Performance Analysis and Optimal Access Class Barring Parameter Configuration in LTE-A Networks with Massive M2M Traffic

Luis Tello-Oquendo, Israel Leyva-Mayorga, Vicent Pla, Jorge Martinez-Bauset, José-Ramón Vidal, Vicente Casares-Giner, and Luis Guijarro

***Abstract*—Over the coming years, it is expected that the number of machine-to-machine (M2M) devices that communicate through LTE-A networks will rise significantly for providing ubiquitous information and services. However, LTE-A was devised to handle human-to-human traffic, and its current design is not capable of handling massive M2M communications. Access class barring (ACB) is a congestion control scheme included in the LTE-A standard that aims to spread the accesses of user equipments (UEs) through time so that the signaling capabilities of the evolved Node B (eNB) are not exceeded. Notwithstanding its relevance, the potential benefits of the implementation of ACB are rarely analyzed accurately. In this paper, we conduct a thorough performance analysis of the LTE-A random access channel (RACH) and ACB as defined in the 3GPP specifications. Specifically, we seek to enhance the performance of LTE-A in massive M2M scenarios by modifying certain configuration parameters and by the implementation of ACB. We observed that ACB is appropriate for handling sporadic periods of congestion. Concretely, our results reflect that the access success probability of M2M UEs in the most extreme test scenario suggested by the 3GPP improves from approximately 30%, without any congestion control scheme, to 100% by implementing ACB and setting its configuration parameters properly.**

***Index Terms*—Access class barring (ACB); cellular-systems; machine-to-machine communications; performance analysis.**

I. Introduction

THE world is moving beyond standalone devices into a new technological age in which everything is connected. Machine-to-Machine (M2M) communication stands for the ubiquitous and automated exchange of information between devices on the edge of networks such as mobile devices, computers, sensors, actuators or cars inside a common network, the so-called Internet-of-Things (IoT). Recognizing the value of the IoT to the industry and the benefits this technological innovation brings to the public, enormous efforts are being made towards its standardization, which includes the development of projects and the organization of events that are directly related to create the environment needed for a vibrant IoT [1]. In coming years, a massive number of interconnected devices will provide ubiquitous access to information and services [2], [3]. These devices, known as user equipments (UEs), are set to exchange information autonomously in M2M applications such as smart metering, e-healthcare, smart transportation, environmental monitoring, among others [4]–[6]. In those scenarios, the network congestion is expected to occur sparingly over time whenever a bulk of UEs transmit in a highly synchronized manner. There is a growing consensus that cellular networks are the best option for UE interconnection, as they provide ubiquitous coverage thanks to a widely deployed infrastructure, global connectivity, high QoS, well-developed charging and security solutions [7]–[9]. Nevertheless, cellular technology was developed to handle human-to-human (H2H) traffic, where few devices (compared to the billions of M2M devices expected by 2020 [3]) communicate simultaneously. Hence, severe congestion is likely to occur when a massive number of M2M devices attempt to access the base stations (known as evolved Node Bs, eNBs, in LTE-A), resulting in performance degradation for both M2M and H2H communications [10], [11].

Recent studies have demonstrated that the current random access procedure deployed in LTE-A networks is not efficient enough for managing massive M2M communications because the random access channel (RACH) suffers from overload in these scenarios [12], [13]. Building on this, the access class barring (ACB) scheme has been included in the LTE-A radio resource control specification [14] as a viable congestion control scheme. In ACB, each UE may randomly delay the beginning of its random access procedure according to a barring rate and a barring time, which are parameters broadcast by the eNB. As a result, ACB spreads the UE accesses through time; hence, ACB may be effective whenever the congestion occurs sparingly and during short periods (in the order of a few seconds). This fact goes in line with the M2M bursty traffic behavior described in [15], [16].

In this paper, we perform a thorough performance analysis of both the LTE-A random access procedure and the ACB congestion control scheme in scenarios with a massive number of M2M UEs that attempt to access the eNB in a highly synchronized manner. Specifically, the main contributions of this article are:

1) Analysis of the steady-state capacity of the LTE-A physical RACH.
2) The identification of the combinations of RACH parameters that enhance the access success probability in scenarios with massive M2M traffic.

Manuscript received xxxx yy, zzzz; revised xxxx yy, zzzz. This research has been supported in part by the Ministry of Economy and Competitiveness of Spain under Grants TIN2013-47272-C2-1-R and TEC2015-71932-REDT. The research of L. Tello-Oquendo was supported in part by Programa de Ayudas de Investigación y Desarrollo (PAID) of the Universitat Politècnica de València. The research of I. Leyva-Mayorga was partially funded by grant 383936 CONACYT-Gobierno del Estado de México 2014.

The authors are with the ITACA Institute. Universitat Politècnica de València, Camino de Vera s/n. 46022 Valencia, Spain (e-mail: {luiteloq, isleyma, vpla, jmartinez, jrvidal, vcasares, lguijar}@upv.es).

3) A thorough analysis of the ACB scheme for properly tuning its parameters according to the network load. We evaluate the performance of LTE-A under the ACB scheme for a wide range of barring rates and barring times. Furthermore, we identify the optimal parameter configuration of ACB for the most congested scenario suggested by the 3GPP [16].
4) The comparison of the KPIs obtained for two possible backoff implementations at UE side:
   a) a uniform backoff (as stated in the LTE-A MAC specification [17]);
   b) an exponential backoff, where the backoff time of each UE depends on the number of transmissions attempted previously.
5) The comparison of the access success probability obtained for two collision models for the LTE-A random access procedure:
   a) Collision model 1: collisions occur only at the transmission of *Msg1*;
   b) Collision model 2: collisions occur only at the transmission of *Msg3*.

   Please refer to Section III for more specific details of the random access procedure.

During this study, we closely follow the 3GPP recommendations, as we have identified that the behavior of ACB is oftentimes misinterpreted [18]. Specifically, we have observed that most studies analyzing the performance of ACB assume a fixed barring time, whereas the 3GPP specifies that this parameter is selected randomly for each barring check (process in which the UE determines its barring status, please refer to Section III-C for specific details of ACB and the barring checks) [14], [19]. Hence, our study is one of the few that evaluates the ACB performance with a randomly selected barring time.

The rest of the paper is organized as follows. In Section II, we conduct a review of the literature regarding the performance analysis of LTE-A and ACB. Then, we describe the random access in LTE-A, the physical RACH capacity, and the ACB scheme in Section III. The selected traffic model, the configuration parameters, and the performance metrics for the RACH evaluation are presented in Section IV. Our most relevant results including the performance analysis of LTE-A and ACB are shown in Sections V and VI, respectively. Finally, we present our conclusions in Section VII.

