Resource-Optimized Recursive Access Class Barring for Bursty Traffic in Cellular IoT Networks

Han Seung Jang, *Member, IEEE*, Hu Jin, *Senior Member, IEEE*, Bang Chul Jung, *Senior Member, IEEE*, and Tony Q. S. Quek, *Fellow, IEEE*

***Abstract*—A massive number of Internet-of-Things (IoT) and machine-to-machine (M2M) communication devices generate various types of data traffic in cellular IoT networks: periodic or nonperiodic, bursty or sporadic, etc. In particular, bursty and nonperiodic traffic may cause an unexpected network congestion and temporary lack of radio resources. In order to effectively accommodate such bursty and nonperiodic traffic, we propose a novel *recursive* access class barring (R-ACB) technique to optimally utilize the available resources associated with the random access procedure (RAP) that consists of multiple steps in cellular IoT networks, while existing ACB schemes only considered the resource of the first step of RAP, i.e., the number of available preambles. The proposed R-ACB technique consists of two main parts: 1) online estimation of the number of active IoT/M2M devices who have data to transmit to an eNodeB and 2) adjustment of the ACB factor that indicates the probability that an active device sends a preamble to eNodeB. It is notable that the estimation and the adjustment recursively affect each other when R-ACB operates. In addition, we also propose mathematical models to analyze the performance of R-ACB in terms of total service time, average access delay, resource efficiency, and energy efficiency (EE). Through extensive computer simulations, we show that the proposed R-ACB technique outperforms the conventional ACB schemes.**

***Index Terms*—Access class barring (ACB), backlog estimation, Internet of Things (IoT), massive IoT, random access (RA).**

Manuscript received March 12, 2020; revised October 7, 2020 and December 26, 2020; accepted February 1, 2021. Date of publication February 11, 2021; date of current version July 7, 2021. This work was supported in part by the NRF grant funded by the Korea Government Ministry of Science and ICT under Grant 2019R1F1A1061023; in part by the 5G-based IoT Core Technology Development Project Grant funded by the Korean Government (MSIT, Core Technologies for Enhancing Wireless Connectivity of Unlicensed Band Massive IoT in 5G+ Smart City Environment) under Grant 2020-0-00167; in part by the NRF through the Basic Science Research Program funded by the Ministry of Science and ICT under Grant NRF-2019R1A2B5B01070697; in part by the MOE ARF Tier 2 under Grant T2EP20120-0006; and in part by the SUTD Growth Plan Grant for AI. *(Corresponding authors: Hu Jin; Bang Chul Jung.)*

Han Seung Jang is with the School of Electrical, Electronic Communication, and Computer Engineering, Chonnam National University, Yeosu 59626, South Korea (e-mail: hsjang@jnu.ac.kr).

Hu Jin is with the Division of Electrical Engineering, Hanyang University, Ansan 15588, South Korea (e-mail: hjin@hanyang.ac.kr).

Bang Chul Jung is with the Department of Electronics Engineering, Chungnam National University, Daejeon 34134, South Korea (e-mail: bcjung@cnu.ac.kr).

Tony Q. S. Quek is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

Digital Object Identifier 10.1109/JIOT.2021.3058808

I. Introduction

EMERGING cellular Internet-of-Things (IoT) and machine-to-machine (M2M) communication services have changed typical wireless network operation paradigms for human-oriented services to more complicated and sophisticated ones in order to efficiently accommodate tens of billions of devices [1], [2]. In general, a massive number of IoT devices generate various types of data traffic. Periodic and sporadic traffic is well supported in cellular IoT networks, but nonperiodic and bursty traffic may cause an unexpected network congestion/overload and temporary lack of radio resources [3].

To be specific, the network congestion may occur at the initial random access procedure (RAP) when a large number of idle IoT devices wake up and attempt random access (RA) simultaneously to enter into the radio resource control (RRC) connected state [4], [5], which is also known as *RA overload* problem. The effect of simultaneous RA attempting of massive devices on physical random access channel (PRACH) was investigated in 3GPP LTE systems [6], [7]. In order to resolve the RA overload problem, various techniques have been proposed for cellular IoT networks in [8]–[10]. Even in the 3GPP LTE standard, several mechanisms have also been proposed to mitigate the RA overload, including access class barring (ACB), separate RA channels, dynamic resource allocation, backoff-based schemes, and group paging [11].

Among the above techniques, the ACB scheme has been considered as a promising technology since it effectively controls bursty traffic in cellular IoT networks [12]. The ACB scheme inherently controls the number of simultaneous accesses by adjusting the *ACB factor*, which denotes the probability that each active device will attempt RA, and it has been actively investigated in the literature [13]–[26]. Cheng *et al.* [13] proposed a prioritized dynamic ACB scheme based on both preallocating PRACH resources and class-dependent backoff times. Duan *et al.* [15] proposed a dynamic ACB scheme to adaptively update the ACB factor to reduce RA delay based on both the number of active devices and the number of preamble collisions at the previous time slots. Oh *et al.* [20] proposed a joint optimal PRACH resource allocation and access control mechanism to maintain RA delay to be smaller than a certain threshold under RA overload situations. Wei *et al.* provided detailed modeling and analysis of

RA channels with bursty arrivals [18] and an extended access barring (EAB) mechanism for M2M communications [19]. In particular, based on the drift approximation in [18], the stochastic process containing the number of devices engaged in their nth transmission at an RA slot was approximated to a deterministic discrete-time system. Consequently, the expected numbers of successful and failed preamble transmissions can be analyzed. Cheng *et al.* [19] introduced the detailed analysis of the EAB algorithm, and proposed a method to optimize the setting of paging cycle and repetition period of system information block. Compared to [18] and [19], this article focuses more on proposing a novel resource-optimized ACB scheme that controls the ACB factor slot by slot by considering all the available resources associated with the RAP so that it can adapt to dynamic network scenarios such bursty traffic.

Wang and Wong [21], [22] proposed an ACB scheme that exploits timing advance (TA) information of stationary or fixed devices. Jin *et al.* [24] proposed a recursive pseudo-Bayesian method for estimating the number of active devices and dynamically updating the ACB factor based on both the number of available preambles and the number of unused (or idle) preambles at the first step of RAP, while many ACB schemes assumed the perfect information on the number of active devices at eNodeB. Leyva-Mayorga *et al.* [25] provided an analytical model for the performance evaluation of LTE-A RAP with the ACB scheme.

As noted in [27], however, most ACB schemes focused on efficiently utilizing the resource associated with the first step of RAP, i.e., the number of available preambles, while less consideration was paid for the resources needed for the random access response (RAR) messages at the second step of RAP and the physical uplink shared channel (PUSCH) resources needed for devices' packet transmissions at the third step of RAP. In general, even though a large number of devices successfully transmit their preambles without collisions at the first step of RAP, much severe bottleneck may occur at the second or third step of RAP due to the lack of resource for the RAR messages [18], [19], [28]–[30] or lack of the PUSCH resources for devices' packet transmissions [31]–[33]. Thus, it is necessary to carefully consider the resources associated with the overall RAP when designing an efficient ACB scheme. In addition, most of the previous ACB schemes were investigated in terms of total service time, access delay, and access throughput, but resource efficiency and energy efficiency (EE) of the ACB schemes have been rarely investigated in the literature [34]. It is worth noting that EE is especially important to prolong lifetime of battery-powered IoT/M2M devices.

In this article, we propose a *recursive* ACB (R-ACB) technique to effectively accommodate bursty and nonperiodic traffic of massive IoT/M2M devices in cellular networks, which is designed in a purpose of optimally utilizing the resources associated with all the steps of RAP. In particular, the proposed R-ACB consists of two main parts. In the first part, a practical online estimation method for the number of active devices is designed. It is worth to mention that in each slot, if the backlog size information, i.e., the number of active devices, is available, we can optimize the performance of RA. In the second part, the ACB factor is adaptively adjusted based on the estimated backlog size and the available resources associated with all the steps of RAP. In addition, when analyzing the optimal ACB factor, the possibility of partially identifying preamble collisions at the eNodeB side is also taken into account. One feature of the proposed R-ACB is that the estimation of the backlog size and the adjustment of the ACB factor recursively affect each other slot by slot when R-ACB operates. Compared to the previous work on ACB in the literature, our main contributions can be summarized by threefold.

1) Due to the consideration of the resources associated with all the steps of RAP, a methodology to obtain the optimal ACB factor is newly developed. We also introduce an approximation for the optimal ACB factor which further facilitates the low-complexity system operation.
2) Based on the optimal ACB factor newly calculated, the update rule of estimating the backlog size is redesigned compared to the previous work. As a result, the offset parameters, which are needed to correct the estimation on the backlog size by observing the preamble utilization, are derived.
3) The phenomenon that eNodeB may possibly identify preamble collisions is reflected in the design of the proposed R-ACB. It is notable that this phenomenon was not considered by the existing ACB schemes in the literature.
4) For R-ACB, we also propose mathematical models to analyze the performance in terms of total service time, average access delay, resource efficiency, and EE. Due to the complicated operation of RAP in cellular IoT networks, there is little work on the performance analysis when ACB is applied. With our proposed mathematical models, it is easy to catch the insight associated with R-ACB and we are also able to predict the system performance in a simpler way.

The remainder of this article is organized as follows. In Section II, we describe the system model considered in this article. In Section III, we explain the proposed R-ACB technique in detail. The mathematical models to analyze the performance of R-ACB is introduced in Section IV and the performance evaluation is presented in Section V. Finally, conclusions are drawn in Section VI.

II. System Model

To introduce our proposed R-ACB, we define the following parameters.

1) M: It is the number of available preambles at step 1 of RAP.
2) Q: It is the maximum number of RAR messages transmittable at step 2.
3) U: It is the maximum number of PUSCH resources at step 3.
4) N: It is the total number of devices.
5) v_i: It is the estimated number of active devices at the ith slot,
6) p_i: It is the ACB factor at the ith slot.

In general, a single RAR message can deliver N_{grant} uplink resource grants for PUSCH resources [35]. If the maximum

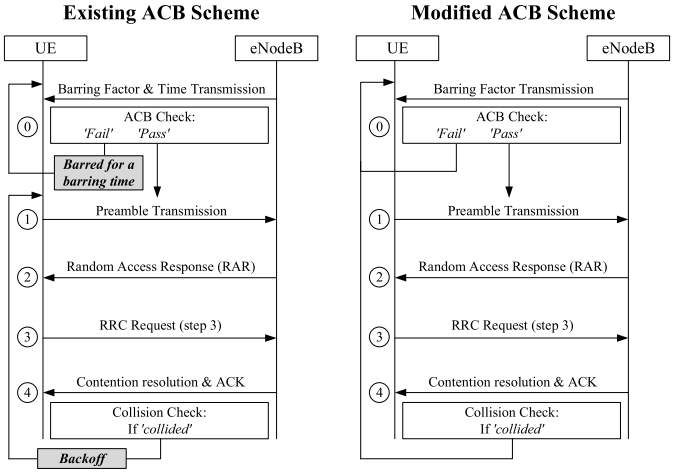


Fig. 1. Procedures of the existing and modified ACB schemes.

number of uplink resource grants transmittable ($Q \cdot N_{\mathsf{grant}}$) at step 2 is larger than the maximum number of PUSCH resources U at step 3, the eNodeB can deliver only U uplink resource grants at step 2, otherwise, it can deliver ($Q \cdot N_{\mathsf{grant}}$) uplink resource grants at step 2. Thus, the number of allocable PUSCH resources K is bounded by

$$K = \min\{Q \cdot N_{\mathsf{grant}}, U\}. \tag{1}$$

Consequently, when designing the ACB technique, we need to consider the number of available preambles M and the number of allocable PUSCH resources K.

