Energy-Efficient and Reliable IoT Access without Radio Resource Reservation

Azari, Amin; Stefanovic, Cedomir; Popovski, Petar; Cavdar, Cicek

Published in:
IEEE Transactions on Green Communications and Networking

DOI (link to publication from Publisher):
10.1109/TGCN.2021.3051033

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3051033, IEEE Transactions on Green Communications and Networking

Energy-Efficient and Reliable IoT Access without Radio Resource Reservation

Amin Azari[1], Čedomir Stefanović[2], Petar Popovski[2], Cicek Cavdar[3]
[1]Ericsson AB, [2]Aalborg University, [3]KTH Royal Institute of Technology
Email: cavdar@kth.se, {cs,petarp}@es.aau.dk

***Abstract*—One of the major challenges for Internet-of-Things applications is that the existing cellular technologies do not support the uplink IoT traffic in an energy-efficient manner. There are two principal ways for serving the uplink IoT traffic: grant-based (i.e. scheduled) and grant-free (i.e. random access). Grant-based access provides fine-grained control of reliability and latency at the cost of energy consumption required for signaling. Grant-free access removes the signaling overhead at the cost of looser control of performance in terms of reliability and latency. However, a precise analysis of reliability, latency and energy performance of grant-free access (GFA) is largely missing. This article focuses on a GFA-type protocol, in which a device transmits several packet replicas, asynchronously with respect to the other devices. Using stochastic geometry, we derive closed-form expressions for reliability, delay, and energy consumption, which can be used to identify the tradeoffs among these performance parameters. In order to improve the performance of the protocol, we develop a receiver that leverages the random timing and frequency offsets among the devices in order to facilitate resolution of collisions. This is complemented by a per-device adaptive scheme that controls the number of transmitted replicas. The evaluation confirms the validity of the analysis and the potential of the proposed solution, identifying operating regions in which GFA outperforms the grant-based access.**

***Index Terms*—grant-free access, random access, 5G, beyond 5G, energy-efficiency**

I. Introduction

The Internet-of-Things (IoT) service category, which includes massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC), is one of the major drivers of 5G and beyond 5G standardization. In most of the reporting applications that comprise mMTC, the devices are battery powered, necessitating long battery lifetimes [1]. Thus, in contrast to the standard cellular traffic, the mMTC traffic requires support for low energy consumption. On the other hand, mMTC use-cases also require a certain level of reliability and feature limited latency budget [2], for which, standard cellular access methods may be inadequate [3, 4].

The state-of-the-art solutions for IoT connectivity over cellular networks can be divided into evolutionary and revolutionary ones [5]. The evolutionary solutions aim at enhancing the legacy grant-based access (GBA) of the existing LTE networks [6]. In this respect, the development of LTE for low-cost mMTC has been initiated in release 12 and continued in release 13 with the introduction of LTE category M (LTE-M) and narrow-band IoT (NB-IoT) [7]. In NB-IoT, bandwidth for communications and data transmission rates is decreased to 200 KHz and 200 Kbps, respectively, in order to improve the link budget, and hence, reduce the energy required for data transmissions [1, 8]. However, the access procedures in both LTE-M and NB-IoT are based on the one of LTE-A, involving listening to the control channel and the exchange of signaling required for random-access, synchronization and resource reservation steps [8, 9], thus presenting the same challenges in terms of reliability and latency. Moreover, battery lifetime of IoT devices will be limited [1, 10]. The revolutionary solutions aim at a fundamental revision of the access procedure in order to provide low energy consumption [11, 12], with a potential to lower access delay. It is expected that in future cellular network GFA schemes will coexist with the legacy GBA solutions [13–16].

This paper focuses on GFA in a setup that features time-frequency offsets among accessing devices, which is a realistic assumption for reporting applications with low-complexity terminals driven by cheap oscillators [17, 18]. Specifically, we develop an analytical approach to assess the reliability, delay and energy efficiency performance of GFA in such scenarios, assuming that devices are able to send multiple replicas of their data packets. We then investigate enhancements of the scheme, proposing a receiver that leverages replica combining, complemented by an approach that optimizes the number of transmitted replicas according to the experienced long-term path loss of devices. We show that the proposed scheme has a potential to significantly outperform the legacy GBA solutions in low to moderate load regions.

The rest of the text is organized as follows. This section is concluded with an overview of GFA and the paper contributions. The system model is described in Section II. The analysis is presented in Section III. The replica control scheme is presented in Section IV. The collision resolving receiver is presented in Section V. The simulation results and concluding remarks are given in Section VI and VII, respectively.

A. Background and Contributions

In GBA, IoT devices usually spend a short fraction of active time in the transmission state due to their short data payloads, and a relatively longer fraction of active time in synchronization, access reservation, and other signaling-related states [8, 9, 19]. This implies that the energy consumption in the signaling states can be much higher than in the transmission state [9, 20]. A potential solution to decrease the overall energy consumption is to use GFA [8, 15]. Another potential benefit of GFA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3051033, IEEE Transactions on Green Communications and Networking

is that the success of the access procedure (i.e. its reliability) is not conditioned on the successful exchange of signaling with the base station (BS) [21]. Furthermore, the access delay, i.e. the delay from a packet arrival at the device until the packet transmission is shortened [22]. Finally, GFA is also a preferable solution to be used by extremely low-complexity transmitters [23, 24]. However, since GFA is essentially a non-orthogonal access method, efficient dealing with multi-user interference is the main challenge of GFA scheme design.

B. Grant-Free Access: State of the Art

GFA is implemented in IoT technologies that operate in the unlicensed spectrum, such as SigFox and LoRaWAN [25]; in both technologies, once a packet arrives at the device, it is transmitted without any handshaking, resource reservation, or authentication process. On the other hand, the mobile cellular standards have been traditionally based on the GBA schemes. As part of the efforts shaping the new radio for 5G, use of non-orthogonal multiple access methods, including GFA, has become a hot topic in recent years [26–28]. In particular, a recurrent element in various 6G visions is to move towards zero-cost zero-energy IoT communications [29]. Hence, GFA that offers reduced complexity and energy consumption for end-devices could be a highly relevant candidate for IoT communications in future wireless access networks, along with the legacy GBA schemes.

GFA was investigated in a number of recent works - here we briefly mention just a handful of the ones most relevant to the work presented here [30–33]. In [30], GFA with massive multiple-input multiple-output (MIMO) was investigated and the analytic expressions of success probability for conjugate and zero-forcing beamforming techniques was derived. In [34], a novel distributed GFA scheme was proposed, where a cell is divided into different layers based on the pre-determined inter-layer received power difference, and transmission of each device is adapted to its respective layer. In [32], performance of a massive grant-free network was investigated and approximate expressions for outage probability and throughput were derived. Leveraging deep learning at the receiver for increasing reliability of GFA systems has been investigated in [35]. Machine learning-powered radio resource provisioning for hybrid grant-free/grant-based ultra-reliable low-latency communication (URLLC) networks has been investigated in [36]. In the context of the 3GPP standardization, the set of radio resources that should be allocated to GFA communications, the choice of modulation and coding scheme, and the impact on grant-based communications were investigated in [37, 38]. Unsourced grant-free multiple access was investigated in [39], presenting finite-length bounds on the performance.

C. Challenges of Grant-Free Access

Reliability and Scalability: Reliability of GFA was investigated in [24, 40] by assuming a very dense deployment of access point in which all devices are able to perform power control via channel inversion, a feature that is hardly the case in practical scenarios [41]. In [42], an advanced compressed sensing technique for massive GFA connectivity was proposed, and its performance in reliable detection of simultaneously active devices investigated. Capacity of wireless channels shared by many users with finite block-lengths and without scheduling, was partially investigated in [43]; however, it is still an open problem, especially when it comes to time-frequency overlapping packet transmissions. The reliability and scalability analysis of GFA are also active study items in the 3GPP standardization [37]; the main focus is on assessing the impact of the number of devices that share radio resources dedicated to grant-free communications on reliability.

Impact of Mutual Interference: Since it is inherently uncoordinated, GFA schemes suffer from collisions among devices' transmissions. To counteract the interference from other devices, each device may transmit several replicas of its packets. In [44], it was proposed to prepend transmitted replicas with preambles that can be used to perform collision resolution in case of overlapping transmissions. Moreover, the diversity in replica transmission could be exploited by the receiver-side through: a) decoding of packets by combining their (partially) interference-free replicas; and b) removal of replicas of the decoded packet through interference cancellation, which also enables potential decoding of replicas of other packets. Such successive interference cancellation (SIC)-based receivers for asynchronous ALOHA systems have been investigated in [23, 45]. The solution in [45] exploits timing offsets and replica diversity, where the proposed receiver requires complete knowledge of replicas' positions, and for this purpose, the use of correlation and robust header encoding has been suggested. Similarly, in [46], time-frequency offsets are used for detection of multiple packets.