II. Related Work

The complexity of the random access procedure and the wide variety of configuration parameters make it challenging to evaluate the performance of LTE-A under M2M traffic. For instance, there is no consensus regarding the moment of the random access procedure in which collisions occur. It is oftentimes assumed that all the collisions occur at the first step of the random access procedure [20]–[22] (at the transmission of *Msg1* as suggested by the 3GPP in [16]). But studies such as [23]–[26], assume that all the collisions occur at the transmission of *Msg3* (the random access procedure will be explained in detail in Section III). It is evident that the performance of LTE-A can be affected by these assumptions, but no study has yet compared them directly. However, regardless of the assumed outcome of the random access procedure, it has been demonstrated that, in its current form, it is not capable of handling massive M2M communications [13], [15], [16], [23], [24].

The 3GPP has provided a list of parameters which describe a typical configuration for the RACH and serve as initial guidelines for its performance analysis [16]. But commonly the performance of the LTE-A RACH is only evaluated with this particular configuration. Hence, the impact of parameters such as the backoff time of UEs and the maximum number of preamble transmissions allowed per UE on the network performance have been largely overlooked. Such is the case of [15], where a thorough mathematical analysis of the random access procedure is performed. Specifically, the authors assess the performance of LTE-A when a bulk of UEs attempt to access the eNB in a highly synchronized manner (as expected in most M2M applications) and obtain several KPIs specified by the 3GPP; however, only the typical RACH configuration is evaluated.

In [27], authors define the capacity of the physical random access channel (PRACH), $c(R)$, as the maximum expected number of successful UE access requests per random access opportunity (RAO), being R the number of available preambles in the system, and propose a dynamic congestion-control solution. The performance of this solution is compared with the implementation of ACB. However, since the ACB analysis is performed for a very limited selection of barring rates and barring times, the advantages of the proposed solution are magnified. Furthermore, the authors assume a constant barring time for all ACB checks, whereas the 3GPP states in [19] that the barring time is calculated randomly for each ACB check. The use of a constant barring time reduces the performance of ACB. The latter is a common problem in ACB analysis which is also present in [21], where a dynamic approach for selecting the optimal barring rate is presented. Here, authors select a constant barring time of one access opportunity, which highly differs from the protocol specification [14], [19]. Besides, it is assumed that the eNB is capable of updating and broadcasting the optimal barring rate at the beginning of each access opportunity, which is clearly not possible because the updating period of the system information blocks is much longer.

The implementation of a static barring scheme affects the access delay of every UE, even in cases of no congestion, when the scheme is not needed at all. In these cases, the dynamic adaptation of barring parameters may be desirable, but its implementation is not straightforward. Specifically, the activation and deactivation of dynamic barring schemes are based on the collection of network congestion statistics (such as the ratio of transmitted preambles to successful accesses), which are dramatically altered whenever the barring scheme is active [21], [29]. This fact, in combination with the lack of knowledge regarding the behavior of ACB, makes it extremely tough to develop an effective adaptive ACB scheme. As such, in this study, we focus on the performance analysis of an ACB scheme whose barring parameters remain static for the entire period in which the accesses of the UEs to the eNB are studied. A major difference with many other studies is that we evaluate the performance of the ACB scheme considering that its parameters can take any values within the whole range suggested by the 3GPP, avoiding the restriction of these

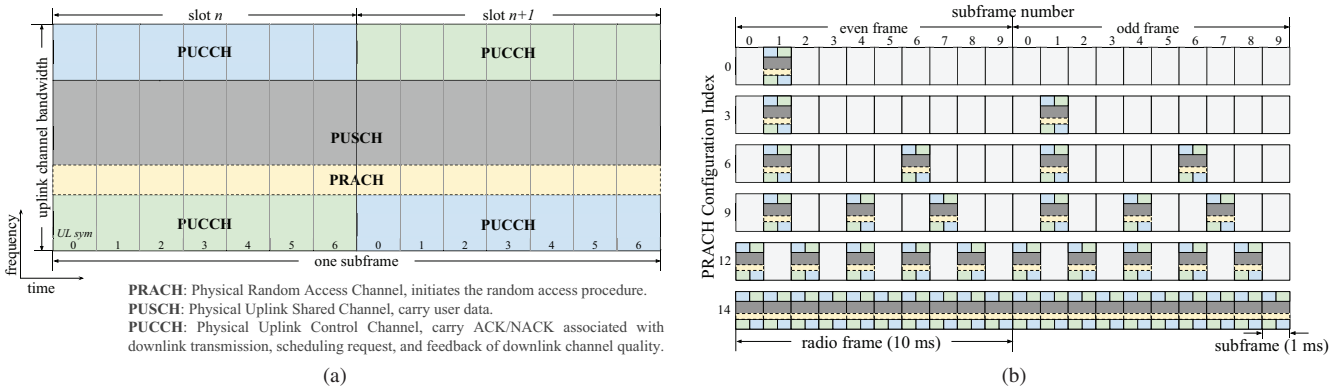


Figure 1. Resource allocation in a random access cycle. (a) Physical uplink resources for initial transmission. (b) Examples of six PRACH configurations, determined by *prach-ConfigIndex*; frame structure type 1 [28].

parameters only to the typical ones. This approach provides us with a wider perspective of the operation of the ACB scheme and enables the selection of optimal parameter configurations.

III. Random Access in LTE-A

This section provides a general overview of the random access procedure in LTE-A networks. Two modes are defined for the random access: contention-free and contention-based. The former is used for critical situations such as handover, downlink data arrival or positioning. The latter is the standard mode for network access; it is employed by UEs to change the radio resource control state from idle to connected, to recover from a radio link failure, to perform uplink synchronization or to send scheduling requests [17].

Random access attempts of UEs are allowed in predefined time/frequency resources herein called RAOs. Two uplink channels are required; namely, the physical random access channel (PRACH) for preamble transmission and the physical uplink shared channel (PUSCH) for data transmission, see Fig. 1a. The PRACH is used to signal a connection request when a UE attempts to access the cellular network. In the frequency domain, the PRACH is designed to fit in the same bandwidth as six resource blocks of normal uplink transmission (6×180 kHz); this fact makes it easy to schedule gaps in normal uplink transmission to allow for RAOs. In the time domain, the periodicity of the RAOs is determined by the parameter *prach-ConfigIndex*, provided by the eNB; a total of 64 PRACH configurations are available, Fig. 1b illustrates some examples [28]. Thus, the periodicity of the RAOs ranges from a minimum of 1 RAO every two frames to a maximum of 1 RAO every subframe, i.e., from 1 RAO every 20 ms to 1 RAO every 1 ms [13], [30], [14], [28].

