Grant-Free Access with Multipacket Reception: Analysis and Reinforcement Learning Optimization

Augustin Jacquelin, Mikhail Vilgelm, Wolfgang Kellerer
Chair of Communication Networks
Technical University of Munich, Germany
Email: {augustin.jacquelin, mikhail.vilgelm, wolfgang.kellerer}@tum.de

***Abstract*—Grant-free access has been identified by 3GPP as a potential solution for Industrial Internet-of-Things applications in 5G networks. It allows to decrease overhead and delay, but it is also prone to collisions in the high-load regime. To reduce the effects of collisions, Non-Orthogonal Multiple Access or other Successive Interference Cancellation (SIC) protocols can be applied, allowing to partially recover collisions. In this paper, we abstract the grant-free access protocols with SIC with a K-Multipacket Reception (K-MPR) model. Based on this abstraction, we analyze its one-frame and steady-state throughput, delay and failure probability under different back-off schemes. Furthermore, we propose a reinforcement learning approach to allocate grant-free resources dynamically in order to maximize the normalized throughput of the protocol. Monte-Carlo simulations are employed to confirm the accuracy of analytical results and to evaluate the throughput, delay, and reliability of the proposed resource allocation approach.**

I. Introduction

5G networks are expected to satisfy diverse requirements of upcoming Internet of Things (IoT) applications. Especially challenging requirements come from a subset of Industial IoT (IIoT) applications, where ultra-reliable and low-latency communication (URLLC) is needed [1]. IIoT communication patterns are typically sporadic and consist of occasional small packet transmissions. For such communication patterns, purely grant-based communication inherited from LTE becomes inefficient due to high overhead and delays associated with acquisition of a scheduling grant [2]. To overcome this issue, grant-free and hybrid operation modes are considered by 3GPP as potential enablers for low-latency IIoT [3]. In a grant-free mode, User Equipments (UEs) are allowed to use a certain fraction of resources for direct transmissions to the next Generation Node B (gNB) without requesting the scheduling grant prior to transmission [4]. Grant-free mode further considers two options: dedicated mode (semi-persistent scheduling), or shared mode (random access). The applications with sporadic communication, such as IIoT, are expected to use shared mode to prevent resource waste.

In essence, grant-free protocols with shared resources are partially coordinated random access protocols, where collisions between UEs on the same grant-free resources might occur. Therefore, grant-free protocols only work well in low to moderate load scenarios. If load rises above a certain level, collision probability becomes high and the delay rapidly grows. To reduce the effects of collisions and provide higher throughput, novel random access protocols based on Successive Interference Cancellation (SIC) have been proposed, including Non-Orthogonal Multiple Access (NOMA). SIC-based grant-free access allows to go beyond the ALOHA-like performance by applying interference cancellation or joint decoding to recover some of the collisions [5], [6]. Analysis of SIC-based protocols is highly complex and often not generalizable, as it depends heavily on the power control, channel and propagation effects, and physical layer techniques in use. A useful Medium Access Control (MAC) layer abstraction of SIC or NOMA physical layer is *K-multipacket reception (K-MPR) model*: A generalization of the collision channel, assuming that up to K collided users can be decoded on a single resource. The model is well-known in the literature on Radio Frequency Identification (RFID) and satellite communication networks [7], only recently gaining attention in the context of 5G networks.

In this paper, we analyze and optimize the grant-free access in 5G by modeling it as a K-MPR protocol. Our novel contributions are twofold. (i) First, we provide analytical results on the throughput, delay, and reliability of a generic grant-free protocol with K-MPR. We combine the single-frame and steady-state analysis using Markov chain approach to derive performance metrics under different back-off and re-transmission policies. (ii) Second, we propose a reinforcement learning approach to dimension grant-free resources, where the number of resources is chosen to maximize expected normalized throughput. We show that reinforcement learning provides close to optimal results with reasonable convergence time, and hence it can be used for dynamic on-line adjustment of grant-free resources.

The structure of the paper is as follows. We review related work in Sec. II, and present the system model in Sec. III. Performance analysis is presented in Sec. IV, and performance optimization in Sec. V. Evaluation and simulation results are provided in Sec. VI. Finally, the paper is concluded in Sec. VII.

II. Related Work

3GPP has suggested grant-free (GF) access to reduce the delay and improve efficiency for IIoT scenarios [3]. As a candidate technology, GF protocols have been actively studied for 5G RAN and its URLLC communication. Complementary to GF access, NOMA and other SIC-based protocols enabling mutlipacket reception capabilities (MPR, also referred to as multi-user detection - MUD - in some works) can be applied for performance boost.

Reliability of transmissions via shared grant-free resources with different re-transmission strategies is simulatively studied

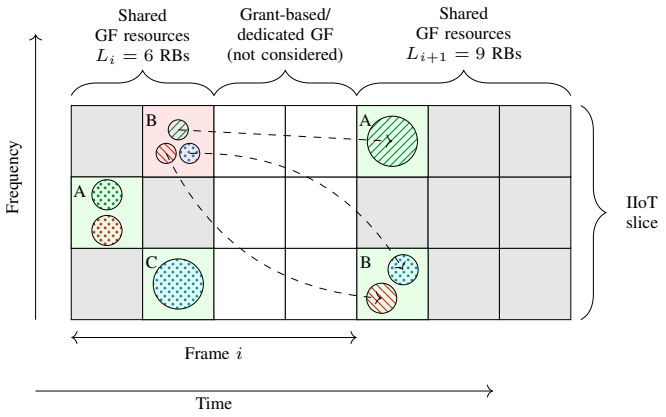

Fig. 1: Illustration of the system model with $K = 2$-MPR. In the first frame i, $L_i = 6$ RBs are available for grant-free access and 6 UEs are contending for it. RBs A, B, C are chosen by 2, 3, and 1 UEs respectively, resulting it three successfully decoded UEs (with A and C), and one collision on B. Collided UEs re-transmit in the next frame, choosing RBs A and B. Since both RBs are occupied by $\leq K = 2$ UEs, all are decoded successfully.

