$ as a set of all combinations of k active GCs on the j^{th} RB which can be mathematically written in terms of GC activation index as $\mathcal{I}_k=\{\mathbf{I}_k=(\mathbb{I}_1^j,\cdots,\mathbb{I}_{N_g}^j)\mid\sum_{i=1}^{N_g}\mathbb{I}_i^j=k\}$ with $\parallel\mathcal{I}_k\parallel=\binom{N_g}{k}$. Lemma 3 provides a mathematical expression for ESR.

Lemma 3. *The effective sum rate of the proposed RA-NOMA MAC protocol is given as (11), where*

$$\epsilon_{l|\mathbf{I}_k}=Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{l|\mathbf{I}_k}\right)}}\left(\mathcal{C}\left(\gamma_{l|\mathbf{I}_k}\right)-\frac{n_d}{n_b}\right)\right), \tag{12}$$

and

$$\gamma_{l|\mathbf{I}_k}=\frac{P_l^{G(l)}}{\sum_{h=l+1}^{N_g}P_h^{G(h)}+\sigma^2}, \tag{13}$$

in which $G(\beta)=\sum_{q=1}^{\beta}\mathbb{I}_\beta^j$

Proof: See Appendix C. ∎

$$ESR^{NM} = R_b \sum_{k=1}^{N_g} \left((\mathcal{P}_{\mathbb{I}})^k (1-\mathcal{P}_{\mathbb{I}})^{N_g-k} \sum_{all\, \mathbf{I}_k \in \mathcal{I}_k} \sum_{i=1}^{N_g} \mathbb{I}_i^j \frac{n_d}{T_f} \prod_{l=1}^{i} \left(1-\epsilon_{l|\mathbf{I}_k}\right) \right), \tag{11}$$

$$\Pr\left(E_i^{succ}\right) = \sum_{k=0}^{N_p-1} \binom{M_g-1}{k} \left(\frac{1}{R_b}\right)^k \left(1-\frac{1}{R_b}\right)^{M_g-k-1} \frac{(N_p-1)!}{(N_p-k-1)!(N_p)^k} \prod_{l=1}^{i} \left(1-\epsilon_{i|l,k+1}\right), \tag{15}$$

$$ESR^{NAL} = R_b \sum_{k=1}^{N_p} \left(\frac{M_g!}{(M_g-k)!} \left(\frac{1}{R_b N_p}\right)^k \left(1-\frac{1}{R_b}\right)^{M_g-k} \sum_{all\, \mathbf{J}_k \in \mathcal{J}_k} \sum_{i=1}^{N_p} \mathbb{J}_i^j \frac{n_d}{LT_f} \prod_{l=1}^{i} \left(1-\epsilon_{l|\mathbf{J}_k}\right) \right), \tag{16}$$

IV. NOMA-ALOHA with Transmission Diversity

In this section, network metrics for NOMA-ALOHA [42] with transmission diversity are derived which helps evaluate the proposed RA-NOMA MAC protocol. In NOMA-ALOHA-TD, all the active nodes randomly select one RB and then start transmitting their data on the selected RB with the channel inversion power control scheme [26,43]. To be consistent with the underlying scheme, it is assumed that there are N_p available power levels to be chosen by the nodes. These predefined power levels are set according to the power control strategy for the proposed scheme. To accommodate transmission diversity scheme, it is assumed that all the nodes transmit each data packet L times in L successive frames.

Note that NOMA-ALOHA-TD can be viewed as M/D/1 queueing model, since the data arrival process is assumed Poisson with rate λ_p, with constant packet service time of LT_f. Then, the average packet latency equals $\overline{\mathcal{D}}^{NAL} = \frac{2-\rho}{2(1-\rho)} LT_f$ [44] in which $\rho = \lambda_p LT_f$. Furthermore, reliability for U_i in NOMA-ALOHA-TD is derived in Lemma 4.

Lemma 4. *The U_i's reliability metric in NOMA-ALOHA-TD is given as,*

$$\mathcal{R}_i^{NAL} = 1 - \left(1 - Pr(E_i^{succ})\right)^L, \tag{14}$$

where $Pr(E_i^{succ})$, the probability that a typical packet is successfully decoded at the BS, is derived as (15). Moreover, the conditional error probability $\epsilon_{l_1|l_2,l_3}$ is defined in (9).

Proof: See Appendix D. ∎

Furthermore, the ESR in NOMA-ALOHA-TD is derived in Lemma 5.

Lemma 5. *The effective sum rate of NOMA-ALOHA-TD is expressed as (16) in which,*

$$\epsilon_{l|\mathbf{J}_k} = Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{l|\mathbf{J}_k}\right)}} \left(\mathcal{C}\left(\gamma_{l|\mathbf{J}_k}\right) - \frac{n_d}{n_b} \right) \right), \tag{17}$$

and

$$\gamma_{l|\mathbf{J}_k} = \frac{P_l^{F(l)}}{\sum_{h=l+1}^{N_p} P_h^{F(h)} + \sigma^2}. \tag{18}$$

and $F(\beta) = \sum_{q=1}^{\beta} \mathbb{J}_\beta^j$. Furthermore, $\mathcal{G}_k$ is a set of all combinations of k active power levels on the j^{th} RB which can be mathematically written in terms of power level activation index as $\mathcal{J}_k = \{\mathbf{J}_k = (\mathbb{J}_1^j, \cdots, \mathbb{J}_{N_p}^j) \mid \sum_{i=1}^{N_p} \mathbb{J}_i^j = k\}$ with $\| \mathcal{J}_k \| = \binom{N_p}{k}$.

Proof: See Appendix E. ∎

V. Numerical Results and Comparison

This section evaluates the theoretical results derived in Section III and compares them with the metrics derived for NOMA-ALOHA-TD in Section IV. Without loss of generality, it is assumed that all the nodes in the i^{th} GC are located at the center of the GC. Particularly, the distance between U_i to the BS is assumed $d_i = \frac{(2i-1)d_{max}}{2N_g}$ $(i \in \{1, \cdots, N_g\})$.

TABLE II
Simulation Parameters

Parameter	Value	Parameter	Value
B	180 KHz	R_b	50
ν	2.5	T_{pre}	80 μs
T_b	15 μs	T_t	0.2 ms
k_d	3	N_0	-174 dBm/Hz
N_g	3	n_d	32 Bytes
P_u	10^{-6} W	ϱ	10 dB
P_{max}	0.3 W	d_{max}	1000 m

Fig. 3 shows the effect of blocklength n_b on the average packet latency for the proposed RA-NOMA scenario and NOMA-ALOHA with $L = 1, 2, 3, 4$. It can be observed that the average packet latency for both RA-NOMA and NOMA-ALOHA increases with the increase in n_b. This is because the higher the blocklength is, the longer the transmission duration, and hence, the greater the average packet latency. Furthermore, the higher the number of replicas used in NOMA-ALOHA, the higher the rate at which the latency increases which is the drawback of systems with re-transmissions. Another interesting observation is that although the average packet latency in RA-NOMA starts at higher values with lower values of n_b compared with that of NOMA-ALOHA, which is due to the extra signaling overhead of contention and feedback phases[10], it then increases with a lower rate so that the latency for $n_b > 55$ stays below the curves of NOMA-ALOHA with $L = 2, 3, 4$. This fulfills the low latency requirements in critical mIoT networks which is defined lower than 1 ms in industrial IoT applications [45].

The average packet latency for RA-NOMA versus the countdown value k_d for different values of M_g is depicted

[10]Note that for RA-NOMA when n_b is very small (near zero), the length of transmission phase is also very small. However, the average packet latency is not near zero and equals 0.33 ms. This is due to the overhead incurred by the contention and feedback phases.

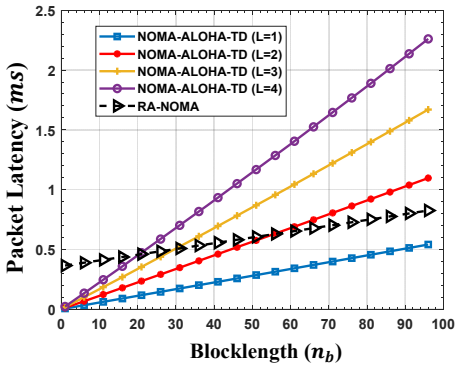

Fig. 3. Average packet latency vs. Blocklength - $M_g = 30$.