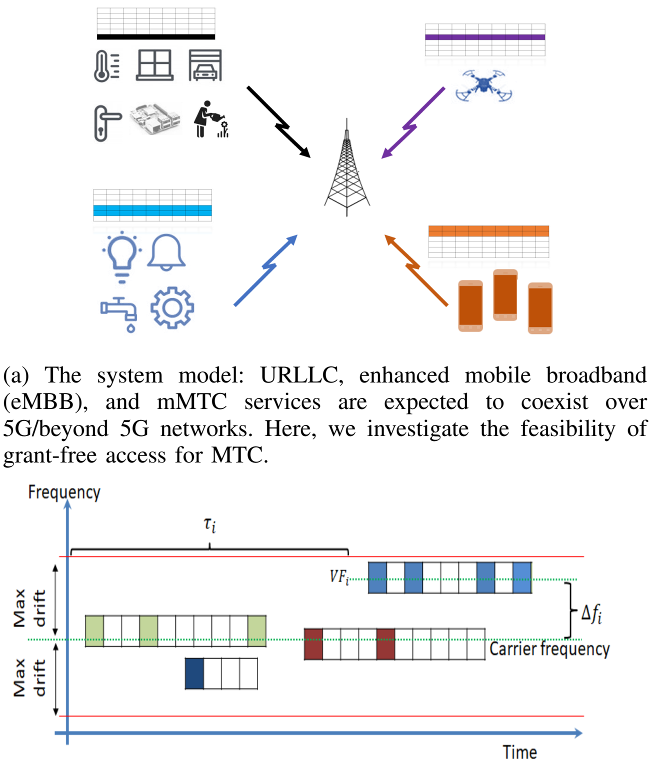
Characterization of Operating Regions: In mobile-cellular standardization, GFA is not considered as a replacement of the legacy GBA approach, but a complementary solution where a load-adaptive access protocol can switch between the grant-free and grant-based modes[40]. In order to optimize the performance of such hybrid access solutions in future networks, the load regions where one access mode outperforms the other should be known. This calls for an investigation of performance evaluation of grant-free and grant-based modes as functions of the traffic-load.

D. Contributions

The aim of this work is to advance state-of-the-art of GFA for IoT communications. Towards this end, we present a novel analytical framework for evaluation of reliability, delay and energy-efficiency performance of GFA in which devices may send multiple replicas of their packets. We then investigate reliability challenges and devise solutions that improve the reliability.

The work presented in the paper builds-up on the conference works presented in [15, 47]. In particular, the initial version of the reliability analysis for GFA, without considering replica transmissions and combining (in contrast to this version), was presented in [47]. Leveraging carrier frequency offset (CFO) for collision decoding was presented in [15], assuming channel-inversion power-control for IoT devices and a constant number of replica transmissions per packet. In this paper, we substantially expand these preliminary results:

- Employing an approach based on stochastic geometry, we derive closed-form statistics for reliability, delay, and energy consumption of the scheme. Moreover, the analytical model from [47] is further extended to incorporate the case in which the number of transmitted replicas is adapted to the long-term path-loss. We validate the derived expressions via simulations and use them to investigate performance tradeoffs in GFA protocol design.
- We develop a receiver that performs decoding of collided replicas by utilizing time-frequency offsets among devices. The proposed receiver is an upgraded version of the one presented in [15], leveraging a novel module for filtering of the detected peaks in the replica recovery process.
- We devise a scheme that optimizes the number of transmitted replicas. The optimization is performed according to the coverage regions in which devices are located, which are created according to the experienced long-term path-loss.
- We compare the proposed scheme with a GBA scheme based on the one of LTE-A and identify operating regions in which GFA shows clear benefits. Specifically, the results indicate that GFA with the proposed optimization scheme and collision decoding can outperform benchmark schemes in energy efficiency and delay at the cost of potential collisions, especially in low-to-medium traffic load regimes.



(a) The system model: URLLC, enhanced mobile broadband (eMBB), and mMTC services are expected to coexist over 5G/beyond 5G networks. Here, we investigate the feasibility of grant-free access for MTC.

(b) Received VFs at the BS with GFA. The replicas with the same color refer to the same packet. The length of packets in bits, the length M and the number N of replicas can be different for different devices.

Fig. 1: Graphical illustration of the system model

II. System Model

We consider a cell with a BS serving a multitude of IoT devices over a shared spectrum of W Hz. Fig. 1a represents such a cell in which, the black-colored time-frequency resources have been allocated to a set of IoT devices for GFA. The devices are battery-powered, of low-complexity, and driven by cheap oscillators; thus, the devices are inherently asynchronous, i.e. their timing and frequency references exhibit offsets around the nominal ones employed by the BS [17]. The devices wake-up intermittently, triggered by packet arrivals. The packet arrival process at each device is modeled as a Poisson process with rate of $1/T_{\text{r}}$, where T_{r} is the reporting period. When a device has a packet to transmit, it assumes a virtual frame (VF) consisting of M slots, each with duration of T_{p} and bandwidth of w, where T_{p} is the time-duration of a packet transmission of the device. Then, one replica of the packet is sent immediately at the first slot of the VF, and the N-1 replicas of the packet are sent in N-1 randomly selected slots out of M-1 remaining slots, as depicted in Fig. 1b. The number of replica transmissions per data packet, N, is upper bounded as $N \leq N_{\max} \leq M$. As per state-of-the-art on low-power IoT devices, see e.g. [48] and the references herein, the transmission power is assumed to be fixed to P_{t} for all devices and all replicas. Furthermore, we assume that there is a feedback link from the BS, and if the device does not receive the ACK after VF transmission, it starts retransmitting after T_{w} backoff time. In case a new data packet arrives to the device after an unsuccessful VF transmission attempt, combination of data from both packets will be sent in one VF. Each packet (i.e. packet replica) includes a preamble of length $\mathbb{N}_{\text{zc}}$, where the preamble is constant for replicas within a VF. The cardinality of the set of available preambles, denoted by $|\mathcal{S}|$, and the length of preambles $\mathbb{N}_{\text{zc}}$ are design parameters. In general, it holds that is $N \in \{1, \cdots, M\}$. If the value of N is chosen completely randomly, or is fixed for all devices, we refer to the scheme as a GFA scheme without replica control. Otherwise, the scheme is called a GFA scheme with replica control, which is discussed in details in Section V. Once N is determined, the choice of N out of M slots of the VF in which replica will be transmitted is random on VF basis. This is similar to the pseudo-noise (PN) code assignment in the code-division multiple access (CDMA).

A quasi-static fading channel model is assumed, which means that the channel gain is constant over a VF of each device. The transmitted signal bandwidth is denoted by w for all devices/packets. As already noted, devices feature carrier-frequency offset (CFO); in ultra-narrowband systems, the maximum level of CFO is $F_{\text{m}} \simeq 0.5(W-w)$ and, by definition, it is expected to be several orders larger than the signal bandwidth w [18, section 3.2.2]. We denote the actual carrier frequency that the i-th transmitter uses for data transmission by f_i, and its offset from the intended carrier frequency f as Δf_i, where $\Delta f_i = f_i - f$. While the devices' CFOs in different wakeup epochs can be assumed to be independent, Δf_i is, in essence, constant during a VF [49]. The timing offsets,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3051033, IEEE Transactions on Green Communications and Networking

CFOs, and the number of transmitted replicas are independent among devices, and the transmissions of different devices are uncoordinated, and hence, overlapping of the packet replicas sent by different devices could happen.

III. Analysis

Assume that the BS is located in the center of a 2D plane and that the devices are distributed in the service area of the cell formed around the BS according to a homogeneous Poisson point process (PPP) with density λ. As noted in Section II, the devices are asynchronous in time and frequency, and hence, the received packets at the BS may overlap in time and frequency domains. A transmitted replica of a packet will be received successfully at the BS if the received signal to interference and noise ratio (SINR) is greater than a threshold value, denoted by γ_{th}. Hence, the performance analysis requires characterizing the received interference at the BS, as investigated in the sequel.

A. Interference Characterization

Assume a reference device located at the point $\mathbf{z}$ in the service area. The received interference from other devices at the BS by can be modeled as a random variable (RV) described via a Laplace functional, which enables investigation of its moments [50]. Towards this end, we introduce a stationary and isotropic process Ψ, which represents the PPP that determines the locations of interfering nodes. For a BS located at the origin, the received power from an interfering device located at point $\mathbf{x}$ is modeled as $P_h(\mathbf{x}) = P_{\text{t}}\, h\, \text{g}(\mathbf{x})$, where P_{t} is the fixed transmission power, h is the power fading coefficient associated with the channel between the device and the BS, and $\text{g}(\mathbf{x})$ represents a distance-dependent path-loss model of the form $G||\mathbf{x}||^{-\delta}$, where δ is the path-loss exponent and G is the system loss coefficient. We further assume that h is a Rayleigh RV whose probability density function (PDF) is given by:

$$\text{p}_h(q) = \mu \exp\left(-\mu q\right). \tag{1}$$

The Laplace functional of the received interference due to the interferer at point $\mathbf{x}$ could be modeled as $L_h(sP_{\text{t}}\text{g}(\mathbf{x}))$. Using Laplace table, $\mathcal{L}_h\big(sP_{\text{t}}\text{g}(\mathbf{x})\big)$ is derived as $\mathcal{L}_h\big(sP_{\text{t}}\text{g}(\mathbf{x})\big) = \frac{1}{1+sP_{\text{t}}\text{g}(\mathbf{x})/\mu}$. Then, the Laplace functional of the aggregated received interference I at the BS is given by [51]:

$$\mathcal{L}_I(s) = \mathbb{E}\big[\exp(-sI)\big] = \mathbb{E}_{\mathbf{x}}\left[\prod\nolimits_{\mathbf{x}\in\Psi} \frac{1}{1+sP_{\text{t}}\text{g}(\mathbf{x})/\mu}\right]. \tag{2}$$

To proceed further, we recall a lemma from [52].