A. Overall Steps of Modified ACB-Based Random Access Procedure

In the existing ACB scheme of 3GPP LTE, the eNodeB broadcasts one of ACB factors in $\{0.05, 0.1, \ldots, 0.90, 0.95\}$ and one of ACB times in $\{4, 8, 16, \ldots, 512\}$ [36], [37]. Then, before the initiation of RAP, each UE determines its barring status with the information provided from the eNodeB as shown in Fig. 1 (left). If it fails in ACB check, it is barred for a random barring time $T_{\mathsf{barring}} = [0.7 + 0.6 \times U[0, 1)] \times T_{\mathsf{ACB}}$ [36], where $U[0, 1)$ and T_{ACB} represent a uniform random variable between 0 and 1, and a broadcast ACB time, respectively. On the other hand, if it passes the *ACB check*, then it starts a four-step contention-based RAP. However, in this article, we consider a modified ACB scheme as shown in Fig. 1 (right), in which the *ACB check* is checked for every RA (re)attempt. In addition, the ACB time and the backoff are not used while the per-RA-based ACB check plays the role of mitigating congestions alternatively. It is notable that this kind of modified ACB scheme has been extensively discussed in the literature [15], [22]–[24], [26], [38]. More specifically, the overall modified RAP consists of five steps, which includes the ACB check (step 0) and the typical four-step contention-based RAP (steps 1–4) [4].

Step 0 (ACB Check): At the beginning of the ith PRACH slot,[1] each activated device generates a random number $q \in [0, 1]$ and compares it with $p_i \in [0, 1]$, which is the ACB factor of the ith PRACH slot notified by eNodeB. If $q \leq p_i$, the device attempts RA in the ith PRACH slot. Otherwise, it defers the RA attempt to the $(i + 1)$th PRACH slot.

[1]One PRACH slot corresponds to a PRACH time period defined by the 3GPP PRACH configuration index.

Algorithm 1 Estimation of the Collision-Identification Probability

1: Initialize $\hat{\theta}_0$ to 0.1 and do the following every RAP.
2: At the first step of RAP in the ith PRACH slot, the eNodeB counts the number of collision-identified preambles as ϑ_i.
3: At the third step of RAP in the ith PRACH slot, the eNodeB counts the number of collided preambles as $C_{u,i}$ and calculates $C_i = \vartheta_i + C_{u,i} \times \max(1, W/K)$.
4: Calculates $C_i^T = \sum_{k=i-T}^{i} C_k$ and $\vartheta_i^T = \sum_{k=i-T}^{i} \vartheta_k$.
5: Update the collision-identification probability as $\hat{\theta}_i = \vartheta_i^T / C_i^T$.

Step 1 (Preamble Transmission and Detection): Each device that passed the ACB check randomly selects one of M available preambles and sends it to eNodeB. Due to the random selection of the preamble, more than one devices may select the same preamble. During the preamble detection procedure at the eNodeB, it can distinguish whether a particular preamble is detected (i.e., transmitted) or not (i.e., idle), while it cannot perfectly identify if a detected preamble is transmitted by a single device (i.e., *collision free*) or multiple devices (i.e., collision). If we denote S and C, respectively, by the number of collision-free preambles and the number of collided preambles, the number of detected preambles D is equal to $S + C$. In addition, the collided preambles can be further divided by two types.

1) *Collision-Identified Preamble:* The eNodeB identifies that this preamble is transmitted by more than one devices through advanced signal processing techniques such as observing the power delay profile and, consequently, does not need to serve this kind of preambles at the consequent RAP steps so that the resource usage can be saved.
2) *Collision-Unidentified Preamble:* While the preamble is transmitted by multiple devices, the eNodeB does not realize this fact and it only knows this preamble is transmitted. Consequently, the eNodeB needs to serve this kind of preambles in later RAP steps.

If we denote the number of collision-unidentified preambles by C_u, the number of preambles to be served by the eNodeB in the later RAP steps W is equal to $S + C_u$. Let θ be the *collision-identification probability*, i.e., the probability that the eNodeB successfully identifies a collided preamble, which can be possibly accomplished by observing power delay profile at step 1 of RAP. As a statistical value, θ can be obtained by eNodeB by observing the collided preambles identified at the first step of RAP and the collided preambles detected at the third step of RAP. Algorithm 1 introduces an online estimation procedure to obtain the collision-identification probability at the eNodeB. The ratio of W/K in line 3 reflects the case when the number of preambles to be served W is larger than the number of allocable PUSCH resources K. The value T in line 4 is the average window size for calculating expectation, and

in this article, we set T by 50. In the special case of $\theta = 0$, we have $C = C_u$.

Step 2 (RAR): After the detection of preambles at the eNodeB, it sends the RAR messages for the W preambles to be served, each of which conveys the identity of a preamble, TA information, and an initial uplink resource grant for packet transmissions at step 3 of RAP. If the number of allocable PUSCH resources K is less than W, the eNodeB randomly chooses K preambles among W in order to allocate PUSCH resources via RAR messages. In this case, the devices transmitted one of the remaining $W - K$ preambles wait for the next PRACH slot and perform the RAP again.

Step 3 (Uplink Data Transmission): Receiving the initial uplink resource grant via the RAR message, the corresponding device transmits a packet on the assigned PUSCH resource in uplink, which convey an RRC connection request, a tracking area update, or a scheduling request. If multiple devices transmit their individual packets on the same PUSCH resource at step 3, which is possible if the resource is granted for a collision-unidentified preamble, then decoding failure occurs at the eNodeB. While 3GPP standard supports hybrid automatic repeat request (HARQ) mechanism for the packet transmissions in step 3, HARQ is not helpful when multiple devices continuously retransmit their packets at the same resource, which causes interference. Hence, in this article, we assume that HARQ retransmission is disabled for those collision-unidentified preambles. Through this decoding failure, eNodeB inherently recognizes the preamble collision.

Step 4 (Contention Resolution Message Transmission): If the eNodeB successfully decodes the uplink packet transmitted by a single device, it sends back the contention resolution message including the device identity (ID) obtained from the decoded packet, otherwise, the eNodeB sends nothing back in order to notify the preamble collision. The devices that received its own ID in the contention resolution message sends back a positive ACK message, while the devices that do not obtain their ID regard the situation as the preamble collision, and they do not send any message back. Based on the existing ACB scheme of 3GPP LTE standard, those devices that did not receive the contention resolution message should reattempt the RAP after backoff for several slots, which are randomly chosen from a predefined backoff window. Since this kind of operation may not efficient for the bursty traffic scenario, in the modified ACB scheme, we assume that those devices start RAP again in the next PRACH slot with a reattempt probability that is identical to the ACB factor. Note that the reattempt probability can also emulate the random backoff while the slot-by-slot control of the ACB factor p_i (or the reattempt probability) further introduces the adaptability to the network dynamics.

B. Bursty Traffic Model

For the traffic model, each of N devices is activated at time $x \in (0, T_{\mathsf{act}})$, and the activation time x follows Beta distribution with parameters $\alpha = 3$ and $\beta = 4$ [39], i.e.,

$$f_X(x) = \frac{x^{\alpha-1}(T_{\mathsf{act}} - x)^{\beta-1}}{(T_{\mathsf{act}})^{\alpha+\beta-1}B(\alpha, \beta)}$$

where $f_X(x)$ and $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ denote the probability density function (PDF) of Beta distribution and the beta function, respectively. There exist total I_{act} PRACH slots within time duration of T_{act}, and the expected number of newly activated devices in the ith PRACH slot is given by $\lambda_i = N \int_{t_{i-1}}^{t_i} f_X(x)dx$ for $i = 1, 2, \ldots, I_{\mathsf{act}}$, where t_{i-1} and t_i represent the end times of $(i-1)$th slot and ith slot, respectively.

III. Recursive Access Class Barring

In this section, we introduce our proposed ACB technique, which mainly consists of two parts: 1) estimation of the backlog size, i.e., the number of active devices, by adopting the concept of Bayes' rule and 2) derivation of the optimal ACB factor, which maximizes the *access throughput*, i.e., the number of devices that eventually gets successful RA, while considering the resource limitation over the whole RAP and the possibility of identifying preamble collisions at the eNodeB. As well known, Bayes' rule, in general, introduces a distribution as the estimation result rather than a specific value. Therefore, when performing estimation, we need to store the resulting distribution for further processing. For simplicity, we approximate the estimated distribution by Poisson, which can be specified by its mean and still gives us a reasonably good estimation on the backlog size. It is notable that Poisson distribution has been widely used for estimating the network size in [24], [40], and [41] and we shall also show the effectiveness of such an approximation in the section of performance evaluation. With our proposed technique, in each slot, the ACB factor is derived based on the previous slot's estimated backlog size. Then, by observing the status of preambles when the derived ACB factor is utilized, the estimation is performed. Hence, the estimation of backlog size and the derivation of ACB factor recursively affect each other.

In Section III-A, we derive the average number of detected, collision-free, and collided preambles at step 1 of RAP when the backlog size has Poisson distribution with mean ν. In Section III-B, we first calculate the access throughput by considering the resource limitations: the number of available preambles M, the number of allocable PUSCH resources K, and the collision-identification probability θ. Then, the ACB factor that maximizes the access throughput in each slot is derived. Finally, in Section III-C, we describe our proposed algorithm in detail.

A. Average Number of Detected, Collision-Free, and Collided Preambles

At step 1 of RAP, each of the activated devices, which passed the ACB check, transmits an arbitrary preamble among M available preambles on PRACH, and then, the eNodeB detects the transmitted preambles. In general, detected preambles are classified into collision-free preambles (transmitted by a single device) and collided preambles (transmitted by two or more devices), but, in practice, the eNodeB cannot identify if the detected preamble is collision-free or collided at the first step of RAP. For the analysis, let us define the following discrete random variables.

1) *D:* It is the number of detected preambles.
2) *S:* It is the number of collision-free preambles.
3) *C:* It is the number of collided preambles.

Then, we have $D = S + C$.

We first consider the following joint probability that when m devices simultaneously attempt RA, among a total of $L(\leq M)$ preambles, randomly chosen $(s+r)$ preambles are of interests where s preambles are transmitted by a single device (i.e., collision-free) and r preambles are not selected by any device (i.e., idle) [3]

$$\Psi^{L}_{s+r}(s|m) = \binom{L}{s+r}\binom{s+r}{s}\binom{m}{s}s!\frac{\{L-(s+r)\}^{m-s}}{L^m} \quad (2)$$

where $\binom{L}{s+r}\binom{s+r}{s}$ represents the total number of cases that we choose $(s+r)$ preambles among L preambles, and then choose s collision-free preambles among the chosen $(s+r)$ preambles. In addition, $\binom{m}{s}s!\{L-(s+r)\}^{m-s}$ represents the total number of cases that among the RA-attempting devices, s devices select exactly s preambles (i.e., collision free) and the remaining $(m-s)$ devices select the preambles in $\{L-(s+r)\}$ out-of-interest preambles. Then, $\binom{L}{s+r}\binom{s+r}{s}\binom{m}{s}s!(d-s)^{m-s}$ is divided by L^m, which represents the total number of possible cases that m RA-attempting devices select any of L preambles.