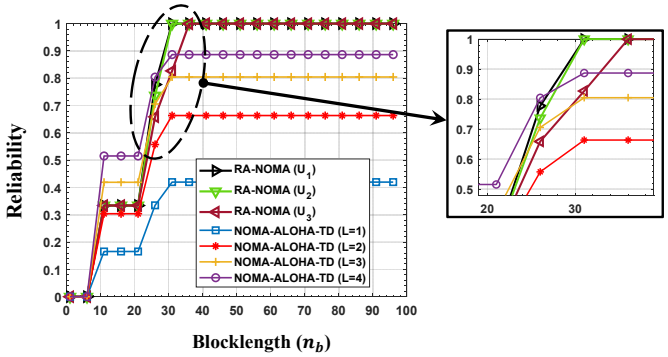

Fig. 5. Reliability vs. Blocklength - $M_g = 30$.

in Fig. 4. As can be seen, the average packet latency starts to decrease with increase in k_d and then reaches to a minimum value at $k_d = 2$ for $M_g = 30, 60$, and $k_d = 3$ for $M_g = 90$, respectively. This is because for small values of k_d, specifically $k_d = 1$, there is a high probability of collision in the contention phase, preventing IoT nodes from transmitting their data packets. Therefore, the average packet latency is high for $k_d = 1$ and starts to decrease with the increase in k_d . This can be observed from Fig. 4 especially for higher M_g where the average packet latency is greater than that for lower M_g, due to higher probability of collision when there are more active nodes in the GCs. However, when $k_d > 2$ for $M_g = 30, 60$ and $k_d > 3$ for $M_g = 90$, the packet latency starts to increase. This is because the excessive increase in k_d does not significantly reduce collision in the contention phase, yet incurs additional overhead in a frame, which raises the packet latency. Note that there is no contention phase in NOMA-ALOHA and hence, such a curve is not plotted for it.

The reliability of IoT nodes versus n_b for U_i $(i = 1, 2, 3)$ in RA-NOMA as well as that for NOMA-ALOHA-TD with different number of packet replicas is illustrated in Fig. 5. Generally speaking, at lower values of n_b, the decoding error probability in FBL regime is significant and hence, reliability is very low for both RA-NOMA and NOMA-ALOHA. The reliability then increases sharply with increase in n_b until it reaches its maximum value and remains almost constant where excessive increase in blocklength does not further improve the decoding error probability in FBL regime. However, it is seen

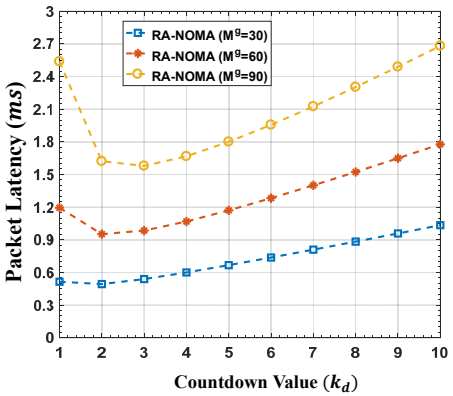

Fig. 4. Average packet latency vs. countdown value- $n_b = 33$.

that $\mathcal{R}_1^{NM} \geq \mathcal{R}_2^{NM} \geq \mathcal{R}_3^{NM}$ which is due to the fact that reliability of each node in a typical GC in RA-NOMA depends on reliability of previously decoded nodes in a NOMA cluster. Another observation is that $\mathcal{R}_1^{NM}$ experiences two level-offs after rising up, one in $10 < n_b < 21$ and another for $n_b > 30$. The reason is in the sporadic behavior of nodes' data transmission in NCs. Due to the stochastic data transmission pattern of the nodes in its NC, U_1 can be the only active node, the first node of the total of two active nodes, and the first node of the total of three active nodes. For each of the mentioned cases, U_1's SINR at the BS differs significantly. Therefore, the decoding error and consequently reliability behaves differently for each case when n_b changes. Since reliability metric reflects all the possible events in the networks, it experiences some rises at some values of n_b. Note that $\mathcal{R}_2^{NM}$ and $\mathcal{R}_3^{NM}$ are also affected by $\mathcal{R}_1^{NM}$, experiencing the same trend. When comparing reliability of both approaches, it can be seen that the maximum reliability achieved with NOMA-ALOHA is 0.9614 for $L = 4$ and is lower for $L < 4$ (0.9128 for $L = 3$, 0.8034 for $L = 2$, and 0.5566 for $L = 1$) which is far from critical IoT requirements, while reliability of the nodes in the proposed RA-NOMA can reach to 0.999999, fulfilling the URLL requirements.

The effect of the number of active nodes M_g on reliability for RA-NOMA and NOMA-ALOHA is plotted in Fig. 6. As can be observed, reliability in NOMA-ALOHA degrades severely with increase in the number of active nodes from 1 to 0.5 when M_g increases from 1 to 90 for $L = 4$ and more degradation for $L < 4$. However, the proposed RA-NOMA degrades less than 5% for the U_3 and less than 1% for U_1 and U_2. Therefore, the proposed protocol is expected to be more reliable in massive critical IoT networks.

The network effective sum rate as a function of blocklength n_b for both RA-NOMA and NOMA-ALOHA is shown in Fig. 7. It can be seen that ESR for NOMA-ALOHA increases with increase in n_b, experiencing its maximum and then starts to decrease. This is because increasing n_b from low values significantly improves the decoding error probability and reliability as mentioned for Fig. 5. Therefore, the effective number of error free data bits drastically increases which helps increasing the ESR. However, excessively increasing n_b does not improve the decoding performance, yet increases the frame duration while the same number of bits n_d is transferred within the frame. This increases the redundancy

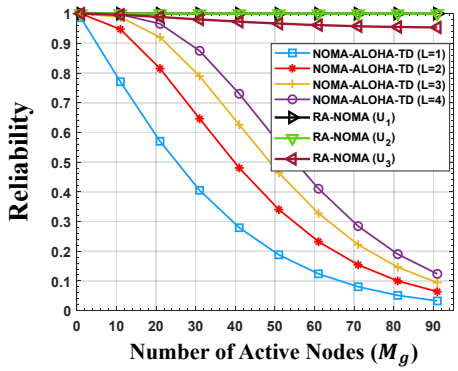

Fig. 6. Reliability vs. Number of active nodes- $n_b = 33$.

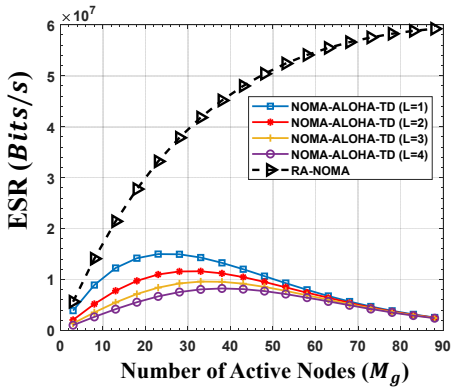

Fig. 8. Effective sum rate vs. Number of active nodes- $n_b = 33$.

and reduces the effective sum rate. Another observation is that NOMA-ALOHA with $L = 1$ has the most ESR than others. This is because although reliability increases with increase in the number of packet replicas, the redundancy imposed by the packet re-transmissions outweighs the improvement in decoding performance, lowering the ESR.

On the other hand, ESR in RA-NOMA experiences two sharp rises each preceded by a fall which happens at $10 < n_b < 20$ and $n_b > 35$. Looking at the nodes' reliability for RA-NOMA scenario in Fig. 5, it is found that the rises in ESR happen at the same positions as the rises in reliability, which means drastic decrease in decoding error and hence, significant improvement in the effective number of delivered data bits. The fall seen after the two rises is due to increasing block-length redundancy and lengthening the frame duration when reliability keeps its level. From the comparison perspective, it is worth noting that when $M_g = 30$ and $R_b = 50$, the proposed protocol has less ESR than NOMA-ALOHA with $L = 1$ for $n_b < 17$. However, two key points must be considered. The first one is that NOMA-ALOHA with $L = 1$ does not meet the critical IoT requirements.

