IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 4, 2021, accepted July 12, 2021, date of publication July 15, 2021, date of current version August 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3097553

Performance Analysis of Grant-Free Random-Access NOMA in URLL IoT Networks

MOHAMMAD REZA AMINI[1], (Senior Member, IEEE), AND MOHAMMED W. BAIDAS[2], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Borujerd Branch, Islamic Azad University, Borujerd 73916-9867, Iran
[2]Department of Electrical Engineering, College of Engineering and Petroleum, Kuwait University, Kuwait City 13060, Kuwait

Corresponding author: Mohammed W. Baidas (m.baidas@ku.edu.kw)

This work was supported in part by the Kuwait Foundation for the Advancement of Sciences (KFAS) under Project PN17-15EE-02.

ABSTRACT Internet-of-Things (IoT) networks have recently emerged to provide massive connectivity for many application scenarios and services. Additionally, developing spectrum-access strategies for a large number of nodes with sporadic data traffic behaviors in IoT networks has attracted much attention recently. However, developing such strategies becomes more challenging when ultra-reliable low-latency (URLL) transmissions are required. As IoT networks entail spectrum-efficient transmission schemes, non-orthogonal multiple-access (NOMA) has emerged as a key enabler for such networks. On the other hand, grant-free random-access (RA) techniques are particularly promising for high spectral-efficiency and massive connectivity, since they reduce signaling overhead, and packet latency. Therefore, in this paper, uplink RA-NOMA IoT networks with clustered IoT devices is studied, where short packet and diversity transmissions are adopted to meet the URLL requirements. To reduce the negative effect of diversity transmission on packet latency, multiple replicas of packets are accommodated within different resource blocks (RBs) in the same transmission time interval (TTI). The analytical expressions of network metrics, namely, average packet latency, reliability, and GoodPut are derived. Furthermore, the effect of the number of packet replicas, blocklength, and cluster size on the network metrics is evaluated. Finally, the analytical derivations are utilized to find the optimal values for the number of packet replicas, blocklength, and power control parameters, such that the network GoodPut is maximized, subject to URLL constraints.

INDEX TERMS Internet-of-Things, low-latency, NOMA, random-access, ultra-reliability.

I. INTRODUCTION

Fifth generation (5G) cellular networks have ignited numerous research areas since its introduction. The fundamental difference between 5G and the previous generations is that 5G is the driver for implementing two generic types of communications, namely, ultra-reliable low-latency communication (URLLC) and massive machine-type communication (mMTC) [1]. The combination of these two communication types along with the explosive and exponential growth in the number of smart devices, applications, and services is the advent of massive Internet-of-Things (mIoT) networks. This is in addition to the extensive emerging mission-critical applications and use cases, such as tactile Internet (involving remote motion control, tele-surgery, etc.), factory automation, Industrial IoT (IIoT), and those under the Industry 4.0 paradigm [2]–[5]. However, reliability, latency, and massive connectivity are inherently conflicting features in such networks, resulting in striking trade-offs between network parameters and performance metrics [3], [6], [7], which calls for intelligent transmission techniques and strategies. To improve spectrum-efficiency, non-orthogonal multiple-access (NOMA) has emerged as a viable solution [8], [9]. To achieve low-latency communications, short packet transmissions in the finite blocklength regime (FBL) have been proposed [10], [11]. On the other hand, ultra-reliability is a feature that can be achieved via diversity transmission techniques [12]–[14]. In mMTCs, characterized by sporadic data traffic behaviors, it is inefficient to allocate dedicated time-frequency resources, since the resources are not constantly utilized. This motivates the use of random-access (RA) techniques. However, the conventional RA techniques (e.g. the current RA-LTE standard) are not directly applicable to low-latency transmissions, which is

The associate editor coordinating the review of this manuscript and approving it for publication was Noor Zaman.

due to the excessive grant acquisition delay and signaling overhead [15]. Hence, applying and analyzing RA-NOMA for IoT networks with URLL requirements is of paramount importance.

A. RELATED WORKS

In general, NOMA transmission can be categorized as grant-based (GB) and grant-free (GF) access schemes. Unlike the conventional GB access schemes, GF access has been put forth as a key medium access control technique to achieve massive connectivity, reduce latency, and minimize signaling overhead, which is crucial for URLL IoT networks [16], [17]. When considering uplink (UL) GF NOMA (via RA techniques), it is noticed that power control—which has significant effect on the NOMA performance—is particularly challenging, since it is performed distributively, without each node knowing the other transmitting nodes or their locations. Signature-based NOMA is a viable option to provide ubiquitous connectivity for the future networks with massive number of nodes [16], [18]. Specifically, using sequences—such as spreading, scrambling and/or interleaving sequences—as device-specific signatures is one way to enable non-orthogonal transmissions. Not only that, but employing signature sequences along with power control in RA-NOMA to increase the capacity (or overload factor) has recently been addressed, as it helps reduce implementation complexity at the base-station (BS), since signature sequences are usually selected from a finite set. For instance, in [19], a new type of NOMA, called multi-user shared access (MUSA) has been proposed, where a set of complex sequences with short length are chosen as spreading sequences. This enables the BS to perform multi-user detection based on successive interference cancellation (SIC), while coping with high transmission load. Particularly, it has been demonstrated that MUSA can yield significant user overloading gain performance in comparison to orthogonal systems, while reducing control overhead. Non-Orthogonal coded access (NOCA) for contention-based (CB) transmission with parallel interference cancellation (PIC) at the BS is investigated in [20], where Zadoff-Chu (ZC) sequences have been used to spread the data bits over orthogonal frequency-division multiplexing (OFDM) symbols. It has been shown that the NOCA-PIC scheme is capable of providing four times concurrent RA sessions compared to OFDM with CB transmission. In [21], the authors propose a random NOMA strategy, where each device attaches its unique ID to its message, and encodes it using a Raptor code, which is followed by transmission of its final message codeword over a selected sub-band. Moreover, the BS broadcasts pilot signals over all sub-bands, and each device estimates its channel to the BS over the randomly selected sub-band. By performing load estimation and SIC over each sub-band, the BS can recover the messages of all active devices. A framework for massive GF NOMA, where devices have stringent latency requirements but no packet re-transmissions, has been proposed in [22]. To be specific, each device chooses a pilot sequence—as its signature—from a pre-defined set. Conventionally, a collision is assumed to occur when at least two devices choose the same pilot sequence for data transmission. However, the authors propose to treat a collision as interference to the remaining signals. Then, by exploiting Poisson point processes and ordered statistics, the authors approximated the outage probability, and system throughput, and validated them via extensive simulations.

There are a few studies that consider reliability and latency in RA-NOMA. For example, in [23], hybrid automatic repeat request (HARQ) with only one re-transmission is exploited to improve the reliability and also reduce latency. By using Markov models, the packet error rate (PER) and throughput of each user has been derived and investigated. A non-orthogonal HARQ (N-HARQ) scheme to support UR communication in the FBL regime for delay-sensitive applications is proposed in [24]. Unlike conventional HARQ, in the proposed scheme, retransmission of a packet is served together with the next arriving packet, which reduces queueing delay. The PER and throughput have been analyzed, and the superiority of the N-HARQ scheme has been demonstrated over the baseline orthogonal HARQ. Dynamic HARQ (D-HARQ) for guaranteed-delay applications is proposed in [25], in which packets can be re-transmitted more than L times, given that the previous packet was delivered with less than L re-transmissions. The authors showed that under the same reliability constraint, D-HARQ achieves a higher throughput than the conventional HARQ with a fixed re-transmission.

B. MOTIVATION AND CONTRIBUTIONS

In realistic cellular networks with UL RA transmissions, the BS needs to dynamically detect the active user equipments (UEs), and decode their signals. Another problem in NOMA-based transmissions is that the BS has to dynamically derive the SIC order, and this may be challenging as the network UEs may be randomly scattered in the network area, and characterized with random channel conditions. To reduce the decoding complexity at the BS, resolve the problem of active UE detection, and facilitate dynamic SIC decoding, this paper utilizes orthogonal preamble transmission in power-domain NOMA. Particularly, this paper focuses on performance analysis of distributed UL RA-NOMA IoT networks with URLL requirements. Specifically, a UL NOMA network with clustered IoT UEs is considered, along with the time-reversal (TR) strategy [26], where the IoT network UEs transmit orthogonal preambles as their signatures. Adopting the TR strategy along with preamble transmissions enables the IoT UEs to distributively adjust their transmit powers, while allowing the BS to perform active UE detection and dynamic SIC in each transmission frame. To satisfy the stringent URLL requirements, short packet and diversity transmissions are adopted. Moreover, multiple replicas of each packet are transmitted within each transmission time interval (TTI), yet over different resource

blocks (RBs) to reduce packet latency. From a mathematical point of view, the moment generating function (MGF) is used to derive the joint probability density function (PDF) of the UEs' transmit power and intra-cluster interference power, based on order statistics [27], [28]. This facilitates the analytical derivation of network performance metrics, such as average packet latency, reliability, and GoodPut. For the average packet latency, both waiting time in a UE's buffer and transmission time are considered. For the reliability, both inter-cluster interference and block decoding error (due to channel impairment and intra-cluster interference)—due to the adopted FBL regime—are considered. The network GoodPut is then straightforwardly determined based on the obtained reliability. Unlike many existing studies that are based on the saturated data traffic model, this paper considers sporadic data traffic behavior, and the IoT UEs are randomly located.[1] Additionally, the effect of the number of packet replicas, blocklength, and cluster size on the network metrics is investigated, and various reliability and packet latency tradeoffs are explored and highlighted. Thus, the main contributions of this paper can be summarized as follows:

- Analyzed the performance of UL RA-NOMA IoT networks with URLL requirements. Particularly, an UL IoT network with clustered UEs exploiting RA-NOMA transmissions is explored. Each UE within each cluster is characterized with sporadic data traffic behavior, and distributively controls its transmit power, while utilizing the FBL regime and diversity transmission to satisfy the URLL requirements.
- Exploited the TR strategy along with orthogonal preamble transmission in order for the UEs to distributively achieve their target received power at BS, while allowing the BS to detect active UEs, and dynamically perform SIC decoding in each transmission frame.
- Considered the UEs' data traffic and their locations as random processes to model a realistic scenario of mobile UEs with stochastic data arrival patterns.
- Derived analytical expressions for critical IoT network performance metrics (i.e. average packet latency, reliability, and GoodPut). The analytical derivations are numerically validated and the effect of number of replicas, blocklength, and cluster size, and their trade-offs on the network metrics are explored and highlighted.
- Formulated the network GoodPut maximization problem, subject to constraints on the URLL requirements. Specifically, the derived analytical expressions for the network metrics are utilized to maximize the network GoodPut by optimizing the number of packet replicas, blocklength, cluster size, and power control parameters.