Lemma 1: If $X \subset \mathbb{R}^2$ is assumed to be a PPP with intensity function $\beta(\mathbf{x})$, for any Borel function b: $\mathbb{R}^2 \to [0,1]$, the following holds:

$$\mathbb{E}_{\mathbf{x}}\left[\prod\nolimits_{\mathbf{x}\in X} b(\mathbf{x})\right] = \exp\left(-\int_{\mathbb{R}^2}(1-b(\mathbf{x}))\beta(\mathbf{x})d\mathbf{x}\right). \tag{3}$$

Substituting (2) into (3) produces:

$$\mathcal{L}_I(s) = \exp\left(\int_{\mathbb{R}^2} \frac{-sP_{\text{t}}\text{g}(\mathbf{x})}{\mu + sP_{\text{t}}\text{g}(\mathbf{x})}\upsilon d\mathbf{x}\right) \tag{4}$$
$$= \exp\left(\int_{\mathbb{R}^2} \frac{-\upsilon d\mathbf{x}}{\frac{\mu}{sP_{\text{t}}G||x||^{-\delta}}+1}\right) = \exp\left(\int_0^\infty \frac{-2\pi\upsilon r dr}{\frac{\mu}{sP_{\text{t}}Gr^{-\delta}}+1}\right)$$

where υ denotes density of interfering devices, i.e. a fraction of λ, which are interfering with the transmission of interest. υ could be modeled as $\lambda\frac{NT_{\text{p}}}{T_{\text{r}}}\frac{w}{2F_{\text{m}}}$, in which the term $\frac{NT_{\text{p}}}{T_{\text{r}}}$ represents fraction of time in which device is transmitting and the term $\frac{w}{2F_{\text{m}}}$ represents the fraction of frequency band in which device is transmitting. We proceed by deriving the probability of successful reception of a replica of the packet (i.e., any of its N replicas) in a VF, assuming that the replicas are not combined at the receiver.

B. Reliability Analysis

Denote by θ the additive noise at the receiver. The probability of successful reception of a replica transmitted by the reference device located at point $\mathbf{z}$ is $\text{p}_{\text{s}}(\mathbf{z}) = \Pr\left(\frac{P_{\text{t}}h\text{g}(\mathbf{z})}{\theta+I} \geq \gamma_{\text{th}}\right)$. Using (1), we get:

$$\text{p}_{\text{s}}(\mathbf{z}) = \int_0^\infty \exp\left(-\frac{\gamma_{\text{th}}q\mu}{P_{\text{t}}\text{g}(\mathbf{z})}\right) d\Pr(I+\theta \leq q) = \mathcal{L}_I(s)\mathcal{L}_\theta(s)\Big|_{s=\frac{\gamma_{\text{th}}\mu}{P_{\text{t}}\text{g}(\mathbf{z})}}$$
$$\overset{\text{(a)}}{=} \exp\left(-\int_0^\infty \frac{2\pi\upsilon}{1+\left(\frac{r}{||\mathbf{z}||}\right)^\delta \frac{1}{\gamma_{\text{th}}}} r dr\right) \exp\left(-\frac{\mu\Theta\gamma_{\text{th}}||\mathbf{z}||^\delta}{P_{\text{t}}G}\right)$$
$$= \exp\left(-\upsilon||\mathbf{z}||^2(\gamma_{\text{th}})^{\frac{2}{\delta}}\frac{2\pi^2}{\delta}\csc\left(\frac{2\pi}{\delta}\right)\right) \exp\left(-\frac{\mu\Theta\gamma_{\text{th}}||\mathbf{z}||^\delta}{P_{\text{t}}G}\right) \tag{5}$$

where Θ denotes the power spectral density of noise, and (a) follows from [53, Eq. 32]. Obviously, $\text{p}_{\text{s}}(\mathbf{z}) = \text{p}_{\text{s}}(||\mathbf{z}||)$, i.e. success probability is a function of distance to the BS due to the isotropy of the path-loss function.

The probability of successful reception of the packet in a VF containing its N replicas, assuming independent processing of the replicas at the BS is given by:

$$\text{P}_{\text{s}}(z) \approx 1 - [1 - \text{p}_{\text{s}}(z)]^N \overset{\text{(a)}}{=} \text{P}_{\text{s}}^{\text{lb}}(z). \tag{6}$$

In case that the receiver is able to combine replicas, as it will be discussed in Section V-C, the expression (6) represents a lower bound on the success probability.

C. Delay Analysis

The average experienced delay from packet arrival at a device located at point z to its successful reception at the receiver is modeled as $\mathcal{D}(z) \approx MT_{\text{p}} + T_{\text{us}}$, where the first term indicates the length of VF, and the second term indicates the time spent in unsuccessful transmissions, i.e.,

$$T_{\text{us}} = \sum\nolimits_{i=1}^{B-1} i(MT_{\text{p}} + T_{\text{w}})(1 - \text{P}_{\text{s}}(z))^i \text{P}_{\text{s}}(z) \tag{7}$$

where the $\text{P}_{\text{s}}(z)$ has been described in (6)-(13), and B represents the maximum allowable number of attempts. Naturally, one should design the access protocol such that the (average)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2021.3051033, IEEE Transactions on Green Communications and Networking

experienced delay is shorter than the time-span over which the data embedded in packet is valid. In the following sections, we take this into account and bound the average experienced delay to the maximum tolerated delay.

D. Device Energy Consumption and Lifetime Analysis

Here, we analyze energy consumption of the scheme and derive the expected lifetime of a device. As mentioned in the system model, the packet arrival process at each device is modeled as a Poisson process with rate of T_r, and hence, energy consumption of a device can be seen as a semi-regenerative process where the regeneration points are located at the end of successful packet transmission epochs [6]. Denote the battery capacity of a device at the reference time as E_0. The power consumption of each device in the active, listening and transmission modes are denoted as P_c, P_l, and $\bar{P}_\text{t} = \alpha P_\text{t} + P_\text{c}$, respectively, where P_c is the power consumed by electronic circuits, P_t is the transmission power, and α is the inverse power amplifier (PA) efficiency. In line with definition of the expected battery lifetime for battery-powered devices used in the state-of-the-art [54–56], we define the expected *battery lifetime* at the regeneration point as the ratio between remaining energy and the rate of energy consumption, as follows:

$$L(z) = \frac{E_0}{\frac{1}{T_\text{r}}\left[E_\text{st} + \frac{1}{\text{p}_\text{s}(z)}\left[\bar{P}_\text{t} N T_\text{p} + P_\text{c}[M-N]T_\text{p} + P_l T_\text{ps}\right]\right]}. \tag{8}$$

In this expression, $\Big[E_\text{st} + \frac{1}{\text{p}_\text{s}(z)}\left[\bar{P}_\text{t} N T_\text{p} + P_\text{c}[M-N]T_\text{p} + P_l T_\text{ps}\right]\Big]$ is the average total energy consumption in each reporting period (T_r), T_ps is the listening time for acknowledgment before making another transmission, and E_st is the average static energy consumption, e.g. for data gathering, in each reporting period. Via $L(z)$, one can derive the *network lifetime* based on the definition of interest. For instance, the network lifetime can be computed as $\mathbb{E}_z[L(z)]$, which is the average device lifetime in the network. On the other hand, the network lifetime in some applications is determined by the outage of a certain fraction of the devices in the area due to energy depletion, which can be also calculated via $L(z)$ as indicated in [55].

E. Analysis of Performance Tradeoffs

In the following, we comment on the tradeoffs between the performance parameters. From (8), it can be observed that the device lifetime increases when the success probability $\text{p}_\text{s}(z)$ is increased, which is intuitively clear. Further, (5) indicates that decreasing distance between the device and the BS (i.e. $\mathbf{z}$) increases the success probability, both through the interference-related term (the first component of (5)), and the noise-related term (the second component of (5)). Hence, by increasing the density of BSs and, thus, shortening this distance, reliability of communications increases significantly; however, this also increases the cost of the access network. Eq. (5) also shows that success probability can be increased, and hence experienced delay can be shortened, by increasing the system bandwidth, as the latter decreases υ, i.e. the chance of collision.

Since the increase of success probability plays an important role, in Sections V and VI we elaborate methods to increase this performance parameter through optimization of the access scheme and the enhancement of the reception algorithm, while avoiding costly solutions in terms of increase of the density of the access network or increase of the system bandwidth.

IV. The Replica Control Scheme

Eq. (6) shows that probability of successful packet reception increases with the number of transmitted replicas transmissions. On the other hand, (5) shows that an increase in the number of transmitted packets increases the υ, and hence, decreases the probability of successful reception of a replica. A way to improve the reliability performance, i.e. the probability of successful packet transmission, is to divide the service area in the coverage regions according to the long-term experienced path-loss, and optimize the number of replica transmissions in each region. We denote such scheme as the adaptive replica control (ARC).

Specifically, we propose to define the coverage regions according to certain thresholds on the long-term path loss, which could be mapped to threshold values of the distance from the BS. Intuitively, the devices located further from the BS should transmit more replicas in order to increase the probability of successful packet decoding, but this comes at the expense of generating more interference to the devices that are closer to the BS. The ARC scheme aims at finding the threshold values and the corresponding number of replicas for each region, such that a minimum QoS per region could be guaranteed.