The second point is that the proposed RA-NOMA achieves a significant improvement in terms of ESR at higher number of active nodes, which is the case of the underlying study and mMTC. This is verified in Fig. 8, where the ESR is plotted in terms of M_g. It can be observed that ESR in RA-NOMA significantly outperforms NOMA-ALOHA when massive number of nodes is considered. Particularly, for $M_g > 7$, $R_b = 50$,

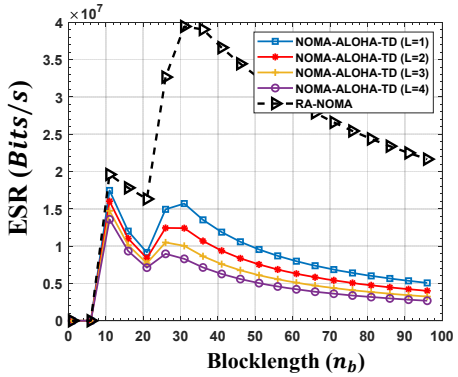

Fig. 7. Effective sum rate vs. Blocklength - $M_g = 30$.

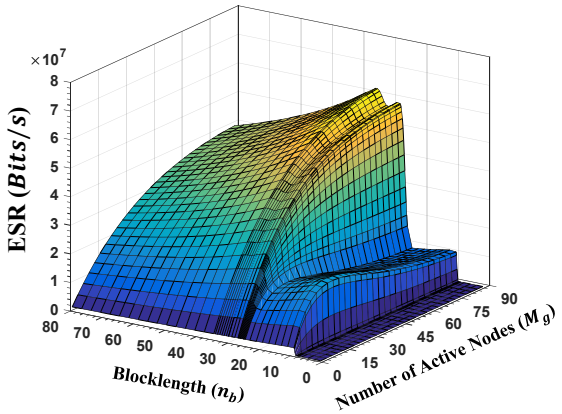

Fig. 9. Effective sum rate for RA-NOMA vs. Blocklength and number of active nodes- M_g=30.

and $n_b = 33$, ESR in the proposed RA-NOMA is greater than NOMA-ALOHA with $L = 1$. Another important observation is that unlike NOMA-ALOHA, ESR in RA-NOMA does not decline with increase in the number of active nodes and follows saturated exponentially increasing trend[11].

Fig. 9 shows the effect of n_b and M_g on ESR for RA-NOMA within the same graph. As explained for Figs. 7 and 8, the rising and falling behavior of ESR in terms of n_b and its increasing trends for massive number of nodes are clearly observed in this figure.

Fig. 10, depicts the delay violation probability, i.e., $\Pr(\mathcal{D} > 5ms)$ as a function of n_b and k_d. Generally speaking, it increases with the increase in n_b and k_d and this is due to the increase in the frame length. The plotted green area shows the zone at which the average packet latency with the probability less than 10^{-5} violates the threshold of $5ms$, which is satisfies the delay constraint in many critical IoT applications.

VI. Discussions

This section considers the effect of the estimation error on the number of active nodes. It is assumed that the digital energy detector is employed at the BS for detecting the number of active nodes. Furthermore, the channel is assumed constant during the preamble interval, the noise distribution is

[11]The ESR curve for the proposed RA-NOMA declines in higher values for the number of active nodes. The figure is not extended to have a clear snap shot of all the curves.

Gaussian with mean zero and variance σ^2 and it is assumed uncorrelated to the preambles. Additionally, the test statistic, $\Xi = \sum_{i=1}^{S} |y(i)|^2$, represents the sum of squares of S Gaussian but arbitrarily distributed samples of $y(i)$ as the i^{th} sample of the received signal. The number of samples should be at least $S = 2WT$ where WT_{pre} stands for the time-bandwidth product. In general, when the BS exploits multiple antennas, the sample covariance matrix of the received signal follows Wishart distribution. In the simple form, we assume only one antenna at the BS. Then, the distribution of the test statistic can be well approximated to normal distribution by using central limit theorem (CLT) for large enough number of samples. To define the number of active nodes, Ξ is compared with some pre-defined threshold values $\delta_{m'}$. Let $\mathcal{H} = \{H_0, \cdots, H_{M_g}\}$ be the set of all possible hypotheses defined to estimate the number of active nodes in which $H_{m'}$ indicates that there are m' active nodes transmitting their preambles to the BS. Accordingly, the conditional distribution of the test statistic under hypothesis $H_{m'}$ follows $\mathcal{N}\left(S(1+\xi_{m'})\sigma^2, 2S(1+\xi_{m'})\sigma^4\right)$, where $\xi_{m'} = \frac{m' P_{pre}}{\sigma^2}$ [46,47] and P_{pre} is the average received preamble power. Finally, Lemma 6 gives the U_i's reliability considering the underlying error.

Lemma 6. *The U_i's reliability under detection error, $\tilde{\mathcal{R}}_i^{NM}$, is obtained as per (17) in which $\tilde{\mathcal{P}}_{\mathbb{I}}$, the probability that the i^{th} GC be active on the j^{th} RB, is given in (18). Furthermore, $\mathcal{Q}_\theta$ in (18) is expressed as,*

$$\mathcal{Q}_\theta = Q\left(\frac{\delta_\theta - \sigma^2(1+\xi_\theta)S}{\sqrt{2S(1+\xi_\theta)\sigma^2}}\right) \tag{16}$$

where $Q(.)$ is the standard Q-function.

Proof: See Appendix F. ■

To investigate the effect of estimation error, Figs. 11 and 12 are plotted. Fig. 11 shows the U_3's reliability as a function of blocklength in the presence of estimation error. Since such error is highly dependent on the preamble transmit power, reliability is plotted for different values of preamble power ($P_{pre} = 1, 50, 100, 150, 200, 250, 300\, mW$). As can be seen, it follows the same trend as in Fig. 5. especially, when $P_{pre} = 200, 250, 300\, mW$, the estimation error is such insignificant that it has almost no effect on reliability, touching almost one

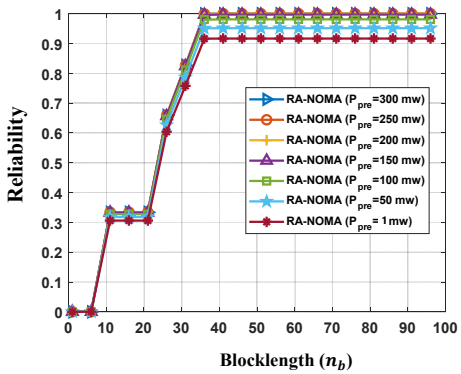

Fig. 11. U_3's Reliability vs. Blocklength for different values of preamble power- $M_g = 30$, $k_d = 3$.

for $n_b > 36$. However, for lower values of preamble transmit power, i.e., $P_{pre} = 150, 100, 50, 1\, mW$ reliability decreases to 0.9971, 0.9809, 0.9515, 0.9168, respectively, for $n_b > 36$. Such a behavior can be observed in Fig. 12 where reliability is plotted versus the number of active nodes for different values of preamble transmit power. U_3's reliability is almost one when $P_{pre} = 200, 250, 300\, mW$ and decreases to the lowest value of 0.9168 for $P_{pre} = 1\, mW$.

In Fig. 13 reliability is demonstrated as a function of k_d. Surprisingly, reliability increases with the increase in k_d. This is due to the fact that higher countdown values is equivalent to wider contention window which in turn, reduces the probability of collisions. This indicates that the low values of reliability can be compensated with increasing countdown value when transmit power is low. For example, when $P_{pre} = 1mW$, reliability increases from 0.78 for $k_d = 1$ to 0.991 for $k_d = 12$, and also from 0.87 for $k_d = 1$ to 0.9994 for $k_d = 12$ when $P_{pre} = 50mW$. However, increasing k_d will increase the average packet latency as shown in Fig. 4, indicating another reliability-latency trade-off for countdown value, similar to the trade-off for blocklength.

To verify our derivations, the Discrete Event Simulation (DES) in MATLAB is adopted. To this aim, all the events in the simulation scenario are time-stamped based on their statistics which include, generating packets for UEs, selecting resource block, choosing countdown value, completing trans-

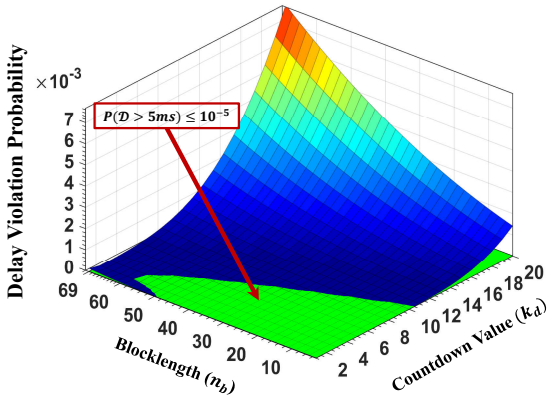

Fig. 10. Delay violation probability vs. Blocklength and countdown value- M_g=30.