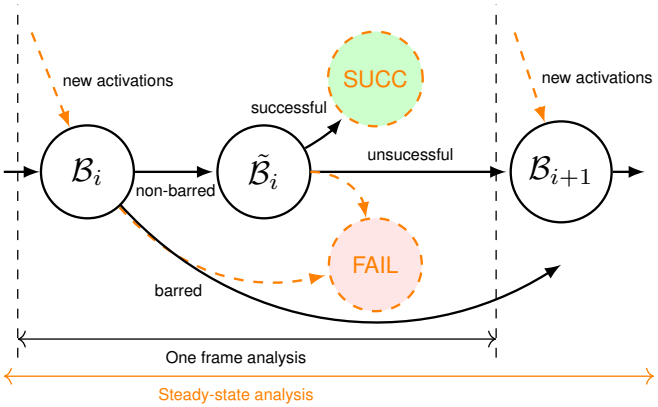

Fig. 2: Network-level illustration of the system model. $\mathcal{B}_i$ denotes the set of ready-to-transmit UEs in the ith frame (after back-off, or newly activated), and $\tilde{\mathcal{B}}_i$ denotes the set of non-barred UEs in the ith frame. One-frame analysis in Sec. IV-A concerns only with the states and transitions depicted in black, whereas steady-state analysis in Sec. IV-B covers full performance assessment.

in [8], [4]. Both articles compare GF with grant-based protocols, pointing out operating regimens and conditions where GF mode becomes beneficial compared to grant-based. In [9], grant-free protocol is analyzed for short packet communication scenario, to achieve low-delay and low-consumption. As a way to increase the reliability, the authors suggest transmissions with packet replicas. Similar to [9], the authors in [10] are optimizing the amount of replicas to meet reliability requirement with a given probability of multi-user detection.

Multi-packet reception has been extensively studied in a parallel line of work [6]. Compressive sensing for MUD in multi-carrier systems have been analyzed in [5]. In [11], the authors study with K-MPR for inhomogeneous CSMA networks. They use K-MPR for analysis and assess performance of such techniques such as SIC and compute-and-forward. In [12], the authors propose a cross-layer (PHY/MAC) decentralized Medium Access Control to coordinate access of UEs, assuming K-MPR. The authors in [13] use K-MPR to analyze the successful reception of a packet from UEs in far-field regions, depending on the channel transmission properties. Slotted ALOHA and its more advanced versions have been studied under K-MPR model in [7] and the references therein.

III. System Model

We consider a gNB serving one cell with N IIoT UEs, where the set of all UEs is denoted by $\mathcal{N}$. IIoT UEs operate in a designated part of the cell's time-frequency resources (depicted as IIoT slice in Fig. 1). IIoT slice is further subdivided into shared grant-free (GF) and grant-based/dedicated GF part, where the second part can be used for re-transmissions or for applications with periodic schedule [8]. In the following, we concentrate solely on shared GF resources. We assume framed time, where a frame also denotes the periodicity of GF resources. As depicted in Fig. 1, each UE contends for L_i GF resources (RBs). The amount of resources L_i is decided prior to frame i by the gNB and is communicated to the UEs via the system information broadcast. Each UE becomes active in a given frame with probability p (i.e., a packet from the application layer arrives into the buffer).

UEs contend for GF resources using a generic K-MPR protocol. We assume that the protocol allows to recover up to K collisions[1], where K is known in advance. I.e., assuming k UEs choose the same RB, all UEs' transmissions are successfully decoded whenever $k \leq K$, whereas all the k UEs are not decoded (collided) whenever $k > K$. To control the load, gNB might use access barring, where the UEs skip current frame with probability p_b. Collided and barred UEs re-transmit in another frame according to a certain back-off policy. We choose to consider two back-off policies. (1) Barring-based back-off, geometric-distributed (denoted as BB): For every frame, UE independently randomly decides whether to skip with probability p_b, or to contend with probability $1-p_b$. This back-off scheme is used for overload control during random access procedure in LTE and NR as a part of the Access Class Barring [14]. (2) Fixed back-off (denoted as FB): After a collided transmission, UE waits fixed number of W frames before re-transmitting. This scheme is considered by 3GPP as a possible scheme for grant-free access in 5G [3], [8]. For stability reasons, number of transmission attempts is limited to M. If UE is not successful after M attempts, the packet is considered dropped and UE goes into inactive mode.

Fig. 2 illustrates queuing point of view on the system. At every frame i, we have a set $\mathcal{B}_i \subseteq \mathcal{N}$ of back-logged UEs ready to transmit, consisting of re-transmitting and newly activating UEs. If barring is applied, the set $\mathcal{B}_i$ is further reduced to $\tilde{\mathcal{B}}_i \subseteq \mathcal{B}_i$ with back-logged UEs which pass the barring. After the contention, UEs from the set $\tilde{\mathcal{B}}_i$ are either successful and become inactive again, or go for the back-off. In the next section, we present analytical results, where we first describe single-frame performance (relations between $\mathcal{B}_i$, $\tilde{\mathcal{B}}_i$, and success probability) in Sec. IV-A, and then extend the analysis towards steady-state in Sec. IV-B, considering activation and back-off schemes.

[1] Although our system model assumptions are common in the literature [7], for completeness, we must remark that the GF RBs do not necessary correspond to the physical resources in the grid, and must be viewed as logical RBs. Depending on the exact protocol in use, there might be multiple physical RBs needed to form a logical RB to enable K-MPR capabilities.

IV. Performance Analysis

In this section, we conduct the performance analysis based on the system model. First, one frame analysis is presented, and then it is extended to the steady-state analysis. The main notations we use are summarized in Table I.

A. One Frame Performance

First, let us assume a frame i with a set of back-logged UEs which already passed access barring, $\tilde{B}_i$, with $\tilde{n}_i \triangleq |\tilde{B}_i|$. Every UE from the set chooses uniformly at random one of L_i RBs to transmit, thus, the number of UEs choosing the same RB is a random variable $\mathbf{k}$. We describe the probability distribution of $\mathbf{k}$ conditioned on $\tilde{n}_i, L_i$ by the following lemma.

Lemma 1: For a given RB, the probability $p\left(k|\tilde{n}_i, L_i\right) = \Pr[\mathbf{k} = k|\tilde{n}_i, L_i]$ that exactly k UEs choose it, is expressed as:

$$p\left(k|\tilde{n}_i, L_i\right) = \binom{\tilde{n}_i}{k} p_L^k (1-p_L)^{\tilde{n}_i - k}, \tag{1}$$

where $p_L \triangleq \frac{1}{L_i}$.