Analysis

Here, we extend the performance analysis from Section III-B to the case in which ARC scheme is in effect, and hence, the number of replicas per packet is varying throughout the service area. We illustrate the approach through a simple, but insightful example with two coverage regions. Consider a 2D deployment of IoT devices in a service area of inner and outer radius of R_0 and R_c, respectively, with a single distance threshold value, i.e. d_th. In this case, the proposed ARC scheme can be denoted as $\text{ARC}(N_1, N_2, d_\text{th})$, where N_1 denotes the number of replicas sent per packet by devices located closer than d_th, and N_2 denotes the number of replicas sent by the devices that are located further than d_th. Consequently, the expression (5) can be adapted as follows:

$$\begin{aligned}\text{p}_\text{s}(z) = & \exp\left(\int_{R_0}^{d_\text{th}} \frac{-2\pi\upsilon_1}{1+(\frac{r}{z})^\delta \frac{1}{\gamma_\text{th}}} r dr\right) \\ & \exp\left(\int_{d_\text{th}}^{R_\text{c}} \frac{-2\pi\upsilon_2}{1+(\frac{r}{z})^\delta \frac{1}{\gamma_\text{th}}} r dr\right) \exp\left(-\frac{\mu\Theta\gamma_\text{th} z^\delta}{P_\text{t} G}\right)\end{aligned} \tag{9}$$

For $\delta = 4$, we have $\text{p}_\text{s}(z) = \exp\big(-\upsilon_1[H(d_\text{th}) - H(R_0)] - \upsilon_2[H(R_\text{c}) - H(d_\text{th})]\big)\exp\left(-\frac{\mu\Theta\gamma_\text{th} z^4}{P_\text{t} G}\right)$. In (9), $\upsilon_1 = \upsilon|_{N=N_1}$, $\upsilon_2 = \upsilon|_{N=N_2}$ (see Section III-B for derivation of υ as a function of N), and $H(x) = \pi\,\text{atan}(\frac{1}{\sqrt{\gamma_\text{th}}}\frac{x^2}{z^2}) / \frac{1}{\sqrt{\gamma_\text{th}} z^2}$. An extension of (9) to a general ARC scheme with multiple distance threshold values is straightforward. Moreover, following the procedure presented in Sections III-C and III-D, one can derive expressions for delay and device lifetime via (9). The expression in (9) could be also used to derive optimized ARC parameters. For example, in the following, we formulate an optimization problem aiming at deriving optimized ARC parameters for minimizing a weighted sum outage probability across the network:

$$\min_{\{N_1,N_2,d_{\text{th}}\}} \omega \int_{R_0}^{d_{\text{th}}} \left[1\text{-}\mathrm{p}_{\mathrm{s}}(z)\right]^{N_1} dz + (1\text{-}\omega) \int_{d_{\text{th}}}^{R_{\text{c}}} \left[1\text{-}\mathrm{p}_{\mathrm{s}}(z)\right]^{N_2} dz$$
$$\text{subject to: } N_1 \leq N_2 \leq N_{\max} \in \mathbb{N},\ R_0 \leq d_{\text{th}} \leq R_{\text{c}} \qquad (10)$$

where ω and $(1-\omega)$ represent the weights associated to reliability of devices sending N_1 and N_2 replicas, respectively. In general, the problem in (10) is not convex due to non-convexity of, e.g. $\mathrm{p}_{\mathrm{s}}(z)$, but could be solved by search over feasible values of N_1, N_2, and d_{th} as described in Algorithm 1.

If we assume that $\gamma_{\text{th}} \gg 1$ and neglect the impact of noise in comparison with the impact of interference, $H(x)$ could be approximated by πx^2, and hence,

$$\mathrm{p}_{\mathrm{s}}(z) = \exp(\pi(\upsilon_2\text{-}\upsilon_1){d_{\text{th}}}^2 + \pi\upsilon_1 R_0^2 = \pi\upsilon_2 R_{\text{c}}^2) = a\exp(b{d_{\text{th}}}^2) \qquad (11)$$

where $a = \exp(\pi\upsilon_1 R_0^2 - \pi\upsilon_2 R_{\text{c}}^2)$ and $b = \pi(\upsilon_2 - \upsilon_1)$. Then,

$$\begin{aligned} d_{\text{th}}^{(i,j)} &= \arg\min_{\{d_{\text{th}}\}} \omega(d_{\text{th}}\text{-}R_0)\left[1\text{-}\mathrm{p}_{\mathrm{s}}(z)\right]^i + (1\text{-}\omega)(R_{\text{c}}\text{-}d_{\text{th}})\left[1\text{-}\mathrm{p}_{\mathrm{s}}(z)\right]^j, \\ &= \arg\min_{\{d_{\text{th}}\}} \omega(d_{\text{th}}\text{-}R_0)\left[1\text{-}a\exp(b{d_{\text{th}}}^2)\right]^i + (1\text{-}\omega)(R_{\text{c}}\text{-}d_{\text{th}}) \times \\ &\quad \left[1\text{-}a\exp(b{d_{\text{th}}}^2)\right]^j, \end{aligned} \qquad (12)$$

where the objective is a quasi-convex function of d_{th}, which implies that problem (10) can be solved using standard optimization toolboxes [57]. For other values of γ_{th}, the integral in step 3 of Algorithm 1 could be solved numerically, e.g. using Romberg method [58].

Algorithm 1 A Method for Solving Problem (10)

```
1: Given $N_{\max}$
2: for $i \in \{1, \cdots, N_{\max}\}$ and $j \in \{1, \cdots, N_{\max}\}$ and $i \leq j$ do
3:     $d_{\text{th}}^{(i,j)} = \arg\min_{\{d_{\text{th}}\}} \omega \int_{R_0}^{d_{\text{th}}} \left[1\text{-}\mathrm{p}_{\mathrm{s}}(z)\right]^i dz + (1-\omega) \int_{d_{\text{th}}}^{R_{\text{c}}} \left[1\text{-}\mathrm{p}_{\mathrm{s}}(z)\right]^j dz$
4: end for
5: Return $\{N_1, N_2\} = \arg\min_{i,j} d_{\text{th}}^{(i,j)}$
6: Return $d_{\text{th}} = \min_{i,j} d_{\text{th}}^{(i,j)}$
```

V. The Collision Decoding Scheme

The received (composite) signal at the BS $X(t)$ is sampled at rate F_{s},[1] searching for the (potentially collided) replicas, see Fig. 2a. Once the presence of energy (potential signal) is detected, to which we refer to as an event, the receiver processes samples organized in a time frame with a length of T_{f}, where we denote by $X_i(n)$ the array of samples corresponding to the ith time frame. The samples in the time frame are processed using a periodogram module, which aims at finding periodic components in the signal, and returns the detected carrier frequencies, as depicted in Fig. 2b. Denote the number of detected carrier frequencies, which are determined by CFOs of the contending devices, as K (see Fig. 2a). Then, the samples from the time frame are demodulated and correlated with the set of preambles K times (in Fig. 2b, for the sake of simplicity, we present the correlation for one preamble). Each of the correlations may return peaks. Denote by $Y_j(n)$ the output of correlation of the preamble with $X_{i,j}$, where $X_{i,j}$ is the demodulated version of $X_i(n)$ by $f + \Delta f_j$. In $Y_j(n)$, the respective peak of j-th CFO is located in the correct timing offset (i.e. at the starting point of the packet replica)[2], while the respective peaks of Y_l, $l \neq j$ are shifted due to the mismatch between f_j and f_l, as depicted in Fig. 3a.[3] The task of peak-detection module in Fig. 2a is to report the set of detected peaks. The task of decision-making module is to make a decision on K time offsets, respective to the K CFOs, based on the set of detected peaks. Then, the respective demodulated sequence of each carrier frequency, e.g. $X_{i,j}(n)$ for f_i, is truncated from $\tau_{i,j}$ to the length of a packet (i.e. a packet replica) and is fed to the SIC module along with its time-frequency offset. In other words, $(Z_{i,j}(n), f_{i,j}, \tau_{i,j})$ tuple for $j \in \{1, \cdots, K\}$ is fed to the SIC module, as depicted in Fig. 2a.