It should be noted that this work is different from our previous NOMA-based URLLC works in [29], [30]. In particular, this study analyzes the IoT network metrics with arbitrary NOMA UE cluster sizes, while [29] analyzes the network metrics in NOMA-based IoT networks with only two energy-harvesting UEs in each NOMA cluster, which makes the analytical derivations fundamentally different. Furthermore, the network model in this work is different, since multiple replicas of each packet are accommodated within each TTI over different RBs to further reduce packet latency. Contrarily, the conventional diversity transmission through multiple successive frames has been adopted in [29], [30], which makes the analysis different, while yielding various trade-offs in the underlying network structure. Additionally, the UEs in this study exploits the TR strategy along with preamble transmissions, which facilitates distributed location-based power control, and helps the BS to detect active UEs, and dynamically perform SIC decoding in each frame. However, in [29], [30], the UEs are assumed to have fixed/pre-defined transmit power, irrespective of their locations.

The novelty of this work can be summarized as follows. Firstly, this study is in the area of performance evaluation, and steady-state analysis of RA UL NOMA-based transmissions, as opposed to other existing works that focus on devising algorithms and transmission schemes for resource allocation. Secondly, to the best of our knowledge, almost all the studies in the area of performance evaluation of NOMA-based transmissions are based on assumption that all the UEs in the network have saturated traffic, which implies that in each transmission frame, all the UEs are active and have data packets for transmission. Another common assumption is that the number of active UEs in the whole network within each cluster is fixed and known, and that the order of SIC decoding is known to BS. Note that for an IoT network with mobile users, the derived analytical expressions in the literature do not reflect the proper behavior of a practical network.[2] The case is even worse when analyzing critical mMTC and mIoT networks, which have been received much attention recently [31]. For instance, [32]–[34] analyzed the performance of RA-NOMA with a fixed number of UEs, having saturated data traffic, and adopting a channel inversion approach. Particularly, a network with a single NOMA cluster—with known SIC order at the BS—has been considered. To address the issue of dynamic SIC decode ordering at the BS for UL NOMA, [35] derived the outage probability for a network consisting of only a single 3-user NOMA cluster, proving how complicated it is to derive performance metrics for a NOMA cluster with more than two users when dynamic SIC ordering considered. The authors in [36] derived closed-form expressions for the rate and outage probability in the FBL regime with diversity transmission, yet for NOMA clusters with two users, and a saturated data traffic model. Also, [37] analyzed the steady-state packet loss and delay violation probability in

[1]This reflects a realistic network scenario, both in terms of network performance metrics as well as network scalability.

[2]In practical networks, UEs are characterized by sporadic data traffic patterns, as opposed to saturated data packet arrivals.

URLL NOMA-based communications with saturated traffic model, and predefined SIC ordering. Furthermore, [38] studied the outage probability and ergodic sum-rate in NOMA 5G systems with nonlinear high-power amplifiers and an arbitrary number of users in a NOMA cluster; however, with saturated data traffic and fixed SIC ordering. A joint user association and decoding order selection scheme for distributed UL NOMA has been studied in [39], where the outage probability has been derived for a 2-user NOMA cluster with saturated data traffic. On the other hand, the studied system model in this paper is based on randomly located UEs, and adopts short packet with diversity transmissions on different RBs over single TTI, which has not been considered previously; and thus, leads to different trade-offs to be investigated. Adopting the TR strategy along with preamble transmissions is another unique part in the studied transmission frame structure, which enables the IoT UEs to distributively adjust their transmit powers, while allowing the BS to perform active UE detection and dynamic SIC in each transmission frame. On top of this, utilizing the MGF to derive the joint PDF of the UEs' transmit powers and intra-cluster interference power based on order statistics for an arbitrary number of clustered UEs in the network is also considered novel. Specifically, the derivations in this work are different from those existing in the literature [27], [28] in that the joint PDF of an order statistic and partial sum of the remaining least order statistics are computed, which facilitates the mathematical analysis, and paves the way to investigate the effect of cluster size and the number of clusters on the network metrics.

The rest of this paper is organized as follows. Section II describes system model. Analytical derivations for the network metrics are presented in Section III, whereas Section IV provides the numerical results. Section V presents the network GoodPut maximization problem based on the derived network metrics. Future research directions of this work are outlined in Section VI. Finally, the conclusions are drawn in Section VII.

II. SYSTEM MODEL

A. IoT NETWORK MODEL

Consider an IoT network consisting of a BS and N transmitting UEs with URLL requirements. The UEs are paired into M clusters, each of which consists of $N_c = \lfloor \frac{N}{M} \rfloor$ UEs transmitting their data packets to the BS in the FBL regime via RA-NOMA. Additionally, there are R_b RBs in each TTI, and each RB has bandwidth B. Each packet is assumed to be transmitted within one RB. However, to transmit one packet and its $K-1$ replicas within each TTI, a typical UE randomly selects a resource unit (RU), which is assumed to be a group of K RBs. Hence, there are $R_u = \left\lfloor \frac{R_b}{K} \right\rfloor$ RUs, which are utilized by the UEs within each cluster for frame-based RA-NOMA transmissions [40]. Briefly, all the UEs in a cluster select an RU, and transmit their packets (and all their $K-1$ replicas) within the K RBs in the selected RU. In turn, the UEs within each cluster exploit diversity transmission in each TTI to improve the reliability, without incurring additional delay.

Remark 1: For each IoT UE, the data generating traffic behavior follows a Poisson arrival process with rate λ_p.[3]

Let $U_{i,m}$ denote IoT UE $i \in \{1, \ldots, N_c\}$ in the m^{th} cluster, for $m \in \{1, \ldots, M\}$. Furthermore, $U_{i,m}$ transmits its data packets over the selected RU with transmit power of $P_{i,m}$, such that $P_{i,m} \leq P_{\max}$, $\forall i \in \{1, \ldots, N_c\}$, $\forall m \in \{1, \ldots, M\}$, where $P_{\max}$ is the maximum transmit power per IoT UE.

There are N communication links, which originate from the N UEs to the BS, and experience independent but not necessarily identically distributed (i.n.n.i.d.) Rayleigh block fading. Moreover, all the links are assumed to be constant during each transmission frame, which is due to the short packet transmission. The channel coefficient between $U_{i,m}$ and the BS is denoted $h_{i,m}$. Therefore, the corresponding channel gain $|h_{i,m}|^2$ follows an exponential distribution with mean $d_{i,m}^{-\nu}$, where $d_{i,m}$ is the distance between IoT UE $U_{i,m}$ and the BS, whereas ν is the path-loss exponent. Furthermore, the background noise in all 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 is the noise spectral density. The distance of each UE at each cluster to the BS is assumed to be uniformly distributed in the interval $[d_{\min}, d_{\max}]$. Suppose that there are N_a active UEs in a typical cluster, where $N_a \leq N_c$. For notational convenience, let $d_{1,m} < d_{2,m} < \cdots < d_{N_a,m}$ be the ordered distances of the UEs in the m^{th} cluster. Thus, $|h_{1,m}|^2 > |h_{2,m}|^2 > \cdots > |h_{N_a,m}|^2$, and hence $P_{1,m} \geq P_{2,m} \geq \cdots \geq P_{N_a,m}$ [43], [44].

Table 1 presents the main symbols used in this paper as well as their descriptions.

B. FRAME STRUCTURE AND CHANNEL ACCESS

The IoT UEs within each cluster transmit their data packets in a frame-based structure over the same RB by exploiting power-domain NOMA [8], [45]. It should be noted that the different clusters access the RUs randomly.[4] Depicted in Fig. 1 is the frame structure of the IoT network. Particularly, each frame consists of three phases, namely, time-reversal (TR), preamble transmission, and data payload transmission of durations T_r, T_p and T_t, respectively. Hence, the whole frame duration is $T_f = T_r + T_p + T_t$.

In the first phase, the BS sends a reference waveform, and the resulting waveforms are received and recorded by the IoT UEs. In turn, the channel state information can be estimated, and then exploited in the preamble and data transmission phases. Specifically, by time-reversing (and conjugating if complex-valued) the received waveform, the IoT UEs can mitigate channel fading and path-loss, where

[3]This model has been extensively used in telecommunication networks due to its simplicity [41], [42].

[4]In practice, this can be achieved by feeding the random number generators with the same seed [46].

TABLE 1. Notations.

Parameter	Description
$U_{i,m}$	i^{th} UE in the m^{th} cluster
N_c	Number of UEs in a cluster (cluster size)
N	Total number of UEs in the network
M	Number of clusters
K	Number of replicas for each packet
R_b	Number of resource blocks
R_u	Number of resource units
$h_{i,m}$	Channel coefficient between $U_{i,m}$ and the BS
$d_{i,m}$	Distance between $U_{i,m}$ and the BS
N_0	Noise spectral density
B	Bandwidth of each RB
N_a	Number of active UEs in a cluster
$P_{i,m}$	Transmit power of $U_{i,m}$
T_f	Duration of transmission frame
T_r	Duration of time-reversal phase
T_p	Duration of preamble transmission phase
T_t	Duration of data transmission phase
$\gamma_{i,m}$	Received SINR of $U_{i,m}$'s signal at the BS
n_d	Number of data bits in a packet
n_b	Blocklength
λ_p	Packet arrival rate

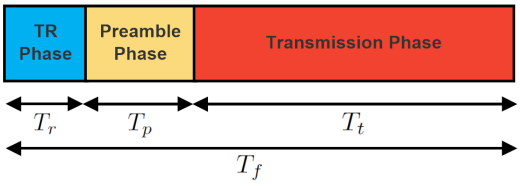


FIGURE 1. Frame structure of IoT network.

the communication channel is assumed to be reciprocal and stationary for at least one frame duration [26].[5]

In the second phase, all IoT UEs in a cluster transmit their orthogonal preamble sequences with the same prescribed power on all K RBs over the randomly selected RU r (for $r \in \{1, \ldots, R_u\}$) to the BS.[6] Transmitting preambles with the same power enables the BS to detect the stronger/weaker active UEs within each cluster, and hence determine the SIC decoding order. Particularly, this allows both active UE detection and dynamic SIC decoding.

In the third phase, each IoT UE transmits its data packet (and its replicas) over the selected RU in the FBL regime. However, in such a case, Shannon's capacity is no longer applicable, since the decoding block error is not negligible. Hence, for a given blocklength of $n_b > 100$ with n_d data bits per data packet, the instantaneous block error rate of decoding the signal of $U_{i,m}$ can be approximated as [51]

$$\Upsilon\left(\gamma_{i,m}, n_b, n_d\right) = Q\left(\sqrt{\frac{n_b}{\chi\left(\gamma_{i,m}\right)}}\left(C\left(\gamma_{i,m}\right) - \frac{n_d}{n_b}\right)\right), \quad (1)$$

[5]These assumptions have been validated in practice through real experiments [47], [48].

[6]Orthogonal preamble transmissions are used in RA-LTE, and defined in 3GPP RA [49], [50].

where $C\left(\gamma_{i,m}\right) = \log_2(1+\gamma_{i,m})$ is the Shannon capacity, while $\chi\left(\gamma_{i,m}\right) = \left(1 - \frac{1}{1+\gamma_{i,m}^2}\right)\left(\log_2 e\right)^2$ is the channel dispersion. Also, $\gamma_{i,m}$ is the received signal-to-interference-plus-noise (SINR) of $U_{i,m}$'s signal at the BS.