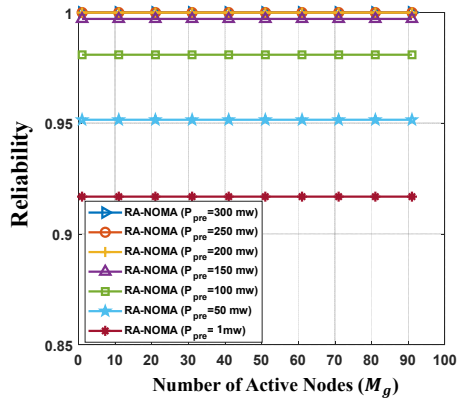

Fig. 12. U_3's Reliability vs. number of active nodes for different values of preamble power- $n_b = 40$, $k_d = 3$.

$$\tilde{\mathcal{R}}_i^{NM} = \sum_{q=1}^{i} \sum_{\alpha=q}^{q+N_g-i} \left[\prod_{l=1}^{q} \left(1-\epsilon_{i|l,\alpha}\right) \binom{i-1}{q-1} \binom{N_g-i}{\alpha-q} \left(\tilde{\mathcal{P}}_{\mathbb{I}}\right)^{\alpha-1} \left(1-\tilde{\mathcal{P}}_{\mathbb{I}}\right)^{N_g-\alpha} \right], \tag{17}$$

$$\tilde{\mathcal{P}}_{\mathbb{I}} = \sum_{m=0}^{M_g} \binom{M_g}{m} \left(\frac{1}{R_b}\right)^{m} \left(1-\frac{1}{R_b}\right)^{M_g-m} \left(\mathcal{Q}_1-\mathcal{Q}_2\right) \left[\sum_{t=0}^{k_d-1} \frac{m}{k_d} \left(\frac{k_d-t-1}{k_d}\right)^{m-1} \right]. \tag{18}$$

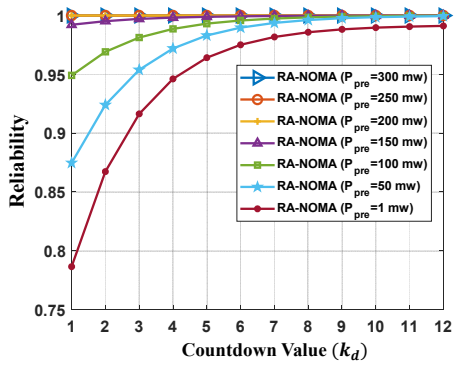

Fig. 13. U_3's Reliability vs. countdown value for different values of preamble power- $n_b = 40$, $M_g = 30$.

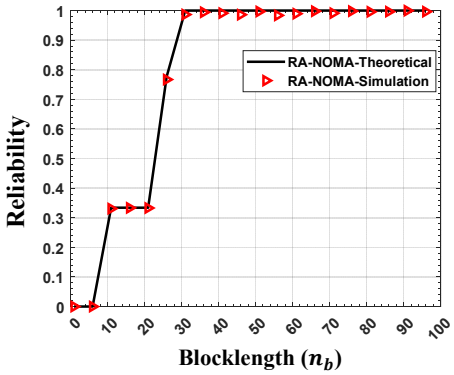

Fig. 14. Simulation and theoretical results for reliability vs. blocklength - $M_g = 30$.

missions and scheduling wait-transmission process. Particularly, starting from the first event, all the necessary metrics are calculated and the simulation time is proceeded to the next scheduled nearest event. After executing the simulation scenario for 50 times each for 30,000 seconds, the results have been averaged and plotted in Figs. 14 and 15. As can be seen, the simulation results agree with those obtained via the theoretical analyses.

Additional results on Root Mean Square Error (RMSE) between the simulation and theoretical results for reliability and packet latency have been plotted in Figs. 16 and 17. These RMSEs have been calculated for three different simulation times, 5×10^3, 1.5×10^4 and 3×10^4 seconds. The curves show that by increasing the simulation time from 5×10^3 seconds to 3×10^4 seconds, the RMSE between the simulation and theoretical results is drastically decreased. Intuitively, by increasing the duration of the simulation, we expect the simulated curves to perfectly coincide with the analytical derivations.

As the last topic in this section, it is worth taking nodes' energy consumption into consideration. Note that the energy consumption of the nodes in the proposed protocol is expected to be more than the regular RA-NOMA protocols since 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 GC index, the modified version in this paper relates the power level to the network's traffic by incorporating q in (1). To make it clear, consider the UEs in all GCs other than the first GC. In regular back-off power strategy, q is replaced with the GC index or some parameters related to UE's distance to the BS. However, in our scenario, q might equal to one when UE_2 is the first active UE in its NC, which means higher transmit power compared to regular strategy. However, from the energy efficiency perspective, the result might be different. 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 is higher. This topic should be investigated comprehensively in another study as energy consumption and energy efficiency is beyond the scope of this work. Furthermore, power back-off step can be optimized for maximum energy efficiency under URLL constraint. The effect of network parameters on the energy efficiency can also be investigated. As an example, energy consumption grows with increase in n_b since more data bits are transmitted. However, there might be an optimum value for n_b to have 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 (high GoodBit). Increasing k_d is expected to increase the energy efficiency since it does not affect energy consumption but it

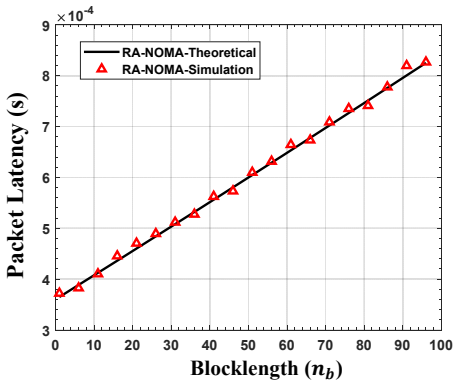

Fig. 15. Simulation and theoretical results for packet latency vs. blocklength - $M_g = 30$.

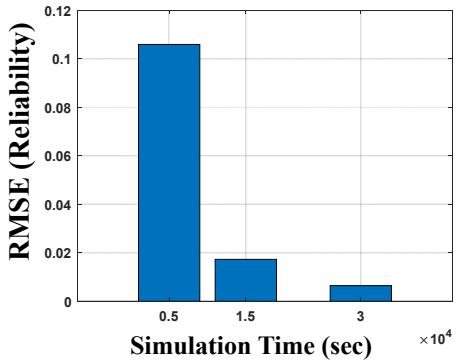

Fig. 16. RMSE between the simulation and theoretical results for reliability.

improves reliability (GoodBit).

VII. Future Research Directions

The proposed protocol is so flexible that can be adapted to other communication scenarios. To exemplify, it can be developed for heterogeneous networks by attributing a simple priority index to each node and relating such index to the node's back-off timer value.

Multi-cell scenario of the proposed scheme is also deserving of study. To overcome the multi-cell challenges in NOMA transmissions, coordinated scheduling/beamforming (CS/CB) and joint processing (JP) can be employed in the proposed protocol. Furthermore, Han-Kobayashi (HK) scheme or fractional frequency reuse (FFR) for the cell-edge regions of the adjacent cells to improve the fairness and avoid intra-cell interference are suggested.

Another idea is to include adaptive-persistent transmissions in the proposed protocol. More specifically, since all the nodes are aware of the status of its NC and other NCs through feedback information, they can transmit their data packets with certain probabilities in the same NC or other vacant NCs if collision is reported by the BS. The transmission probability can be devised as a function of the number of collided nodes in the current NC which is known by adopting active user detection.

Proposing machine learning techniques to be employed by the nodes to adjust their transmit power in distributed fashion based on the observed NC deployment of active nodes in the feedback phase is the other idea to be studied so that the energy-efficient protocol is achieved.

The final future research line related to this study is incorporating distributed queuing (DQ) into the proposed NOMA-based protocol such that the spectrum is fully utilized by the massive devices without any inter-nodes collision.

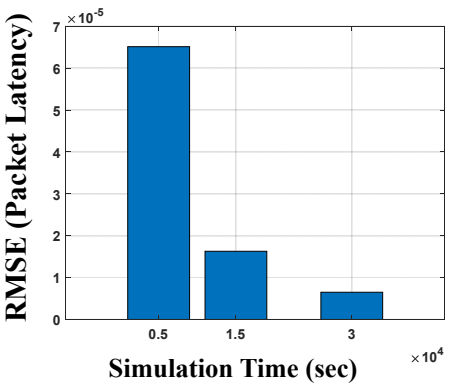

Fig. 17. RMSE between the simulation and theoretical results for average packet latency.