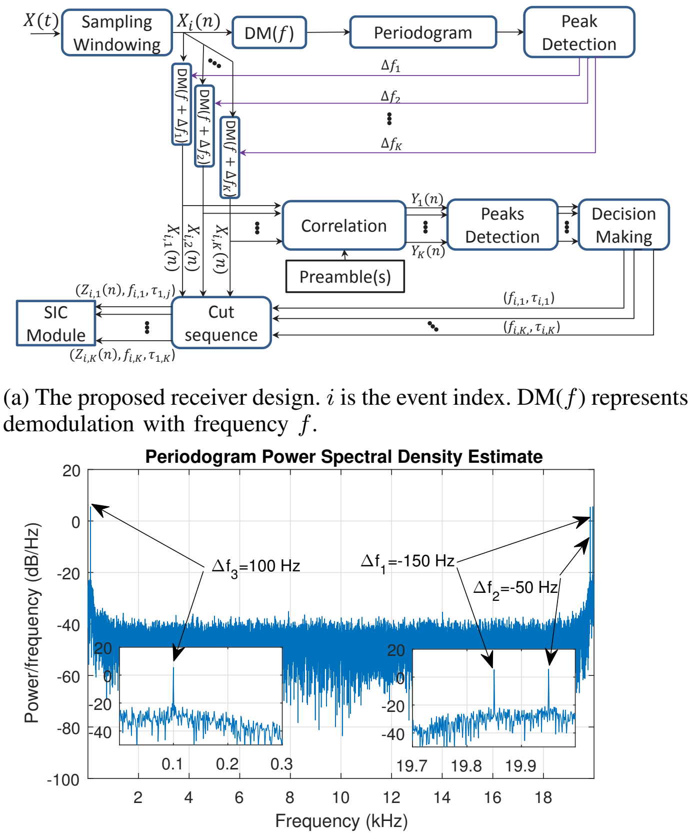
A. The SIC Module

The SIC module continuously receives and saves demodulated sequences related to processed events, including their respective carrier frequencies and time offsets. Then, it tries to decode each sequence. If the sequence, i.e. the supposedly contained packet replica, is decoded successfully, the locations of the other replicas become known, and hence, these are removed from the contributing collision events. If the packet replica cannot be decoded correctly, the SIC module tries to find replicas related to the same packet by performing a search in the set of previously processed and stored events containing the same CFO. If the other replicas are also in collision, the SIC module attempts to combine them, as follows. Using the previously obtained CFOs and time offsets, the SIC module reconstructs the time/frequency map of collisions. Then the module assesses the level of interference that each replica is suffering from, and uses a combining scheme, e.g. selection combining (SC) or equal gain combining (EGC), in order to improve the performance. EGC combines replicas with equal gain, while SC merges interference-free parts of the received replicas in order to construct the original packet. After combining the replicas that belong to a single packet, the SIC module attempts decoding. If the decoding succeeds, the receiver removes all replicas of the decoded packet, thus

[1] The choice of F_{s} incurs a system performance tradeoff, since a higher sampling rate increases the collision decoding capability, as it will be shown in this section, but also the the receiver cost.

[2] The output of the correlation with the preamble is expected to produce a peak located at the starting point of the packet replica in the received signal. However, some lower peaks are also observed due to existence of other CFOs.

[3] The example in Fig. 3a is discussed in more details in Section V-B.



(a) The proposed receiver design. i is the event index. DM(f) represents demodulation with frequency f.

(b) Leveraging CFO for collision decoding. This figure represents output of periodigram when 3 frequency components remain after demodulation.

Fig. 2: The proposed receiver design and motivation for leveraging CFO for collision decoding.

lowering the interference level for other transmissions, and enabling potential decoding of other packets. If decoding fails, the demodulated sequence is stored for further processing, while the receiver triggers decoding of newly arrived events. In case that the subsequent decoding successes lower the level of interference in the previously stored events, new decoding attempts could be made. The length of the period during which a stored event can be reprocessed depends on the time-validity of the contained data, and can differ from one IoT application to another.

B. Processing of Detected Peaks

The shifts of peak locations when correlating preambles with the demodulated signals are inevitable in case of CFO. The result of correlation of a preamble sequence $P(n), n \in \{0, \cdots, \mathbb{N}_{\text{zc}}\text{-}1\}$, with itself, is a sequence of length $2\mathbb{N}_{\text{zc}}-1$, i.e. $P(m), m \in \{1, \cdots, 2\mathbb{N}_{\text{zc}}-1\}$, with a peak at $\mathbb{N}_{\text{zc}}+1$. Now, assume $P(n)$ has been transmitted with carrier frequency f, and has been demodulated with $f+\Delta f$, i.e. the CFO is Δf. Denote by $\tilde{P}(n)$ the demodulated signal, i.e. $\tilde{P}(n) = P(n)e^{j2\pi\Delta f n T_{\text{b}}}$, $\forall n \in \{0, \mathbb{N}_{\text{zc}}-1\}$, where T_{b} is the bit duration. When correlating $P(n)$ with $\tilde{P}(n)$, the peak location moves periodically between $-\lfloor \mathbb{N}_{\text{zc}}/2 \rfloor$ and $\lfloor \mathbb{N}_{\text{zc}}/2 \rfloor$, as discussed in [59]. Given T_{b} and $\mathbb{N}_{\text{zc}}$ as characteristics of the system, the level of shift in the position of peak, denoted by Q, can be derived as a function of Δf, as depicted in Fig. 3.

The decision-making module (shown in Fig. 2a), decides which subset of detected peaks represents the correct time offsets of the replicas. In this module, a successive peak detection (SPD) function is used, explained in the following. Denote the set of received peaks from the peak detection module as $\{\mathcal{T}_1, \cdots, \mathcal{T}_K\}$, where $\mathcal{T}_j$ is the subset of detected peaks in $Y_j(n)$, and $Y_j(n)$ is the result of correlation of $X_j(n)$ with the preamble, as in Fig. 2a. The SPD function searches over subsets $\mathcal{T}_j$, $j \in \{1, \cdots, K\}$ and makes a map of peaks and their shifted positions. For example, given p_j as a candidate peak position[4] in $\mathcal{T}_j$, SPD checks $\mathcal{T}_{j+k}$ to see whether it contains a peak at $p_j - Q\big(\Delta f_{j+k} - \Delta f_j\big)$, $\forall k \in \{1, \cdots, K\} \setminus j$, or not. If p_j's repetitions can be found in $\mathcal{T}_k$, $kk \in \{1, \cdots, K\} \setminus j$, then p_j is validated, else it is removed from $\mathcal{T}_j$.

An example of SPD procedure is illustrated in Fig. 3a for $K = 3$, where the set of CFOs is $\{-150, -50, 100\}$ Hz, and the set of time-index offsets from a reference time is $\{1000, 1100, 1160\}$ indices. In this example, $\mathcal{T}_1 = \{1000, 1180, 1200, 1380, 1640\}$ and $\mathcal{T}_2 = \{920, 1100, 1300\}$. In $\mathcal{T}_2$, 1100 represents the correct starting time of signal with $\Delta f_2 = -50$ Hz; however, 920 and 1380 represent the shifted positions of peaks in $Y_1(n)$, which has been shifted -80 indices. Referring to Fig. 3, one sees that the expected shift in position of peak is $Q(\Delta f_2 - \Delta f_1) = Q(100)$, which is equal to $0.088\,\mathbb{N}_{\text{zc}}\,F_{\text{s}}\,T_{\text{b}} = 80$ indices. Furthermore, the peaks in $Y_3(n)$ are the shifted versions of peaks in $Y_2(n)$, where the level of shift is $Q(150) = 140$. In order to get rid of side-peak processing, one may increase the length of preambles, i.e. $\mathbb{N}_{\text{zc}}$, which eliminates the side peaks at the cost of longer packet transmissions [60]. On the other hand, increase in $\mathbb{N}_{\text{zc}}$ may increase energy consumption of devices in transmission of longer preambles. Hence, $\mathbb{N}_{\text{zc}}$ regulates a tradeoff between energy consumption of transmitting devices and complexity of receiver. Investigation of this tradeoff is beyond the scope of this work, and the interested reader can refer to [61].

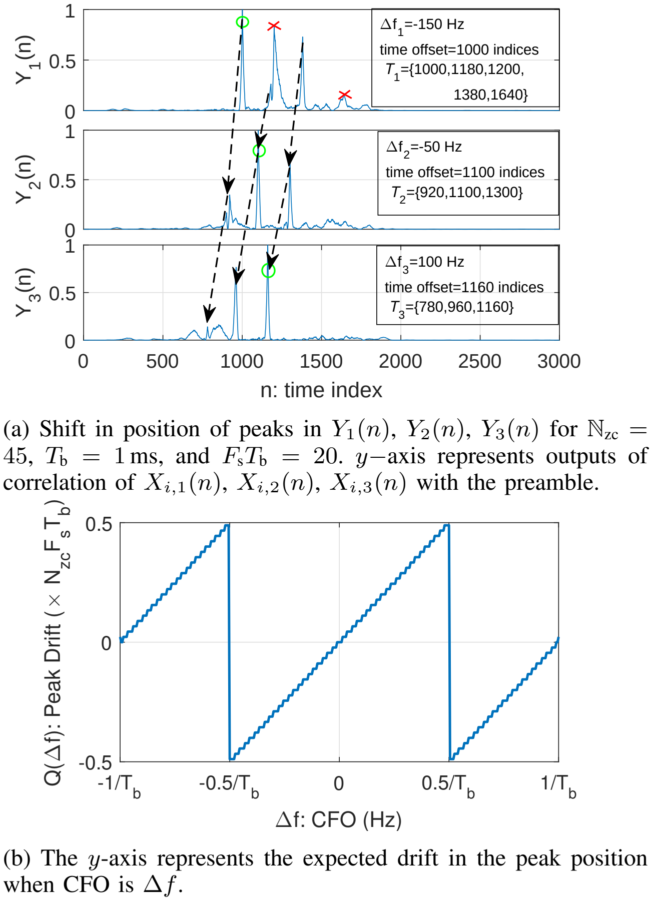
C. Analysis of the Impact on the Performance

When deriving the expression in (6), i.e. success probability in GFA, we assumed independent processing of replicas, and, thus, (6) can be assumed to be a lower bound on the probability of successful packet reception with respect to the case in which the receiver that performs replica combining is used. In the proposed collision decoding scheme, the probability of successful packet reception could be upper bounded as:

$$\mathrm{P}_{\text{s}}^{\text{ub}}(z) \approx \mathrm{p}_{\text{s}}(z)\big|_{\gamma_{\text{th}} \leftarrow \frac{\gamma_{\text{th}}}{N}}. \tag{13}$$

In Section VI, we show that the achieved success probability when selection combining is used indeed falls in the interval $[\mathrm{P}_{\text{s}}^{\text{lb}}(z), \mathrm{P}_{\text{s}}^{\text{ub}}(z)]$.