Remark 2: The inter-cluster interference occurs when UEs in at least two clusters simultaneously transmit their data packets over the same RU. In turn, all data packets transmitted by all the UEs over that RU collide, and thus are lost.

As explained earlier, to mitigate channel impairments, the TR strategy is adopted in the TR phase. By employing power control along with TR strategy, each UE is able to cancel the effect of channel fading and path-loss. To this aim, a distributed power control strategy is utilized, in which each UE adjusts its transmit power according to[7]

$$P_{i,m}\left(d_{i,m}\right) = d_{i,m}^{v} P_0 \ln\left(\beta \frac{d_{\min}}{d_{i,m}^{\eta}}\right), \quad (2)$$

where $\eta > 1$, P_0, and β can be selected to control a UE's transmit power, and thus, yield distinct received powers at the BS.[8] Particularly, η determines the slope at which the UE's received power at the BS decreases with respect to the UE's distance. This in turn provides more distinct received powers at the BS for the UEs with small distance difference, ultimately reducing the decoding error probability at BS. Additionally, P_0 is chosen such that $P_{i,m}(d_{i,m}) < P_{\max}$. Moreover, β is used to prevent $P_{i,m}$ from being negative (i.e. $\beta \frac{d_{\min}}{d_{i,m}^{\eta}} > 1$), and thus $\beta > \frac{d_{\max}^{\eta}}{d_{\min}}$ ensures that $P_{i,m} > 0$, $\forall d_{i,m} \in [d_{\min}, d_{\max}]$. Furthermore, $d_{i,m}^{v}$ in (2) is obtained via the TR strategy for path-loss cancellation. Consequently, $U_{i,m}$'s received power at BS is written as

$$\bar{P}_{i,m}(d_{i,m}) = P_0 \ln\left(\beta \frac{d_{\min}}{d_{i,m}^{\eta}}\right). \quad (3)$$

Since the IoT UEs are randomly located in the network area, the probability density function (PDF) of the received power at the BS can be obtained as per **Lemma 1**.

Lemma 1: The PDF of an IoT UE's received power at the BS, assuming uniformly located in $[d_{\min}, d_{\max}]$, is given by

$$f_{\bar{P}}(\bar{p}) = \alpha_1 e^{-\alpha_2 \bar{p}}, \quad (4)$$

where $\alpha_1 = \frac{\sqrt[\eta]{\beta d_{\min}}}{P_0 \eta (d_{\max} - d_{\min})}$ and $\alpha_2 = \frac{1}{P_0 \eta}$.

Proof: See Appendix A. ■

III. DERIVATION OF PERFORMANCE METRICS

In this section, the IoT network metrics, namely average packet latency, reliability and GoodPut are derived. However, a few definitions must first be stated [29].

[7]Other transmit power functions can be used. In this work, (2) is adopted for mathematical tractability and to gain some insights from the derived analytical expressions of the network metrics.

[8]The aforementioned parameters can be incorporated into a network optimization problem with URLL constraints for optimal transmit power, and as will be shown in Section V.

A. DEFINITIONS

Definition 1 (Average Packet Latency): The average packet latency $\mathcal{L}$ is defined as the mean delay of receiving a typical data packet (and all its replicas) at the BS, which includes both the data transmission delay, and the waiting time in a UE's buffer.

Definition 2 (Reliability): The reliability $\mathcal{R}$ is the probability that a data packet is delivered successfully to the BS, without any inter-cluster collision or channel distortion error.

Definition 3 (GoodPut): The GoodPut $\mathcal{G}$ is defined as the average error-free effective rate (i.e. non-redundant data bits per time unit) at the BS.

B. AVERAGE PACKET LATENCY

To derive the average packet latency, queueing theory is employed. However, it should first be noted that since short packet transmission is adopted, the data transmission duration equals $T_t = \frac{n_b}{B}$. Hence, by considering the number of packets in $U_{i,m}$'s buffer as a state variable, the arrival and departure of the data packets processes can be modeled as a $M/D/1$ queueing system. Specifically, the packet arrival rate is represented by λ_p, while the deterministic service time duration equals T_f. In turn, T_f can be written as

$$T_f = T_r + T_p + \frac{n_b}{B}. \tag{5}$$

Therefore, the average number of packets in the buffer of a typical IoT UE's is $\overline{\mathcal{D}} = \frac{\rho}{1-\rho}\left(1-\frac{\rho}{2}\right)$ [52], in which $\rho = \lambda_p(T_r + T_p + \frac{n_b}{B})$ is $U_{i,m}$'s link utilization (i.e. a measure of buffer length stability).[9] Thus, according to Little's formulae [52], the average packet latency is determined as

$$\mathcal{L} = \left(T_r + T_p + \frac{n_b}{B}\right)\left(\overline{\mathcal{D}} + 1\right). \tag{6}$$

Furthermore, the probability that a typical IoT UE does not transmit data in a frame (e.g. when it does not have any data packet in its buffer) is obtained as

$$\begin{aligned}\Pi_0 &\triangleq 1 - \rho \\ &= 1 - \lambda_p\left(T_r + T_p + \frac{n_b}{B}\right).\end{aligned} \tag{7}$$

C. RELIABILITY

To determine $U_{i,m}$'s reliability $\mathcal{R}_{i,m}$, one must obtain the PDF of the SINR $\gamma_{i,m}$. According to the principle of UL NOMA, $\gamma_{i,m}$ is given by[10]

$$\gamma_{i,m}\left(\bar{P}_{i,m}, \bar{Q}_{i+1,N_a}\right) = \frac{\bar{P}_{i,m}}{\bar{Q}_{i+1,N_a} + \sigma^2}, \tag{8}$$

where $\bar{Q}_{i+1,N_a} \triangleq \sum_{j=i+1}^{N_a} \bar{P}_{j,m}$ is the inter-user interference from UEs in the same cluster with decoding order less than that of $U_{i,m}$. Since both $\bar{P}_{i,m}$ and $\bar{Q}_{i+1,N_a}$ are random variables, their joint PDF is the metric of interest, and is determined as per **Lemma 2**.

Lemma 2: Let $\mathcal{Z}_{i,m}$ be a vector containing random variables $\bar{P}_{i,m}$ and $\bar{Q}_{i+1,N_a}$ (i.e. $\mathcal{Z}_{i,m} \triangleq \left(\bar{P}_{i,m}, \bar{Q}_{i+1,N_a}\right)$). Then, the joint PDF of $\bar{P}_{i,m}$ and $\bar{Q}_{i+1,N_a}$ for $U_{i,m}$ (i.e. $f_{\mathcal{Z}_{i,m}}(\bar{p}_i, \bar{q}_i)$) is given by (9), as shown at the bottom of the page, where $\bar{F}_{\bar{P}}(\bar{p}_i)$ is the complementary CDF of $\bar{P}$, obtained as

$$\begin{aligned}\bar{F}_{\bar{P}}(\bar{p}_i) &\triangleq 1 - F_{\bar{P}}(\bar{p}_i) \\ &= \frac{\alpha_1}{\alpha_2}\left(e^{-\alpha_2 \bar{p}_i} - e^{-\alpha_2 \bar{p}_{\max}}\right),\end{aligned} \tag{10}$$

while $\bar{p}_{\min} = P_0 \ln\left(\beta \frac{d_{\min}}{d_{\max}^{\eta}}\right)$ and $\bar{p}_{\max} = P_0 \ln\left(\beta \frac{d_{\min}}{d_{\min}^{\eta}}\right)$.

Proof: See Appendix B. ∎

Based on **Lemma 2**, the reliability $\mathcal{R}_{i,m}$ of a UE $U_{i,m}$ is obtained in **Lemma 3**.

Lemma 3: The reliability of a UE $U_{i,m}$ determined as

$$\mathcal{R}_{i,m} = 1 - \left(1 - \Pr\left(E^{NIC}\right)\Pr\left(E_{i,m}^{NDE}\right)\right)^K, \tag{11}$$

in which $\Pr\left(E^{NIC}\right)$ and $\Pr\left(E_{i,m}^{NDE}\right)$ are the probabilities of no inter-cluster collision (NIC), and no decoding error (NDE), respectively. Specifically, $\Pr\left(E^{NIC}\right)$ and $\Pr\left(E_{i,m}^{NDE}\right)$ are given by

$$\Pr\left(E^{NIC}\right) = \sum_{l=0}^{M-1} \omega_l \binom{M-1}{l}\left(1-\Pi_0^{N_c}\right)^l \left(\Pi_0^{N_c}\right)^{M-l-1}, \tag{12}$$

and

$$\Pr\left(E_{i,m}^{NDE}\right) = \sum_{k=i}^{N_c}\prod_{j=1}^{i}(1-\bar{\Upsilon}_{i,j,k,m})\binom{N_c}{k}(1-\Pi_0)^k\,\Pi_0^{N_c-k}, \tag{13}$$

respectively. Furthermore, ω_l and $\bar{\Upsilon}_{i,j,k,m}$ are given as

$$\omega_l = \begin{cases} 1, & l = 0, \\ \left(\frac{R_u - 1}{R_u}\right)^l, & \text{otherwise}, \end{cases} \tag{14}$$

and

$$\bar{\Upsilon}_{i,j,k,m} = \int_{p_{\min}}^{p_{\max}} \int_{(k-i)\bar{p}_{\min}}^{(k-i)\bar{p}_i} \Upsilon_{j,k,m} f_{\mathcal{Z}_{i,m}}(\bar{p}_i, \bar{q}_i) d\bar{q}_i d\bar{p}_i, \tag{15}$$

where $\Upsilon_{j,k,m} \triangleq \Upsilon\left(\gamma_{j,m}(\bar{P}_{j,m}, \bar{Q}_{j+1,k}), n_b, n_d\right)$.

Proof: See Appendix C. ∎

[9] Intuitively, the finite latency requirement necessitates the stability condition, which implies that $\rho < 1$.

[10] In this work, perfect SIC is assumed.

$$\begin{aligned} &f_{\mathcal{Z}_{i,m}}(\bar{p}_i, \bar{q}_i) \\ &= \frac{\alpha_1^{N_a-i+1} N_a! e^{(N_a-i+1)\bar{p}_i + (1-\alpha_2)i - 1} e^{-\alpha_2 \bar{q}_i}}{(N_a-i)!(N_a-i-1)!(i-1)!}\left(\bar{F}_{\bar{P}}(\bar{p}_i)\right)^{i-1} \sum_{j=0}^{N_a-i}(-1)^{N_a-i+j}\binom{N_a-i}{j}\left[\bar{q}_i - (N_a-i-j)\bar{p}_i - j\bar{p}_{\min}\right]^{N_a-i-1} \end{aligned} \tag{9}$$

D. GoodPut

Recall that the successfully delivered non-redundant data bits per time unit at the BS by $U_{i,m}$ is defined as the GoodPut $\mathcal{G}_{i,m}$. Specifically, in a frame of duration T_f, the number of effective successfully delivered bits is $n_d\mathcal{R}_{i,m}$, and thus $\mathcal{G}_{i,m} = \frac{n_d\mathcal{R}_{i,m}}{T_f}$. In turn, the network GoodPut is written as

$$\mathcal{G}_N = \sum_{m=1}^{M}\sum_{i=1}^{N_c}\mathcal{G}_{i,m} = \sum_{m=1}^{M}\sum_{i=1}^{N_c}\frac{n_d}{T_f}\mathcal{R}_{i,m}. \tag{16}$$

IV. NUMERICAL RESULTS

In this section, the effect of the number of packet replicas K, blocklength n_b, and number of UEs in a cluster (i.e. cluster size) on the network GoodPut, reliability, and average packet latency is evaluted.[11] The simulated network parameters are set according to Table 2[12],[13] unless stated otherwise [54].