Let E_l denote the event that among a total of $L(\leq M)$ preambles, the lth preamble is selected by at most one device (none or one device). Then, the probability of the event E_l when m devices attempt RA is given by

$$\Pr\{E_l|m\} = \left(1-\frac{1}{L}\right)^m + \binom{m}{1}\frac{1}{L}\left(1-\frac{1}{L}\right)^{m-1} \quad (3)$$

for $l = 1, \ldots, L$, where $(1/L)$ denotes the probability that a device selects the lth preamble among L possibilities.

Once E_l is defined, $\{\cup_{l=1}^{L}E_i\}$ can represent the event that none of the total L preambles is selected by more than one device. Consequently, its complement $\overline{\{\cup_{l=1}^{L}E_i\}}$ represents the event that all the L preambles are selected by at least two devices (i.e., collided). Based on the inclusion-exclusion principle [42], we can first obtain the probability of $\{\cup_{l=1}^{L}E_l\}$ when m devices attempt RA as

$$\begin{aligned}\Pr\left\{\cup_{l=1}^{L}E_l|m\right\} &= \sum_{k=1}^{L}(-1)^{k+1}\sum_{j=0}^{k}\Psi^{L}_{k}(j|m)\\ &= \sum_{k=1}^{L}\sum_{j=0}^{k}(-1)^{k+1}\binom{L}{k}\binom{k}{j}\binom{m}{j}j!\frac{(L-k)^{m-j}}{L^m}\end{aligned} \quad (4)$$

where $\sum_{j=0}^{k}\Psi^{L}_{k}(j|m)$ represents the probability that each of the randomly chosen k preambles is selected by at most one device.

Based on (4), the complementary probability to $\Pr\{\cup_{l=1}^{L}E_l|m\}$ can be further derived as

$$\begin{aligned}\Phi(L|m) &= 1-\Pr\left\{\cup_{l=1}^{L}E_l|m\right\}\\ &= 1-\sum_{l=1}^{L}\sum_{j=0}^{l}(-1)^{l+1}\binom{L}{l}\binom{l}{j}\binom{m}{j}j!\frac{(L-l)^{m-j}}{L^m}\\ &= \sum_{l=0}^{L}\sum_{j=0}^{i}(-1)^{l}\binom{L}{l}\binom{l}{j}\binom{m}{j}j!\frac{(L-l)^{m-j}}{L^m}.\end{aligned} \quad (5)$$

Note that it is the probability that when m devices attempt RA, all the L preambles are selected by at least two devices, i.e., collided.

Finally, the probability that when m devices attempt RA with a total of M preambles, d preambles are detected (each of the d preambles is selected by at least one device) among which s preambles are collision free and the remaining $(d-s)$ preambles are all collided that can be obtained as

$$\begin{aligned}\Pr\{D=d, S=s|m\} &= \Psi^{M}_{s+(M-d)}(s|m)\Phi(d-s|m-s)\\ &= \binom{M}{M-d+s}\binom{M-d+s}{s}\binom{m}{s}\\ &\times s!\frac{(d-s)^{m-s}}{M^m}\\ &\times \sum_{l=0}^{d-s}\sum_{j=0}^{l}(-1)^{i}\binom{d-s}{l}\binom{l}{j}\\ &\times \binom{m-s}{j}j!\frac{(d-s-l)^{m-s-j}}{(d-s)^{m-s}}\\ &= \sum_{l=0}^{d-s}\sum_{j=0}^{l}\binom{M}{d}\binom{d}{s}\binom{m}{s}\\ &\times s!\frac{(d-s)^{m-s}}{M^m}\\ &\times (-1)^{l}\binom{d-s}{l}\binom{l}{j}\binom{m-s}{j}\\ &\times j!\frac{(d-s-l)^{m-s-j}}{(d-s)^{m-s}}\end{aligned} \quad (6)$$

where s, $(M-d)$, and $(d-s)$ in the first line represent the numbers of collision-free, idle, and collided preambles, respectively. $\Psi^{M}_{s+(M-d)}(s|m)$ represents the probability that among the total of M preambles, $[s+(M-d)]$ are randomly chosen in which s preambles are collision free. Note that $\Phi(d-s|m-s)$ indicates the probability that the remaining $M-[s+(M-d)] = (d-s)$ preambles are all collided. For the derivation in (6), we applied the identity of $\binom{M}{M-d+s}\binom{M-d+s}{s} = \binom{M}{d}\binom{d}{s}$.

As mentioned in the first of this section, we presume that the backlog size n is estimated with a Poisson distribution with mean of ν as [40]

$$\mathbb{P}(n|\nu) = \frac{\nu^n}{n!}e^{-\nu} \quad (7)$$

where ν could be considered as the estimated backlog size. Then, given an ACB factor p and mean ν, the joint probability that n devices are active, m devices attempt RA, d preambles are detected, and s preambles are collision free can be expressed

$$\Pr\{d,s,n,m|p,\nu\} = \Pr\{d,s|m\}\underbrace{\Pr\{m|p,n\}}_{\mathbb{B}^{n}_{m}(p)}\underbrace{\Pr\{n|\nu\}}_{\mathbb{P}(n|\nu)} \quad (8)$$

where $\mathbb{B}^{n}_{m}(p) = \binom{n}{m}p^m(1-p)^{(n-m)}$ denotes the binomial distribution with an ACB factor p and the number of active devices n. By summing up with regard to n and m, we have the joint probability that s preambles are collision free and d preambles

are detected as

$$\begin{aligned}\Pr\{D=d,S=s|p,\nu\} &= \sum_{n=0}^{\infty}\sum_{m=0}^{n}\Pr\{d,s,n,m|p,\nu\}\\ &= \binom{M}{d}\binom{d}{s}e^{-p\nu}\left(\frac{p\nu}{M}\right)^{s}\\ &\quad\times\left(-1-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{d-s}. \end{aligned} \quad (9)$$

The detailed derivation of (9) can be referred to [23] and [24]. From this, we have $\Pr\{D=d|p,\nu\}$ and $\Pr\{S=s|p,\nu\}$, respectively, as follows:

$$\begin{aligned}\Pr\{D=d|p,\nu\} &= \sum_{s=0}^{d}\Pr\{D=d,S=s|p,\nu\}\\ &= \binom{M}{d}e^{-p\nu}\sum_{s=0}^{d}\binom{d}{s}\left(\frac{p\nu}{M}\right)^{s}\\ &\quad\times\left(-1-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{d-s}\\ &= \binom{M}{d}e^{-p\nu}\left(-1+e^{\frac{p\nu}{M}}\right)^{d} \end{aligned} \quad (10)$$

and

$$\begin{aligned}\Pr\{S=s|p,\nu\} &= \sum_{d=0}^{M}\Pr\{D=d,S=s|p,\nu\}\\ &= e^{-p\nu}\left(\frac{p\nu}{M}\right)^{s}\sum_{d=0}^{M}\binom{M}{d}\binom{d}{s}\\ &\quad\times\left(-1-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{d-s}\\ &= \binom{M}{s}e^{-p\nu}\left(\frac{p\nu}{M}\right)^{s}\sum_{d=0}^{M}\binom{M-s}{d-s}\\ &\quad\times\left(-1-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{d-s}\\ &= \binom{M}{s}e^{-p\nu}\left(\frac{p\nu}{M}\right)^{s}\left(-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{M-s}. \end{aligned} \quad (11)$$

The number of collided preambles C is easily obtained by $C=D-S$, and then, we have $\Pr\{C=c,S=s|p,\nu\}$ by substituting $s+c$ for d in (9), and $\Pr\{C=c|p,\nu\}$ by $\sum_{s=0}^{M-c}\Pr\{C=c,S=s|p,\nu\}$, i.e.,

$$\begin{aligned}\Pr\{C=c|p,\nu\} &= \binom{M}{c}e^{-p\nu}\left(-1-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{c}\\ &\quad\times\left(1+\frac{p\nu}{M}\right)^{M-c}. \end{aligned} \quad (12)$$

Then, for given p and ν, the expected values of D, S, and C are calculated as

$$\begin{aligned}\mathbb{E}[D|p,\nu] &= \sum_{d=0}^{M}d\binom{M}{d}e^{-p\nu}\left(-1+e^{\frac{p\nu}{M}}\right)^{d}\\ &= Me^{-p\nu}\left(-1+e^{\frac{p\nu}{M}}\right)\sum_{d=0}^{M}\binom{M-1}{d-1}\left(-1+e^{\frac{p\nu}{M}}\right)^{d-1}\\ &= M\left(1-e^{-\frac{p\nu}{M}}\right) \end{aligned} \quad (13)$$

$$\begin{aligned}\mathbb{E}[S|p,\nu] &= \sum_{s=0}^{M}s\binom{M}{s}\left(\frac{p\nu}{M}\right)^{s}e^{-p\nu}\left(-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{M-s}\\ &= p\nu e^{-p\nu}\sum_{s=0}^{M}\binom{M-1}{s-1}\left(\frac{p\nu}{M}\right)^{s-1}\left(-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)^{M-s}\\ &= p\nu e^{-\frac{p\nu}{M}} \end{aligned} \quad (14)$$

and

$$\mathbb{E}[C|p,\nu] = M\left(-1-\frac{p\nu}{M}+e^{\frac{p\nu}{M}}\right)e^{-\frac{p\nu}{M}} \quad (15)$$

respectively.

B. Optimal ACB Factor Maximizing Access Throughput

As explained in Section II-A, the collided preambles can be divided by two types: 1) collision-identified preambles and 2) collision-unidentified preambles. In general, the eNodeB does not need to serve those collision-identified preambles in the later RAP steps to save resource usage. Therefore, the eNodeB only needs to serve collision-unidentified preambles and the collision-free preambles. If W denotes the number of preambles to be served in the remaining RAP steps, we have

$$W = S + C_u \quad (16)$$

where S is the number of collision-free preambles and C_u is the number of collision-unidentified preambles. If θ denotes the collision-identification probability of the eNodeB, with $(1-\theta)$ probability, it fails in identifying a collided preamble that is denoted by collision unidentified. Then, we have special cases $W=S+C=D$ for $\theta=0$ and $W=S$ for $\theta=1$. When the allocable PUSCH resources are insufficient to support all the W preambles, some devices cannot receive grant for PUSCH resource at step 2 of RAP. Note that a successful access can be achieved only when a preamble is transmitted by a single device while the device also receives the PUSCH resource grant.

For the calculation of the access throughput A, i.e., the number of devices, which eventually get successful RA, we need to consider two cases: $W \le K$ and $W > K$. When $W \le K$, the eNodeB can allocate resources to all W preambles. On the other hand, when the eNodeB identifies that W preambles require resource grants at the first step, which is greater than K, i.e., $W > K$, it cannot allocate resources to all the W preambles. In this case, the eNodeB randomly selects K preambles among W preambles and allocates resources to them, while the unselected $(W-K)$ preambles are not allocated by resources.