VIII. Conclusions

In this paper, a collision free NOMA-based random access (RA) transmission scheme for critical mMTC has been proposed. To realize URLL requirements, short packet transmissions with timer-power back-off strategy has been adopted in which the frame structure exploits back-off timer along with back-off power control technique to transmit short-length packets in order to meet the target reliability and latency requirements when there is a massive number of nodes in the network. Furthermore, the proposed frame structure enables BS to perform dynamic successive interference cancellation (SIC) ordering where NOMA clustering is dynamically done to resolve the clustering problems that exist in the networks with sporadic data traffic behavior. The delay distribution, reliability and effective sum rate have been analytically derived and the effects of blocklength, countdown value and the number of active nodes has been well investigated. It has been shown that although the proposed frame structure for RA-NOMA exploits additional signaling phases compared to conventional RA techniques like NOMA-ALOHA, it can fulfill the target reliability and latency requirements of URLL-MTC with massive number of nodes by setting the appropriate network parameters. Such URLL requirements cannot be achieved through NOMA-ALOHA with transmission diversity. Furthermore, the analyses show that the proposed RA-NOMA significantly outperforms the NOMA-ALOHA-TD in terms of ESR and reliability for a large number of active nodes.

Acknowledgement

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

Appendix A
Proof of Lemma 1

To derive the delay distribution, it should be noted that U_i's packet transmission time $\mathcal{T}$ lasts for T_f seconds if U_i succeeds to send its packet by its first attempt. This is equivalent to the event that U_i selects the lowest back-off timer value among all the values selected by the other active nodes in its GC choosing j^{th} RB for their transmission. Suppose there are $A_{i,j}$ active nodes– except the underlying node– in the i^{th} GC that selects the j^{th} RB ($A_{i,j} \leq M_g - 1$). Also, assume that the back-off timer value selected by U_i be $t_{i,j} \in \{0, \ldots, k_d - 1\}$. Defining $E_{A_{i,j}}^{t_{i,j}^+}$ as the event that the back-off timer values selected by other $A_{i,j}$ nodes in the i^{th} GC selecting the j^{th} RB is greater than $t_{i,j}$, then the probability that U_i succeeds to send its packet by its first attempt equals

$$\mathcal{P}_{T_f} \triangleq \Pr\left(\mathcal{T}=T_f\right)=\sum_{a=0}^{M_g-1}\sum_{t=0}^{k_d-1}\sum_{n_a=a}^{M_g-1}\binom{N_a-1}{a}\binom{M_g}{n_a}\frac{1}{k_d}\left(\frac{k_d-t-1}{k_dR_b}\right)^a\left(1-\frac{1}{R_b}\right)^{M_g-a-1}\times\left(1-e^{-\lambda_pT_f}\right)^{n_a}\left(e^{-\lambda_pT_f}\right)^{M_g-n_a-1}. \tag{A.6}$$

$$\begin{aligned}\Pr\left(\mathcal{T}=T_f\right)&=\Pr\left(E_{A_{i,j}}^{t_{i,j}^+}\right)\\&=\sum_{a=0}^{M_g-1}\Pr\left(E_{A_{i,j}}^{t_{i,j}^+}\mid A_{i,j}=a\right)\Pr(A_{i,j}=a)\\&=\sum_{a=0}^{M_g-1}\sum_{t=0}^{k_d-1}\Bigg[\Pr\left(E_{A_{i,j}}^{t_{i,j}^+}\mid A_{i,j}=a,t_{i,j}=t\right)\\&\qquad\Pr(t_{i,j}=t\mid A_{i,j}=a)\Pr(A_{i,j}=a)\Bigg].\end{aligned} \tag{A.1}$$

$\Pr\left(A_{i,j}=a\right)$ in (A.1) is written as,

$$\Pr\left(A_{i,j}=a\right)=\sum_{n_a=a}^{M_g-1}\Pr\left(A_{i,j}=a\mid N_a=n_a\right)\Pr\left(N_a=n_a\right)\binom{M_g-1}{a}\left(\frac{1}{R_b}\right)^a\left(1-\frac{1}{R_b}\right)^{M_g-a-1}, \tag{A.2}$$

where N_a shows the total number of active UEs in the i^{th} GC- except U_i-. Due to Poisson packet arrival assumption during a frame, $\Pr\left(N_a=n_a\right)$ is easily found as,

$$\Pr\left(N_a=n_a\right)=\binom{M_g-1}{n_a}\left(1-e^{-\lambda_pT_f}\right)^{n_a}\left(e^{-\lambda_pT_f}\right)^{M_g-n_a-1} \tag{A.3}$$

Furthermore, $\Pr\left(A_{i,j}=a\mid N_a=n_a\right)$ in (A.2) can be written as,

$$\Pr\left(A_{i,j}=a\mid N_a=n_a\right)=\binom{n_a}{a}\left(\frac{1}{R_b}\right)^a\left(1-\frac{1}{R_b}\right)^{n_a-a}. \tag{A.4}$$

Note that $\Pr\left(t_{i,j}=t\mid A_{i,j}=a\right)=\frac{1}{k_d}$. Moreover, $\Pr\left(E_{A_{i,j}}^{t_{i,j}^+}\mid A_{i,j}=a,t_{i,j}=t\right)$ can be simply written as,

$$\Pr\left(E_{A_{i,j}}^{t_{i,j}^+}\mid A_{i,j}=a,t_{i,j}=t\right)=\left(\frac{k_d-t-1}{k_d}\right)^a. \tag{A.5}$$

Finally, the probability that U_i's packet transmission time $\mathcal{T}$ lasts for T_f seconds, $\mathcal{P}_{T_f}\triangleq\Pr\left(\mathcal{T}=T_f\right)$, is given as (A.6).

Since the selection of the RB and back-off timer value is performed independently at each frame, the probability that U_i's packet transmission time $\mathcal{T}$ lasts for $n\times T_f$ is geometrically distributed with parameter $\mathcal{P}_{T_f}$ given as,

$$\Pr\left(\mathcal{T}=nT_f\right)=\mathcal{P}_{T_f}\left(1-\mathcal{P}_{T_f}\right)^{n-1}. \tag{A.7}$$

The process of arriving-transmitting packets of a typical node can be modeled as M/G/1 queue with Poisson process arrival and geometrically distributed service time. However, to derive delay distribution, we can approximate geometrical distribution with exponential as in [4,48][12]. Therefore, the distribution of the packet transmission time is approximated by an exponential distribution with the mean $\left[\ln(\frac{1}{1-\mathcal{P}_{T_f}})\right]^{-1}T_f$. Specifically,

$$F_{\mathcal{T}}(\tau)=1-e^{-\ln(\frac{1}{1-\mathcal{P}_{T_f}})T_f^{-1}\tau} \tag{A.8}$$

Then, according to the analysis for the M/M/1 queue, the cumulative distribution function (CDF) of the packet delay can be derived as [49],

$$F_{\mathcal{D}}(t)=\Pr(\mathcal{D}<t)=1-e^{-\varpi t} \tag{A.9}$$

where $\varpi=\ln(\frac{1}{1-\mathcal{P}_{T_f}})T_f^{-1}-\lambda_p$. Therefore, the mean packet delay or average packet latency equals $\bar{\mathcal{D}}=\left[\ln(\frac{1}{1-\mathcal{P}_{T_f}})T_f^{-1}-\lambda_p\right]^{-1}$. Since short packet transmission are adopted, the transmission duration equals $T_t=\frac{n_b}{B}$. Hence, the frame duration is $T_f=T_p+T_b+\frac{n_b}{B}$.