TABLE 2. Simulation Parameters.

Parameter	Value	Parameter	Value
N	60	R_b	210 RBs
λ_p	100 pkt/s	T_p	0.2 ms
B	180 KHz	N_0	-174 dBm/Hz
$d_{\min}$	100 m	$d_{\max}$	1000 m
T_r	1 ms	n_d	32 Bytes
η	3.2	β	4×10^7
P_0	0.0407	N_c	$1, 2, 3$

Fig. 2 shows the effect of blocklength n_b and number of replicas K on the network GoodPut $\mathcal{G}_N$ when $N_c = 2$. It can be observed that $\mathcal{G}_N$ increases sharply as n_b increases beyond 75, peaks at some values, and then gradually starts to decrease. For low values of n_b, $\mathcal{G}_N$ is low, which is due to the excessively high decoding error at the BS in FBL regime. Moreover, increasing the blocklength increases the number of successfully decoded data bits (or equivalently lowers the decoding error), and improves $\mathcal{G}_N$. However, excessively increasing n_b lowers $\mathcal{G}_N$. This is because the excessive increase in blocklength does not result in further improvement in the decoding error; on the contrary, it increases the frame duration, leading to a decrease in the number of effective and non-redundant data bits transmitted per time unit by each IoT UE. A similar trend is observed for K. Clearly, increasing the number of packet replicas results in higher successful data delivery rate, and $\mathcal{G}_N$ peaks at $K = 4$. However, for excessively high values of K (e.g. when $K > 4$), the $\mathcal{G}_N$ starts to decrease. The reason for this phenomenon is that the higher the number of replicas is, the lower the number of RUs R_u available for each cluster, and hence, the higher the inter-cluster interference and packet collisions. Thus, $\mathcal{G}_N$ decreases when transmitting higher number of packet replicas.

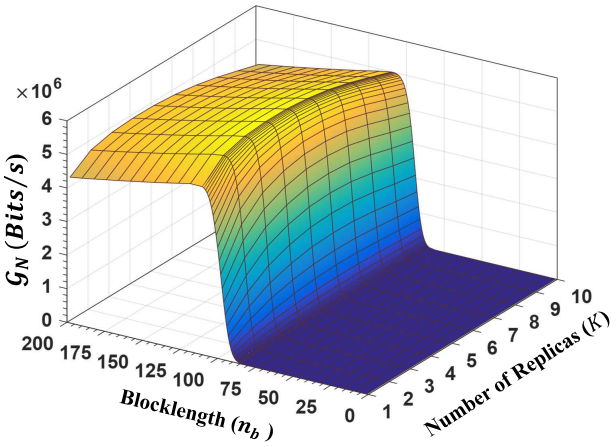


FIGURE 2. Network GoodPut vs. number of replicas K and blocklength n_b - $N_c = 2$.

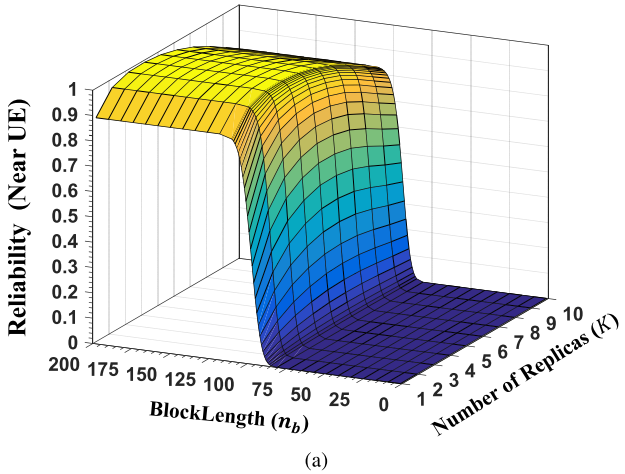
To explore the reliability metric, the reliability of the near and far UEs in a typical cluster with $N_c = 2$ is plotted as a function of the number of replicas K and blocklength n_b in Figs. 3a and 3b, respectively. One can see from Figs. 3a and 3b that the higher the blocklength is, the less the decoding error, and consequently, the higher the reliability. Moreover, increasing the number of replicas initially improves the reliability, since the packet success delivery rate becomes high when multiple replicas of a packet are sent. However, transmitting an excessively higher number of replicas ($K > 4$) reduces the number of available RUs for each cluster, which in turn increases inter-cluster interference and collisions. Therefore, the reliability starts to decrease. Note that due to the SIC decoding, and according to the derivation of (11), the reliability of the far UE in the cluster is lower or equal to the reliability of the other UEs (near UE herein) in the same cluster. Hence, Fig. 3b is the minimum reliability experienced by the UEs in the cluster.

The reliability of the far UE in a typical cluster in terms of K for different cluster sizes (i.e. $N_c = 1$, 2 and 3) is plotted in Fig. 4. It should be noted that since the total number of UEs in the network set to 60, each cluster size corresponds to a specific number of clusters, which are labeled in Fig. 4 as $M = 20$, 30, and 60. Now, it can be seen that for all values of N_c, the reliability peaks at some values of K, and then starts to decrease for the reasoning provided for Figs. 3a and 3b. However, when $N_c = 1$, the reliability of the only UE in the cluster is lower than the case in which $N_c = 2$ and $N_c = 3$. Note that $N_c = 1$ is equivalent to the OMA scenario, in which all the UEs contend for the limited RUs randomly. Hence, the inter-cluster interference is higher than when $N_c = 2$

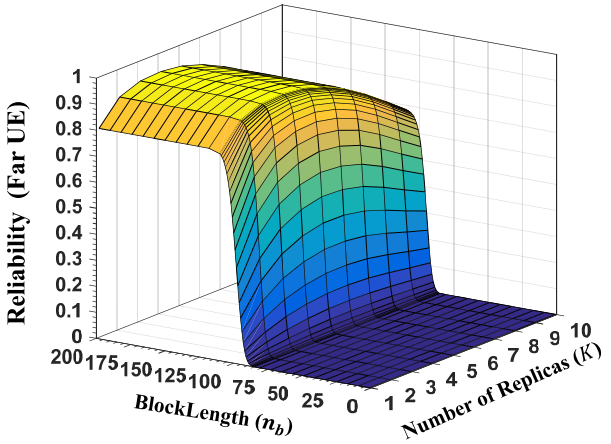
[11] Note that the total number of UEs in the network is assumed constant. Thus, changing the cluster size is equivalent to changing the number of clusters. Intuitively, the greater the cluster size is, the smaller the number of clusters.

[12] The preamble time T_p is adopted from RA-LTE as PRACH preamble format #4 for UL-TX [53].

[13] In order to have integer numbers of UEs in each cluster, the number of UEs in the network is set to $N = 60$. Furthermore, the curves have been interpolated for the missing values of cluster size to obtain the smooth plot.



(a)



(b)

FIGURE 3. UE Reliability vs. number of replicas K and blocklength n_b - $N_c = 2$, for: (a) Near UE, and (b) Far UE.

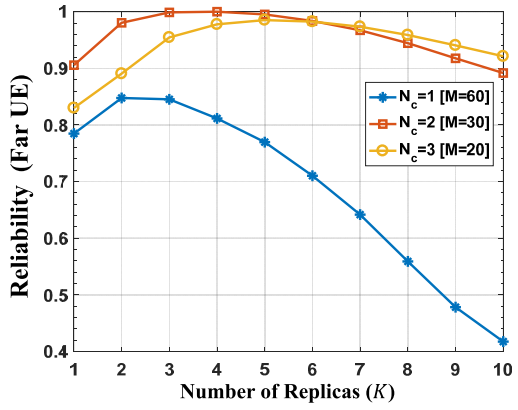


FIGURE 4. Far UE reliability vs. number of replicas K - $n_b = 120$.

and $N_c = 3$, which in turn highlights the merits of NOMA in IoT networks. Another observation is that for $K < 6$, the reliability of the far UE when $N_c = 2$ is greater than two other scenarios, and for $K > 6$, the reliability of the far UE in $N_c = 3$ is higher than in the cases of $N_c = 1$ and $N_c = 2$. To see this, note that increasing N_c has an adverse effect on the reliability of the far UE. On one side, increasing N_c can increase the reliability by decreasing the number of clusters

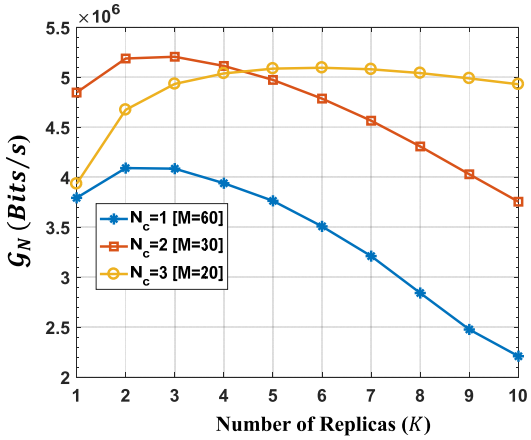


FIGURE 5. Network GoodPut vs. number of replicas K - $n_b = 120$.

contending for the available RUs; ultimately, lowering the inter-cluster interference. On the other side, the higher the number of UEs in a cluster is, the lower the far UE reliability, which is due to the block error resulting from decoding the UEs at the previous stages of the SIC. This also reflects the interplay between K and N_c, since each value of K affects the inter-cluster interference of all clusters. Finally, such an interplay manifests itself in $K = 6$, in which the reliability for $N_c = 2$ and $N_c = 3$ intersect.

Fig. 5 depicts the network GoodPut as a function of K for different cluster sizes N_c. Generally speaking, by increasing the number of replicas, $\mathcal{G}_N$ for all cluster sizes increases, peaks at some values, and then decreases for the same reasons given for Fig. 2. Furthermore, similar trends to Fig. 4 can be seen for $\mathcal{G}_N$. Particularly, $\mathcal{G}_N$ when $N_c = 1$ is the lowest compared to other values of N_c. This is because when $N_c = 1$, there are more clusters contending for the RUs than in $N_c = 2$ and $N_c = 3$. Thus, the inter-cluster interference is so high that the number of successfully transferred bits is much lower than that for $N_c = 2$ and $N_c = 3$. Another observation is that for $K < 4$, $\mathcal{G}_N$ for $N_c = 2$ is greater than for $N_c = 3$; while for $K > 5$, $\mathcal{G}_N$ for $N_c = 3$ becomes higher than for $N_c = 2$.