As mentioned before, the PRACH carries a preamble (signature) for initial access to the network; up to $R = 64$ orthogonal preambles are available per cell. In the contention-free mode, collision is avoided through the coordinated assignment of preambles, but eNBs can only assign these preambles during specific slots to specific UEs. In the contention-based mode, preambles are selected in a random fashion by the UEs, so there is a risk of collision, i.e., multiple UEs in the cell might pick the same preamble signature in the same RAO; therefore, contention resolution is needed. In the sequel, we focus on the analysis of the contention-based random access procedure.

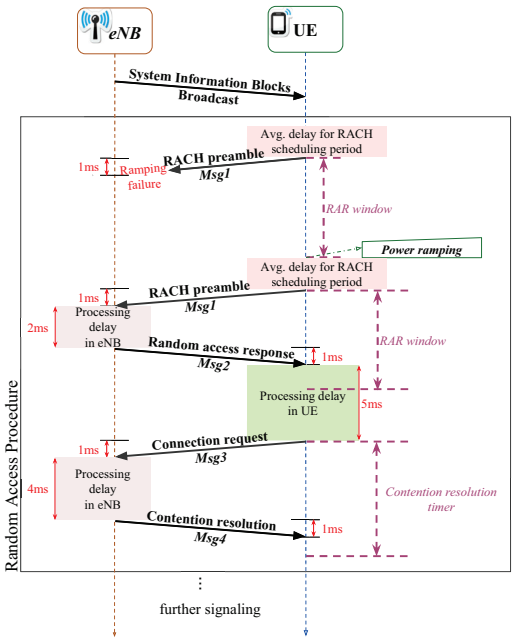


Figure 2. LTE-A contention-based random access procedure.

A. Contention-Based Random Access Procedure

Before initiating the random access procedure, the UEs must first obtain some basic configuration parameters such as the slots in which the transmission of preambles is allowed (RAOs). The eNB broadcasts this information periodically through *Master Information Block* (*MIB*) and *System Information Blocks* (*SIBs*). Once the UE has acquired this information, it may proceed with the four-message handshake illustrated in Fig. 2. Next, we describe the four-message handshake of the contention-based random access and the backoff procedures. The interested reader is referred to [14], [17], [31], [32] for further details.

RACH preamble (Msg1): Whenever a UE attempts transmission, it sends a randomly chosen preamble in a RAO, *Msg1*. Due to the orthogonality of the different preambles, multiple UEs can access the eNB in the same RAO, using different preambles. The eNB can, without a doubt, decode a preamble transmitted (with sufficient power) by exactly one

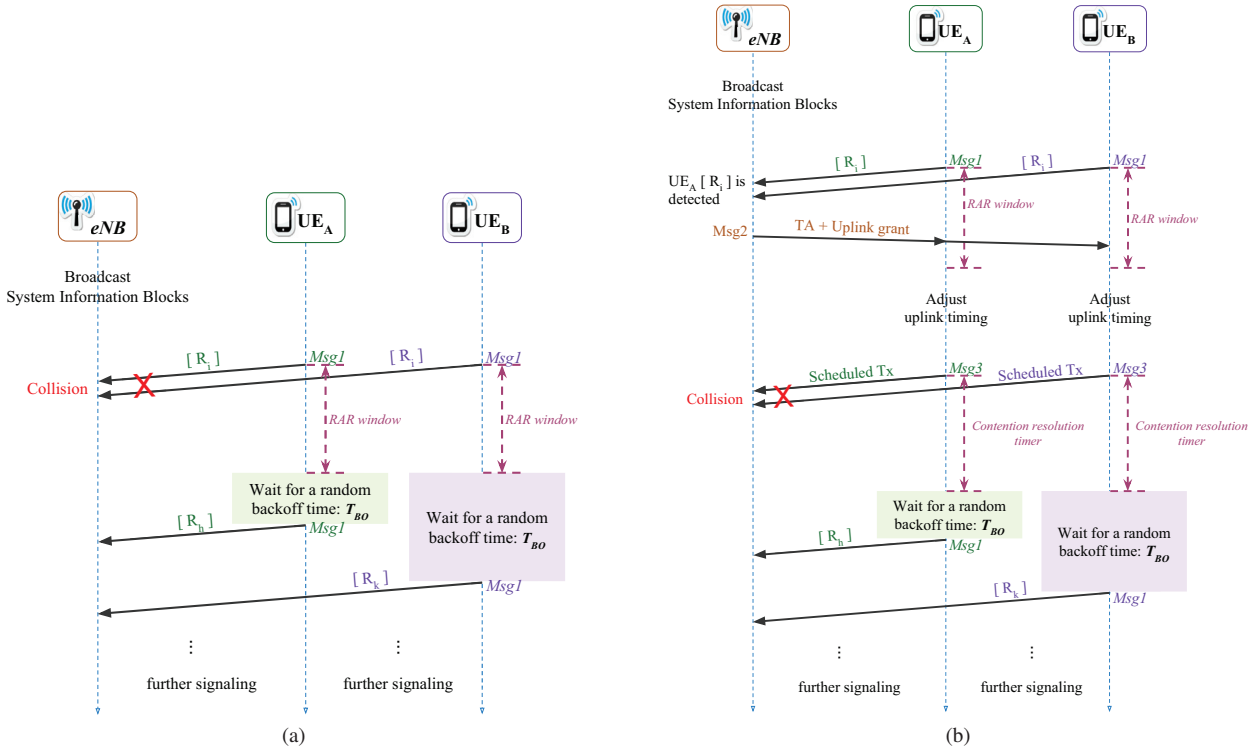


Figure 3. Collision outcomes in the LTE-A contention-based random access procedure. (a) Collision at the transmission of *Msg1*. (b) Collision at the transmission of *Msg3*. R_i, R_h, and R_k are the preambles transmitted at the ith, hth, and kth RAO, respectively, *TA* represents the time alignment provided by the eNB, and *Uplink grant* is the reserved physical uplink shared channel (PUSCH) resources for *Msg3* transmission.