■

APPENDIX B
PROOF OF LEMMA 2

. By definition, reliability is the probability that a typical packet is delivered successfully to the BS. There are only one sources of impairment or error for data packets which comes from the channel distortion. Defining E_i^{ne} as the event of occurring no error at decoding the U_i's signal, $\mathcal{R}_i^{NM}$ can be written as,

$$\mathcal{R}_i^{NM}=\Pr\left(E_i^{ne}\right). \tag{B.1}$$

To derive $\Pr\left(E_i^{ne}\right)$, consider that U_i is the winner of the contention for the j^{th} RB. Furthermore, assume that U_i be the Q^{th} active node in its NC in which Q is a random variable that $Q\in\{1,\ldots,i\}$. Then, $\Pr\left(E_i^{ne}\right)$ can be written as,

$$\Pr\left(E_i^{ne}\right)=\sum_{q=1}^{i}\Pr\left(E_i^{ne}\mid Q=q\right)\Pr(Q=q). \tag{B.2}$$

Let $\mathcal{A}_j$ ($\mathcal{A}_j\in\{0,\ldots,N_g-1\}$) be the random variable indicating the number of active nodes– except the node of interest U_i – in the j^{th} NC. Then, (B.2) can be further conditioned on $\mathcal{A}_j$ as,

[12]In fact, the geometrical distribution values are only the sampling of the exponential one

$$\Pr\left(E_i^{ne}\right)=\sum_{q=1}^{i}\sum_{\alpha=q}^{q+N_g-i}\Big[\Pr\left(E_i^{ne}\mid\mathcal{A}_j=\alpha,Q=q\right)\times\Pr(\mathcal{A}_j=\alpha\mid Q=q)\Pr(Q=q)\Big]. \tag{B.3}$$

The conditional probability of being $\mathcal{A}_j=\alpha$ active nodes in the j^{th} NC, $\Pr(\mathcal{A}_j=\alpha\mid Q=q)$, equals the probability that exactly α GCs from the remaining N_g-1 have the winner node on j^{th} RB. Since U_i is the q^{th} active node, it is sufficient that exactly $(\alpha-q)$ GCs from (N_g-i) GCs (GCs that farther than i^{th} GC) have the winner nodes on j^{th} RB. Therefore, $\Pr(\mathcal{A}_j=\alpha\mid Q=q)$ in (B.3) can be written as,

$$\Pr(\mathcal{A}_j=\alpha\mid Q=q)=\binom{N_g-i}{\alpha-q}\Big(\Pr(\mathbb{I}_i^j=1)\Big)^{\alpha-q}\times\Big(1-\Pr(\mathbb{I}_i^j=1)\Big)^{N_g+q-i-\alpha}, \tag{B.4}$$

in which $\Pr(\mathbb{I}_i^j=1)$ is the probability that a typical GC–say the i^{th} GC– be active on the j^{th} RB. By defining M^{rb} as the random variable indicating the number of nodes in a typical GC that selects the j^{th} RB from all R_b RBs and considering that M_g active nodes are assumed at each GC, $\Pr\left(\mathbb{I}_i^j=1\right)$ can be expressed as,

$$\mathcal{P}_{\mathbb{I}}\triangleq\Pr(\mathbb{I}_i^j=1)=\sum_{m=0}^{M_g}\Pr\left(\mathbb{I}_i^j=1\mid M^{rb}=m\right)\Pr\left(M^{rb}=m\right) \tag{B.5}$$

Note that $\Pr\left(M^{rb}=m\right)$ can be readily written as,

$$\Pr\left(M^{rb}=m\right)=\binom{M_g}{m}\left(\frac{1}{R_b}\right)^m\left(1-\frac{1}{R_b}\right)^{M_g-m}. \tag{B.6}$$

To derive $\Pr\left(\mathbb{I}_i^j=1\mid M^{rb}=m\right)$ in (B.5), define $E_m^{t^+}$ as the event that only one node from all m nodes in a typical GC selects t as the back-off timer value and the remaining nodes select values bigger than t ($t\in\{0,\ldots,k_d-1\}$). Then, $\Pr\left(\mathbb{I}_i^j=1\mid M^{rb}=m\right)$ can be written as,

$$\begin{aligned}\Pr\left(\mathbb{I}_i^j=1\mid M^{rb}=m\right)&=\Pr\left(\bigcup_{all\,t}E_m^{t^+}\right)=\sum_{t=0}^{k_d-1}\Pr\left(E_m^{t^+}\right)\\&=\sum_{t=0}^{k_d-1}\binom{m}{1}\left(\frac{1}{k_d}\right)\left(\frac{k_d-t-1}{k_d}\right)^{m-1}\\&=\sum_{t=0}^{k_d-1}\frac{m}{k_d}\left(\frac{k_d-t-1}{k_d}\right)^{m-1}.\end{aligned} \tag{B.7}$$

Therefore, $\mathcal{P}_{\mathbb{I}}$ in (B.5) can be obtained as (8).

Furthermore, $\Pr(Q=q)$ in (A.3) equals the probability that $(q-1)$ nodes in the previous $(i-1)$ geographical clusters are the winners of contention for transmitting data over the j^{th} RB. Therefore, based on $\mathcal{P}_{\mathbb{I}}$ derived in (8), $\Pr(Q=q)$ can be given as,

$$\Pr(Q=q)=\binom{i-1}{q-1}\left(\mathcal{P}_{\mathbb{I}}\right)^{q-1}\left(1-\mathcal{P}_{\mathbb{I}}\right)^{i-q},\quad \text{for}\quad q\in\{1,\ldots,i\} \tag{B.8}$$

To derive the decoding error probability, remember that U_i's signal experiences no decoding error at the BS if its signal and the signals of all previous active nodes in its NOMA cluster are decoded without error. Hence, $\Pr\left(E_i^{ne}\mid\mathcal{A}_j=\alpha,Q=q\right)$ can be obtained as,

$$\Pr\left(E_i^{ne}\mid\mathcal{A}_j=\alpha,Q=q\right)=\prod_{l=1}^{q}\left(1-\epsilon_{i|l,\alpha}\right). \tag{B.9}$$

where $\epsilon_{i|l,\alpha}$ is the conditional block error rate of decoding the signal of node U_i at the BS– provided that U_i is the l^{th} node from the total α active nodes in the j^{th} NC specified as,

$$\epsilon_{i|l,\alpha}=Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{i|l,\alpha}\right)}}\left(\mathcal{C}\left(\gamma_{i|l,\alpha}\right)-\frac{n_d}{n_b}\right)\right), \tag{B.10}$$

$\gamma_{i|l,\alpha}$ is the SINR for U_i's signal at the BS which is given as,

$$\gamma_{i|l,\alpha}=\frac{P_i^l}{\sum_{h=l+1}^{\alpha}P_i^l+\sigma^2}. \tag{B.11}$$

Finally, considering (B.9), (B.8), (B.7), (B.4) and (B.3), reliability in (B.1) can be written as (7).

■

APPENDIX C
PROOF OF LEMMA 3

To derive ESR, the sum of effective rate for all nodes at all NCs has to be considered. Because the number of NCs equals the number of RBs, ESR can be expressed as

$$ESR^{NM}=\mathbb{E}\left[\sum_{j=1}^{R_b}r_j\right]=R_b\,\mathbb{E}\left[r_j\right], \tag{C.1}$$

where $\mathbb{E}[.]$ represents the expected value and r_j is the effective sum rate on the j^{th} NC. Note that (C.1) has been yielded because all the RBs (or equivalently NCs) have similar expected rates since the nodes select them randomly without any priority. Define $\mathcal{I}_k$ as a set of all combinations of k active GCs on the j^{th} RB which can be mathematically written in terms of GC activation index as $\mathcal{I}_k=\{\mathbf{I}_k=(\mathbb{I}_1^j,\cdots,\mathbb{I}_{N_g}^j)\mid\sum_{i=1}^{N_g}\mathbb{I}_i^j=k\}$ with $\parallel\mathcal{I}_k\parallel=\binom{N_g}{k}$. Therefore, conditioning on all combinations in $\mathcal{I}_k$ for all possible k, (C.1) can be rewritten as,

$$ESR^{NM}=R_b\sum_{k=1}^{N_g}\sum_{all\,\mathbf{I}_k\in\mathcal{I}_k}\mathbb{E}\left[r_j\mid\mathbf{I}_k\right]\Pr(\mathbf{I}_k), \tag{C.2}$$

Based on $\mathcal{P}_{\mathbb{I}}$ derived in (8), $\Pr(\mathbf{I}_k)$ can be derived as,

$$\begin{aligned}\Pr\left(\mathbf{I}_k\right)&=\Pr\left(\sum_{i=1}^{N_g}\mathbb{I}_i^j=k\right)\\&=\left(\mathcal{P}_{\mathbb{I}}\right)^k\left(1-\mathcal{P}_{\mathbb{I}}\right)^{N_g-k}.\end{aligned} \tag{C.3}$$