Consequently, given p and ν, we can write $\Pr\{A=a|p,\nu\}$ as follows:

$$\begin{aligned}\Pr\{A=a|p,\nu\} &= \Pr\{A=a,W\le K|p,\nu\}\\ &\quad+\Pr\{A=a,W>K|p,\nu\}\\ &= \Pr\{S=a,W\le K|p,\nu\}\\ &\quad+\Pr\{S\ge a,W>K|p,\nu\}\mathbb{B}_a^s\left(\frac{K}{W}\right) \end{aligned} \quad (17)$$

where $\mathbb{B}_a^s(K/W) = \binom{s}{a}(K/W)^a(1-[K/W])^{s-a}$ is resulted from the eNodeB's random assignment of PUSCH grants for

K among W preambles. Then, the expected value of A is calculated by

$$\mathbb{E}[A|p,\nu] = \sum_{a=0}^{K}\sum_{w=0}^{K} a\Pr\{W=w, S=a|p,\nu\} + \sum_{a=0}^{K}\sum_{w=K+1}^{M}\sum_{s=a}^{w} a\Pr\{W=w, S=s|p,\nu\} \times \mathbb{B}_a^s\left(\frac{K}{w}\right). \tag{18}$$

In order to find the optimal ACB factor that maximizes the expected access throughput, we need to solve the following optimization problem:

$$p^* = \arg\max_{0\le p\le 1} \mathbb{E}[A|p,\nu]. \tag{19}$$

By observing (18), we can find that two cases should be discussed: $K \ge M$ and $K < M$.

First, when $K \ge M$, in the right-hand side of (18), only the first term appears and we have $\mathbb{E}[A|p,\nu] = \mathbb{E}[S|p,\nu] = p\nu e^{-(p\nu/M)}$. Then, the optimal p^* is obtained by equating the first derivative of $\mathbb{E}[A|p,\nu]$ with respect to p to be 0, i.e.,

$$p^* = \frac{M}{\nu}, \quad \text{if} \quad K \ge M \tag{20}$$

which is equivalent to the result when the constraint of PUSCH resources is ignored as in [24]. Once p^* in (20) is applied, we have

$$\mathbb{E}\left[A\middle|p^* = \frac{M}{\nu}, \nu\right] = Me^{-1}. \tag{21}$$

One interesting phenomenon is that $\mathbb{E}[A|p^*,\nu]$ is independent of ν. Hence, we can conclude that no matter what the backlog size is, we can always obtain the expected access throughput as Me^{-1}, if we use the optimal ACB factor as $p^* = (M/\nu)$. Such a nice property enables us to further derive the mathematical model to analyze the performance of R-ACB later.

Second, when $K < M$, in fact, it is difficult to find the closed-form expression for the optimal value of p^*. Hence, we approximate $\mathbb{E}[A|p,\nu]$ as

$$\mathbb{E}[A|p,\nu] \approx \mathbb{E}[S|p,\nu]\cdot\min\left\{1, \frac{K}{\mathbb{E}[W|p,\nu]}\right\}. \tag{22}$$

Underlying logic of the above approximation is that when the average number of collision-free preambles plus collision-unidentified preambles is larger than K, i.e., $\mathbb{E}[W|p,\nu] > K$, among the collision-free preambles, the portion of $\mathbb{E}[S|p,\nu]/\mathbb{E}[W|p,\nu]$ succeeds on the average. To check the closeness of the approximation, as an example, we plot Fig. 2 showing the access throughput over varying p when $K = 20$, $\nu = 80$, and $\theta = 0$. The solid line shows the access throughput of (18) while the dashed line shows its approximation in (22). As a reference, the average number of detected preambles is also plotted with dashed-dot line, which is much larger than the expected access throughput. We can observe that the approximated access throughput is quite close to the exact one. In particular, if we define $\tilde{p}$ as the ACB factor that maximizes (22), we can further observe that the difference

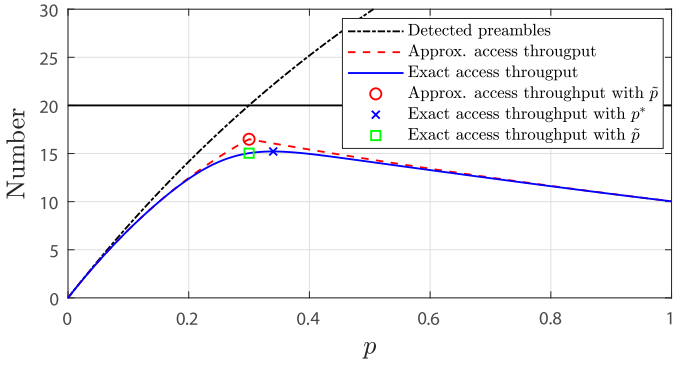


Fig. 2. Number of detected preambles, the exact access throughput, and the approximated access throughput when $K = 20$, $\nu = 80$, and $\theta = 0$.

between the expected access throughput with $\tilde{p}$, i.e., $\mathbb{E}[A|\tilde{p},\nu]$ and that with the optimal ACB factor p^*, i.e., $\mathbb{E}[A|p^*,\nu]$ is almost negligible. Therefore, we can conclude that $\tilde{p}$ is a reasonably good approximation for the optimal ACB factor p^*.

Now, we are in a position to find a closed-form expression of $\tilde{p}$ that would serve as the approximation for p^*. Plugging (14) and (15) into (22) and applying the fact that

$$\mathbb{E}[W|p,\nu] = \mathbb{E}[S|p,\nu] + (1-\theta)\mathbb{E}[C|p,\nu]$$

we can obtain

$$\mathbb{E}[A|p,\nu] \approx p\nu e^{-\frac{p\nu}{M}}\cdot\min\left\{1, \frac{K}{\chi+\theta p\nu e^{-\frac{p\nu}{M}}}\right\} \tag{23}$$

where $\chi = (1-\theta)M(1-e^{-(p\nu/M)})$. For the simplicity of the following description, we define $\hat{p}$ that satisfies $[K/(\chi+\theta p\nu e^{-(p\nu/M)})] = 1$. After some manipulation, we can obtain $\hat{p}$ as

$$\hat{p} = \Theta\frac{M}{\nu} \tag{24}$$

where

$$\Theta = \begin{cases} -\ln\left(1-\frac{K}{M}\right), & \text{if } \theta = 0 \\ \dfrac{\left\{(1-\theta)-\theta\mathcal{W}_0\left(\frac{(1-\theta)M-K}{M\theta}e^{\frac{(1-\theta)}{\theta}}\right)\right\}}{\theta}, & \text{if } 0<\theta<1 \\ -\mathcal{W}_0\left(-\frac{K}{M}\right), & \text{if } \theta = 1 \end{cases} \tag{25}$$

and $\mathcal{W}_0(x)$ represents the principle (upper) branch Lambert W function with $\mathcal{W}_0(x) > -1$ [43]. It is notable that when $K \ll M$, we have $-\ln(1-[K/M]) \approx (K/M)$. As $\chi+\theta p\nu e^{-(p\nu/M)}$ is an increasing function of p, we have

$$\mathbb{E}[A|p,\nu] \approx \begin{cases} p\nu e^{-\frac{p\nu}{M}}, & \text{if } p \le \hat{p} \\ p\nu e^{-\frac{p\nu}{M}}\cdot\frac{K}{\chi+\theta p\nu e^{-\frac{p\nu}{M}}}, & \text{if } p > \hat{p}. \end{cases} \tag{26}$$

In searching the maximum value of $\mathbb{E}[A|p,\nu]$, we have the following discussion.

1) When $p > \hat{p}$, the term $p\nu e^{-(p\nu/M)}\cdot[K/(\chi+\theta p\nu e^{-(p\nu/M)}]$ is a decreasing function of p and, hence, its maximum is achieved with $p = \hat{p}$.
2) When $p \le (M/\nu) \le \hat{p}$, the maximum of $p\nu e^{-(p\nu/M)}$ is achieved with $p = (M/\nu)$. Note that the condition of $\hat{p} \ge (M/\nu)$ is equivalent to $K \ge (1-\theta)M + Me^{-1}(2\theta-1)$, which can be derived from the definition of $\hat{p}$.

Algorithm 2 Calculation Algorithm for the ACB Factor

1: Given M preambles and K allocable PUSCH resources.
2: Estimate the backlog size ν.
3: **if** $K \geq (1-\theta)M + Me^{-1}(2\theta - 1)$ **then**
4: $\quad p = \min\{1, \frac{M}{\nu}\}$
5: **else**
6: $\quad p = \min\left\{1, \Theta\frac{M}{\nu}\right\}$.
7: **end if**

3) When $p \leq \hat{p} < (M/\nu)$, the maximum of $p\nu e^{-(p\nu/M)}$ is achieved with $p = \hat{p}$ as it is an increasing function in $[0, \hat{p}]$. Note that the condition of $\hat{p} < (M/\nu)$ is equivalent to $K < (1-\theta)M + Me^{-1}(2\theta - 1)$.

By concluding the above-mentioned three cases, we can obtain the optimal ACB factor as follows:

$$p^* \approx \tilde{p} = \begin{cases} \frac{M}{\nu}, & \text{if } (1-\theta)M + Me^{-1}(2\theta-1) \leq K < M \\ \Theta\frac{M}{\nu}, & \text{if } K < (1-\theta)M + Me^{-1}(2\theta-1). \end{cases} \tag{27}$$

For the description purpose, we define $p_1^* = (M/\nu)$ and $p_2^* = \Theta(M/\nu)$. Although $K < M$, if K is larger than or equal to $(1-\theta)M + Me^{-1}(2\theta-1)$, we can still use the optimal ACB factor by (M/ν), which is similar to the case of no upper bound on K, i.e., $K \geq M$. Then, the expression of the expected access throughput is identical to (21). On the other hand, if $p^* = p_2^*$ is applied when $K < (1-\theta)M + Me^{-1}(2\theta-1)$, the access throughput can be obtained as

$$\mathbb{E}\big[A|p^*, \nu\big] = M\Theta e^{-\Theta} \tag{28}$$

which is still independent of ν.

One more interesting observation is that when the optimal p^* in (27) is applied, the formula of $\min\{1, [K/(\chi + \theta p\nu e^{-(p\nu/M)})]\}$ in (23) is always equal to 1. In the first case of $p^* = p_1^*$, $\min\{1, [K/(\chi + \theta p\nu e^{-(p\nu/M)})]\}$ goes to $\min\{1, [K/((1-\theta)M + Me^{-1}(2\theta-1))]\}$, which is equal to 1 due to the minimum operation. In the second case of $p^* = p_2^*$, it becomes $\min\{1, (K/K)\} = 1$. Therefore, once the optimal p^* is applied, we can always write $\mathbb{E}[A|p^*, \nu] = p^*\nu e^{-(p^*\nu/M)}$. In the next Section, we will apply this property to mathematically analyze the EE of R-ACB.

By combining (20) and (27), we can derive the calculation algorithm for the ACB factor, which is shown in Algorithm 2. When $K \geq (1-\theta)M + Me^{-1}(2\theta-1)$, the optimal ACB factor is suggested as $p_1^* = (M/\nu)$ while if $K < (1-\theta)M + Me^{-1}(2\theta-1)$, it is suggested by $p_2^* = \Theta(M/\nu)$. Once the ACB factor is calculated, which is the function of the mean backlog size ν, we are ready to explain the estimation method of ν in the next subsection.

C. Estimation and Update of the Backlog Size

As we can observe from the previous subsections, the calculation of the optimal ACB factor needs the information about the backlog size ν. Therefore, in this subsection, we introduce online estimation algorithm for the backlog size. We exploit the Bayesian estimation algorithm introduced in [24], which was designed based on the number of undetected (idle) preambles. Consequently, our proposed estimation algorithm needs not to consider those collision free and collided preambles, i.e., S and C. After devices attempt RA via transmitting preambles with the ACB factor p, which is a function of ν, K, and M, the eNodeB counts the number of undetected preambles r for estimating the backlog size.