UE and estimate the transmission timing of the terminal. However, if two or more UEs transmit the same preamble, two outcomes are possible: in the first one, the transmitted preamble cannot be decoded by the eNB, i.e., a collision occurs at the transmission of *Msg1* (see Fig. 3a) and, in the second one, the transmitted preambles are correctly decoded by the eNB. The main reason behind this second outcome is that the received power from one of the transmitted preambles may be much higher than the others (capture effect [33] whose quantitative evaluation is out of the scope of this study); hence, the different signals may appear as a single transmission going through multiple fading paths. The preamble transmission may also fail because the UE is too far away from the eNB (insufficient transmission power).

Random access response (Msg2): The eNB computes an identifier for each successfully decoded preamble, ID $= f(\text{preamble}, \text{RAO})$, and sends the random access response (RAR) *Msg2* through the physical downlink control channel (PDCCH). It includes, among other data, information about the identification of the detected preamble (ID), time alignment (TA), uplink grants (reserved PUSCH resources) for the transmission of *Msg3*, the backoff indicator (BI), and the assignment of a temporary identifier.

Exactly two subframes after the preamble transmission has ended (this is the time needed by the eNB to process the received preambles), the UE begins to wait for a time window, RAR window (W_{RAR}), to receive an uplink grant from the eNB.

There can be up to one RAR message in each subframe, but it may contain up to three uplink grants. Each uplink grant is associated to a successfully decoded preamble. The length of the W_{RAR} (in subframes) is broadcast by the eNB through the SIB Type 2 (SIB2) [14]. Hence, there is a maximum number of uplink grants that can be sent within the W_{RAR}. Only the UEs that receive an uplink grant can transmit the *Msg3*. In case the eNB is not capable of decoding the preambles transmitted by multiple UEs, these UEs will not receive an uplink grant (failed UEs).

Connection request (Msg3): After receiving the corresponding RAR, the UE adjusts its uplink transmission time according to the received TA and transmits a scheduled message, *Msg3*, to the eNB using the reserved PUSCH resources; hybrid automatic repeat request (HARQ) is used to protect the message transmission. Recall that, if the eNB correctly decoded the preambles transmitted by multiple UEs, these UEs will transmit their *Msg3* over the same physical resources, thus generating a collision at this point (see Fig. 3b). Therefore, the eNB will not be able to decode the transmitted messages.

Contention resolution (Msg4): The eNB transmits *Msg4* as an answer to *Msg3*. The eNB also applies an HARQ process to send *Msg4* back to the UEs. If a UE does not receive *Msg4* within the contention resolution timer, then it declares a failure in the contention resolution and schedules a new access attempt according the considerations detailed in the next paragraph.

If an access failure occurs at any of the steps previously described (due to insufficient transmission power or to a collision or to the expiration of the contention resolution timer), then the failed UEs ramp up their power and re-transmit a new randomly chosen preamble in a new RAO, based on a

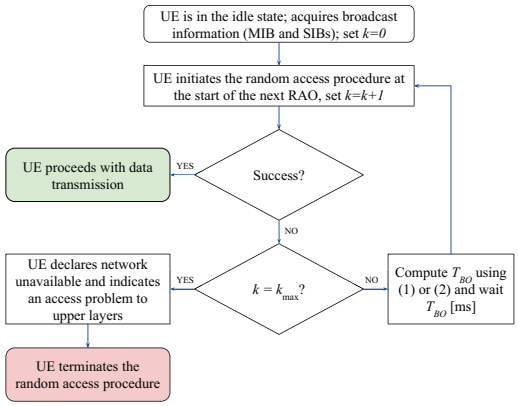


Figure 4. Backoff procedure performed by the failed UEs.

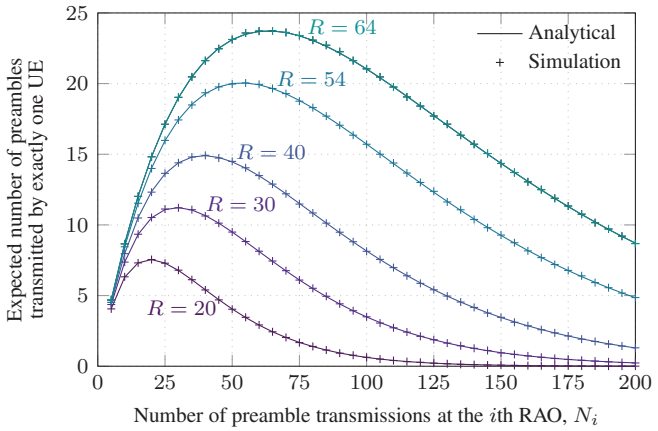


Figure 5. Expected number of preambles selected by exactly one UE at the ith RAO for the given number of available preambles, R, and the number of preamble transmissions, N_i [27, Fig. 3].

uniform backoff scheme (explained next) that uses the BI. Note that each UE keeps track of its preamble transmissions. When a UE has transmitted a certain number of preambles without success, *preambleTransMax* (notified by the eNB through a SIB), the network is declared unavailable by the UE, a random access problem is indicated to upper layers, and the random access procedure is terminated.

Backoff procedure: According to the LTE-A standard [17], if the random access attempt of a UE fails, regardless of the cause, the UE has to start the random access procedure all over again. For doing so, the UE should first perform a backoff procedure as illustrated in Fig. 4. In this process, the UE waits for a random time, T_{BO} [ms], until it can attempt a new preamble transmission as follows

$$T_{BO} = \mathcal{U}(0, BI), \tag{1}$$

where $\mathcal{U}(\cdot)$ stands for uniform distribution, BI is the backoff indicator defined by the eNB, and its value ranges from 0 to 960 ms. The value of BI is sent in the RAR (*Msg2*), which is read by all the UEs that sent a RACH preamble in the previous RAO. This means that every UE that did not get a *Msg2*, i.e., failed attempt, receives the BI.

Herein, we also studied the potential benefits of implementing an exponential backoff scheme, where the backoff time, T_{BO}, depends on the number of preamble transmissions of each UE, $k \in \{1, 2, \ldots, k_{\max}\}$, as follows

$$T_{BO} = \mathcal{U}(0, 10 \times 2^{k-1}), \tag{2}$$

where the value of $k_{\max}$ is defined by the parameter *preambleTransMax*, broadcast by the eNB through the SIB2 [14].

B. RACH Capacity

The capacity of the LTE-A RACH for the support of M2M communications is determined by two network parameters:

1) Number of available preambles: According to the LTE-A physical layer standard [28], preambles are constructed using Zadoff-Chu (ZC) sequences [34]. These sequences possess good periodic correlation properties, i.e., a negligible time is required to calculate its correlation, which allows the LTE-A system to efficiently support a large number of users per cell. Nevertheless, ZC sequences are difficult to generate in real-time due to the nature of their construction [35], [36] and storing them requires a significant amount of memory (around 4.9 Mbits for a pool of 64 preambles).