Since a typical active node in a GC– say U_i– transmits n_d data bits within frame duration with the successful decoding

probability of $\prod_{l=1}^{i}\left(1-\epsilon_{l|\mathbf{I}_k}\right)$, its conditional effective rate equals $\frac{n_d}{T_f}\prod_{l=1}^{i}\left(1-\epsilon_{l|\mathbf{I}_k}\right)$. Then, $\mathbb{E}\left[r_j \mid \mathbf{I}_k\right]$ in (C.2) can be written as,

$$\mathbb{E}\left[r_j \mid \mathbf{I}_k\right] = \sum_{i=1}^{N_g} \mathbb{I}_i^j \frac{n_d}{T_f} \prod_{l=1}^{i}\left(1-\epsilon_{l|\mathbf{I}_k}\right) \tag{C.4}$$

where

$$\epsilon_{l|\mathbf{I}_k} = Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{l|\mathbf{I}_k}\right)}}\left(\mathcal{C}\left(\gamma_{l|\mathbf{I}_k}\right) - \frac{n_d}{n_b}\right)\right), \tag{C.5}$$

and

$$\gamma_{l|\mathbf{I}_k} = \frac{P_l^{G(l)}}{\sum_{h=l+1}^{N_g} P_h^{G(h)} + \sigma^2}\,. \tag{C.6}$$

where $G(\beta) = \sum_{q=1}^{\beta} \mathbb{I}_\beta^j$. Finally, the ESR is obtained as (11) by substituting (C.3) and (C.4) into (C.2)

■

APPENDIX D
PROOF OF LEMMA 4

To draw $\mathcal{R}_i^{NAL}$, it should be noted that a typical packet is received and decoded successfully if at least one replica from the total L transmitted packet replicas is received and decoded correctly. Defining E_i^e as the event that each packet replica cannot be decoded successfully at the BS, $\mathcal{R}_i^{NAL}$ can be expressed as,

$$\mathcal{R}_i^{NAL} = 1 - \Pr(E_i^e)^L = 1 - \left(1 - \Pr(E_i^{succ})\right)^L, \tag{D.1}$$

in which E_i^{succ} is the event that each packet replica is successfully decoded at the BS. Note that for each replica to be decoded correctly, it is necessary to have no collision between nodes with the same power levels at the j^{th} RB and no block decoding error at the BS. For a typical replica to experience no inter-node collision on a specific RB and a specific power level, it is necessary that all the nodes selecting j^{th} RB select different power levels (no two or more nodes with the same power level exist in the j^{th} RB.). Given the number of nodes selecting the j^{th} RB– except U_i, n_j^{rb}, equals to k, E_i^{succ} can be written as,

$$\Pr\left(E_i^{succ}\right) = \sum_{k=0}^{N_p-1} \Pr\left(E_i^{succ} \mid n_j^{rb}=k\right)\Pr\left(n_j^{rb}=k\right), \tag{D.2}$$

where $\Pr\left(n_j^{rb}=k\right) = \binom{M_g-1}{k}\left(\frac{1}{R_b}\right)^k\left(1-\frac{1}{R_b}\right)^{M_g-k-1}$. To derive $\Pr\left(E_i^{succ} \mid n_j^{rb}=k\right)$, we define E_i^{dec} and E_i^{coll} as the event that U_i's packet is decoded without any error at the BS and the event that U_i experiences no collision in its NC, respectively. Accordingly, we can write,

$$\begin{aligned}\Pr\left(E_i^{succ} \mid n_j^{rb}=k\right) = \Pr\left(E_i^{coll} \mid n_j^{rb}=k\right) \\ \times\ \Pr\left(E_i^{dec} \mid n_j^{rb}=k\right),\end{aligned} \tag{D.3}$$

in which $\Pr\left(E_i^{coll} \mid n_j^{rb}=k\right) = \mathbb{P}\left(N_p-1,k\right)/N_p^k = \frac{(N_p-1)!}{(N_p-k-1)!N_p^k}$ where $\mathbb{P}(.)$ is the permutation function. Furthermore, $\Pr\left(E_i^{dec} \mid n_j^{rb}=k\right)$ can be written as,

$$\Pr\left(E_i^{dec} \mid n_j^{rb}=k\right) = \prod_{l=1}^{i}\left(1-\epsilon_{i|l,k+1}\right). \tag{D.4}$$

where the conditional error probability $\epsilon_{l_1|l_2,l_3}$ is defined in (9). Finally, $\mathcal{R}_i^{NAL}$ can be derived as (14) and (15).

■

APPENDIX E
PROOF OF LEMMA 5

The proof of ESR^{NAL} is similar to ESR^{NM}. The main difference is that we have N_p power levels instead of N_g GCs. Therefore, ESR can be expressed as,

$$ESR^{NAL} = \mathbb{E}\left[\sum_{j=1}^{R_b} r_j\right] = R_b\,\mathbb{E}\left[r_j\right], \tag{E.1}$$

Now, define $\mathcal{J}_k$ as a set of all combinations of k active power levels on the j^{th} RB which can be mathematically written in terms of power level activation index as $\mathcal{J}_k = \{\mathbf{J}_k = (\mathbb{J}_1^j, \cdots, \mathbb{J}_{N_p}^j) \mid \sum_{i=1}^{N_p}\mathbb{J}_i^j = k\}$ with $\parallel \mathcal{J}_k \parallel = \binom{N_p}{k}$. Therefore, conditioning on all combinations in $\mathcal{J}_k$ for all possible k, (E.1) can be rewritten as,

$$ESR^{NAL} = R_b \sum_{k=1}^{N_p} \sum_{all\ \mathbf{J}_k \in \mathcal{J}_k} \mathbb{E}\left[r_j \mid \mathbf{J}_k\right]\Pr(\mathbf{J}_k), \tag{E.2}$$

$\Pr(\mathbf{J}_k)$ can be derived as,

$$\begin{aligned}\Pr\left(\mathbf{J}_k\right) &= \binom{M_g}{k}\left(\frac{1}{R_b}\right)^k\left(1-\frac{1}{R_b}\right)^{M_g-k}\frac{k!}{(N_p)^k} \\ &= \frac{M_g!}{(M_g-k)!}\left(\frac{1}{R_b N_p}\right)^k\left(1-\frac{1}{R_b}\right)^{M_g-k}.\end{aligned} \tag{E.3}$$

Since a typical active node with a specific power level– say U_i– transmits n_d data bits within frame duration with the successful decoding probability of $\prod_{l=1}^{i}\left(1-\epsilon_{l|\mathbf{J}_k}\right)$, its conditional effective rate equals $\frac{n_d}{LT_f}\prod_{l=1}^{i}\left(1-\epsilon_{l|\mathbf{J}_k}\right)$. Then, $\mathbb{E}\left[r_j \mid \mathbf{J}_k\right]$ in (E.2) can be written as,

$$\mathbb{E}\left[r_j \mid \mathbf{J}_k\right] = \sum_{i=1}^{N_p} \mathbb{J}_i^j \frac{n_d}{LT_f} \prod_{l=1}^{i}\left(1-\epsilon_{l|\mathbf{J}_k}\right), \tag{E.4}$$

where

$$\epsilon_{l|\mathbf{J}_k} = Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{l|\mathbf{J}_k}\right)}}\left(\mathcal{C}\left(\gamma_{l|\mathbf{J}_k}\right) - \frac{n_d}{n_b}\right)\right), \tag{E.5}$$

and

$$\gamma_{l|\mathbf{J}_k} = \frac{P_l^{F(l)}}{\sum_{h=l+1}^{N_p} P_h^{F(h)} + \sigma^2}\,. \tag{E.6}$$

where $F(\beta) = \sum_{q=1}^{\beta} \mathbb{J}_\beta^j$. Finally, the ESR is obtained as (16) by substituting (E.3) and (E.4) into (E.2)

■

APPENDIX F
PROOF OF LEMMA 6

To investigate the effect of active node detection error on reliability, the miss-detection probability should be firstly specified. Since this is a multi hypothesis test and there exist M_g hypotheses, different miss-detection probabilities are defined. Let N_i^a be the random variable indicating the number of detected active nodes by the BS in the i^{th} GC on the j^{th} RB. Then, we are interested on specifying $\Pr(N_i^a = 1|H_{m'})$ for $m' \geq 2$. The BS measure the test statistic Ξ and then compare it to some threshold values $\delta_{m'}$ to determine the

number of active nodes. if $\delta_{m'} < \Xi < \delta_{m'+1}$, then the BS concludes that there are m' active nodes in the i^{th} GC on the j^{th} RB. The values of $\delta_{m'}$ should be set for optimum detector performance. Hence, considering the conditional distribution of the test statistic, $\Pr(N_i^a = 1|H_{m'})$ can be written as,

$$\Pr(N_i^a = 1|H_{m'}) = Q\left(\frac{\delta_{m'} - \sigma^2(1+\xi_{m'})S}{\sqrt{2S(1+\xi_{m'})\sigma^2}}\right) - Q\left(\frac{\delta_{m'+1} - \sigma^2(1+\xi_{m'})S}{\sqrt{2S(1+\xi_{m'})\sigma^2}}\right), \tag{F.1}$$

where $Q(.)$ is the standard Gaussian Q function.