Proof: For given k UEs, the probability of choosing the RB is $\left(\frac{1}{L_i}\right)^k (1-\frac{1}{L_i})^{\tilde{n}_i-k}$, where the first term reflects the probability that k UEs choose the RB, and the second term – the probability that other $\tilde{n}_i - k$ UEs do not choose the RB. As there are $\binom{\tilde{n}_i}{k}$ ways to pick k UEs from the set, resulting probability is expressed as $p\left(k|\tilde{n}_i, L_i\right) = \binom{\tilde{n}_i}{k} p_L^k (1-p_L)^{\tilde{n}_i-k}$. ∎

Lemma 1 defines per-RB load given $\tilde{n}_i$. On the other hand, $\tilde{n}_i$ is a function of n_i and barring probability p_b. We consider this relationship in the following corollary.

Corollary 1: For a given RB, the probability $q\left(k|p_b, L_i, n_i\right) = \Pr[\mathbf{k} = k|n_i, L_i, p_b]$ that exactly k UEs choose it given n_i, barring probability p_b and L_i available RBs is:

$$q\left(k|p_b, L_i, n_i\right) = \binom{n_i}{k} (1-p_b)^k p_L^k (1 - p_L + p_b p_L)^{n_i - k} \tag{2}$$

Proof: The proof follows the same logic as in Lemma 1. ∎

TABLE I: Summary of main analysis and system model notations.

Notation	Description
$\mathcal{N}$, N	Set and number of UEs in the cell
$\mathcal{B}_i$, $\tilde{\mathcal{B}}_i$	Sets of back-logged and non-barred UEs in i^{th} frame
n_i, $\tilde{n}_i$	Number of back-logged and non-barred UEs in i^{th} frame
p, p_b	Activation / barring probability of a UE
L_i	Number of available grant-free RBs at time i
K	Maximum collision size which can be recovered by the protocol (multipacket reception capabilities)
$T, \bar{T}$	Expected throughput and expected normalized throughput per RB in one frame
p_u, p_c	Unsuccessful transmission and collision probabilities
M	Maximum number of transmission attempts
W	Back-off window size (frames)
b_s	Steady-state probability of a UE being in the state s
R, D	Average per-packet reliability / delay (steady-state)

We further define the throughput per frame as the number of successfully decoded UEs. Throughput $\mathbf{T}_i$ and normalized throughput per RB $\bar{\mathbf{T}}_i$ are random variable with expectations $T \triangleq \mathbb{E}\left[\mathbf{T}_i\right]$ and $\bar{T} \triangleq \mathbb{E}\left[\bar{\mathbf{T}}_i\right]$, respectively.

The results of Lemma 1 and its Corollary 1 can be directly applied to obtain the throughput expectations.

Corollary 2: The expected throughput per frame normalized per RB $\bar{T}(n_i, p_b, L_i, K)$ and the expected total throughput per frame $T(n_i, p_b, L_i, K)$ given n_i, p_b, L_i and K are expressed as:

$$\bar{T}(n_i, p_b, L_i, K) = \sum_{k=1}^{K} k\, q\left(k|p_b, L_i, n_i\right) \tag{3}$$

$$T(n_i, p_b, L_i, K) = L_i\, \bar{T}(n_i, p_b, L_i, K) \tag{4}$$

Proof: Note that, in a given RB, throughput equals to the amount of UEs choosing the RB if $0 < k \leq K$ and 0 otherwise. So, the normalized throughput expectation $\mathbb{E}_k\left[\bar{\mathbf{T}}\right]$ conditioned on k, is defined as:

$$\mathbb{E}_k\left[\bar{\mathbf{T}}\right] = \begin{cases} k, & \text{if } 0 < k \leq K, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The probability distribution of k is given by Corollary 1. Applying law of total probability to Eqns. (5) and (2), we obtain the result in Eqn. (3). Since the expectation of a sum is equal to the sum of expectations, the Eqn. (4) readily follows from (3). ∎

An illustration of Corollary 2 is presented in Fig. 3, where throughput and normalized throughput are plotted against L_i for different multipacket reception capability K. We observe that analytical results (denoted "ana") closely match the Monte-Carlo simulations ("sim"). The particular case $K = 1$ is the legacy collision channel model that considers any collision as unrecoverable. The analysis shows that $T(n_i, p_b, L_i, K)$ asymptotically goes to $n_i\, p_b$ with increasing L_i, for any K. On the contrary, normalized throughput $\bar{T}$ achieves a different maximum value depending on K and it holds that $\bar{T} \leq K$. The number of available RBs maximizing normalized throughput $L_i^\star = \arg\max_{L_i} \bar{T}$ is inversely dependent on K: $L_i^\star$ is decreasing with increasing K, for a given back-log n_i. This directly suggests that dynamic adaptation of L_i according to the back-log can be used to keep the throughput at maximum. We will return to the question of throughput maximization problem later in Sec. V.

B. Steady-State Analysis

In the previous section, we studied single-frame performance. This analysis does not allow to predict expected delay and reliability of the protocols, since the amount of back-logged UEs in a frame depends on the new arrivals and on the respective back-off or re-transmission scheme. To account for these, we study the steady-state performance by the means of Markov-chain analysis [15], for barring-based back-off (BB) and fixed back-off (FB), and described in Sec. III.