In [27], it is found that the capacity of the PRACH, $c(R)$, defined as the maximum expected number of preambles selected by exactly one UE in a RAO, i.e., the maximum value of the expected number of UEs that access successfully in a RAO, approximately coincides with the maximum number of stationary UE arrivals per RAO that the PRACH can handle efficiently, $\hat{c}(R)$. In other words, the performance of the PRACH drops whenever the number of UEs that begin the random access procedure at each and every RAO is $N \geq \hat{c}(R)$. If R is the number of available preambles and N_i is the number of UEs accessing at the ith RAO, it can be easily shown [27] that the expected number of preambles selected by exactly one UE is $N_i \left(1 - 1/R\right)^{N_i - 1}$ (see Fig. 5) and its maximum, $c(R)$, is achieved when $N_i = \left[\log\left(R/\left[R-1\right]\right)\right]^{-1} \approx R$, given as follows

$$c(R) = \left[\log\left(\frac{R}{R-1}\right)\right]^{-1} \left(1 - \frac{1}{R}\right)^{\left[\log\left(\frac{R}{R-1}\right)\right]^{-1} - 1}, \tag{3}$$

which, for instance, when $R = 54$ yields $c(54) = 20.05$ successfully transmitted preambles per RAO, see Fig. 6. Furthermore, $c(R)$ can be approximated as follows

$$c(R) \approx R\left(1 - \frac{1}{R}\right)^{R-1} \approx \frac{R}{e}. \tag{4}$$

The first approximation is highly accurate for practical values of R, and both of them turn out to be lower bounds of $c(R)$ as well. Please refer to the Appendix for more details on this matter.

Hence, assuming a typical PRACH configuration (*prachConfigIndex 6*, in conformance to the LTE-A specification [16], [17]), the PRACH can handle a maximum of $\hat{c}(R) \approx 20.05$ stationary UE arrivals per RAO and, given that RAOs occur every $T_{\text{RAO}} = 5$ ms, a maximum of $\hat{c}(R) = \hat{c}(R)/T_{\text{RAO}} = 4010$ stationary UE arrivals per second.

2) Number of available uplink grants per RAO: Up to $N_{\text{RAR}} = 3$ uplink grants can be sent at each subframe in a RAR message, as the length of a downlink control message is 16 control channel elements (CCEs), the size of uplink grant and

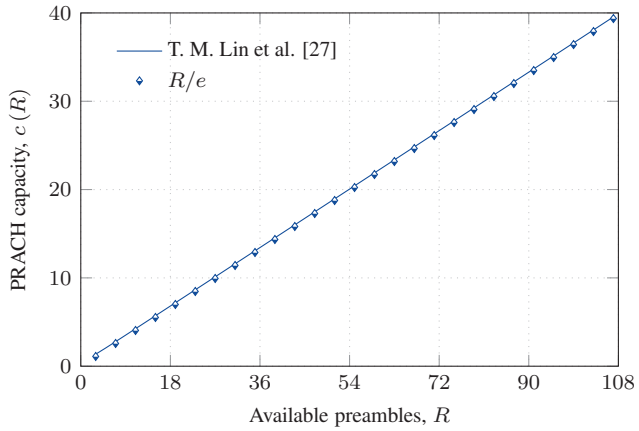


Figure 6. Maximum expected number of UEs that access successfully in a RAO, $c(R)$, calculated as [27] and the R/e approximation for the given number of available preambles, R.

contention resolution messages is 4 CCEs and, at least, 4 CCEs are reserved in each subframe for a contention resolution message, *Msg4*. In the *prach-ConfigIndex 6*, RAOs occur every 5 ms (subframes) and the RAR window size (the time a UE is set to wait for the RAR) is set to $W_{\text{RAR}} = 5$ subframes. As a result, the maximum number of uplink grants that can be sent within the selected W_{RAR} is $N_{UL} = N_{\text{RAR}} \times W_{\text{RAR}} = 15$.

The performance of LTE-A plummets whenever the number of UE arrivals per RAO, N, exceeds either the PRACH capacity, $c(R)$, or the number of uplink grants that the eNB can send between two consecutive RAOs, N_{UL}. Thus, the main objective of congestion control schemes should be to spread UE arrivals through time to maintain the number of UE arrivals per RAO, N, below N_{UL} and $c(R)$, i.e., $N \leq \min\{N_{UL}, c(R)\}$.

C. Access Class Barring

Access Class Barring (ACB) is a congestion control scheme designed for limiting the number of simultaneous access attempts from certain UEs according to their traffic characteristics. For doing so, all UEs are assigned to 16 mobile populations, defined as access classes (ACs) 0 to 15 (see Table I). The population number is stored in UE's SIM/USIM. Each UE belongs to one out of the first 10 ACs (from ACs 0 to 9) and can also belong to one or more out of the five special categories (ACs 11 to 15). Thus, M2M devices may be assigned an AC between 0 and 9, and if a higher priority is needed, other classes may be used. In particular embodiments, AC 10 is used for an emergency call, while AC 11 to AC 15 are special high priority classes [37], [38]. Under the ACB scheme, the network operator may prevent certain UEs from making access attempts or responding to paging messages in specific areas of a public land mobile network (PLMN) based on the corresponding AC [19], [39].

The main purpose of ACB is to redistribute the access requests of UEs through time to reduce the number of access requests per RAO. This fact helps to avoid massive-synchronized accesses demands to the PRACH, which might jeopardize the accomplishment of QoS objectives. Fig. 7 illustrates the ACB process [14], [19]. Note that ACB is

Table I
ACCESS CLASSES DEFINED BY 3GPP [19]

Access class numbers	M2M device
0-9	Normal UEs
10	Indicates network access for Emergency Calls
11-15	Higher priority UEs

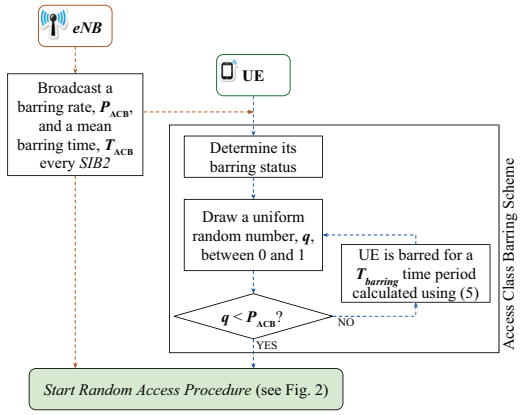


Figure 7. Access class barring scheme.

applied to the UEs before they perform the random access procedure explained in Section III-A.