Algorithm 3 Estimation and Update Algorithm of ν

1: Initialize $\nu_0 = M$ and $k_0 = 0$ and do the following every slot.
2: **if** $p_{i-1} = p_1^*$ **then**
3: $\quad \Delta\nu = \frac{Me^{-1} - r}{1 - e^{-1}}$
4: **else if** $p_{i-1} = p_2^*$ **then**
5: $\quad \Delta\nu = \Theta\left(\frac{Me^{-\Theta} - r}{1 - e^{-\Theta}}\right)$
6: **end if**
7: $\nu_{i-1} = \nu_{i-1} + \Delta\nu$ ▷ Update of ν
8: **if** $\Delta\nu > 0$ **then**
9: $\quad k_i = k_{i-1} + 1$ and $\nu_{i-1} = \nu_{i-1} + k_i \cdot \Delta\nu$ ▷ Boosting
10: **else**
11: $\quad k_i = 0$ and $\nu_{i-1} = \nu_{i-1}$
12: **end if**
13: $\nu_i = \max(1, \nu_{i-1} - c_{i-1})$ ▷ New estimation of ν for slot i

Given the observation of r undetected preambles, we can correct ν as $\nu + \Delta\nu$ by the estimation offset [24]

$$\Delta\nu = \mathbb{E}[n|r, p, \nu] - \nu = \nu p\left(\frac{e^{-\frac{p\nu}{M}} - \frac{r}{M}}{1 - e^{-\frac{p\nu}{M}}}\right). \tag{29}$$

As we have two optimal ACB factors depending on situations, the estimation offsets with p_1^* and p_2^* are

$$\Delta\nu|_{p=p_1^*} = \frac{Me^{-1} - r}{1 - e^{-1}} \tag{30}$$

$$\Delta\nu|_{p=p_2^*} = \Theta\left(\frac{Me^{-\Theta} - r}{1 - e^{-\Theta}}\right) \tag{31}$$

where $p_1^* = (M/\nu)$ and $p_2^* = \Theta(M/\nu)$.

Algorithm 3 summarizes the estimation and update algorithm of ν by observing the number of undetected preambles r in each PRACH slot. As an initialization at time slot 0, the eNodeB first estimates ν_0 as the number of preambles M. Following [24], we also introduce a boosting factor k_i in order to consider bursty traffic. It implies that when we observe that the traffic continuously increases based on $\Delta\nu > 0$, we boost the estimation ν. In line 13, the new estimation of ν_i is calculated by $\nu_i = \nu_{i-1} - c_{i-1}$, where c_{i-1} denotes the number of devices succeed in RA on the $(i-1)$th PRACH slot. Note that lines 3 and 5 are introduced to accommodate the resource limitations in the whole RAP.

IV. Performance Analysis

As our proposed R-ACB works in a recursive way, the parameters associated the algorithm, such as the ACB factor, are adaptively changed according to the system situation, in general, it is hard to analyze the corresponding system performance. Accordingly, although there have been many ACB algorithms introduced in the literature, little work proposed mathematical models for the performance analysis. In contrast, in this article, we propose mathematical models to analyze the performance of the proposed algorithm in terms of total service time, average access delay, resource efficiency,

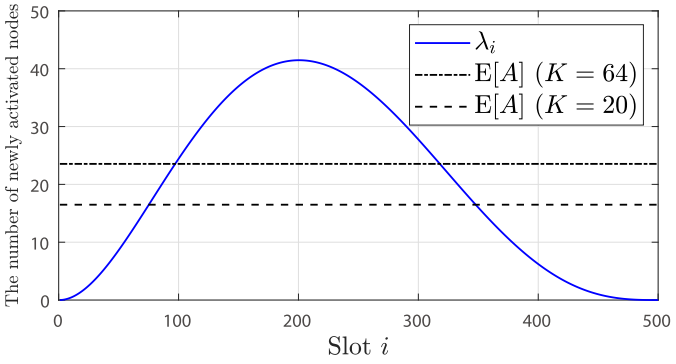


Fig. 3. Numbers of newly activated devices λ_i and $\mathbb{E}[A]$ with $K = 64$ and $K = 20$ when $N = 10000$ and $I_{\mathsf{act}} = 500$ slots.

and EE. Obviously, through the mathematical models, we are able to catch the idea of R-ACB clearly and can easily expect the system performance without complicated simulations.

A. Total Service Time

Total service time T_{service} is defined as the total spent time for all N devices to successfully complete the overall RAP. Here, N can be expressed as $N = \sum_{i=1}^{I_{\mathsf{act}}} \lambda_i$ since λ_i is the number of activated devices on the ith PRACH slot, which was defined in Section II-B. Moreover, as discussed in Section III-B, the average number of devices that succeed in RA with the optimal ACB factor, i.e., the expected access throughput, can be obtained as

$$\mathbb{E}[A] = \begin{cases} Me^{-1}, & \text{if } K \geq (1-\theta)M + Me^{-1}(2\theta - 1) \\ M\Theta e^{-\Theta}, & \text{if } K < (1-\theta)M + Me^{-1}(2\theta - 1). \end{cases}$$

Then, one can readily expect that the total service time can be approximated by $(N/\mathbb{E}[A])$ (slots). However, we should take into account the fact that $\mathbb{E}[A]$ can be achieved when the number of active devices in the system is large enough. In order to obtain more accurate total service time, we introduce δ and μ, which are, respectively, defined as the number of initial nonoverloaded slots and the marginal service time, which is needed to complete the service for the very last device. Reflecting those two parameters, the total service time in terms of PRACH slots can be approximately calculated by

$$T_{\mathsf{service}} \approx T_{\mathsf{interval}} \left\{ \delta + \left\lceil \frac{N - \sum_{i=1}^{\delta} \lambda_i}{\mathbb{E}[A]} \right\rceil \right\} + \mu \tag{32}$$

where we approximated that with R-ACB, the activated devices in the initial nonoverloaded slots can be served completely within one PRACH slot.

As one candidate method, δ can be obtained by finding the smallest slot index where the number of activated nodes λ_i is close to the expected access throughput $\mathbb{E}[A]$, i.e.,

$$\delta = \underset{i \in \{1,2,\ldots,I_{\mathsf{act}}\}}{\arg\min} |\mathbb{E}[A] - \lambda_i|. \tag{33}$$

Fig. 3 shows the number of newly activated devices over 500 slots. As shown in this figure, the number of nonoverloaded slots δ is found to be approximately 97 and 76 slots with $K = 64$ and $K = 20$, respectively.

To obtain the marginal service time μ, we first focus on the fact that when the backlog size is smaller than the following values:

$$\begin{cases} M, & \text{if } K \geq (1-\theta)M + Me^{-1}(2\theta - 1) \\ M\Theta, & \text{if } K < (1-\theta)M + Me^{-1}(2\theta - 1) \end{cases}$$

controlling of the ACB factor shows little effect as the optimal ACB factor always reaches the maximum value 1. Therefore, we take the marginal service time as the time duration to serve the last M or $M\Theta$ devices depending on cases. If a preamble collision occur, a device takes some time to recognize the preamble collision at step 4 of RAP, and then attempt RA again. Hence, we need to calculate the expected marginal service time that remaining M or $M\Theta$ active devices complete access. With $K = 64$, at the first round, among M devices, $M(1 - 1/M)^{M-1}$ devices succeed in RA on average, and by continually subtracting the average number of successful devices, we can calculate the marginal service time. For example, when $M = 64$ and $K = 64$, the backlog size is sequentially reduced as $64 \to 40 \to 18 \to 4 \to 0$ on average, and it takes four rounds.

B. Average Access Delay

Average access delay is defined as the average time difference between the access completion time and the activation time of a device, i.e.,

$$T_{\mathsf{access}} = \mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X] \tag{34}$$

where Y and X represent the random variables of the access completion time and the activation time, respectively. First, $\mathbb{E}[X]$ is calculated according to the traffic model as

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta} T_{\mathsf{act}} \tag{35}$$

where T_{act} denotes the activation duration time, and α and β represent the parameters for the Beta distribution traffic model explained in Section II-B.

To compute $\mathbb{E}[Y]$, we separately consider the nonoverloaded period of $[1, \delta]$ and the bursty period of $[\delta + 1, \delta + \lceil[(N - \sum_{i=1}^{\delta} \lambda_i)/(\mathbb{E}[A])]\rceil]$. During the nonoverloaded period, the activated devices almost complete their accesses right after the activation, i.e., the number of devices completing the RA in the ith PRACH slot is λ_i. Therefore, the summation of the access completion times for the devices activated in this period can be expressed as $\sum_{i=1}^{\delta} i\lambda_i$. On the other hand, during the bursty period, $\mathbb{E}[A]$ devices can successfully compete their access on the average in each slot. By considering this fact, we can approximate $\mathbb{E}[Y]$ by

$$\mathbb{E}[Y] \approx \frac{\sum_{i=1}^{\delta} i\lambda_i + \sum_{i=\delta+1}^{\delta + \left\lceil \frac{N - \sum_{i=1}^{\delta} \lambda_i}{\mathbb{E}[A]} \right\rceil} i\mathbb{E}[A]}{N}. \tag{36}$$

Inserting (35) and (36) into (34), we can obtain the average access delay.

C. PUSCH Resource Efficiency

PUSCH resource efficiency is defined as the ratio of the number of PUSCH resources allocated to the collision-free preambles to the total number of allocated PUSCH resources [34]. When the number of collision-free preambles plus unidentified preambles W is not larger than K, i.e., $W \leq K$, all of them can be assigned with PUSCH resource. On the other hand, if $W > K$, each preamble included in W can have the PUSCH resource assignment with the probability of (K/W). By considering this, we can obtain the PUSCH resource efficiency as

$$\begin{aligned}\eta_{\mathsf{PUSCH}} &= \sum_{w=0}^{K}\sum_{s=0}^{w}\left(\frac{s}{w}\right)\Pr\{W=w,S=s|p,\nu\} \\ &\quad+\sum_{w=K+1}^{M}\sum_{s=0}^{w}\sum_{a=0}^{s}\left(\frac{a}{K}\right) \\ &\quad\times\ \Pr\{W=w,S=s|p,\nu\}\mathbb{B}_a^s\left(\frac{K}{w}\right) \\ &= \sum_{w=0}^{M}\sum_{s=0}^{w}\left(\frac{s}{w}\right)\Pr\{W=w,S=s|p,\nu\} \\ &= \mathbb{E}\left[\frac{S}{W}|p,\nu\right] \\ &= \frac{p\nu e^{-\frac{p\nu}{M}}}{(1-\theta)M\left(1-e^{-\frac{p\nu}{M}}\right)+\theta p\nu e^{-\frac{p\nu}{M}}} \end{aligned} \tag{37}$$

where $\mathbb{B}_a^s(K/w)$ represents the probability that among the s devices who sent collision-free preambles, a devices are granted the PUSCH resources while the remaining $(s-a)$ devices are not granted the PUSCH resources. Then, the conventional scheme in [24], which was designed to have $p^* = (M/\nu)$, achieves the PUSCH resource efficiency as

$$\eta_{\mathsf{PUSCH}}^{\mathsf{conv}} = \frac{Me^{-1}}{(1-\theta)M+Me^{-1}(2\theta-1)}. \tag{38}$$