The Markov chain depicted in Fig. 4 represents all possible states of any UE. A UE starts in the inactive state OFF, and with probability p becomes active in the subsequent frame. The

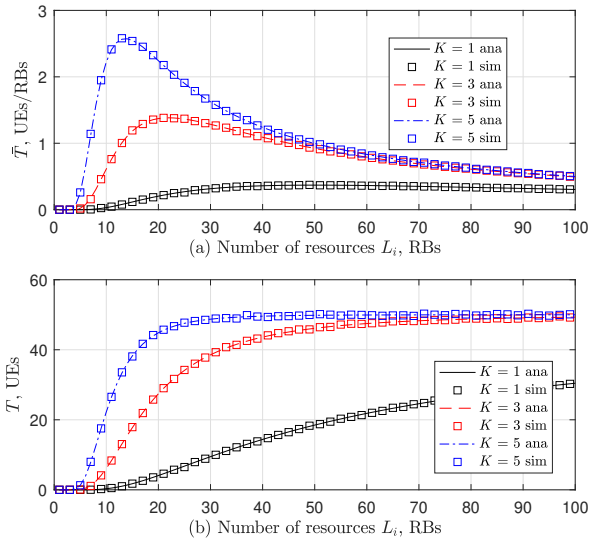

Fig. 3: Results of the Corollary 2. Evolution of $\bar{T}$ and T as a function of L_i for $n_i = 100$, $p_b = 0.5$, $K \in \{1, 3, 5\}$. The case with $K = 1$ represents the legacy collision channel without multipacket reception.

probability p can be used to approximate any packet generation pattern by a Bernoulli process. Upon activation, UE is placed directly in a back-off stage $i = 1$ and state (i, x), where x depends on the back-off scheme used ($x = 0$ for BB, and $x = W - 1$ for FB). We denote the back-off stage i with j frames to wait as (i, j) state. For FB, UE waits for W slots before preforming a transmission. Once UE is in state $(i, 0)$, it can attempt a transmission. For *barring-based BO* (BB), no back-off window is considered, i.e., $W = 0$. At every back-off stage, UE attempts a transmission with probability $1-p_b$ and it is barred with probability p_b. If the transmission is successful, UE goes in the successful state "SUCC" and then goes to the inactive state "OFF". We denote the probability $1 - p_u$ of a transmission to be successful as $1 - p_u$. In case of an unsuccessful or barred transmission, UE goes to the next back-off stage $i + 1$. If the transmission is unsuccessful after M attempts, the UE goes into the FAIL state and the packet is dropped. For the notations, we refer the reader to Table I.

The probability of a transmission to be unsuccessful p_u depends on the collision probability p_c and barring probability p_b. For FB, $p_b = 0$, hence:

$$p_u = \begin{cases} p_c & \text{fixed BO (FB),} \\ p_b(1 - p_c) + p_c & \text{barring-based BO (BB).} \end{cases} \tag{6}$$

Transition probabilities between the states are depicted in Fig. 4. The steady-state probabilities $b_s, \forall s \in \mathcal{S}$, where $\mathcal{S}$ is the set of all states, are expressed as a function of b_{off} using the global balance equations:

$$b_{\text{on}} = p\, b_{\text{off}} \tag{7a}$$
$$b_{i,0} = p_u^{i-1} b_{\text{off}} p \quad \forall i \in [1, M], \tag{7b}$$
$$b_{i,j} = b_{i-1,0}, \quad \forall i \in [2, M], \forall j \in [0, W - 1] \tag{7c}$$
$$b_{\text{fail}} = p\, b_{\text{off}}\, p_u^M \tag{7d}$$
$$b_{\text{succ}} = (1 - p_u) \sum_{i=1}^{M} b_{i,0} = p\, b_{\text{off}} (1 - p_u^M) \tag{7e}$$

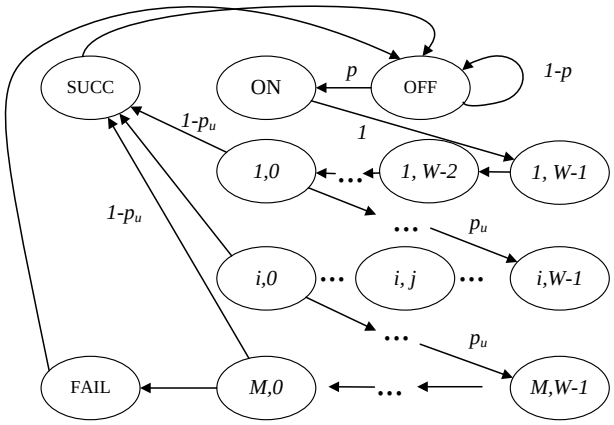

Fig. 4: Markov chain depicting states of a UE and the transition probabilities between them. Both back-off schemes are illustrated here: for barring-based back-off (BB) $W = 0$, and for fixed back-off (FB) $p_u = p_c$.

Finally, by employing the identity condition $\sum_{s \in \mathcal{S}} b_s = 1$ we obtain:

$$b_{\text{off}} = \left(1 + p\left(1 + W \frac{p_u^M - 1}{p_u - 1}\right)\right)^{-1} \tag{8}$$

C. Probability of a successful transmission p_s

Considering any frame i, the probability of collision p_c is approximated using the total throughput T, divided by the average effective number of transmitting users $\lambda = \mathbb{E}[\tilde{n}_i]$:

$$p_c = 1 - \frac{T(\lambda, L_i, K)}{\lambda} \tag{9}$$

On the other hand, λ is expressed via Markov chain analysis as the expected number of UEs in the states $b_{i,0}\ \forall i$:

$$\lambda = \begin{cases} N \sum_{i=1}^{M} b_{i,0} = N\, p\, b_{off} \frac{1-p_c^M}{1-p_c} & \text{for FB,} \\ N(1 - p_b) \sum_{i=1}^{M} b_i = N(1 - p_b) p\, b_{off} \frac{1-p_u^M}{1-p_u} & \text{for BB.} \end{cases} \tag{10}$$

Equations (9) and (10) form a system of two independent equations with two unknowns p_c and λ. The solution is obtained by any non-linear numerical solver based on root finding.

D. Performance Metrics

Having obtained λ and p_c from the Markov chain analysis, we can now use them to predict average steady-state performance of the protocols in terms of throughput, delay, and reliability. We define the steady state reliability R as the probability that a packet is successfully decoded at gNB within $\leq M$ transmission attempts. Using the global balance equations (7), we obtain reliability as:

$$R = \frac{b_{\text{succ}}}{b_{\text{succ}} + b_{\text{fail}}} = 1 - p_u^M. \tag{11}$$

The expected steady-state delay D is expressed as:

$$D = W \frac{\sum_{r=1}^{M} r\, p_u^{r-1}}{\sum_{k=1}^{M} p_u^{k-1}} = W \left(p_u \frac{1 + (M-1)p_u^M - M p_u^{M-1}}{(1-p_u)(1-p_u^M)} + 1 \right). \quad (12)$$

Average throughput per RB $\bar{T}$ is given by Eqn. (3) where $\tilde{n}_i$ is substituted by its expectation λ, obtained from Markov chain analysis.