If ACB is not implemented, all ACs are allowed to access the PRACH. When ACB is implemented, the eNB broadcasts (through SIB2) mean barring times, $T_{\text{ACB}} \in \{4, 8, 16, \ldots, 512\,\text{s}\}$, and barring rates, $P_{\text{ACB}} \in \{0.05, 0.1, \ldots, 0.3, 0.4, \ldots, 0.7, 0.75, 0.8, \ldots, 0.95\}$, that are applied to ACs 0-9. Then, at the beginning of the random access procedure, each UE determines its barring status with the information provided from the eNB. For this, the UE generates a random number between 0 and 1, $\mathcal{U}[0, 1)$. If this number is less than or equal to P_{ACB}, the UE selects and transmits its preamble. Otherwise, the UE waits for a random time calculated as follows

$$T_{barring} = [0.7 + 0.6 \times \mathcal{U}[0, 1)] \times T_{\text{ACB}}. \quad (5)$$

It is worth noting that ACB is only useful for relieving sporadic periods of congestion, i.e., when a massive number of UEs attempt transmission at a given time but the system is not continuously congested. In other words, ACB spreads the load offered to the system through time, but the total offered load is not affected.

IV. RACH EVALUATION

Comparing novel congestion control schemes is not straightforward due to the large number of variables and test scenarios. For that reason, 3GPP TR 37.868 [16] has defined two different traffic models, see Table II, and five key performance indicators (KPIs) to assess the efficiency of the LTE-A random access procedure with M2M communications. These directives allow for a fair comparison of novel congestion solution proposals.

Regarding the traffic models for M2M communications, traffic model 1 can be considered as a typical scenario in which the arrivals of N_M M2M UEs are uniformly distributed over

Table II
M2M TRAFFIC MODELS FOR RACH EVALUATION [16]

Characteristics	Traffic model 1	Traffic model 2
Number of M2M UEs (N_M)	1000, 3000, 5000, 10000, 30000	1000, 3000, 5000, 10000, 30000
Arrival distribution over T	Uniform	*Beta*$(3, 4)$
Distribution period, T	60 seconds	10 seconds

Table III
RACH CONFIGURATION

Parameter	Setting
PRACH Configuration Index	*prach-ConfigIndex* $= 6$
Periodicity of RAOs	5 ms
Subframe length	1 ms
Available preambles for contention-based random access	$R = 54$
Maximum number of preamble transmissions	*preambleTransMax* $= 10$
RAR window size	$W_{\text{RAR}} = 5$ subframes
Maximum number of uplink grants per subframe	$N_{\text{RAR}} = 3$
Maximum number of uplink grants per RAR window	$N_{UL} = W_{\text{RAR}} \times N_{\text{RAR}} = 15$
Preamble detection probability for the kth preamble transmission	$P_d = 1 - \frac{1}{e^k}$ [16]
Backoff Indicator	$BI = 20$ ms
Re-transmission probability for *Msg3* and *Msg4*	0.1
Maximum number of *Msg3* and *Msg4* transmissions	5
Preamble processing delay	2 subframes
Uplink grant processing delay	5 subframes
Connection request processing delay	4 subframes
Round-trip time (RTT) of *Msg3*	8 subframes
RTT of *Msg4*	5 subframes

a period, i.e., in a non-synchronized manner. Traffic model 2 can be seen as an extreme scenario in which a vast number of M2M UE arrivals occur in a highly synchronized manner, e.g., after an application alarm that activates them.

A. Simulation Assumptions, PRACH Configuration, and Performance Metrics

A single cell environment is assumed to evaluate the network performance. The system accommodates both H2H and M2M UEs with different access request intensities. The access attempts of H2H UEs are distributed uniformly over time with an arrival rate of $\lambda_H = 1$ arrivals/s. Regarding the M2M UEs, $N_M = 30000$ UEs (unless otherwise stated) access the eNB as described in traffic model 2 (see Table II). As such, we evaluate the performance of the RACH in the most congested scenario suggested by the 3GPP.

In this study, we assume a typical PRACH configuration, *prach-ConfigIndex 6*, where the subframe length is 1 ms and the periodicity of RAOs is 5 ms. $R = 54$ out of the 64 available preambles are used for contention-based random access and the maximum number of preamble transmissions of each UE, *preambleTransMax*, is set to 10. Table III lists the parameters used throughout our analysis (unless otherwise stated).

The five KPIs for the purpose of RACH capacity evaluation are the following [16]:

1) Collision probability, defined as
$$P_c = \frac{\text{Number of preambles transmitted by multiple UEs}}{R \times N_{\text{RAOs}}}, \quad (6)$$
where N_{RAOs} is the number of consecutive RAOs that compose the measurement period.
2) Access success probability, P_s, defined as the fraction of UEs that successfully complete the random access procedure.
3) Statistics of the number of preamble transmissions for the UEs that successfully complete the random access procedure. We assess this KPI in terms of its mean value, $\mathbb{E}[k]$.
4) Statistics of the access delay, i.e., the time elapsed between the first access attempt (preamble transmission or ACB check) and the successful completion of the random access procedure. To assess this KPI we obtain its cumulative distribution function (CDF) and the 10th, 50th and 95th percentile, D_{10}, D_{50} and D_{95}, respectively.
5) Statistics of the simultaneous preamble transmissions. We assess this KPI in terms of the maximum number of total preamble transmissions per RAO.

To obtain these KPIs, we developed two independent discrete-event simulators that allow us to corroborate our results. The first one is coded in Matlab and the second one is C-based. In each simulation, N_M UE arrivals are distributed within a period of T seconds (see Table II), and the contention-based random access procedure described in Section III-A is replicated with the parameters listed in Table III. Simulations are run j times until each and every one of the cumulative KPIs obtained at the jth simulation differed from those obtained at the $(j-1)$th simulation by less than 0.1 percent; different simulation seeds are used.

B. Collision Model

As mentioned in Section III-A, if two or more UEs transmit the same preamble simultaneously, two outcomes are possible. In the first one, see Fig. 3a, a collision occurs at the transmission of *Msg1* and, in the second one, see Fig. 3b, a collision occurs at the transmission of *Msg3*. To evaluate the impact of these two possible outcomes on the network performance, we have defined two collision models, namely collision model 1 and collision model 2. In collision model 1, all the collisions occur at the transmission of *Msg1*, i.e., the eNB is not capable of decoding any of the preambles transmitted by multiple UEs, so the uplink grants are only sent to the preambles transmitted by exactly one UE. In collision model 2, *Msg1* is always correctly decoded (the eNB successfully decodes the preambles transmitted by multiple UEs), and all the collisions occur at the transmission of *Msg3*. Note that, in practice, both types of collisions might occur. However, our interest is to study and compare the behavior of the RACH in these extreme operation scenarios. Then, the performance of real scenarios will be bounded by that of the extreme ones.