The reliability of the far UE as a function of blocklength n_b is shown in Fig. 6. As can be seen, for a fixed number of K and N_c, the higher the blocklength is, the lower the FBL decoding error, and hence, the higher the reliability. It can also be observed that for a target reliability, the greater the cluster size is, the greater the blocklength. For instance, $\mathcal{R} = 0.8$ can be achieved with $n_b = 72$, $n_b = 100$, and $n_b = 116$ for $N_c = 1$, $N_c = 2$, and $N_c = 3$, respectively. This is because the received SNR of far UE (third UE) at the BS in a network with $N_c = 3$ is lower than the that of far UE (second UE) in a network with $N_c = 2$.[14] Thus, for a low decoding error,

[14]Note that the UEs distances to the BS are assumed to be uniformly distributed. Thus, the average distance (or equivalently the received power) for the far UE when $N_c = 3$ is greater (lower) than that of the far UE when $N_c = 2$. This can be concluded from (3).

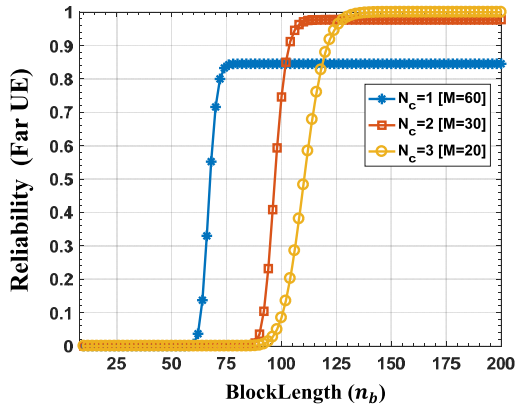


FIGURE 6. Far UE reliability vs. blocklength n_b - $K = 2$.

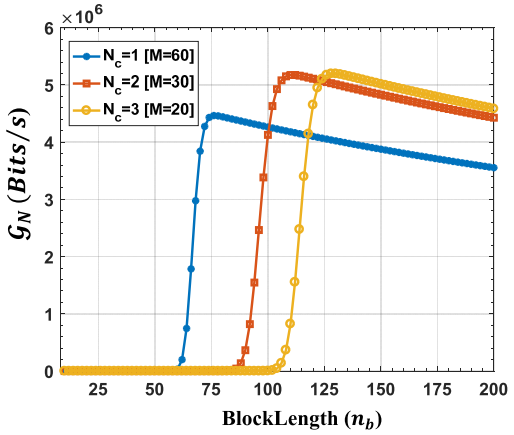


FIGURE 7. Network GoodPut vs. blocklength n_b - $K = 2$.

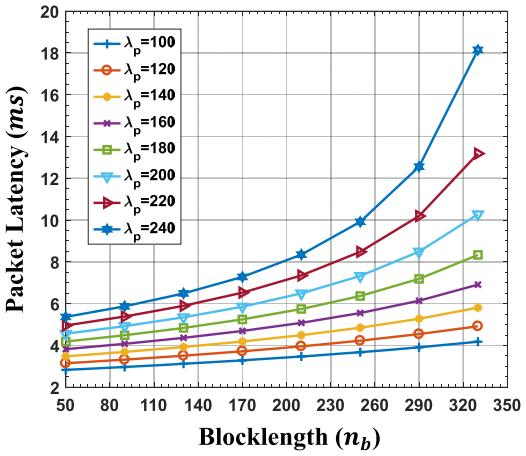


FIGURE 8. Average packet latency for different λ_p vs. blocklength n_b.

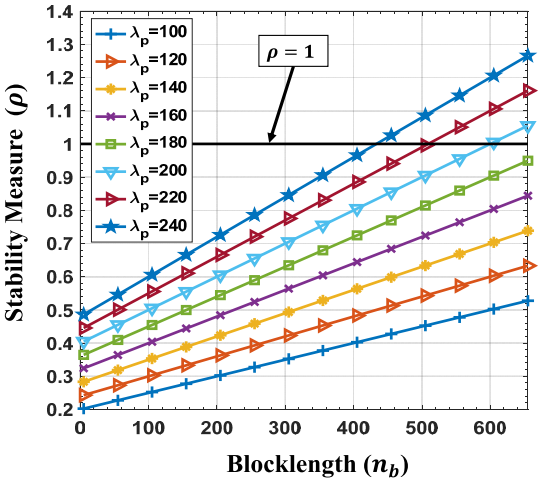


FIGURE 9. ρ for different λ_p vs. blocklength n_b - $K = 2$.

a longer blocklength is needed. Another observation is that when n_b is sufficiently high (e.g. $n_b > 130$), the reliability for $N_c = 3$ is higher than the other two cases of $N_c = 1$ and $N_c = 2$. This is because when higher blocklength is considered, the FBL decoding error is sufficiently small, and hence, the inter-cluster interference plays the main role in the reliability value, which is lower for $N_c = 3$ than for $N_c = 1$ and $N_c = 2$.

Fig. 7 presents the network GoodPut as a function of the blocklength n_b for different cluster sizes. Intuitively, the higher the blocklength is, the higher the successfully received bits and hence, the higher the $\mathcal{G}_N$. However, excessively increasing the blocklength results in a longer frame duration, which is counterproductive in terms of network GoodPut (i.e. reduces $\mathcal{G}_N$). Hence, n_b critically affects the network GoodPut.

Finally, the average packet latency $\mathcal{L}$ for a typical UE as a function of n_b for different values of λ_p is demonstrated in Fig. 8. As can be seen, the higher the blocklength is, the higher the average packet latency. This is due to the fact that the frame duration increases by increasing n_b. Additionally, the rate at which the average packet latency increases is higher when the packet arrival rate λ_p of the UEs is high, resulting in higher latencies at the same blocklength values. The reason is that for high values of λ_p, the average packet waiting time in a UE's buffer is high, which threatens the network stability. To see this, Fig. 9 shows a UE's link utilization ρ, which is the measure of UE's buffer stability in terms of n_b, and for different values of λ_p. As explained earlier, the higher the values of n_b and λ_p are, the higher the link utilization. Note that the black line (i.e. for $\rho = 1$) is the stability threshold line; below which is the stability area. For example, the UE's buffer is unstable for $n_b > 430$ when $\lambda_p = 240$, $n_b > 505$ when $\lambda_p = 220$, $n_b > 607$ when $\lambda_p = 200$.

To summarize, the following trade-offs can be stated for the network metrics. Both reliability and network GoodPut experience a peak by varying the number of replicas K. Furthermore, $\mathcal{G}_N$ also experiences another peak by varying the blocklength n_b. Additionally, both reliability and $\mathcal{G}_N$ are significantly affected by cluster size N_c (or number of clusters M), which should be carefully set by network operators.

A direct trade-off is observed between reliability and packet latency through n_b, where increasing the blocklength improves the reliability, but increases the packet latency. An indirect trade-off between the two is seen via K, where the number of replicas must be carefully set (neither too low nor too high), such that the reliability requirement is met with the lowest possible blocklength, which in turn helps to achieve the average packet latency requirement. Hence, by carefully selecting the network parameters, $\mathcal{G}_N$ can be maximized subject to packet latency and reliability requirements.

V. NETWORK GoodPut MAXIMIZATION

The analytical derivations of the different network metrics can be utilized to maximize the network GoodPut, subject to constraints on the average packet latency and reliability. Specifically, the network GoodPut maximization (NGP-MAX) problem can be formulated as

NGP-MAX:

$$\max_{n_b,K,\beta,\eta,P_0} \quad \mathcal{G}_N \tag{17a}$$

$$\text{s.t. } \mathcal{L} \leq \delta_{th}^{L}, \tag{17b}$$

$$\mathcal{R}_{i,m} \geq \delta_{th}^{R}, \quad \forall i, m, \tag{17c}$$

$$P_{i,m} \leq P_{\max}, \quad \forall i, m, \tag{17d}$$

$$\beta \geq \frac{d_{\max}^{\eta}}{d_{\min}}, \tag{17e}$$

$$\eta \geq 1, \tag{17f}$$

$$n_b, K \in \{1, 2, \ldots\}. \tag{17g}$$

In problem **NGP-MAX**, Constraint (17b) ensures that the average packet latency does not exceed δ_{th}^{L}, while Constraint (17d) ensures that the reliability is at least δ_{th}^{R}. Moreover, Constraint (17d) enforces the maximum transmit power per UE, whereas Constraint (17e) ensures that $P_{i,m} > 0$, as discussed below (2). The remaining constraints define the range of values the decision variables take. Notably, optimizing the values n_b, K, β, η, and P_0 can maximize the network GoodPut, while satisfying the stringent URLL requirements for IoT applications, as per 3GPP and ITU specifications [55], [56].

Remark 3: Problem **NGP-MAX** is a nonlinear mixed integer programming problem, which is non-convex and computationally-intensive [57]. This is evident from the nonlinear analytical expressions of $\mathcal{L}$, $\mathcal{R}_{i,m}$, and $\mathcal{G}_N$ as well as the integer-valued decision variables. Despite the non-convexity of problem **NGP-MAX**, the incurred computational delay is irrelevant, which is due to the steady-state analysis.

Now, problem **NGP-MAX** is solved via a genetic algorithm (GA) for an IoT network with $N = 60$ UEs, and clusters of size $N_c = 3$. Table 3 shows the GA parameters.[15] Fig. 10 illustrates how the GA converges to the optimal network GoodPut value between different generations. Specifically, Fig. 10 shows the best and the mean values of the network GoodPut among 2000 individuals at each generation. The optimal $\mathcal{G}_N$ value is 5.13 Mbits/s, while the optimal values of the decision variables are $(n_b, K, \beta, \eta, P_0) = (155, 4, 3.1 \times 10^7, 3.43, 0.03)$ when $\delta_{th}^{R} = 0.99999$, $\delta_{th}^{L} = 10$ ms, and $P_{max} = 0.3$ W.

[15]Problem **NGP-MAX** is solved via the Genetic Algorithm Toolbox in MATLAB [58].

TABLE 3. Genetic Algorithm Parameters.

Parameter	Value	Parameter	Value
Population Size	2000	Generation	1000
Fitness Limit	10^{-30}	Fun Tolerance	10^{-20}
Crossover Fraction	0.8	Elite Count	120

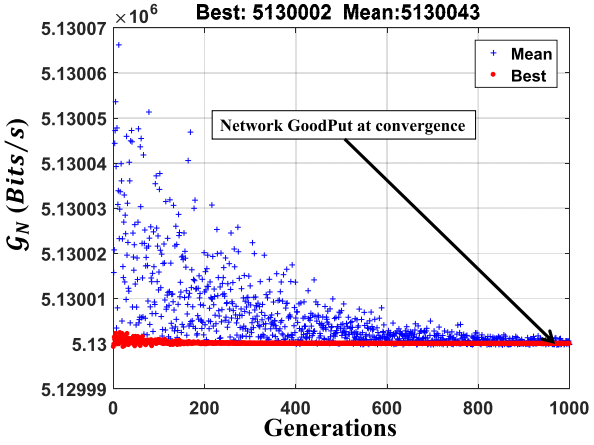


FIGURE 10. Convergence to optimal solution for network GoodPut.