The results of the steady-state analysis are displayed in Fig. 5: failure probability, delay, and throughput per RB for two BO schemes. We compare Monte-Carlo simulations with Markov-chain analysis. We observe that the simulations largely match the analytical results, with the exception of throughput and delay for small L_i under FB policy. We expect the results to match, if the simulations were to be run longer: Since FB policy causes large delays in the low L_i regime, same simulation runs produce less samples, and therefore less accuracy. Additionally, analysis and simulation results are shown for two values of K: $K = 1$ corresponds to the legacy case of collision channel model, and $K = 4$ to the protocols with high MPR capabilities. As expected, K-MPR outperforms the case $K = 1$ in terms of throughput per preamble, delay and reliability. FB presents a higher delay than BB. We also remark from Eqn .(11), that the failure probability $(1 - R)$ of FB does not depend on W, but only on p_c. The case FB with $W = 1$ is thus equivalent to BB with $p_b = 0$. Therefore, the failure probability of FB is the same as BB with $p_b = 0$, and the delay is W times larger.

We note that the presented methodology can be applied to obtain expected reliability for URLLC applications with strict deadlines. That is, if the back-off policy is configured such that MW equals to the application deadline, reliability R directly reflects expected probability of deadline violation.

V. Performance Optimization

As we observe from the analysis, performance of GF access is heavily dependent on the configuration parameters: back-off scheme, back-off setting, and, above all, on the number of allocated resources L_i. We also observe that the delay is low and the reliability stays high as long as the amount of allocated resources is sufficient to maximize the normalized throughput. Further increasing the number of resources behind this point decreases the efficiency of the protocol. Thus, the question is: What number of RBs is necessary to maximize the normalized throughput? It can be formalized as an optimization problem:

$$\underset{L_i}{\text{maximize}} \quad \bar{T}(L_i) \quad (13a)$$

$$\text{subject to} \quad W, M, p_b \quad (13b)$$

$$p, N, n_i, K \quad (13c)$$

The constraints (13b) represent chosen back-off policy and its configuration[2], whereas the constraints (13c) are given system parameters: multipacket reception capabilities K, number of UEs and their activation pattern. The optimization problem can be approached as a static (off-line) allocation problem, where steady-state throughput is maximized given the estimate of the expected back-log λ as input, or as a dynamic (on-line) allocation problem, where expected throughput is optimized given estimate of the current back-log n_i as input.

[2]Clearly, back-off policy configuration can be an optimization variable as well. Nevertheless, we restrict ourselves by only optimizing the number of RBs, and leave joint optimization for future work.

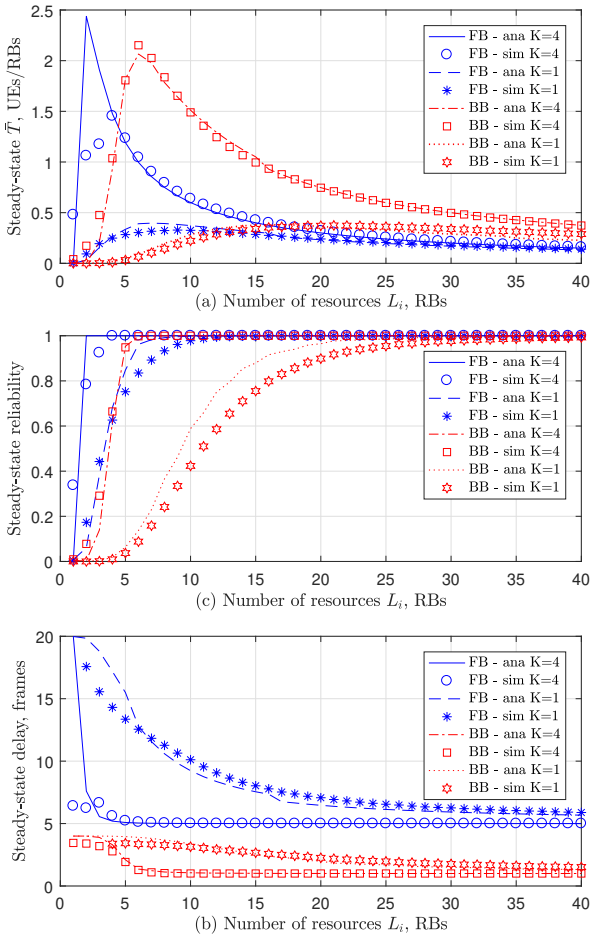

Fig. 5: Steady-state (a) $\bar{T}(n_i, p_b, L_i, K)$, (b) reliability, and (c) delay for $N = 40$, $p = 0, 8$, $\mathcal{K} = 4$, $M = 8$, $W = 5$ for FB and $p_b = 0.3$.

Given the back-log estimates, the solution to the static allocation problem can be obtained via the Markov chain analysis. For dynamic allocation, we can obtain $L_i^\star = \arg\max_{L_i} \bar{T}(n_i, p_b, L_i, K)$ using numerical root-finding algorithm for maximization of the Eqn. (3). While the function $\bar{T}(L_i | n_i, p_b, L_i, K)$ seems to have one maximum (from the plots), it is non-concave and it is not proven that there is only single maximum, therefore, the worst-case complexity of maximization is not guaranteed to be polynomial. We plot the numerically obtained $L_i^\star$ as a function of $\tilde{n}_i$ for different K in Fig. 6. We observe that for large enough $\tilde{n}_i \geq 4$ the dependency becomes linear. Interestingly, corresponding maximum throughput is the largest at the moment where the dependency becomes linear, and saturates to a lower value with increasing $\tilde{n}_i$.