We have simulated the random access procedure with the selected traffic characteristics (traffic model 2 and $N_M = 30000$ M2M UEs) using, on the one hand, the collision model 1 and on the other hand, the collision model 2. The obtained access success probability, P_s, of both M2M and H2H UEs is shown

Table IV
COMPARISON OF THE ACCESS SUCCESS PROBABILITY, P_s, FOR COLLISION MODEL 1 AND COLLISION MODEL 2

UEs	Collision model 1	Collision model 2
M2M	31.305%	16.426%
H2H	61.335%	48.091%

in Table IV. It can be clearly observed that the P_s obtained under collision model 2 is much lower than the one obtained under collision model 1. This drastic reduction in P_s is mainly because in collision model 2 some uplink grants are sent in response to the transmission of a given preamble by multiple UEs, which will cause a collision during the transmission of *Msg3* and leads to (i) the waste of the limited uplink grants, and (ii) the increase of the number of contending UEs in future RAOs.

Hereafter, we select collision model 1 to conduct the performance analysis of LTE-A as suggested by the 3GPP [16] because selecting collision model 2 would magnify the increase in the performance provided by the implementation of ACB. Please note that if a different collision model is used, the performance of the RACH will differ from the one presented in this study.

V. PERFORMANCE ANALYSIS OF LTE-A

In this section, we present some relevant results derived from our performance analysis of the LTE-A random access procedure. We begin our analysis by evaluating the capacity of the PRACH. For this, we generate a stationary distribution of $N \in \{1, 2, \ldots, 40\}$ new UE arrivals per RAO and study the effect of the number of available preambles, R, on the access success probability of UEs, P_s. To overcome the limitations of the PDCCH and evaluate the PRACH on its own, we assume that $N_{UL} = R$. Fig. 8 illustrates the evolution of P_s for $R \in \{20, 30, 40, 54, 64\}$. It can be observed that for each R, $P_s \approx 1$ up to a maximum value of N and then plummets. For example, when $R = 54$, $P_s \approx 1$ until $N \approx 20$, then P_s drops rapidly as N increases. Note that, $\hat{c}(R) = \{\max(N) | P_s \approx 1\}$ in a complex real scenario like the one studied is close to the PRACH capacity per RAO, $c(R)$, defined by (3), that was obtained using relatively simple arguments. Hence, there is a maximum stationary UE arrival rate, $\hat{c}(R) \approx c(R)$, for which UEs can efficiently access the PRACH.

Once we have studied the behavior of the PRACH in steady state, we proceed to investigate the performance of the LTE-A random access procedure according to the assumptions and the simulation parameters detailed in Section IV-A and Table III, respectively. As a baseline, Fig. 9 illustrates the expected number of UE arrivals per RAO (number of UEs that begin its random access procedure at the ith RAO), preambles with collision (collided preambles), successful accesses (UEs that complete the random access procedure successfully), and total preamble transmissions per RAO. Note that when $N_M = 30000$, traffic model 2 leads to network congestion, as the *Beta*$(3, 4)$ distribution of UE arrivals exceeds the PRACH capacity ($c(54) = 20.05$ UE arrivals per RAO as calculated using (3) and $N_{UL} = 15$) from the 343rd to the 1329th RAO. This massive number of UE arrivals results in

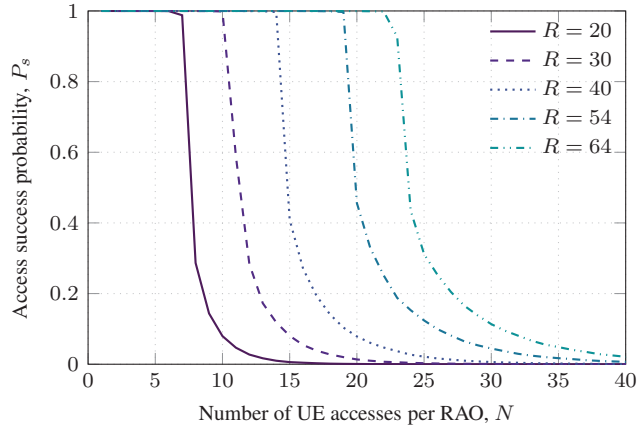


Figure 8. Access success probability of UEs, P_s, given the number of UE accesses per RAO, $N \in \{1, 2, \ldots, 40\}$, and the number of available preambles, $R \in \{20, 30, 40, 54, 64\}$.

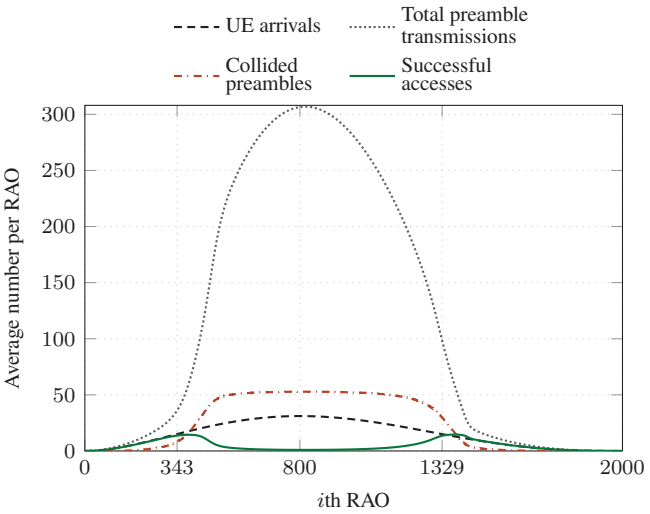


Figure 9. Temporal distribution of M2M UE arrivals, total preamble transmissions, collided preambles, and successful accesses; traffic model 2, $N_M = 30000$.

a congestion period of $T_c = 4.93$ s, where up to 300 average preamble transmissions per RAO occur at the 800th RAO. As a result, the average number of successful accesses sharply decreases during this period, and the access success probability is severely affected: $P_s = 31.305\,\%$.