In contrast, the proposed R-ACB technique, which selects the optimal ACB factor depending on the value of K, achieves the PUSCH resource efficiency as

$$\eta_{\mathsf{PUSCH}}^{\mathsf{prop}} = \begin{cases} \frac{Me^{-1}}{(1-\theta)M+Me^{-1}(2\theta-1)} \\ \quad \text{if } K \geq (1-\theta)M+Me^{-1}(2\theta-1) \\ \frac{M\Theta e^{-\Theta}}{(1-\theta)M\left(1-e^{-\Theta}\right)+\theta M\Theta e^{-\Theta}} \\ \quad \text{if } K < (1-\theta)M+Me^{-1}(2\theta-1). \end{cases} \tag{39}$$

D. Energy Efficiency

EE is defined as the ratio of the energy consumption for transmitting a preamble (E_{S1}) at step 1 and transmitting a packet in the assigned PUSCH resource (E_{S3}) at step 3 to the average energy consumption of the whole preamble and packet transmissions for the final successful RAP, which is denoted by $\overline{E}$. Note that $E_{S1}+E_{S3}$ is the minimum energy consumption for a device to complete RAP. Then, EE can be written as

$$\eta_{\mathsf{EE}} = \frac{E_{S1}+E_{S3}}{\overline{E}}. \tag{40}$$

TABLE I
SIMULATION PARAMETERS AND VALUES

Parameters	Values
The number of preambles, M	40 and 64
The number of allocable PUSCH resources, K	$20-64$
The number of activation slots, I_{act}	500 slots
Total number of RA-attempting devices, N	10000
Traffic parameters, α and β	3 and 4
The maximum number of reattempts	10
Broadcasting period of ACB factor,	$1-100$ slots
Collision identification probability, θ	0 and 0.3

Considering the average number of RA-attempting devices $p\nu$ and the average number of collision-free preambles $\mathbb{E}[S|p,\nu]$, EE is approximated as

$$\eta_{\mathsf{EE}} \approx \frac{(E_{S1}+E_{S3})\mathbb{E}[S|p,\nu]\cdot\min\left(1,\frac{K}{\mathbb{E}[W|p,\nu]}\right)}{p\nu\left\{E_{S1}+E_{S3}\cdot\min\left(1,\frac{K}{\mathbb{E}[W|p,\nu]}\right)\right\}} \tag{41}$$

where $\min(1,\mathbb{E}[(K/W)|p,\nu])$ represents the probability that a PUSCH resource is randomly assigned to a detected preamble. The conventional scheme in [24] achieves the following EE:

$$\eta_{\mathsf{EE}}^{\mathsf{conv}} \approx \begin{cases} e^{-1} \\ \quad \text{if } K \geq (1-\theta)M+Me^{-1}(2\theta-1) \\ \frac{(E_{S1}+E_{S3})Ke^{-1}}{\left\{(1-\theta)M+Me^{-1}(2\theta-1)\right\}E_{S1}+KE_{S3}} \\ \quad \text{if } K < (1-\theta)M+Me^{-1}(2\theta-1). \end{cases} \tag{42}$$

As introduced at the end of Section III-B, R-ACB always makes $\min(1,[K/(\mathbb{E}[W|p,\nu])])$ be equal to 1 in both regions $K \geq (1-\theta)M+Me^{-1}(2\theta-1)$ and $K < (1-\theta)M+Me^{-1}(2\theta-1)$. Therefore, we have the following approximation of $\eta_{\mathsf{EE}}^{\mathsf{prop}}$

$$\eta_{\mathsf{EE}}^{\mathsf{prop}} \approx \frac{(E_{S1}+E_{S3})\mathbb{E}[S|p,\nu]}{p\nu(E_{S1}+E_{S3})} = e^{-\frac{p\nu}{M}}. \tag{43}$$

Interestingly, (43) is independent of E_{S1} and E_{S3}. In particular, with p_1^* and p_2^*, R-ACB achieves the following EE:

$$\eta_{\mathsf{EE}}^{\mathsf{prop}} \approx \begin{cases} e^{-1}, & \text{if } K \geq (1-\theta)M+Me^{-1}(2\theta-1) \\ e^{-\Theta}, & \text{if } K < (1-\theta)M+Me^{-1}(2\theta-1). \end{cases} \tag{44}$$

V. PERFORMANCE EVALUATION

For the performance evaluation, we prepared the computer simulations with MATLAB software to emulate the behavior of the RAP as well as the traffic arrivals. In order to focus on the effect of the proposed algorithm in controlling the ACB factor, we assumed that at the physical layer, each preamble and step 3 data can be perfectly decoded at the eNodeB as long as they are not collided. When a preamble is transmitted by multiple devices, we introduced the collision-identification probability θ for the eNodeB. The detailed simulation parameters are summarized in Table I. Moreover, when a statistical result is needed in the performance evaluation, we run each simulation for 10 000 iterations for meaningful results.

We assume that total 10 000 machine devices try to connect to the eNodeB through RAP. Various numbers of allocable PUSCH resources K are considered in the simulations, and

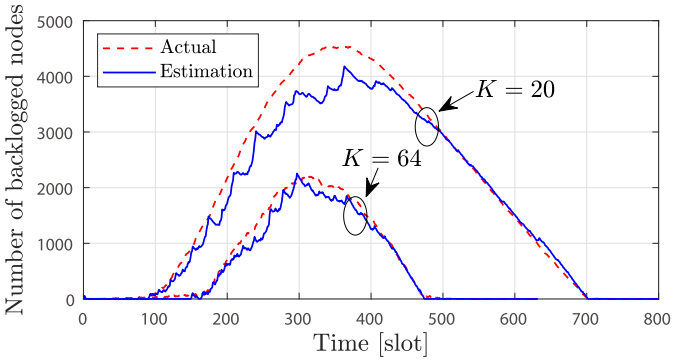


Fig. 4. Estimation of the backlog size for $N = 10000$ machine devices when $K = 20$ and $K = 64$.

we consider two cases: 1) K is constant on every PRACH slot and 2) K is varying over PRACH slots. We evaluate the performance of the proposed R-ACB in terms of total service time, average access delay, PUSCH resource efficiency, and EE. For comparison, we also evaluate the performance of the conventional ACB scheme introduced in [24] and the ideal scheme, which follows the work flow of R-ACB while it is assumed to know the backlog size exactly in each slot. It is notable that the ideal scheme gives us a guideline on the upper-bound performance of R-ACB. We will first show the performance results when the collision-identification probability θ is 0 and the eNodeB broadcasts the updated ACB factor at each PRACH slot, and then, in the last part of this Section, we will show the the performance results for $\theta > 0$ and a longer broadcasting period for updating the ACB factor.

Fig. 4 shows the Bayesian estimation result for the backlog size in the cases of sufficient PUSCH resources ($K = 64$) and insufficient PUSCH resources ($K = 20$), respectively. There exist a gap between the estimation and actual values, especially during the bursty period. However, the estimation algorithm keeps track of the actual values quite well during the traffic descent period. As the estimation is not the exact one, R-ACB may show the degraded performance compared to the ideal scheme while later we would observe the resulting performance gap is minimal.

First, let us compare the performance of the proposed R-ACB scheme to that of the existing ACB scheme of 3GPP LTE standard [37], in which the optimal ACB factor and the optimal ACB time are obtained by finding the minimum value of ACB time for a given ACB factor leading that an access success probability is higher than 0.95. Note that those optimal values are obtained with the methodology introduced in [37]. Fig. 5 compares the total service time of the existing ACB scheme and the proposed R-ACB scheme with $M = 64$ over varying K. We can observe that the proposed R-ACB shows much better performance than the existing ACB scheme in particular for the small values of K. It is mainly because R-ACB has the capability of controlling RA by considering the effect of the allocable PUSCH resources K.

Fig. 6 shows the total service time for varying K. R-ACB generates the ACB factors by considering both of the available preambles and the allocable PUSCH resources, while the conventional scheme only considers the available preambles.

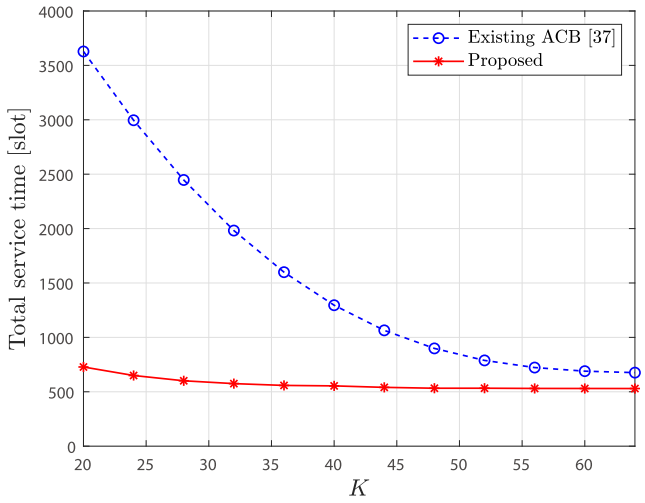


Fig. 5. Comparison between the existing ACB scheme and the proposed R-ACB scheme in terms of total service time versus K.

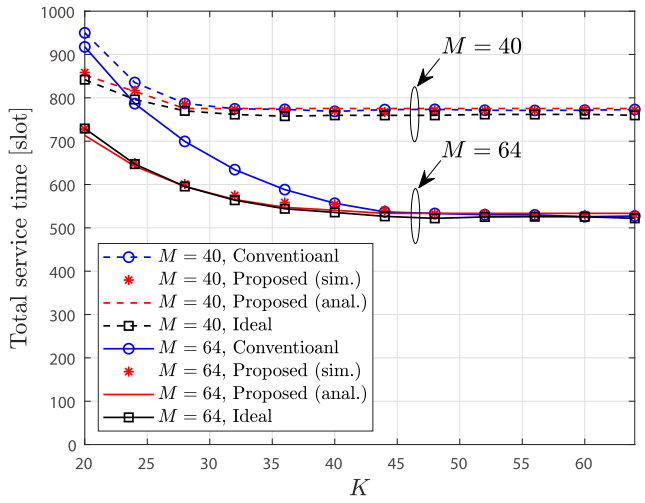


Fig. 6. Total service time versus K.

Thus, when the allocable PUSCH resources are insufficient compared to the number of detected preambles in the case of $M = 64$, e.g., $K < 40$, R-ACB takes shorter time to complete the access service, compared to the conventional scheme. In the case of $M = 40$, we can observe that the allocable PUSCH resources are insufficient when $K < 30$, and R-ACB also takes shorter time to complete the access service, compared to the conventional scheme. In terms of the total service time, R-ACB shows the similar result to that of the ideal scheme, from which we can confirm the effectiveness of R-ACB.

Fig. 7 shows the average access delay for varying K. We can observe a difference of average access delay between R-ACB and the conventional ACB scheme. When $K = 20$, R-ACB shows the average access delay of 168 slots, while the conventional ACB scheme shows the average access delay of 268 slots.