Note that neither n_i nor $\tilde{n}_i$ are known to the gNB, thus, we cannot explicitly use it to maximize $\bar{T}(n_i, p_b, L_i, K)$ given by Eqn. (3). Thus, an additional step of estimation algorithm is needed, which will inherently introduce extra computational

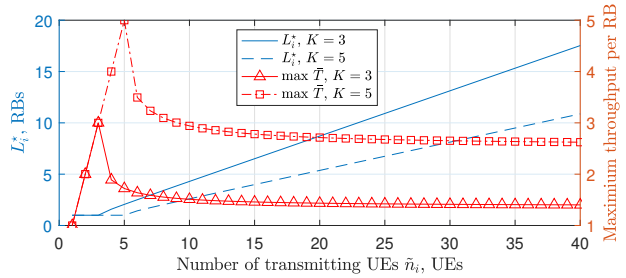

Fig. 6: Evolution of $L_i^\star$ and $\max_L \bar{T}(\tilde{n}_i, L_i, K) = \bar{T}(\tilde{n}_i, L_i^\star, K)$ against the number of active UEs $\tilde{n}_i$, for $N = 400$, $p = 0.5$ and $K = \{3, 5\}$.

time and inaccuracy. The problem of back-log estimation is already a complex task for collision channel model [16], and the complexity is clearly higher for K-MPR model. To circumvent the estimation and complexity issues, we propose a reinforcement learning approach as the main solution to the dynamic allocation problem (13) in the next section.

A. Dynamic Resource Allocation with Q-Learning

Given a fixed back-off policy, the system can be viewed as a Markov Decision Process (MDP), where the system state comprises current states of all UEs and the amount of available resources. The actions in MDP are decisions for the amount of RBs to be allocated in the next frame, and the reward is the normalized throughput achieved. Transition probabilities between the states are dependent on the back-off policy and activity pattern of the UEs. The resulting MDP, however, is not fully observable: gNB cannot directly know the state of every UE, moreover, it cannot even directly observe current back-log. Typically, gNB can only observe outcomes of the contention in previous frame: the number of RBs with collisions $L_i^{(c)}$ and the number of successful UEs $N_i^{(s)}$.

Therefore, we approximate the original MDP with its simplified version based on the contention outcome. The approximation can be viewed as state aggregation technique [17]. Contention outcome is implicitly reflecting the current back-log and it is typically used as a basis of back-log estimation algorithms [16] in the literature on random access. We thus define the *state* space of the MDP as:

$$s_i \in \mathcal{S} \triangleq \{L_i, N_i^{(s)}, L_i^{(c)}\}. \tag{14}$$

Action space corresponds to the optimization variable L_i, $a_i \in \mathcal{A} \triangleq \{L_i | 0 \leq L_i \leq L_{\max}\}$. The *reward* is the quantity that we want to optimize, i.e. the normalized throughput in a given frame $r_i \triangleq \mathbf{T}_i$. We apply off-policy temporal-difference learning based on the computation of the value function of a state-action pair $Q(s_i, a_i)$:

$$Q(s_i, a_i) = \mathbb{E}\left[r_i + \gamma r_{i+1} + \gamma^2 r_{i+2} + ... | s_i, a_i\right]. \tag{15}$$

where $\gamma \in [0, 1)$ is the discount factor for future rewards.

Temporal-difference learning is iteratively updating the value function $Q(\cdot)$, by taking actions and observing rewards.

Algorithm 1 Dynamic Allocation via Q-Learning.

1: Initialize $L_1 \in \mathbb{N}$ randomly, and get $N_1^{(s)}$ and $L_1^{(c)}$
2: Observe current state: $s_i \leftarrow [L, N_1^{(s)}, L_1^{(c)}]$
3: **for** each frame i **do**
4: Adjust $\alpha \in [0, 1]$, $\gamma \in [0, 1]$, and $\epsilon \in [0, 1]$.
5: Choose an action a_i according to the ϵ-greedy policy (17)
6: Set and broadcast L_{i+1} corresponding to action a_i to all UEs
7: Observe contention results: $s_{i+1} \leftarrow [L_{i+1}, N_{i+1}^{(s)}, L_{i+1}^{(c)}]$
8: Observe the reward $r_i = N_{i+1}^{(s)}/L_{i+1}$
9: Update Q table according to (16)
10: Adjust $L_{\max}$ (if needed)
11: **end for**

The values in the table are updated as [17]:

$$Q^{\text{new}}(s_i, a_i) = \tag{16}$$
$$= (1-\alpha) Q^{\text{old}}(s_i, a_i) + \alpha \left(r_{i+1} + \gamma \max_{a_{i+1}} Q(s_{i+1}, a_{i+1})\right).$$

where $\alpha \in [0, 1]$ is the learning rate. We choose to apply Q-learning method for updating the value function: using immediate updates after observing the reward.

The *policy* is a strategy used by the agent (gNB) to take actions. We use the ϵ-greedy policy. Being in state s_i, upon a decision to take, the gNB either chooses a random action (denoted rand) with probability ϵ (to explore the state-action space), or chooses action a_i maximizing $Q(s_i, a_i)$ with probability $1 - \epsilon$ ("to exploit" the reward):

$$a_i = \begin{cases} \text{rand } \{a_i \in \mathcal{A}\} & \text{with probability } \epsilon, \\ \arg\max_{a_i} Q(s_i, a_i) & \text{with probability } 1-\epsilon. \end{cases} \tag{17}$$

The value of ϵ must decay over time, in such a way that there is an exploration phase while the state-action space is not yet know, and a performance phase, where the obtained knowledge is exploited. To get better convergence, α and γ also needs to decrease and increase over time, respectively [18]. The pseudocode of the resulting learning algorithm is summarized in Algorithm 1.

Since the size of state and action space is very large, we apply a form of guided exploration based on constraint restriction. For that, we use a heuristic to compute a bound on possible optimal $L_{\max} \geq L_i$ as follow. It is clear that current back-log is bounded by $\tilde{n}_i \geq N_i^{(s)} + (K+1)L_i^{(c)}$. Also, $L_i^\star \geq \tilde{n}_i/K$, so $L_i^\star \geq (N_i^{(s)} + (K+1)L_i^{(c)})/K$. Let us call $L_{\text{est}} = (N_i^{(s)} + (K+1)L_i^{(c)})/K$, we can then take $L_{\max} = \mu L_{\text{est}}$, where μ is a constraint reduction coefficient. In our algorithm, we take $\mu = 4$ to be sure that $L_{\max} >> L_i^\star$. The value of $L_{\max}$ is updated every 1000 steps. By using $L_{\max}$ we limit the action space by filtering out values of L_i which are not likely to be relevant.