VI. FUTURE RESEARCH DIRECTIONS

One of the main advantages of performing steady-state analyses is to gain insight into the behaviour and characteristics of various network metrics in response to different values of network parameters. Such analyses can be incorporated into learning techniques and algorithms to fine-tune the network configurations and transmission policies [59], which is due to potential model mismatches and uncertainties in practical communication networks. Since the number of UEs, and their data traffic patterns have been considered in deriving the network metrics, the results can be used as labeled training data sets for off-line training of deep neural networks (DNNs). Consequently, each UE in the network can distributively control its transmission parameters (e.g. transmit power), while integrating the results into admission control policies adopted by the network control center [60]. Furthermore, since the UEs deployment is assumed to be random, with the number of UEs and cell radius as network variables, performing scalability analysis is of the essence. Another possible extension of this work is to integrate the proposed frame structure into a distributed queueing (DQ) algorithm [61], and perform performance evaluations for various IoT network metrics.

VII. CONCLUSION

This paper has considered grant-free RA-NOMA in URLL IoT networks with clustered UEs. Short packets with transmission diversity over multiple resource blocks have been adopted to realize URLL within each transmission time interval. By exploiting the time-reversal strategy along with preamble transmission, the IoT UEs within each cluster are able to distributively adjust their transmit powers to achieve the target received power at the base-station. Furthermore, active UE detection at the BS is achieved through the received preambles. Network metrics, namely average packet latency, reliability, and GoodPut have been mathematically derived. Furthermore, the effect of the number of packet replicas, and blocklength on the IoT network metrics have been explored numerically. More importantly, various tradeoffs between the different network metrics and parameters have been highlighted to shed light on the importance of carefully selecting the number of packet replicas and blocklength. Particularly, both reliability and network GoodPut experience a peak by varying the number of packet replicas. Moreover, network GoodPut also experiences another peak by varying the blocklength. Additionally, both reliability and GoodPut are significantly affected by cluster size (or number of clusters), which should be carefully set by network operators. Direct and indirect trade-offs are also observed between reliability and packet latency through blocklength and number of packet replicas, where the number of replicas must be carefully set (neither too low nor too high), such that the reliability requirement is met with the lowest possible blocklength, which in turn helps to achieve the average packet latency requirement. Finally, the derived expressions have then been utilized to maximize the network GoodPut subject to URLL as well as transmit power constraints.

APPENDIX A
PROOF OF LEMMA 1

Proof: To derive the PDF of the received power of an IoT UE located at distance d from the BS, let $\bar{P} \triangleq g(d) = P_0 \ln\left(\beta \frac{d_{\min}}{d^{\eta}}\right)$. Thus, according to probability theory [62], $f_{\bar{P}}(\bar{p})$ is obtained as

$$f_{\bar{P}}(\bar{p}) = f_d(g^{-1}(\bar{p}))\left|\frac{d}{d\bar{p}} g^{-1}(\bar{p})\right|, \tag{A.1}$$

where $g^{-1}(\bar{p}) = \sqrt[\eta]{\beta d_{\min}} e^{-\frac{\bar{p}}{P_0 \eta}}$ and $f_d(\bar{d}) = \frac{1}{d_{\max} - d_{\min}}$. In turn, $f_{\bar{P}}(\bar{p})$ can be simplified as

$$f_{\bar{P}}(\bar{p}) = \alpha_1 e^{-\alpha_2 \bar{p}}, \tag{A.2}$$

where $\alpha_1 = \frac{\sqrt[\eta]{\beta d_{\min}}}{P_0 \eta (d_{\max} - d_{\min})}$ and $\alpha_2 = \frac{1}{P_0 \eta}$. ∎

APPENDIX B
PROOF OF LEMMA 2

Proof: Let $\bar{P}_j$'s (for $j = 1, \ldots, k$) refer to k i.i.d. random variables,[16] with common PDF $f_{\bar{P}}(p)$ and cumulative distribution function (CDF) $F_{\bar{P}}(p)$. Also, let $\bar{P}_{(j)}$ represent their order statistics, such that $\bar{p}_{\max} \geq \bar{P}_{(1)} > \bar{P}_{(2)} > \ldots > \bar{P}_{(k)} \geq \bar{p}_{\min}$, where $\bar{p}_{\min} = P_0 \ln\left(\beta \frac{d_{\min}}{d^{\eta}_{\max}}\right)$ and $\bar{p}_{\max} = P_0 \ln\left(\beta \frac{d_{\min}}{d^{\eta}_{\min}}\right)$ are obtained from (3) for $d_i = d_{\max}$ and $d_i = d_{\min}$, respectively. Then, the $(k-i+1)$-dimensional joint PDF of $\{\bar{P}_{(j)}\}_{j=i}^{k}$ can be inferred as

$$\begin{aligned} &f_{\bar{P}_{(i)},\ldots,\bar{P}_{(k)}}(\bar{p}_i, \ldots, \bar{p}_k) \\ &= \binom{k}{k-i+1}(k-i+1)! \prod_{j=i}^{k} f_{\bar{P}_j}(\bar{p}_j)\left(1 - F_{\bar{P}}(\bar{p}_i)\right)^{i-1} \\ &= \frac{k!}{(i-1)!} \prod_{j=i}^{k} f_{\bar{P}}(\bar{p}_j)\left(1 - F_{\bar{P}}(\bar{p}_i)\right)^{i-1}, \end{aligned} \tag{B.1}$$

where $F_{\bar{P}}(\bar{p}_i) = \int_{\bar{p}_{\min}}^{\bar{p}_i} f_{\bar{P}}(\bar{p}) d\bar{p}$ for $\bar{p}_i \leq \bar{p}_{\max}$. Note that $f_{\bar{P}_{(i)},\ldots,\bar{P}_{(k)}}(\bar{p}_i, \ldots, \bar{p}_k) = \Pr\left(\bigcap_{l=i}^{k} \bar{p}_l - \Delta\bar{p} \leq \bar{P}_{(l)} \leq \bar{p}_l + \Delta\bar{p}\right)$ for small $\Delta\bar{p}$, which is equal to choosing $k-i+1$ random variables from the k variables (related to $\binom{k}{k-i+1}$) with $(k-i+1)!$ permutations. In turn, their values must be in the range $\left[\bar{p}_l - \Delta\bar{p}, \bar{p}_l + \Delta\bar{p}\right]$ (for $l = i, \ldots, k$), with probability $\prod_{j=i}^{k} f_{\bar{P}_j}(\bar{p}_j)$; whereas all the remaining $i-1$ variables must be greater than them, with probability $\left(1 - F_{\bar{P}}(\bar{p}_i)\right)^{i-1}$.

Now, to derive the joint PDF of $\bar{P}_{(i)}$ and $\bar{Q}_{(i+1),k} = \sum_{j=i+1}^{k} \bar{P}_{(j)}$, the MGF is exploited. Particularly, the second order MGF of $\mathcal{Z}_{i,m} = \left(\bar{P}_{(i)}, \bar{Q}_{(i+1),k}\right)$ is defined as[17]

$$\mathcal{M}_{\mathcal{Z}_{i,m}}(s_1, s_2) = \mathbb{E}\left[e^{\left(s_1 \bar{P}_{(i)} + s_2 \bar{Q}_{(i+1),k}\right)}\right]. \tag{B.2}$$

By using (B.1), (B.2) can be expressed as given in (B.3), as shown at the bottom of the next page.

Based on [63, Eq. (9) and (10)], (B.3) can be further simplified as (B.4), as shown at the bottom of the next page, in which $\mathcal{C}(\bar{p}_i, s_2) = \int_{\bar{p}_{\min}}^{\bar{p}_i} f_{\bar{P}}(\bar{p}) e^{s_2 \bar{p}} d\bar{p}$. By utilizing $f_{\bar{P}}(\bar{p})$ in (4), $\mathcal{C}(\bar{p}_i, s_2)$ is obtained as

$$\begin{aligned} \mathcal{C}(\bar{p}_i, s_2) &= \int_{\bar{p}_{\min}}^{\bar{p}_i} \alpha_1 e^{-\alpha_2 \bar{p}} e^{s_2 \bar{p}} d\bar{p} \\ &= \frac{\alpha_1 e^{(s_2 - \alpha_2)\bar{p}_i}}{s_2 - \alpha_2}\left(1 - e^{(s_2 - \alpha_2)(\bar{p}_{\min} - \bar{p}_i)}\right). \end{aligned} \tag{B.5}$$

Hence, by substituting (B.5), as shown at the bottom of the next page into (B.4), $\mathcal{M}_{\mathcal{Z}_{i,m}}(s_1, s_2)$ can be derived as (B.6), as shown at the bottom of the next page. Therefore, the joint PDF of $\mathcal{Z}_{i,m}$ can be derived by taking the inverse Laplace transform as $f_{\mathcal{Z}_{i,m}}(\bar{p}_i, \bar{q}_i) = \mathcal{L}^{-1}_{s_1,s_2}\left\{\mathcal{M}_{\mathcal{Z}_{i,m}}(-s_1, -s_2)\right\}$, as given in (B.7), as shown at the bottom of the next page. However, directly applying the inverse Laplace transform with respect to s_2 is difficult. Alternatively, the binomial expansion is

[16] For notational convenience, the subscript m referring to the cluster index is dropped.

[17] Such a technique has previously been applied to determine the joint statistics of sums of ordered random variables in special cases [27], [28]. The derivations here are different from the existing in the literature in that the joint PDF of the i^{th} order statistic and partial sum of the remaining $(N_a - i + 1)$ least order statistics from the total N_a order statistics is computed.

employed as

$$\left(1-e^{-\alpha_2(\bar{p}_{\min}-\bar{p}_i)}e^{(\bar{p}_{\min}-\bar{p}_i)s_2}\right)^{k-i} = \sum_{j=0}^{k-i}(-1)^j\binom{k-i}{j}e^{-j\alpha_2(\bar{p}_{\min}-\bar{p}_i)}e^{j(\bar{p}_{\min}-\bar{p}_i)s_2}. \quad \text{(B.8)}$$

Finally, (B.9) (shown at the bottom of the page) gives the joint PDF of $f_{\mathcal{Z}_{i,m}}(\bar{p}_i,\bar{q}_i)$, in which $\bar{F}_{\bar{P}}(\bar{p}_i) \triangleq 1-F_{\bar{P}}(\bar{p}_i) = \frac{\alpha_1}{\alpha_2}\left(e^{-\alpha_2\bar{p}_i}-e^{-\alpha_2\bar{p}_{\max}}\right)$. The last equality $\overset{\dagger}{=}$ in (B.9) comes from the fact that $\hbar(\bar{p}_i) = \mathcal{L}_{s_1}^{-1}\left\{\int_{\bar{p}_{\min}}^{\bar{p}_{\max}} e^{-s_1\bar{p}_i}\hbar(\bar{p}_i)d\bar{p}_i\right\}$, and (B.9) holds true for $\bar{p}_{\min} < \bar{p}_i < \bar{p}_{\max}$ and $(k-i)p_{\min} < \bar{q}_i < (k-i)\bar{p}_i$.