[4] p_j represents time offset of a peak w.r.t. the reference time in processing of the respective event.



(a) Shift in position of peaks in $Y_1(n)$, $Y_2(n)$, $Y_3(n)$ for $\mathbb{N}_{\text{zc}} = 45$, $T_{\text{b}} = 1\,\text{ms}$, and $F_{\text{s}}T_{\text{b}} = 20$. $y-$axis represents outputs of correlation of $X_{i,1}(n)$, $X_{i,2}(n)$, $X_{i,3}(n)$ with the preamble.

(b) The y-axis represents the expected drift in the peak position when CFO is Δf.

Fig. 3: Impact of CFO on the position of peaks after correlation with the preamble.

TABLE I: Simulation Parameters

Parameters	*Default value*
Cell out/inner radius [m]	3000 / 50
Number of devices	20000
Path-loss	$133 + 38.3\log(\frac{d}{1000})$
Thermal noise power [dBm/Hz]	-174
$W; w; F_{\text{m}}; F_{\text{s}}$ [Hz]	600; 200; 200; 4000
ξ	1
P_{c}; P_l; αP_{t} [mW]	40; 80; 139
$T_{\text{b}}; T_{\text{p}}; T_{\text{ps}}; T_{\text{w}}$ [s]	0.01; 0.5; 1; 2
Threshold SINR (γ_{th})	1
$D; D_{\text{oh}}$ [bit]	100; 50
$\mathcal{D}_{\text{syn-ti}}; \mathcal{D}_{\text{syn-fr}}$ [s]	2; 1
$\mathcal{E}_{\text{syn-ti}}; \mathcal{E}_{\text{syn-fr}}; E_{\text{st}}$ [J]	$2P_l; P_l; 5P_{\text{c}}$
E_0 [J]	5000
$\mathbb{N}_{\text{zc}}; \lvert\mathcal{S}\rvert$	23; 1
Grant-based: $\lvert\mathcal{S}\rvert_{\text{RA}}$	20
Grant-based: $T_{\text{RA}}; T_{\text{Res}}$ [s]	2; 1

VI. Performance Evaluation

The system model implemented in this section assumes 20000 IoT devices distributed according to a spatial PPP in a service area with a BS in its center [62, 63]. The simulation parameters are chosen according to the following principles: (i) The bandwidth of each narrow-band signal is a fraction of shared bandwidth for communications, (ii) CFO could be comparable to the bandwidth of the narrow-band signal [17, 18], as discussed in Section I, (iii) The traffic load, expressed in terms of time/frequency occupation, changes from low to high in different simulations, (iv) The transmit power is higher than the listening power, and listening power is higher than the power in the idle/sleep mode. An example set of values is given by Table I. In this table, $\mathcal{D}_{\text{syn-ti}}$, $\mathcal{D}_{\text{syn-fr}}$, $E_{\text{syn-ti}}$, and $\mathcal{E}_{\text{syn-fr}}$ represent delay and energy spent in time and frequency synchronization, respectively. In our simulations, the number of replica transmissions N is a design parameter and the goal is to find its optimal value. Furthermore, the length of VF, i.e. M, is considered to be twice the number of replica transmissions (joint optimization of M and N is omitted due to the constrains on manuscript length). In order to compare GFA to GBA, we have implemented a benchmark grant-based scheme that closely follows the one of LTE-A, such that: (i) Devices become synchronized by receiving system information over the broadcast channel (BCH); (ii) Devices contend over the random-access channel (RACH) for resource reservation from a pool of $\lvert\mathcal{S}\rvert_{\text{RA}}$ orthogonal preambles; (iii) Responses on detected preambles are sent within T_{Res} seconds over the downlink control channel (DCCH); (iv) The scheduled devices transmit their packets in the dedicated resources in the uplink shared channel (USCH); (v) The uplink and downlink channels are separated in the frequency domain; and (vi) The uplink RA and data channels are multiplexed in time.[5]

A. Validation of the Analytical Results

Fig. 4a compares the analytical expressions derived in (6) and (13) with the simulations results, for the aggregated arrival rate of new packets in the system (i.e. without considering replicas' transmissions) of $\Lambda = 4.5$ and $\gamma_{\text{th}} = 3$. For $N = 1$ replicas per packet (i.e. when no replica combining is possible), the analytical expression for probability of successful packet reception P_{s} in (6) is tightly matching the simulation results. For $N = 2$, $\text{P}_{\text{s}}(z)$ obtained by simulations is indeed between the lower and upper bounds derived in (6) and (13), respectively. Further, increasing number of replica transmissions from $N = 1$ to $N = 2$ increases $\text{P}_{\text{s}}(z)$ up to a certain distance from the BS, after which $\text{P}_{\text{s}}(z)$ decreases. This is due to the fact that increase in N increases the traffic load; the adverse effect of an increased traffic load on the performance of remote devices that experience a higher level of path-loss is shown in (5).

B. Performance Evaluation of the ARC Scheme

The ARC scheme, proposed in Section IV aims at balancing between the reward from replica transmission in providing diversity and the regret due to collisions, for enhancing the overall reliability of communications in the network, as shown in (10). In this subsection, we compare performance of the

[5] The simulator has been implemented in Matlab, and is available in the repository at github.com/AminAzari/Grant-free-IoT.

GFA with ARC against the benchmarks, which are GBA and basic GFA scheme (i.e. without replica control).

Fig. 4b represents the average energy efficiency [bit/J] in the network[6] as a function of the aggregated packet arrival rate Λ. The depicted curves have a concave shape. It can be observed that the bit-per-joule metric increases by increasing the arrival rate of data packets up to a saturation point beyond which the energy consumption in collisions and retransmissions degrades the performance. Obviously, the GFA schemes are more energy efficient than GBA at this traffic load. The decrease in energy efficiency of basic GFA with the increase in the number of replicas N is a consequence of the fact depicted in Fig. 4a that probability of successful packet reception for devices located beyond 1200 m from the BS, which comprises 83% of devices, is decreased when N increases. The figure also shows that how the value of d_{th} in ARC(1,2, d_{th}) affects the energy efficiency, where it can be observed that that ARC(1,2,2500) has the superior performance among the depicted curves at this traffic load.

Fig. 4c depicts the evaluation of the average service delay as a function of the aggregate traffic load. When Λ is low, waiting for random access opportunities in the GBA imposes a high access delay in comparison with the GFA schemes. As Λ increases, GBA start to perform better then GFA schemes; this is due to the increased number of collisions and consequent backoffs that affect the GFA schemes more. Nevertheless, ARC(1,2,2500) scheme consistently outperforms GBA, e.g. it features 40%-78% decrease in the average delay, in the medium to low traffic-load regimes respectively.

Fig. 4d investigates the device lifetime performance. As it could be expected from the results presented in Fig. 4b, this figure shows that the device lifetime is significantly improved by leveraging the proposed scheme, in particular when Λ is low. There is a 49%-92% improvement in the battery lifetime of IoT devices could be achieved in the medium to low traffic-load regimes by leveraging the ARC scheme.

Finally, comparison of Fig. 4a and Fig. 4d shows that for IoT devices located close to the BS, with a high probability of success in transmission, sending a single replica per data packet and listening for ACK outperforms sending multiple replicas per packet and waiting for their ACKs. On the other hand, for IoT devices located far from the BS, with a low probability of success in transmission, sending multiple replicas per packet and listening for a potential ACK after all of them is more energy efficient compared to listening for ACK after each replica transmission.

We now turn to investigation of the impact of optimization of d_{th} and N_2 on the ARC(N_1,N_2,d_{th}). Fig. 5a shows how the device lifetime changes with GFA with ARC (i.e. the distance threshold), where the benchmark depicts the basic GFA scheme. Obviously, significant gains could be achieved by tuning the value of d_{th}, i.e. by partitioning devices into classes with optimal sizes. Also, the performance relative to the benchmark increases with the increase in the aggregate traffic load, e.g. from 11% for $\Lambda = 1.7$ to 32% for $\Lambda = 4.9$. This observation confirms the potential of GFA with ARC.

Fig. 5b shows how the device lifetime in GFA with ARC changes with N_2. Obviously, the scheme with $N_2 = 2$ achieves the highest battery lifetime for both values of Λ for the given d_{th}. Fig. 5c shows probability of successful packet reception $\mathrm{P}_{\mathbf{s}}(z)$ as a function of z, i.e. the communication distance from the BS. While the curves related to the basic GFA are monotonously decreasing, the curves related to the GFA with ARC are not. Specifically, $\mathrm{P}_{\mathbf{s}}(z)$ for GFA with ARC abruptly increases at the respective distance threshold d_{th} of the scheme, i.e. at 1000 meters for ARC(1,2,1000), which is due to an additional replica transmissions. It is clear that the ARC(1,2,2500) can significantly increase the minimum experienced success probability in the cell, i.e. 100-200% improvement for regions beyond d_{th}, which constitutes 30% of the service area.