VI. Evaluation Results

A. Evaluation Set-up

In this section, we present the performance evaluation of the proposed algorithm. Algorithm 1 has been implemented

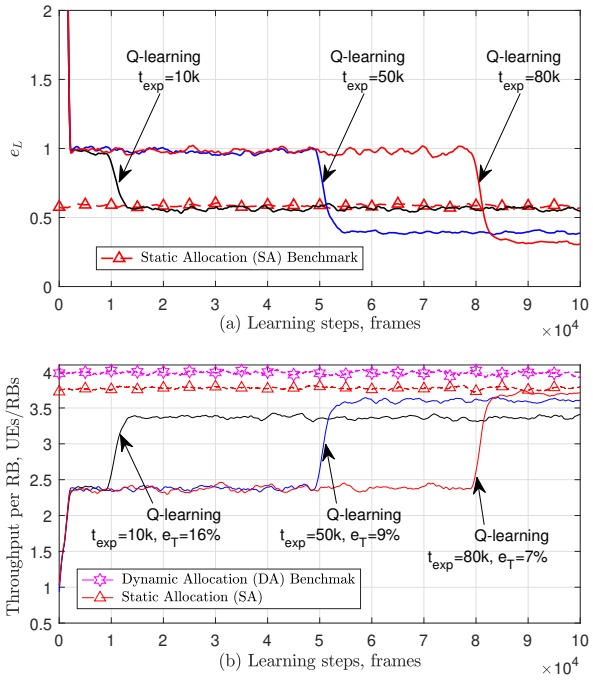

Fig. 7: Evolution of the relative difference e_L, and throughput per RB over the learning steps, for different exploration phase duration, t_{exp}, with $N = 100$, $K = 7$ $p = 0.5$ and geometric back-off parameters $p_b = 0.3$ and $M = 8$. The results have been smoothed with a window of 2000 steps for the purpose of readability. The error to the benchmark throughput, e_T is referred.

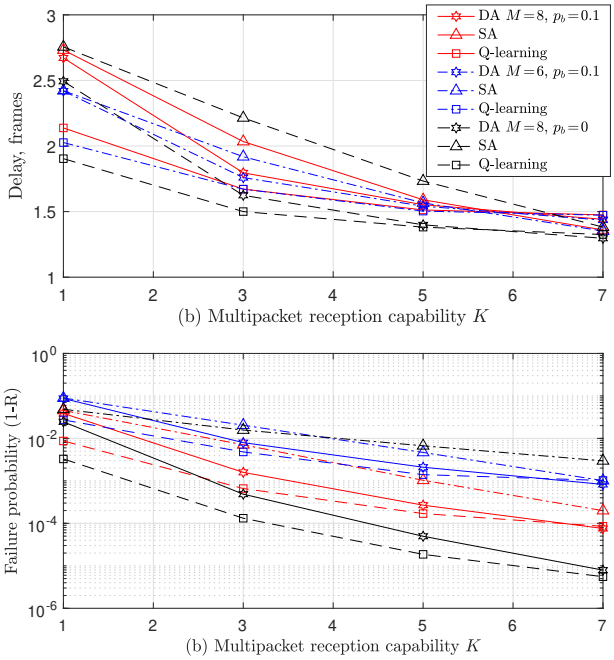

Fig. 8: Average delay and failure probability obtained with Q-learning algorithm with $t_{\text{exp}} = 80000$ steps, in comparison with DA and SA, for $N = 100$ and $p = 0.5$, and different values of p_b and M.

with progression composed of two phases: an exploration phase where $\epsilon = 1$, i.e. actions are taken completely randomly, and a performance phase where $\epsilon = 0$, i.e. actions are taken to maximize the expected cumulative reward $Q(s_i, a_i)$, see Eqn. (17). The duration of the exploration phase is further denoted t_{exp}. The performance of learning algorithms is generally very sensitive to the evolution of ϵ, α and γ over the learning steps. In our implementation, we take an polynomial evolution as in [18]. The evaluation set-up has been implemented for $N = 100$, $p = 0.5$, $K = 7$, $p_b = 0.3$, $M = 8$.

B. Benchmarks

The performance of learning algorithm is compared with two idealistic benchmarks, *dynamic* (DA) and *static* (SA) allocation. DA chooses the optimal number of GF resources $L_i^\star$ *at every frame*, by numerically solving $\bar{T}(n_i, p_b, L_i, K)$ given by (3), and assuming that the amount of transmitting UEs $\tilde{n}_i$ is known. The second benchmark, SA, statically assigns the amount of resources based on the known λ, where $L_i^\star$ is chosen to maximize the steady-state throughput per RB.

Note that both DA and SA are idealistic and serve only as a reference point for the QL performance. They can only be implemented together with estimation for n_i and $\tilde{n}_i$. Estimation will introduce certain performance penalty because of estimation error, and run-time penalty because of high complexity. The run-time for DA and SA is further increased due to numerical computation of $L_i^\star$, and might be infeasible for short frame length. Therefore, we leave load estimation under K-MPR model as a future work.

To quantify the performance of the algorithms, we define relative error in the choice of L_i between the optimal (chosen by DA benchmark) $L_i^\star$, and by the Q-learning algorithm, $L_i^{(\star,Q)}$, as $e_L = |L_i^\star - L_i^{(\star,Q)}|/L_i^\star$, and its evolution over the learning steps, where one step is equal to one frame. The lower e_L, the closer is the learning algorithm to the optimum. In addition, we observe the normalized throughput compared with the SA and DA benchmarks and its relative throughput difference e_T.

C. Simulation Results

The results are presented in Fig. 7 for different exploration phase durations. The length of a learning step (i.e., periodicity of GF access) can vary in 5G: Assuming a standard LTE frame of 10 ms, the exploration phase is varied between $t_{\text{exp}} \in \{100, 500, 800\}$ s. From Fig. 7b, we observe that as we increase the exploration phase, the error e_L for QL decreases from 0.5 (for $t_{\text{exp}} = 10^4$ frames) to ≈ 0.3 (for $t_{\text{exp}} = 10^8$ frames). In terms of throughput, QL achieves 16% lower throughput than DA after around 130 s, and 7% after 830 s, matching the SA performance in the latter case. Evidently, there is a trade-off been performance and exploration time, which needs to be taken into account: For more static systems, where the traffic pattern and number of UEs does not change often, longer exploration phases are advised; For dynamic systems, shorter exploration is recommended. In any case, to account for possible traffic changes, exploration phase can be re-initiated again every now and then.