Fig. 8(a) shows the PUSCH resource efficiency for varying K. Since the conventional ACB scheme always uses the ACB factors of $p = (M/\nu)$, it shows the fixed PUSCH resource efficiency of $(-1+e)^{-1} = 0.5820$, regardless of K. However, R-ACB takes into account of the allocable PUSCH resources

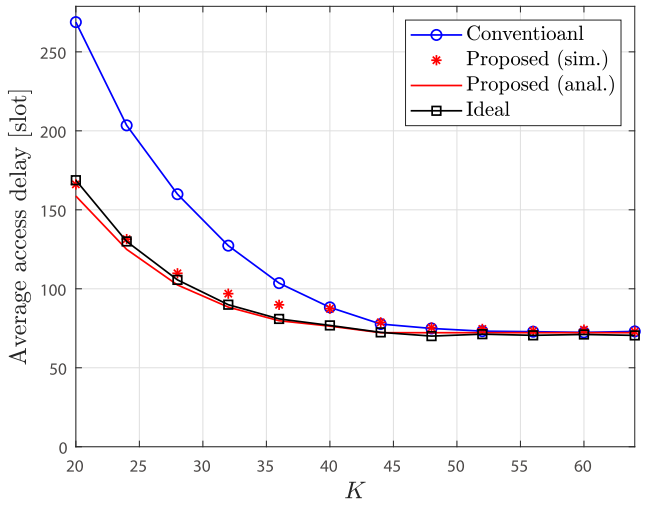


Fig. 7. Average access delay versus K.

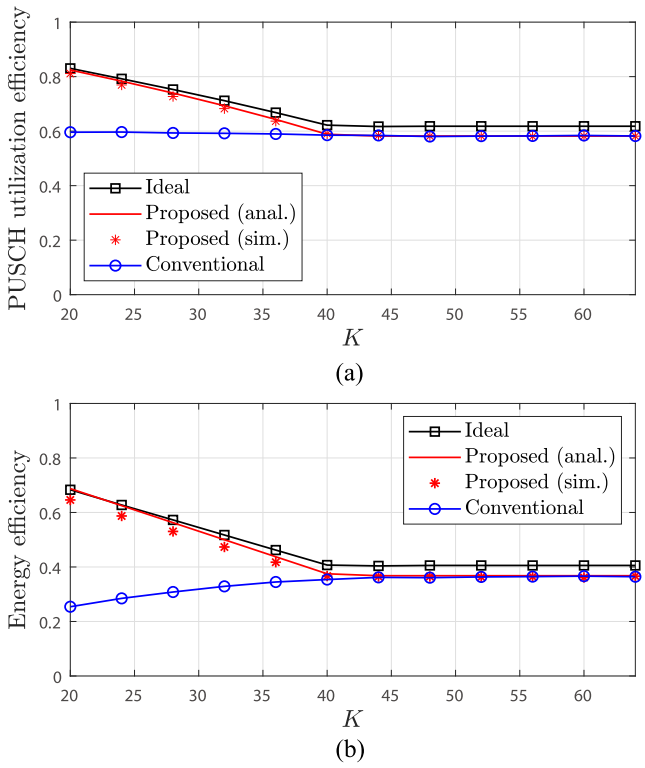


Fig. 8. PUSCH and EE versus K. (a) PUSCH resource efficiency. (b) EE when $E_{S1} = E_{S3}$.

when controlling RA, and it shows a higher PUSCH resource efficiency when PUSCH resources are insufficient ($K < 40$). While R-ACB and the conventional scheme show the same PUSCH resource efficiency with sufficient PUSCH resources ($K \geq 40$), the ideal scheme shows 30% higher efficiency due to the perfect information about the backlog size.

Fig. 8(b) shows the EE for varying K, where we assumed that a device's transmission energy for RA-step 1 and 3 is identical, i.e., $E_{S1} = E_{S3}$. The EE of R-ACB decreases from 0.65 to 0.36 approximately, while the EE of the conventional scheme increases from 0.25 to 0.36 approximately. This result implies that when designing a ACB scheme, considering the

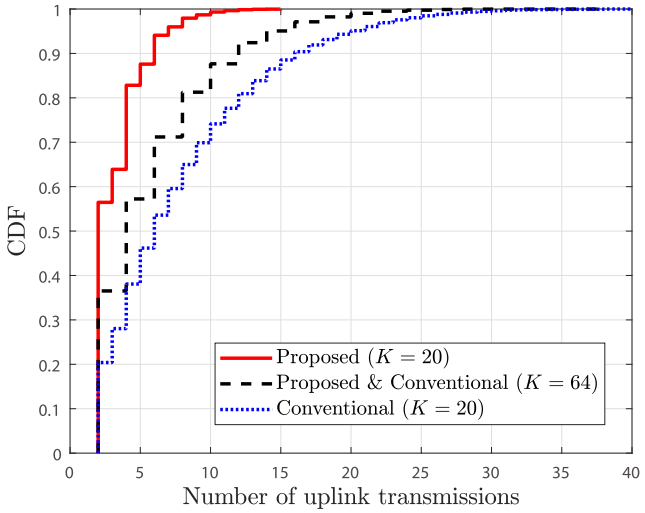


Fig. 9. CDF of the number of uplink transmissions.

allocable PUSCH resources is essential for controlling massive access from battery-powered IoT devices. Fig. 9 shows the CDF of the number of uplink transmissions (preamble transmissions plus packet transmissions) per device when $K = 20$ and $K = 64$. Basically, each device requires at least two uplink transmissions, i.e., the transmissions in the first and third steps of RAP. At the 90 percentile, R-ACB requires 6 ($K = 20$) and 12 ($K = 64$) uplink transmissions for a successful access, while the conventional scheme requires 16 ($K = 20$) and 12 ($K = 64$) uplink transmissions. Especially, more than 50% of devices can succeed in RA with a single attempt with R-ACB when $K = 20$.

From Figs. 6, 7, 8(a), and 8(b), we can observe that the analytical results obtained form equations shown in Section IV match well with the simulations results from which we can confirm the accuracy of our proposed mathematical models.

Next, we show the performance of R-ACB when the allocable PUSCH resources K vary over time.[2] As an example, we set K randomly in each slot by

$$K = \max\left(\bar{K} + \lceil G \rceil, 1\right) \tag{45}$$

where $\bar{K}$ represents the average number of PUSCH resources and G represents the Gaussian random variable with mean of 0 and standard deviation of ϵ. Figs. 10 and 11, respectively, show the average access delay and the EE, with various standard deviation of ϵ when $\bar{K} = 20$ and $\bar{K} = 30$. When ϵ increases, K varies more rapidly and, as a result, the performance of R-ACB degrades. However, it still shows much better performance than the conventional ACB scheme. For instance, when $\bar{K} = 20$ and $\epsilon = 4$, compared to the conventional ACB, R-ACB reduces the access delay by 35% and improves the EE almost by 100%.

For the last part of this section, we investigate the effect of the collision-identification probability θ, and the eNodeB's broadcasting period for updating the ACB factor. First, Fig. 12

[2]Variation of K over time is mainly caused by U, which is the maximum number of PUSCH resources at the third step of the RA procedure since the maximum number of RAR messages transmittable at step 2 (Q) is commonly fixed. In addition, we assume that $U < Q \cdot N_{\text{grant}}$ in these experiments.

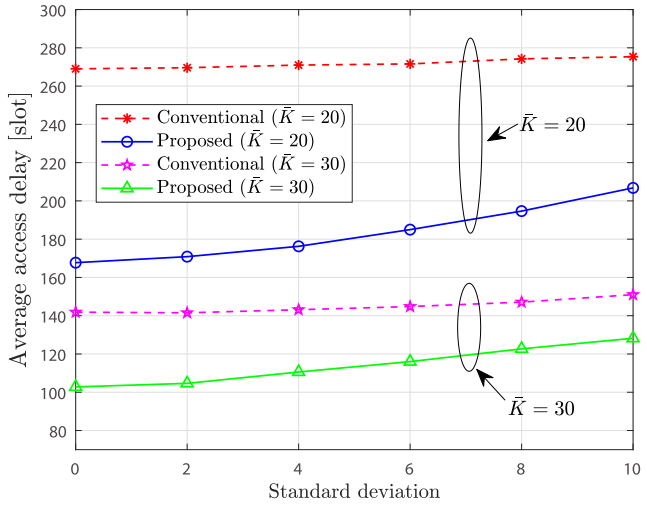


Fig. 10. Average access delay with varying PUSCH resources on each slot.

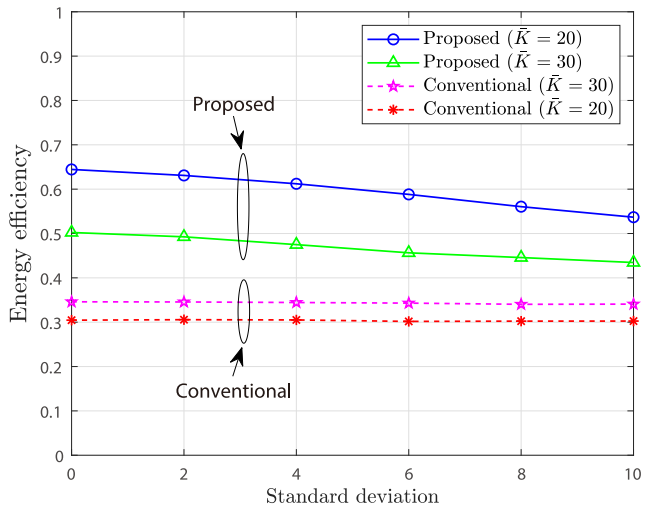


Fig. 11. EE with varying PUSCH resources on each slot.

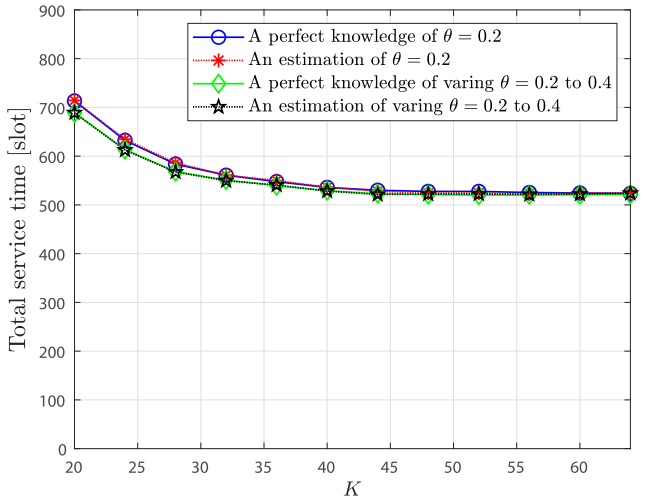


Fig. 12. Total service time with different θ.

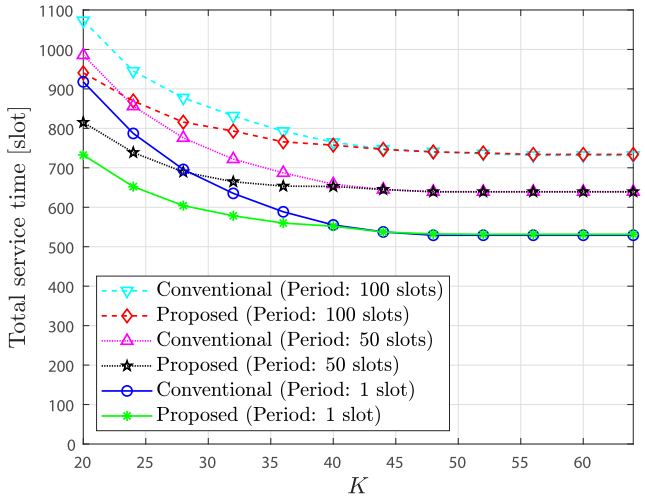


Fig. 13. Total service time with different broadcasting period of ACB factor.

shows the total service time when the collision-identification probability θ has the value of 0.2. In addition, we consider one more scenario where θ changes from 0.2 to 0.4 at the middle point of the service time. For both the scenarios, our proposed R-ACB implements Algorithm 1 to estimate the collision-identification probability. For comparison, we also plot the curves where the collision-identification probability is perfectly known to the eNodeB. From Fig. 12, we can observe that R-ACB with an estimated θ and that with a known θ show similar curves, which confirms that the proposed scheme can also work even the collision-identification varies over time. Moreover, we can also conclude that the total service time becomes smaller when the eNodeB has a better collision-identification capability, i.e., a higher collision-identification probability.