Finally, Fig. 5d illustrates the impact of selection combining of the replicas, as described in Section V-A. For ARC(1,2,2500) scheme, around 8% of the average probability of successful packet reception is achieved by selection combining. In this scheme, 30% of devices transmit 2 replicas per data packet (i.e., 30% of devices are located beyond the d_{th}). Then, replica combining improves the average probability of successful packet reception by 27% for the devices located beyond d_{th}.

VII. Conclusions

In this work, we investigated grant-based and grant-free radio access schemes for uplink-oriented IoT communications. Closed-form expressions of the key performance indicators have been derived for GFA, which help to identify the regions of the traffic load in which grant based/free access performs favorably. The derived results have been subsequently employed in identifying the switchover traffic load values up to which, GFA outperforms the other solutions. A case study has been also introduced to investigate the potential of the proposed solutions. The performance evaluation results show significant advantages of the approach based on GFA both in terms of battery lifetime and latency in low to medium traffic load regimes. The results further indicate that the GFA augmented by replica control can outperform the benchmark schemes in a wide range of traffic loads. These promising results promote integration of the GFA in future cellular networks, along with the legacy grant-based schemes used in the existing LTE and NB-IoT systems.

References

[1] J. Xu, J. Yao, L. Wang, Z. Ming, K. Wu, and L. Chen, “Narrowband internet of things: Evolutions, technologies, and open issues,” *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1449–1462, 2018.

[2] J. Lorca, B. Solana, R. Barco *et al.*, “Deliverable d2. 1: Scenarios, KPIs, use cases and baseline system evaluation,” *E2E-aware Optim. Advancements Netw. Edge 5G New Radio (ONE5G), Tech. Rep. D*, vol. 2, 2017.

[3] A. Laya, L. Alonso, and J. Alonso-Zarate, “Is the random access channel of LTE and LTE-A suitable for M2M communications? a survey of alternatives,” *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 4–16, 2014.

[6] The ratio between number of transmitted data bits and the energy consumed for all communications, including signaling and replica transmissions.

[4] X. Liu, X. Zhai, W. Lu, and C. Wu, "QoS-guarantee resource allocation for multibeam satellite industrial internet of things with NOMA," *IEEE Transactions on Industrial Informatics*, 2019.

[5] Ericsson *et al.*, "A choice of future M2M access technologies for mobile network operators," Tech. Rep., 2014.

[6] A. Azari *et al.*, "Lifetime-aware scheduling and power control for M2M communications in LTE," in *IEEE VTC*, 2015.

[7] 3GPP TS 45.820, "Cellular system support for ultra-low complexity and low throughput internet of things (CIoT)," Tech. Rep., 2015, (Rel. 13).

[8] P. Andres-Maldonado, M. Lauridsen, P. Ameigeiras, and J. M. Lopez-Soler, "Analytical modeling and experimental validation of NB-IoT device energy consumption," *IEEE Internet of Things Journal*, 2019.

[9] P. A. Maldonado *et al.*, "Narrowband IoT data transmission procedures for massive machine-type communications," *IEEE Network*, vol. 31, no. 6, pp. 8–15, Nov 2017.

[10] A. Azari, C. Stefanovic, P. Popovski, and C. Cavdar, "On the latency-energy performance of NB-IoT systems in providing wide-area IoT connectivity," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 1, pp. 57–68, 2020.

[11] É. Morin, M. Maman, R. Guizzetti, and A. Duda, "Comparison of the device lifetime in wireless networks for the internet of things," *IEEE Access*, vol. 5, pp. 7097–7114, April 2017.

[12] C. Bockelmann *et al.*, "Towards Massive Connectivity Support for Scalable mMTC Communications in 5G networks," *IEEE Access*, May 2018.

[13] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sept. 2018.

[14] L. Liu *et al.*, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 88–99, Sept 2018.

[15] A. Azari *et al.*, "Grant-Free Radio Access for Short-Packet Communications over 5G Networks," in *IEEE Globecom*, 2017.

[16] N. El Rachkidy, A. Guitton, and M. Kaneko, "Collision resolution protocol for delay and energy efficient LoRa networks," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 2, pp. 535–551, 2019.

[17] N. M. Balasubramanya *et al.*, "Low SNR uplink CFO estimation for energy efficient IoT using LTE," *IEEE Access*, vol. 4, pp. 3936–3950, 2016.

[18] M. T. Do, "Ultra-narrowband wireless sensor networks modeling and optimization," Ph.D. dissertation, Lyon, INSA, 2015.

[19] B. Al Homssi, A. Al-Hourani, S. Chandrasekharan, K. M. Gomez, and S. Kandeepan, "On the bound of energy consumption in cellular IoT networks," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 2, pp. 355–364, 2020.

[20] M. Lauridsen, "Studies on mobile terminal energy consumption for LTE, 5G," Ph.D. dissertation, Aalborg University, 2015.

[21] P. Popovski *et al.*, "Ultra-Reliable Low-Latency Communication (URLLC): Principles and Building Blocks," *arXiv preprint arXiv:1708.07862*, 2017.

[22] R. Karaki *et al.*, "Performance of autonomous uplink transmissions in unlicensed spectrum," in *IEEE Globecom*, 2017.

[23] R. D. Gaudenzi *et al.*, "Asynchronous contention resolution diversity ALOHA: Making CRDSA truly asynchronous," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6193–6206, Nov 2014.

[24] Z. Li, S. Zozor, J. M. Drossier, N. Varsier, and Q. Lampin, "2D time-frequency interference modelling using stochastic geometry for performance evaluation in Low-Power Wide-Area Networks," in *IEEE ICC*, 2017.

[25] W. Yang *et al.*, "Narrowband wireless access for low-power massive internet of things: A bandwidth perspective," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 138–145, May 2017.

[26] T. Lv, Y. Ma, J. Zeng, and P. T. Mathiopoulos, "Millimeter-wave NOMA transmission in cellular M2M communications for internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1989–2000, 2018.

[27] H. Tabassum *et al.*, "Non-orthogonal multiple access (NOMA) in cellular uplink and downlink: Challenges and enabling techniques," *arXiv preprint arXiv:1608.05783*, 2016.

[28] Z. Ding *et al.*, "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, July 2017.

[29] O. L. A. López, H. Alves, R. D. Souza, S. Montejo-Sánchez, E. M. G. Fernández, and M. Latva-aho, "Massive wireless energy transfer: Enabling sustainable IoT towards 6G era," *arXiv preprint arXiv:1912.05322*, 2019.

[30] J. Ding, D. Qu, H. Jiang, and T. Jiang, "Success probability of grant-free random access with massive MIMO," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 506–516, 2019.

[31] X. Jiang *et al.*, "Low-latency networking: Where latency lurks and how to tame it," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 280–306, Feb 2019.

[32] R. Abbas *et al.*, "A novel analytical framework for massive grant-free noma," *IEEE Trans. on Commun.*, 2019.

[33] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free noma towards reliable low-latency access in tactile internet of things," *IEEE Transactions on Industrial Informatics (Early Access)*, 2019.

[34] H. Jiang *et al.*, "Distributed layered grant-free non-orthogonal multiple access for massive MTC," in *IEEE PIMRC*, 2018.

[35] N. Ye *et al.*, "Deep learning aided grant-free NOMA towards reliable low-latency access in tactile internet of things," *IEEE Transactions on Industrial Informatics*, 2019.

[36] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 42–48, March 2019.

[37] R1-1808304, "Discussion on the reliability enhancement for grant-free transmission," Tech. Rep., Aug. 2018, 3GPP TSG-RAN1 Meeting 94, Gothenburg, Sweden.

[38] R1-163510 , "Candidate NR multiple access schemes," Tech. Rep., April 2016, 3GPP TSG RAN WG1 Meeting 84, Korea.

[39] Y. Polyanskiy, "A perspective on massive random-access," in *IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[40] M. Gharbieh *et al.*, "Spatiotemporal model for uplink IoT traffic: Scheduling and random access paradox," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8357–8372, 2018.

[41] M. Suciu *et al.*, "Discrete transmit power devices in dense wireless networks: methodology and case study," *IEEE Access*, vol. 5, pp. 1762–1778, 2017.

[42] P. Popovski *et al.*, "Ultra-Reliable Low-Latency Communication (URLLC): Principles and Building Blocks," *arXiv preprint arXiv:1708.07862*, 2017.

[43] X. Chen, T. Chen, and D. Guo, "Capacity of gaussian many-access channels," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, Feb 2017.

[44] S. Kim *et al.*, "Novel transceiver architecture for an asynchronous grant-free IDMA system," *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4491–4504, 2019.

[45] F. Clazzer *et al.*, "Exploiting combination techniques in random access MAC protocols: Enhanced contention resolution ALOHA," *arXiv preprint arXiv:1602.07636*, 2016.

[46] K. Fyhn *et al.*, "Multipacket reception of passive UHF RFID tags: A communication theoretic approach," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4225–4237, June 2011.

[47] A. Azari and C. Cavdar, "Performance evaluation and optimization of LPWA IoT networks: A stochastic geometry approach," in *IEEE Globecom*, Dec. 2018.