To derive the U_i's reliability, the same procedure as provided in Lemma 2 is followed.

$$\tilde{\mathcal{R}}_i^{NM} = \sum_{q=1}^{i}\sum_{\alpha=q}^{q+N_g-i}\Big[\Pr\left(E_i^{ne} \mid \mathcal{A}_j = \alpha, Q = q\right) \times \Pr(\mathcal{A}_j = \alpha \mid Q = q)\Pr(Q = q)\Big], \tag{F.2}$$

where

$$\Pr(Q = q) = \binom{i-1}{q-1}\left(\tilde{\mathcal{P}}_{\mathbb{I}}\right)^{q-1}\left(1-\tilde{\mathcal{P}}_{\mathbb{I}}\right)^{i-q}, \quad \text{for} \quad q \in \{1,\ldots,i\} \tag{F.3}$$

and $\Pr(\mathcal{A}_j = \alpha \mid Q = q)$ is as,

$$\Pr(\mathcal{A}_j = \alpha \mid Q = q) = \binom{N_g - i}{\alpha - q}\left(\tilde{\mathcal{P}}_{\mathbb{I}}\right)^{\alpha-q} \times \left(1-\tilde{\mathcal{P}}_{\mathbb{I}}\right)^{N_g+q-i-\alpha}, \tag{F.4}$$

where $\tilde{\mathcal{P}}_{\mathbb{I}}$ is the probability that a typical GC–say the i^{th} GC– be active on the j^{th} RB but with only one transmitting node. Conditioning on M^{rb} as in (B.5), we have

$$\begin{aligned}\tilde{\mathcal{P}}_{\mathbb{I}} &= \sum_{m=1}^{M_g}\Pr(N_i^a = 1|M^{rb} = m)\Pr\left(M^{rb} = m\right)\\ &= \sum_{m=0}^{M_g}\Pr(N_i^a = 1|M^{rb} = m, E_m^{t^+}) \times \Pr\left(E_m^{t^+} \mid M^{rb} = m\right)\Pr\left(M^{rb} = m\right)\end{aligned} \tag{F.5}$$

where $\Pr\left(M^{rb} = m\right)$ is derived in (B.6) and $E_m^{t^+}$ is the event that only one node from all m nodes in a typical GC selects t as the back-off timer value and the remaining nodes select values bigger than t ($t \in \{0,\ldots,k_d-1\}$). $\Pr\left(E_m^{t^+} \mid M^{rb} = m\right)$ can be obtained as,

$$\begin{aligned}\Pr\left(E_m^{t^+} \mid M^{rb} = m\right) &= \sum_{t=0}^{k_d-1}\binom{m}{1}\left(\frac{1}{k_d}\right)\left(\frac{k_d-t-1}{k_d}\right)^{m-1}\\ &= \sum_{t=0}^{k_d-1}\frac{m}{k_d}\left(\frac{k_d-t-1}{k_d}\right)^{m-1}.\end{aligned} \tag{F.6}$$

According to (F.1) $\Pr(N_i^a = 1|M^{rb} = m, E_m^{t^+})$ equals

$$\Pr(N_i^a = 1|M^{rb} = m, E_m^{t^+}) = \Pr(N_i^a = 1|H_1) = Q\left(\frac{\delta_1 - \sigma^2(1+\xi_1)S}{\sqrt{2S(1+\xi_1)\sigma^2}}\right) - Q\left(\frac{\delta_2 - \sigma^2(1+\xi_1)S}{\sqrt{2S(1+\xi_1)\sigma^2}}\right) \tag{F.7}$$

Now, we turn to derive $\Pr\left(E_i^{ne} \mid \mathcal{A}_j = \alpha, Q = q\right)$. Note that U_i's signal experiences no decoding error at the BS if its signal and the signals of all previous active nodes in its NOMA cluster are decoded without error. Note that the condition "$Q = q$" ensures that only one active node is correctly detected. Therefore, $\Pr\left(E_i^{ne} \mid \mathcal{A}_j = \alpha, Q = q\right)$ can be written as,

$$\Pr\left(E_i^{ne} \mid \mathcal{A}_j = \alpha, Q = q\right) = \prod_{l=1}^{q}\left(1 - \epsilon_{i|l,\alpha}\right). \tag{F.8}$$

substituting (F.8), (F.7), (F.6), and (F.5) into (F.2) gives reliability as (17).

REFERENCES

[1] ETSI TR 138 913 V14.2.0, "5g; Study on Scenarios and Requirements for Next Generation Access Technologies3rd Generation Partnership Project;," 2017. [Online]. Available: www.etsi.org

[2] D. Feng, C. She, K. Ying, L. Lai, Z. Hou, T. Q. S. Quek, Y. Li, and B. Vucetic, "Toward ultrareliable low-latency communications: Typical scenarios, possible solutions, and open issues," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 94–102, 2019.

[3] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mmtc: A review of urllc within mmtc," *IEEE Access*, vol. 8, pp. 131 796–131 813, 2020.

[4] 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.

[5] M. R. Amini and M. W. Baidas, "Random-access NOMA in URLL energy-harvesting IoT networks with short packet and diversity transmissions," *IEEE Access*, vol. 8, pp. 220 734–220 754, Dec. 2020.

[6] 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.

[7] 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.

[8] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 148–156, 2020.

[9] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5G networks for the Internet of Things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2017.

[10] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 16–32, 2020.

[11] 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.

[12] 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.

[13] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, "How urllc can benefit from noma-based retransmissions," 2020.

[14] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124–130, 2018.

[15] Y. Liu, Z. Qin, and Z. Ding, *Non-Orthogonal Multiple Access for Massive Connectivity*. Springer International Publishing, 2020.

[16] 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.

[17] M. Shirvanimoghaddam, M. S. Mohammadi, R. Abbas, A. Minja, C. Yue, B. Matuz, G. Han, Z. Lin, W. Liu, Y. Li, S. Johnson, and B. Vucetic, "Short block-length codes for ultra-reliable low latency communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 130–137, 2019.

[18] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2488–2524, 2019.

[19] J. Choi, "Re-transmission diversity multiple access based on sic and harq-ir," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4695–4705, 2016.

[20] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with short packets," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[21] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2018.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] V. Gupta, S. K. Devar, N. H. Kumar, and K. P. Bagadi, "Modelling of IoT traffic and its impact on LoRaWAN," in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, Dec. 2017, pp. 1–6.

[41] T. Hoßfeld, F. Metzger, and P. E. Heegaard, "Traffic modeling for aggregated periodic IoT data," in *Proc. of 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, Feb. 2018, pp. 1–8.

[42] J. Choi, "Noma-based random access with multichannel aloha," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2736–2743, 2017.

[43] J.-B. Seo, H. Jin, and B. C. Jung, "Non-orthogonal random access with channel inversion for 5g networks," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, 2017, pp. 117–119.

[44] U. N. Bhat, *An Introduction to Queueing Theory, Modeling and Analysis in Applications.* Birkhäuser Basel, 2015.

[45] ETSI TR 103 515 V1.1.1, "Digital enhanced cordless telecommunications (dect); study on urllc use cases of vertical industries for dect evolution and dect-2020," 2020. [Online]. Available: www.etsi.org

[46] S. Ciftci and M. Torlak, "A comparison of energy detectability models for spectrum sensing," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, 2008, pp. 1–5.

[47] J. P. R. E. L. Lehmann, *Testing Statistical Hypotheses*. Springer New York, NY, 2005.

[48] L. Alonso, R. Agusti, and O. Sallent, "A near-optimum mac protocol based on the distributed queueing random access protocol (dqrap) for a cdma mobile communication system," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 9, pp. 1701–1718, 2000.

[49] S. M. Ross, *Introduction to Probability Models, 10th Ed.* Elsevier Academic Press, 2010.