■

APPENDIX C
PROOF OF LEMMA 3

Proof: Recall that reliability is defined as the probability that a typical packet is successfully received at the BS. and that transmission diversity is utilized to enhance the reliability. Then, a packet is delivered successfully if at least one packet among the K transmitted replica packets is received successfully. In turn, let E_s be the event that

$$\mathcal{M}_{\mathcal{Z}_{i,m}}(s_1,s_2) = \mathbb{E}\left[e^{\left(s_1\bar{P}_{(i)}+s_2\sum_{j=i+1}^{k}\bar{P}_{(j)}\right)}\right] = \frac{k!}{(i-1)!}\int_{\bar{p}_{\min}}^{\bar{p}_{\max}} f_{\bar{P}}(\bar{p}_i)e^{s_1\bar{p}_i}d\bar{p}_i\int_{\bar{p}_{\min}}^{\bar{p}_i} f_{\bar{P}}(\bar{p}_{i+1})e^{s_2\bar{p}_{i+1}}d\bar{p}_{i+1}\cdots\int_{\bar{p}_{\min}}^{\bar{p}_{k-1}} f_{\bar{P}}(\bar{p}_k)e^{s_2\,p_k}\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}d\bar{p}_k \quad \text{(B.3)}$$

$$\mathcal{M}_{\mathcal{Z}_{i,m}}(s_1,s_2) = \frac{k!}{(i-1)!}\int_{\bar{p}_{\min}}^{\bar{p}_{\max}} f_{\bar{P}}(\bar{p}_i)e^{s_1\bar{p}_i}\left[\frac{1}{(k-i)!}\left(\mathcal{C}(\bar{p}_i,s_2)\right)^{k-i}\right]\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}d\bar{p}_i \quad \text{(B.4)}$$

$$\mathcal{M}_{\mathcal{Z}_{i,m}}(s_1,s_2) = \frac{k!}{(k-i)!(i-1)!}\int_{\bar{p}_{\min}}^{\bar{p}_{\max}}\alpha_1^{k-i+1}e^{-\alpha_2\bar{p}_i}\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}e^{s_1\bar{p}_i}\frac{e^{(k-i)(s_2-\alpha_2)\bar{p}_i}\left(1-e^{(s_2-\alpha_2)(\bar{p}_{\min}-\bar{p}_i)}\right)^{N_a-i}}{(s_2-\alpha_2)^{k-i}}d\bar{p}_i \quad \text{(B.6)}$$

$$f_{\mathcal{Z}_{i,m}}(\bar{p}_i,\bar{q}_i) = \mathcal{L}_{s_1,s_2}^{-1}\left\{\mathcal{M}_{\mathcal{Z}_{i,m}}(-s_1,-s_2)\right\} = \frac{\alpha_1^{k-i+1}k!}{(k-i)!(i-1)!} \times\mathcal{L}_{s_1}^{-1}\left\{\int_{\bar{p}_{\min}}^{\bar{p}_{\max}}e^{-\alpha_2\bar{p}_i}\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}e^{-s_1\bar{p}_i}e^{-(k-i)\alpha_2\bar{p}_i}\mathcal{L}_{s_2}^{-1}\left\{\frac{e^{(k-i)\bar{p}_is_2}\left(1-e^{(s_2-\alpha_2)(\bar{p}_{\min}-\bar{p}_i)}\right)^{N_a-i}}{(-s_2-\alpha_2)^{k-i}}\right\}d\bar{p}_i\right\} \quad \text{(B.7)}$$

$$\begin{aligned}
f_{\mathcal{Z}_{i,m}}(\bar{p}_i,\bar{q}_i) &= \frac{\alpha_1^{k-i+1}k!}{(k-i)!(i-1)!}\mathcal{L}_{s_1}^{-1}\left\{\int_{\bar{p}_{\min}}^{\bar{p}_{\max}}e^{-\alpha_2\bar{p}_i}\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}\right.\\
&\quad\left.\times e^{-s_1\bar{p}_i}e^{-(k-i)\alpha_2\bar{p}_i}\sum_{j=0}^{k-i}\binom{k-i}{j}e^{-j\alpha_2(\bar{p}_{\min}-\bar{p}_i)}\mathcal{L}_{s_2}^{-1}\left\{\frac{e^{-((k-i-j)\bar{p}_i+j\bar{p}_{\min})s_2}}{(-s_2-\alpha_2)^{k-i}}\right\}d\bar{p}_i\right\}\\
&= \frac{\alpha_1^{k-i+1}k!}{(k-i)!(k-i-1)!(i-1)!}\mathcal{L}_{s_1}^{-1}\left\{\int_{\bar{p}_{\min}}^{\bar{p}_{\max}}e^{-\alpha_2\bar{p}_i}\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}\right.\\
&\quad\left.\times e^{-s_1\bar{p}_i}e^{-(k-i)\alpha_2\bar{p}_i}\sum_{j=0}^{k-i}(-1)^{k-i+j}\binom{k-i}{j}\left[\bar{q}_i-(k-i-j)\bar{p}_i-j\bar{p}_{\min}\right]^{k-i-1}e^{-\alpha_2(\bar{q}_i-(k-i)\bar{p}_i)}d\bar{p}_i\right\}\\
&= \frac{\alpha_1^{k-i+1}k!}{(k-i)!(k-i-1)!(i-1)!}\mathcal{L}_{s_1}^{-1}\left\{\int_{\bar{p}_{\min}}^{\bar{p}_{\max}}e^{-s_1\bar{p}_i}\left[e^{-\alpha_2\bar{p}_i}\left(1-F_{\bar{P}}(\bar{p}_i)\right)^{i-1}\right.\right.\\
&\quad\left.\left.\times e^{-(k-i)\alpha_2\bar{p}_i}\sum_{j=0}^{k-i}(-1)^{k-i+j}\binom{k-i}{j}\left[\bar{q}_i-(k-i-j)\bar{p}_i-j\bar{p}_{\min}\right]^{k-i-1}e^{-\alpha_2(\bar{q}_i-(k-i)\bar{p}_i)}\right]d\bar{p}_i\right\}\\
&\overset{\dagger}{=} \frac{\alpha_1^{k-i+1}k!e^{(k-i+1)\bar{p}_i+(1-\alpha_2)i-1}e^{-\alpha_2\bar{q}_i}}{(k-i)!(k-i-1)!(i-1)!}\left(\bar{F}_{\bar{P}}(\bar{p}_i)\right)^{i-1}\sum_{j=0}^{k-i}(-1)^{k-i+j}\binom{k-i}{j}\left[\bar{q}_i-(k-i-j)\bar{p}_i-j\bar{p}_{\min}\right]^{k-i-1}
\end{aligned} \quad \text{(B.9)}$$

one packet replica is received successfully, and $\bar{E}_s$ is its complement. In turn, the reliability for an IoT UE $U_{i,m}$ is determined as

$$\begin{aligned}\mathcal{R}_{i,m} &= 1 - \left(\Pr\left(\bar{E}_s\right)\right)^K \\ &= 1 - \left(1 - \Pr(E_s)\right)^K . \end{aligned} \tag{C.1}$$

Now, a packet is delivered successfully if no inter-cluster collision (NIC) and no decoding error (NDE) occur. To this end, define E^{NIC} as the event of NIC, and $E^{NDE}_{i,m}$ as that of NDE for the $U_{i,m}$'s transmitted packet. Hence, $\Pr(E_s)$ in (C.1) can be written as

$$\Pr(E_s) = \Pr\left(E^{NIC}\right) \Pr\left(E^{NDE}_{i,m}\right). \tag{C.2}$$

Now, let the number of IoT clusters with at least one UE having at least one data packet in its buffer—except the underlying cluster m—be denoted $\bar{N}_m$. Then, $\Pr\left(E^{NIC}\right)$ can be conditioned on $\bar{N}_m$ as

$$\Pr\left(E^{NIC}\right) = \sum_{l=0}^{M-1} \Pr\left(E^{NIC} \,|\, \bar{N}_m = l\right) \Pr\left(\bar{N}_m = l\right). \tag{C.3}$$

Since it is assumed that all the UEs follow the same data arrival process, and there is no preemptive UE (i.e. all the UEs have the same channel access priority), then $\Pr\left(\bar{N}_m = l\right)$ can be written as

$$\Pr\left(\bar{N}_m = l\right) = \binom{M-1}{l} \left(1 - \Pi_0^{N_c}\right)^l \left(\Pi_0^{N_c}\right)^{M-l-1}, \tag{C.4}$$

where Π_0 is defined in (7). On the other hand, $\Pr\left(E^{NIC} \mid \bar{N}_m = l\right)$ can be obtained as

$$\begin{aligned}\omega_l &\triangleq \Pr\left(E^{NIC} \mid \bar{N}_m = l\right) \\ &= \begin{cases} 1, & l = 0, \\ \left(\dfrac{R_u - 1}{R_u}\right)^l, & \text{otherwise.} \end{cases} \end{aligned} \tag{C.5}$$

Particularly, $l = 0$ is related to the case of no collision, since no IoT cluster among the other clusters transmits any data packets. On the other hand, there is no collision on the selected RU if all l IoT UEs select their RUs among the other $(R_u - 1)$ RUs. In turn, by substituting (C.4) and (C.5) into (C.3), $\Pr\left(E^{NIC}\right)$ is determined as

$$\Pr\left(E^{NIC}\right) = \sum_{l=0}^{M-1} \omega_l \binom{M-1}{l} \left(1 - \Pi_0^{N_c}\right)^l \left(\Pi_0^{N_c}\right)^{M-l-1}. \tag{C.6}$$

To obtain $\Pr\left(E^{NDE}_{i,m}\right)$ in (C.2), the decoding error probability expression in (1) is used. However, the number of active UEs N_a in the cluster of interest must be conditioned on, and thus, $\Pr\left(E^{NDE}_{i,m}\right)$ can be written as

$$\Pr\left(E^{NDE}_{i,m}\right) = \sum_{k=i}^{N_c} \Pr\left(E^{NDE}_{i,m} \mid N_a = k\right) \Pr\left(N_a = k\right), \tag{C.7}$$

where $\Pr\left(N_a = k\right) = \binom{N_c}{k} \left(1 - \Pi_0\right)^k \Pi_0^{N_c - k}$. Also, recall that $U_{i,m}$'s signal is successfully decoded if the BS successfully decodes all the $i-1$ previous UEs' signals as well as $U_{i,m}$'s signal via SIC. Hence, by defining $\Upsilon_{j,k,m} \triangleq \Upsilon\left(\gamma_{j,m}(\bar{P}_{j,m}, \bar{Q}_{j+1,k}), n_b, n_d\right)$ based on (1) and (8), $\Pr\left(E^{NDE}_{i,m} \mid N_a = k\right)$ can be written as

$$\Pr\left(E^{NDE}_{i,m} \mid N_a = k\right) = \prod_{j=1}^{i} (1 - \bar{\Upsilon}_{i,j,k,m}), \tag{C.8}$$

in which $\bar{\Upsilon}_{i,j,k,m}$ is obtained using

$$\bar{\Upsilon}_{i,j,k,m} = \int_{p_{\min}}^{p_{\max}} \int_{(k-i)p_{\min}}^{(k-i)\bar{p}_i} \Upsilon_{j,k,m} f_{\mathcal{Z}_{i,m}}(\bar{p}_i, \bar{q}_i) d\bar{q}_i d\bar{p}_i. \tag{C.9}$$

Therefore, $\Pr\left(E^{NDE}_{i,m}\right)$ can be written as

$$\Pr\left(E^{NDE}_{i,m}\right) = \sum_{k=i}^{N_c} \prod_{j=1}^{i} \left(1 - \bar{\Upsilon}_{i,j,k,m}\right) \binom{N_c}{k} \left(1 - \Pi_0\right)^k \Pi_0^{N_c - k}. \tag{C.10}$$

Finally, substituting (C.6) and (C.10) into (C.2) and then into (C.1) gives the reliability expression in (11). ■

REFERENCES

[1] *5G; Study on Scenarios and Requirements for Next Generation Access Technologies*, document ETSI TR 138 913 V14.2.0, 3rd Generation Partnership Project, 2017. [Online]. Available: https://www.etsi.org

[2] G. Hampel, C. Li, and J. Li, "5G ultra-reliable low-latency communications in factory automation leveraging licensed and unlicensed bands," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 117–123, May 2019.