[48] 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*, 2020.

[49] J. Fang *et al.*, "Fine-grained channel access in wireless LAN," *IEEE/ACM Trans. on Networking*, vol. 21, no. 3, pp. 772–787, Aug 2013.

[50] V. Suryaprakash, J. Moller, and G. Fettweis, "On the modeling and analysis of heterogeneous radio access networks using a poisson cluster process," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 2, pp. 1035–1047, Oct 2015.

[51] M. Haenggi, R. K. Ganti *et al.*, "Interference in large wireless networks," *Foundations and Trends in Networking*, vol. 3, no. 2, pp. 127–248, 2009.

[52] J. Moller and R. P. Waagepetersen, *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.

[53] R. K. Ganti and M. Haenggi, "Interference and outage in clustered wireless ad hoc networks," *IEEE Trans. on Inf. Theory*, vol. 55, no. 9, pp. 4067–4086, Aug 2009.

[54] X. Fafoutis, A. Elsts, A. Vafeas, G. Oikonomou, and R. Piechocki, "On predicting the battery lifetime of IoT devices: experiences from the sphere deployments," in *Proceedings of the 7th International Workshop on Real-World Embedded Wireless Systems and Networks*, 2018, pp. 7–12.

[55] A. Azari and G. Miao, "Network lifetime maximization for cellular-based M2M networks," *IEEE Access*, vol. 5, pp. 18 927–18 940, 2017.

[56] G. Miao *et al.*, "E^2 -MAC: Energy efficient medium access for massive M2M communications," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4720–4735, Nov 2016.

[57] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2009.

[58] W. Romberg, "Vereinfachte numerische integration," *Norske Vid. Selsk. Forh.*, vol. 28, pp. 30–36, 1955.

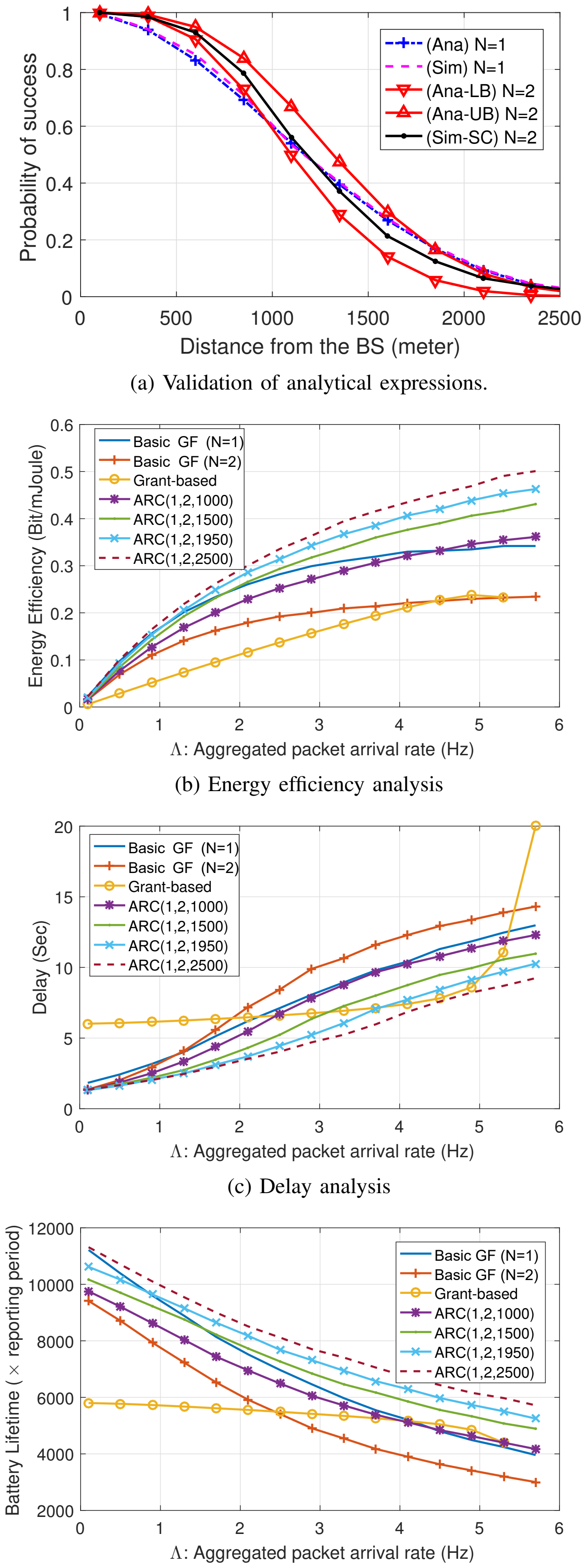
[59] M. Hua *et al.*, "Analysis of the frequency offset effect on Zadoff–Chu sequence timing performance," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4024–4039, Oct 2014.

[60] A. Azari, S. N. Esfahani, and M. G. Roozbahani, "A new method for timing synchronization in OFDM systems based on polyphase sequences," in *2010 IEEE 71st VTC*, May 2010.

[61] A. Bana *et al.*, "Short packet structure for ultra-reliable machine-type communication: Tradeoff between detection and decoding," *arXiv preprint arXiv:1802.10407*, 2018.

[62] S. Rostami and V. T. Vakili, "Three analytical approaches to combine ACB scheme and timing advance information for stationary MTC devices," *IET Communications*, vol. 12, no. 9, pp. 1055–1063, 2018.

[63] X. Jian, Y. Liu, Y. Wei, X. Zeng, and X. Tan, "Random access delay distribution of multichannel slotted ALOHA with its applications for machine type communications," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 21–28, 2016.



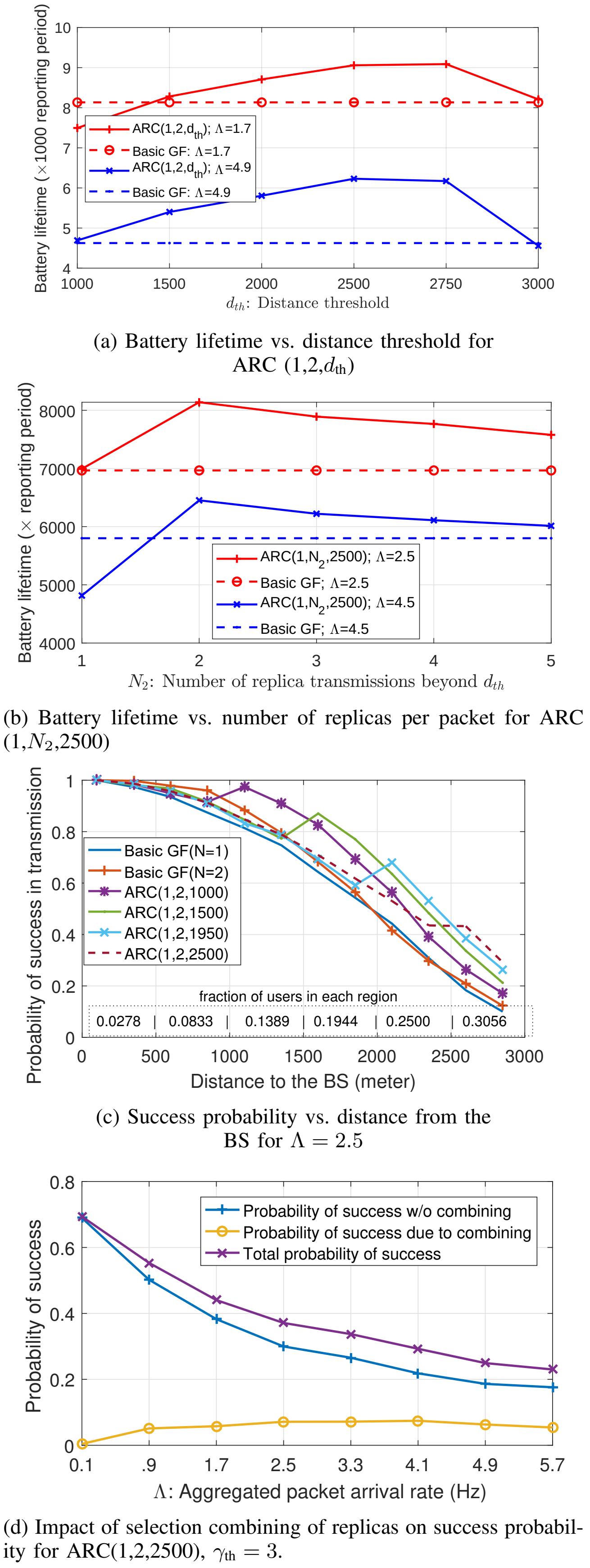
(a) Validation of analytical expressions.

(b) Energy efficiency analysis

(c) Delay analysis

(d) Battery lifetime analysis

Fig. 4: Performance analysis of GFA with the replica control scheme



(a) Battery lifetime vs. distance threshold for ARC (1,2,d_{th})

(b) Battery lifetime vs. number of replicas per packet for ARC (1,N_2,2500)

(c) Success probability vs. distance from the BS for $\Lambda = 2.5$

(d) Impact of selection combining of replicas on success probability for ARC(1,2,2500), $\gamma_{\text{th}} = 3$.

Fig. 5: Tuning design parameters for ARC$(N_1, N_2, d_{\text{th}})$