Another practical problem to be considered is the broadcasting period of ACB factor since the eNodeB may not be able to broadcast the ACB factor on every PRACH slot. Fig. 13 shows the total service time for different broadcasting periods of ACB factor 0, 50, and 100 slots. For the result shown in Fig. 13, Algorithm 2 is invoked every broadcasting period. As the broadcasting period increases, the total service time increases since an identical ACB factor is applied within the broadcasting period, which causes a looser access control. However, the broadcasting period of ACB factor is a system parameter selected by network operators, and the result shows that the proposed R-ACB can adapt to any broadcasting periods of ACB factor.

VI. Conclusion

In this article, we proposed a resource-optimized R-ACB technique to effectively accommodate bursty traffic in massive cellular IoT networks. We exploited Bayesian algorithm to estimate the number of active devices in each slot and derived the optimal ACB factor that maximizes the access throughput while the resource limitation possibly occurred in the whole RAP is considered. Furthermore, when obtaining the optimal ACB factor, the possibility of partially identifying collisions at step 1 of RAP was also considered. Through extensive computer simulations, we evaluated the performance of the proposed R-ACB in terms of total service time, average

access delay, PUSCH resource efficiency, and EE. It is shown that R-ACB outperforms the conventional ACB scheme especially when the allocable PUSCH resources are not sufficient. In addition, R-ACB technique yields higher EE than the conventional ACB scheme, where the EE is significantly important for battery-powered IoT devices. We also proposed mathematical models to analyze the performance of R-ACB and showed the accuracy of the models by comparing with the simulations results.

REFERENCES

[1] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[2] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani, "Internet-of-Things-based smart cities: Recent advances and challenges," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 16–24, Sep. 2017.

[3] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, "Congestion control for bursty M2M traffic in LTE networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 5815–5820.

[4] S. Sesia, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*. Chichester, U.K.: Wiley, 2009.

[5] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. New York, NY, USA: Academic, 2013.

[6] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-advanced random-access mechanism for massive machine-to-machine (M2M) communications," in *Proc. 27th World Wireless Res. Forum (WWRF) Meeting*, Dusseldorf, Germany, 2011, pp. 1–5.

[7] R. C. D. Paiva, R. D. Vieira, and M. Saily, "Random access capacity evaluation with synchronized MTC users over wireless networks," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, Budapest, Hungary, 2011, pp. 1–5.

[8] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[9] G.-Y. Lin and H.-Y. Wei, "Auction-based random access load control for time-dependent machine-to-machine communications," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 658–672, Oct. 2016.

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

[11] *Study on RAN Improvements for Machine-Type Communications*, 3GPP Standard TR 37.868 V11.0.0, Oct. 2011.

[12] 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 Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.

[13] J.-P. Cheng, C.-H. Lee, and T.-M. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *Proc. IEEE GLOBECOM Workshops*, Houston, TX, USA, Dec. 2011, pp. 368–372.

[14] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 27–32, Jan. 2012.

[15] S. Duan, V. Shah-Mansouri, and V. W. S. Wong, "Dynamic access class barring for M2M communications in LTE networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 4747–4752.

[16] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, Jun. 2014.

[17] H. He, Q. Du, H. Song, W. Li, Y. Wang, and P. Ren, "Traffic-aware ACB scheme for massive access in machine-to-machine networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 617–622.

[18] C.-H. Wei, G. Bianchi, and R.-G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.

[19] R.-G. Cheng, J. Chen, D.-W. Chen, and C.-H. Wei, "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956–2968, Jun. 2015.

[20] C.-Y. Oh, D. Hwang, and T.-J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4182–4192, Aug. 2015.

[21] Z. Wang and V. W. S. Wong, "Joint access class barring and timing advance model for machine-type communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 2357–2362.

[22] Z. Wang and V. W. S. Wong, "Optimal access class barring for stationary machine type communication devices with timing advance information," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5374–5387, Oct. 2015.

[23] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[24] H. Jin, W. T. Toor, B. C. Jung, and J.-B. Seo, "Recursive pseudo-Bayesian access class barring for M2M communications in LTE systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, Sep. 2017.

[25] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, Dec. 2017.

[26] H. S. Jang, B. C. Jung, and D. K. Sung, "Dynamic access control with resource limitation for group paging-based cellular IoT systems," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5065–5075, Dec. 2018.

[27] D. T. Wiriaatmadja and K. W. Choi, "Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 33–46, Jan. 2015.

[28] P. Osti, P. Lassila, S. Aalto, A. Larmo, and T. Tirronen, "Analysis of PDCCH performance for M2M traffic in LTE," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4357–4371, Nov. 2014.

[29] T. P. de Andrade, C. A. Astudillo, and N. L. da Fonseca, "Allocation of control resources for machine-to-machine and human-to-human communications over LTE/LTE-A networks," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 366–377, Jun. 2016.

[30] T. P. C. de Andrade, C. A. Astudillo, L. R. Sekijima, and N. L. da Fonseca, "The random access procedure in long term evolution networks for the Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 124–131, Mar. 2017.

[31] H. S. Jang, H.-S. Park, and D. K. Sung, "A non-orthogonal resource allocation scheme in spatial group based random access for cellular M2M communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4496–4500, May 2017.

[32] H. S. Jang, S. M. Kim, H.-S. Park, and D. K. Sung, "An early preamble collision detection scheme based on tagged preambles for cellular M2M random access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 5974–5984, Jul. 2017.

[33] H. S. Jang, S. M. Kim, H.-S. Park, and D. K. Sung, "A preamble collision resolution scheme via tagged preambles for cellular IoT/M2M communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1825–1829, Feb. 2018.

[34] M. Vilgelm, S. R. Liñares, and W. Kellerer, "On the resource consumption of M2M random access: Efficiency and Pareto optimality," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 709–712, Jun. 2019.

[35] *Medium Access Control (MAC) Protocol Specification (Release 14)*, 3GPP Standard TS 36.321 V14.6.0, Mar. 2018.

[36] *Radio resource control (RRC), Protocol Specification*, 3GPP Standard TS 36.331, Sep. 2017.

[37] L. Tello-Oquendo *et al.*, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3505–3520, Apr. 2018.

[38] H. S. Jang, H. Jin, B. C. Jung, and T. Q. S. Quek, "Versatile access control for massive IoT: Throughput, latency, and energy efficiency," *IEEE Trans. Mobile Comput.*, vol. 19, no. 8, pp. 1984–1997, Aug. 2020.

[39] "[70bis#11]-LTE: MTC LTE simulations," in *3GPP TSG-RAN WG2 Meeting# 71 R2-104663*, 3GPP, Sophia Antipolis, France, Aug. 2010.

[40] R. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. 33, no. 3, pp. 323–328, May 1987.

[41] H. Jin, J.-B. Seo, and V. C. M. Leung, "Cooperative pseudo-bayesian backoff algorithms for unsaturated CSMA systems with multi-packet reception," *IEEE Trans. Mobile Comput.*, vol. 14, no. 2, pp. 302–315, Feb. 2015.

[42] W. Feller, *An Introduction to Probability Theory and Its Application*. New York, NY, USA: Wiley, 1968.

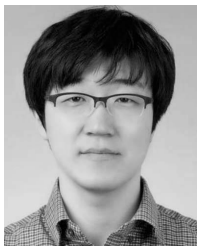
[43] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the LambertW function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.

Han Seung Jang (Member, IEEE) received the B.S. degree in electronics and computer engineering from Chonnam National University, Yeosu, South Korea, in 2012, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute for Science and Technology, Daejeon, South Korea, in 2014 and 2017, respectively.

From May 2018 to February 2019, he was a Postdoctoral Fellow with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore. He was a Postdoctoral Fellow with Chungnam National University, Daejeon, from September 2017 to April 2018. He is currently an Assistant Professor with the School of Electrical, Electronic Communication, and Computer Engineering, Chonnam National University. His research interests include cellular Internet-of-Things/machine-to-machine communications, machine learning, smart grid, and energy ICT.

Hu Jin (Senior Member, IEEE) received the B.E. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2006 and 2011, respectively.

From 2011 to 2013, he was a Postdoctoral Fellow with the University of British Columbia, Vancouver, BC, Canada. From 2013 to 2014, he was a Research Professor with Gyeongsang National University, Tongyeong, South Korea. Since 2014, he has been with the Division of Electrical Engineering, Hanyang University, Ansan, South Korea, where he is currently an Associate Professor. His research interests include medium-access control and radio resource management for random access networks and scheduling systems considering advanced signal processing and queueing performance.



Bang Chul Jung (Senior Member, IEEE) received the B.S. degree in electronics engineering from Ajou University, Suwon, South Korea, in 2002, and the M.S. and Ph.D. degrees in electrical and computer engineering from Korea Advanced Institute for Science and Technology (KAIST), Daejeon, South Korea, in 2004 and 2008, respectively.

He was a Senior Researcher/Research Professor with KAIST Institute for Information Technology Convergence, Daejeon, from January 2009 to February 2010. From March 2010 to August 2015, he was a Faculty of Gyeongsang National University, Tongyeong, South Korea. He is currently an Associate Professor with the Department of Electronics Engineering, Chungnam National University, Daejeon. His research interests include wireless communication systems, IoT communications, statistical signal processing, information theory, interference management, random access, radio resource management, cooperative relaying techniques, in-network computation, and mobile computing.

Dr. Jung was a recipient of the Fifth IEEE Communication Society Asia-Pacific Outstanding Young Researcher Award in 2011 and he has been selected as a winner of Haedong Young Scholar Award in 2015, which is sponsored by the Haedong foundation and given by Korea Institute of Communication and Information Science, the Bronze Prize of Intel Student Paper Contest in 2005, the First Prize of KAIST's Invention Idea Contest in 2008, the Bronze Prize of Samsung Humantech Paper Contest in 2009, the Best Paper Award of Spring Conference of Korea Institute of Information and Communication Engineering in 2015, the Best Paper Awards of the Institute of Electronics and Information Engineering Symposium in 2017/2018, the Best Paper Award of KICS Journal in 2018, and several Best Paper Awards of KICS Conferences in 2017/2018. He has been serving as an Associate Editor of *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences* since 2018.

Tony Q. S. Quek (Fellow, IEEE) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008.

He is currently a Cheng Tsang Man Chair Professor with the Singapore University of Technology and Design (SUTD), Singapore. He also serves as the Head of Information Systems Technology and Design Pillar, Singapore, Sector Lead of the SUTD AI Program, and the Deputy Director of the SUTD-ZJU IDEA. His current research topics include wireless communications and networking, network intelligence, IoT, URLLC, and big data processing.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards—Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, the 2020 Nokia Visiting Professor, and the 2016–2020 Clarivate Analytics Highly Cited Researcher. He has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is currently serving as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and an elected member of the IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS. He is a Distinguished Lecturer of the IEEE Communications Society.