[3] 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 Things J.*, vol. 6, no. 5, pp. 7946–7970, Oct. 2019.

[4] 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, p. 981, Sep. 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/9/981

[5] 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. 110880–110896, Jun. 2020.

[6] 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 Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[7] P. Popovski, C. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[8] 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 J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[9] M. Vaezi, G. A. Aruma Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

[10] 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 Commun. Mag.*, vol. 57, no. 2, pp. 130–137, Feb. 2019.

[11] 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 Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.

[12] J. Choi, "Re-transmission diversity multiple access based on SIC and HARQ-IR," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4695–4705, Nov. 2016.

[13] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency trade-off in ultra-reliable low-latency communication with short packets," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

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

[15] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sep. 2017.

[16] R. Abbas, T. Huang, B. Shahab, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Grant-free non-orthogonal multiple access: A key enabler for 6G-IoT," 2020, *arXiv:2003.10257*. [Online]. Available: http://arxiv.org/abs/2003.10257

[17] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[18] M. Mohammadkarimi, M. A. Raza, and O. A. Dobre, "Signature-based nonorthogonal massive multiple access for future wireless networks: Uplink massive connectivity for machine-type communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 40–50, Dec. 2018.

[19] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for Internet of Things," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[20] Q. Wang, Z. Zhao, D. Miao, Y. Zhang, J. Sun, M. Liu, and Z. Zhong, "Non-orthogonal coded access for contention-based transmission in 5G," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–6.

[21] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

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

[23] F. Ghanami, G. A. Hodtani, B. Vucetic, and M. Shirvanimoghaddam, "Performance analysis and optimization of NOMA with HARQ for short packet communications in massive IoT," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4736–4748, Mar. 2021.

[24] F. Nadeem, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Non-orthogonal HARQ for delay sensitive applications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[25] M. Shirvanimoghaddam, H. Khayami, Y. Li, and B. Vucetic, "Dynamic HARQ with guaranteed delay," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[26] Y. Chen, F. Han, Y. Yang, H. Ma, Y. Han, C. Jiang, H. Lai, D. Claffey, Z. Safar, and K. J. R. Liu, "Time-reversal wireless paradigm for green Internet of Things: An overview," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 81–98, Feb. 2014.

[27] A. H. Nuttall and P. M. Baggenstoss. (2002). Joint distributions for two useful classes of statistics, with applications to classification and hypothesis testing. Defense Technical Information Center. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA477219

[28] S. S. Nam, H.-C. Yang, M.-S. Alouini, and D. I. Kim, "An MGF-based unified framework to determine the joint statistics of partial sums of ordered i.n.d. random variables," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4270–4283, Aug. 2014.

[29] 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. 220734–220754, Dec. 2020.

[30] M. R. Amini and M. W. Baidas, "Performance analysis of URLL random-access NOMA-enabled IoT networks with short packet and diversity transmissions," in *Proc. Int. Conf. Commun., Signal Process., Their Appl. (ICCSPA)*, Mar. 2021, pp. 1–6.

[31] 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. 131796–131813, 2020.

[32] J.-B. Seo, B. C. Jung, and H. Jin, "Performance analysis of NOMA random access," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2242–2245, Nov. 2018.

[33] J.-B. Seo, H. Jin, and B. C. Jung, "Non-orthogonal random access with channel inversion for 5G networks," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2017, pp. 117–119.

[34] J.-B. Seo, B. C. Jung, and H. Jin, "Nonorthogonal random access for 5G mobile communication systems," *IEEE Trans. Veh. Technol*, vol. 67, no. 8, pp. 7867–7871, Aug. 2018.

[35] 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 Commun. Lett.*, vol. 21, no. 10, pp. 2246–2249, Oct. 2017.

[36] E. Dosti, M. Shehab, H. Alves, and M. Latva-Aho, "On the performance of non-orthogonal multiple access in the finite blocklength regime," *Ad Hoc Netw.*, vol. 84, pp. 148–157, Mar. 2019.

[37] M. Amjad and L. Musavian, "Performance analysis of NOMA for ultra-reliable and low-latency communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–5.

[38] O. B. H. Belkacem, M. L. Ammari, and R. Dinis, "Performance analysis of NOMA in 5G systems with HPA nonlinearities," *IEEE Access*, vol. 8, pp. 158327–158334, 2020.

[39] P. D. Diamantoulakis and G. K. Karagiannidis, "Performance analysis of distributed uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 788–792, Mar. 2021.

[40] P. Brown and S. E. Elayoubi, "Semi-distributed contention-based resource allocation for ultra reliable low latency communications," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Jul. 2020, pp. 1172–1180.

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

[42] T. Hoßfeld, F. Metzger, and P. E. Heegaard, "Traffic modeling for aggregated periodic IoT data," in *Proc. 21st Int. Conf. Innov. Clouds, Internet Netw. (ICIN)*, Mar. 2018, pp. 1–8.

[43] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[44] M. Zeng, A. Yadav, O. Dobre, and H. V. Poor, "Energy-efficient joint user-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5119–5131, Feb. 2019.

[45] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[46] H. Tang, T. Qin, Z. Hui, P. Cheng, and W. Bai, "Design and implementation of a configurable and aperiodic pseudo random number generator in FPGA," in *Proc. IEEE 2nd Int. Conf. Circuits, Syst. Simulation (ICCSS)*, Jul. 2018, pp. 47–51.

[47] R. C. Qiu, C. Zhou, N. Guo, and J. Q. Zhang, "Time reversal with MISO for ultrawideband communications: Experimental results," *IEEE Antennas Wireless Propag. Lett.*, vol. 5, no. 1, pp. 269–273, Dec. 2006.

[48] B. Wang, Y. Wu, F. Han, Y. H. Yang, and K. J. R. Liu, "Green wireless communications: A time-reversal paradigm," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1698–1710, Sep. 2011.

[49] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document TS 36.211 V13.2.0, 3GPP, 2016. [Online]. Available: https://www.portal.3gpp.org

[50] S. Sesia, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.
[51] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[52] U. N. Bhat, *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Basel, Switzerland: Birkhäuser, 2015.
[53] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document ETSI TS 136 211 V11.0.0, 3rd Generation Partnership Project, 2012. [Online]. Available: https://www.portal.3gpp.org
[54] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) System Scenarios*, document ETSI TR 136 942 V13.0.0, 3rd Generation Partnership Project, 2016. [Online]. Available: https://www.portal.3gpp.org
[55] *Definitions of Terms Related to Quality of Service*, document Rec. ITU-T E.800, ITU, Jul. 2009. [Online]. Available: http://www.itu.int/rec/T-REC-E.800-200809-I/en
[56] *Technical Specification Group Radio Access Network; Study on NR Industrial Internet of Things (IoT)*, document TR 38.825 V16.0.0, 3GPP, 2019. [Online]. Available: http://www.portal.3gpp.org
[57] G. L. Nemhauser, *Integer and Combinatorial Optimization*. Hoboken, NJ, USA: Wiley, 1988.
[58] MathWorks. *MATLAB r2021a—Genetic Algorithm*. Accessed: Jul. 4, 2021. [Online]. Available: https://www.mathworks.com/help/gads/genetic-algorithm.html
[59] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proc. IEEE*, vol. 109, no. 3, pp. 204–246, Mar. 2021.
[60] Z. Hou, C. She, Y. Li, T. Q. Quek, and B. Vucetic, "Burstiness-aware bandwidth reservation for ultra-reliable and low-latency communications in tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2401–2410, Nov. 2018.
[61] W. Wu, Y. Li, Y. Zhang, B. Wang, and W. Wang, "Distributed queueing-based random access protocol for LoRa networks," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 763–772, Jan. 2020.
[62] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation*. Princeton, NJ, USA: Princeton Univ. Press, 2009.
[63] S. S. Nam, M.-S. Alouini, and H.-C. Yang, "An MGF-based unified framework to determine the joint statistics of partial sums of ordered random variables," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5655–5672, Nov. 2010.

MOHAMMAD REZA AMINI (Senior Member, IEEE) received the B.Eng. degree in electrical and communication system engineering from the Isfahan University of Technology (IUT), Isfahan, Iran, the M.Sc. degree in electrical and communication system engineering from the Malek-Ashtar University of Technology, Iran, and the Ph.D. degree in telecommunications from the IUT, in 2018. He is currently an Assistant Professor with the Department of Electrical Engineering, Islamic Azad University, Borujerd Branch, Iran. He is also an Inspector with the Iran's Standard Institute and accepted the National Foundation of Elites. His research interests include cognitive radio networks, system implementation, and green and energy-harvesting networks. He was a recipient of the Outstanding Teaching Award and the Outstanding Researcher Award in 2012, 2016, 2017, and 2018.

MOHAMMED W. BAIDAS (Senior Member, IEEE) received the B.Eng. degree (Hons.) in communication systems engineering from The University of Manchester, Manchester, U.K., in 2005, the M.Sc. degree (Hons.) in wireless communications engineering from the University of Leeds, Leeds, U.K., in 2006, the M.S. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2009, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2012. He was a Visiting Researcher with The University of Manchester in the academic years from 2015 to 2016 and from 2018 to 2019. He is currently an Associate Professor with the Department of Electrical Engineering, Kuwait University, Kuwait, where he has been on the faculty, since May 2012. He is a frequent reviewer of several IEEE journals and international journals and conferences, with over 80 publications. His research interests include resource allocation and management in cognitive radio systems, game theory, cooperative communications and networking, and green and energy-harvesting networks. He serves as a technical program committee member for various IEEE and international conferences. He was a recipient of the Outstanding Teaching Award from Kuwait University in the academic year from 2017 to 2018.