Delay and failure probability are depicted in Fig. 8 as a function of K for the same configuration setup and varying values of M and p_b. For Q-learning, only performance phase is considered. When $M = 8$ and $p_b = 0$, the delay is less than 4 frames and failure probability is less than 10^{-5}. Interestingly, we observe that the delay and reliability of Q-learning is slightly higher than the benchmarks, despite the lower throughput. A possible explanation is that the guided exploration makes the algorithm prone to over-provision the

resources (see Fig. 7), allocating more RBs than needed. This hints to an important observation that normalized throughput maximization might not be optimal strategy choice, if IIoT application requires strict deadline and reliability.

VII. Conclusions and Discussion

In this paper, we studied the performance of grant-free access in 5G under K-MPR model, which is a generalization of the conventional collision channel, accounting for advanced receiver techniques, such as successive interference cancellation or joint decoding. We analyzed the one frame throughput for arbitrary access barring probability, and then extended the analysis to obtain average steady-state throughput, delay, and reliability via Markov-chain analysis. Furthermore, we have formulated dynamic adaptation of the grant-free resources to maximize the normalized throughput as an optimization problem, and proposed a reinforcement learning algorithm based on Q-learning to solve it. The evaluation results demonstrate that Q-learning approach delivers high throughput with low penalty compared to optimal yet unfeasible benchmarks, at the expense of sub-optimal throughput during exploration phase. The results we provided here can be viewed as first steps, both for optimization of K-MPR protocols, and for applications of reinforcement learning for MAC. They raise multiple questions and discussions point, which we leave for future work.

Back-log estimation. We have provided here the performance analysis, and suggested idealistic optimization approaches for static and dynamic allocation. Making idealistic protocols feasible and load-adaptive requires precise and efficient back-log estimation algorithms, which are not yet available in the literature for MPR. Potentially, collision channel results [16] can be adjusted to account for MPR capabilities.

The choice of objective function. As the results in Sec. V and in particular Fig. 8 suggest, normalized throughput might not be the best choice of the objective function: Sub-optimal throughput can still have higher reliability and lower delays, if the system is over-provisioned. For IIoT applications, reliability and delay play an important role, therefore, choosing reliability as an objective can increase the application performance at the cost of extra resources. The trade-off between performance and resource consumption should also be studied as a multi-objective optimization [19].

Reinforcement learning for MAC. The amount of states and actions in the underlying Markov Decision Process makes the exploration phase duration critical, and convergence of the algorithms long. We have presented here results with relatively small action space, where only amount of resources is adjusted. The action space can be extended, and the back-off configuration can be set as additional optimization variable. In that case, however, convergence times explode, and on-line application of the algorithm becomes hardly feasible. A potential solution here would be to partially pre-train the algorithm off-line, to reduce the search space for on-line application. An additional problem is that the true MDP of the system is not fully observable and has to represent all different back-off stages of a UEs. In the paper, we approximate this process with the observed states, which leads to sub-optimality. For more precise approximation, especially if back-off configuration is also dynamic, deep reinforcement learning with recurrent neural networks [20] can be applied, to capture the memory hidden in the approximation.

References

[1] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. Int. Conf. on 5G for Ubiquitous Connectivity (5GU)*, pp. 146–151, IEEE, 2014.

[2] 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, pp. 4–16, First 2014.

[3] Third Generation Partnership Project, "TR 38.802, Study on New Radio Access Technology - Physical Layer Aspects (Release 14)," tech. rep., 3GPP, 09 2017.

[4] C. Wang, Y. Chen, Y. Wu, and L. Zhang, "Performance evaluation of grant-free transmission for uplink urllc services," in *Proc. IEEE Vehicular Technology Conference (VTC Spring)*, pp. 1–6, June 2017.

[5] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communication," in *Proc. IEEE Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2015.

[6] J.-L. Lu, W. Shu, and M.-Y. Wu, "A Survey on Multipacket Reception for Wireless Random Access Networks," *Journal of Computer Networks and Communications*, 2012.

[7] Č. Stefanović, E. Paolini, and G. Liva, "Asymptotic performance of coded slotted aloha with multipacket reception," *IEEE Communications Letters*, vol. 22, no. 1, pp. 105–108, 2018.

[8] G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovcs, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23602–23611, 2018.

[9] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5g networks," *CoRR*, vol. abs/1709.02179, 2017.

[10] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Communications Letters*, vol. 7, pp. 182–185, April 2018.

[11] S. Ashrafi, C. Feng, and S. Roy, "Performance analysis of csma with multi-packet reception: The inhomogeneous case," *IEEE Transactions on Communications*, vol. 65, pp. 230–243, Jan 2017.

[12] A. Furtado, R. Oliveira, L. Bernardo, and R. Dinis, "Optimal cross-layer design for decentralized multi-packet reception wireless networks," in *Proc. IEEE Vehicular Technology Conference (VTC Spring)*, pp. 1–5, June 2018.

[13] A. Furtado, D. Vicente, R. Oliveira, L. Bernardo, and R. Dinis, "Performance analysis of multi-packet reception wireless systems in far-field region," in *Proc. International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 2045–2049, June 2017.

[14] Third Generation Partnership Project, "Technical Specification 38.331: NR; Radio Resource Control (RRC); Protocol specification;," tech. rep., 3GPP, 2017.

[15] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 535–547, March 2000.

[16] A. Zanella, "Estimating Collision Set Size in Framed Slotted Aloha Wireless Networks and RFID Systems," *IEEE Communications Letters*, vol. 16, pp. 300–303, March 2012.

[17] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[18] V. François-Lavet, R. Fonteneau, and D. Ernst, "How to discount deep reinforcement learning: Towards new dynamic strategies," *CoRR*, vol. abs/1512.02011, 2015.

[19] M. Vilgelm, S. Rueda Liñares, and W. Kellerer, "On the Resource Consumption of M2M Random Access: Efficiency and Pareto Optimality," *IEEE Wireless Communications Letters*, pp. 1–1, 2018.

[20] Z. Chen and D. B. Smith, "Heterogeneous machine-type communications in cellular networks: Random access optimization by deep reinforcement learning," in *Proc. IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2018.