For the remainder of this paper, we focus on increasing the performance of the LTE-A random access procedure (assessed in terms of the KPIs defined in Section IV-A) when a massive number of M2M UEs, $N_M = 30000$, access the eNB according to traffic model 2. In the following, we investigate:

1) The number of available preambles, R, required to achieve a $P_s \approx 1$.
2) The impact of the implementation of an exponential backoff scheme instead of the standard uniform backoff scheme on the network performance.
3) The impact of the manipulation of *preambleTransMax* on the network performance.

Next, we detail the analysis and the results of modifying the three configuration parameters mentioned above.

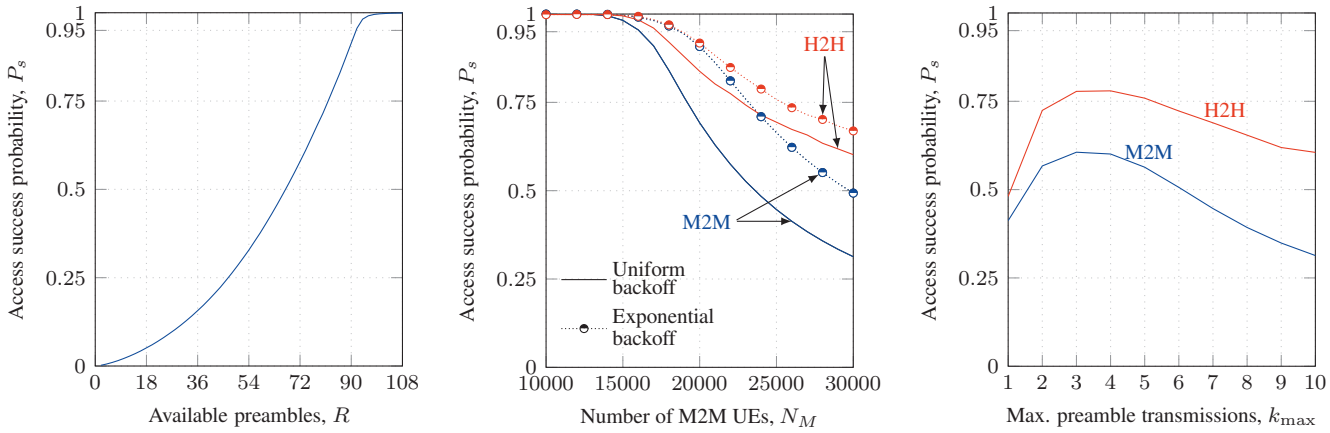


Figure 10. Access success probability, P_s, of M2M and H2H UEs ($\lambda_H = 1$ arrivals/s). (a) P_s of M2M UEs only given the number of available preambles, R, (b) P_s of M2M and H2H UEs given the number of M2M UEs, N_M, and (c) P_s of M2M and H2H UEs given the maximum number of preamble transmissions, $k_{\max} =$ *preambleTransMax*. In (a) and (c) the number of M2M UEs is $N_M = 30000$ and the uniform backoff is used.

A. Impact of Increasing the Number of Available Preambles

To investigate whether increasing the number of available preambles can relieve congestion, we obtained the P_s of M2M UEs for several values of R, see Fig. 10a. Note that we assume $N_{UL} = R$. Here we observe that $P_s \geq 0.9$ is only achieved when $R \geq 90$. In other words, a dramatic increase in the number of available preambles, R, is needed to avoid PRACH congestion considering the most severely congested test scenario suggested by the 3GPP.

As mentioned in Section III-B, preambles are constructed using Zadoff-Chu sequences that are difficult to generate in real-time due to the nature of their construction and storing them requires a significant amount of memory. Hence, such a dramatic increase in the number of available preambles ($R \geq 90$) may not be possible. Instead of increasing the number of available preambles, R, studies such as [40]–[43] propose schemes for a more efficient utilization of preambles as a better solution for relieving PRACH congestion.

B. Impact of Modifying the Backoff Scheme

According to the standard [17], UEs perform a uniform backoff, $T_{BO} = \mathcal{U}(0, BI = 20)$ ms, after a collision. We have previously observed that the use of this backoff scheme is not sufficient for relieving the congestion in the random access channels. On this basis, we investigate the use of an exponential backoff scheme, where the backoff time of each M2M UE depends on the number of preamble transmission being attempted by that specific UE, $k \leq$ *preambleTransMax*, and is given by (2). As mentioned in Section IV-A, the H2H UE arrivals are distributed uniformly over time, with an arrival rate of $\lambda_H = 1$ arrivals/s.

Fig. 10b shows the P_s of M2M and H2H UEs when implementing the uniform and the exponential backoff schemes. On the one hand, it can be observed that the maximum number of M2M UEs that leads to $P_s \geq 0.95$ is approximately $N_M \leq 16000$ given the implementation of the uniform backoff and is $N_M \leq 19000$ when implementing the exponential backoff scheme. Hence, the use of an exponential backoff increases the number of UEs that can efficiently access the eNB. Nevertheless, the use of an exponential backoff is insufficient in cases of severe congestion, e.g., when $N_M \geq 20000$. On the other hand, it can also be observed from Fig. 10b that, in most cases, H2H UEs obtain a higher P_s than M2M UEs; this fact occurs because H2H UEs are distributed uniformly through time whereas the arrivals of M2M UEs are highly concentrate in a short time interval, i.e., between the 343rd and the 1329th RAOs. As a result, most of the H2H UEs begin its random access procedure in RAOs with a low number of preamble transmissions, where the access success probability is high. On the contrary, most of the M2M UEs begin its random access procedure in RAOs with a high number of preamble transmissions, where the access success probability is low.

C. Impact of Modifying the Maximum Number of Preamble Transmissions

In Section V, we have observed that severe congestion occurs when $N_M = 30000$ UEs attempt to access the eNB according to traffic model 2. Specifically, during the period of congestion, up to 300 preamble transmissions per RAO occur, see Fig. 9. Such a high number of preamble transmissions is the consequence of the fact that the higher the number of preamble transmissions in a RAO, the lower the probability of a successful preamble transmission. This fact, in turn, increases the probability of preamble re-transmissions in the following RAOs, hence the probability of a successful preamble transmission is further reduced. Therefore, during periods of congestion, the total number of preamble transmissions per RAO is highly influenced by the parameter *preambleTransMax* (maximum number of preamble transmissions). Hence, we now evaluate whether the congestion of the LTE-A random access channels can be reduced by the modification of this parameter. In Fig. 10c we show the P_s of M2M and H2H UEs when *preambleTransMax* $\in \{1, \ldots, 10\}$. Note that the highest P_s for both M2M and H2H UEs is achieved when *preambleTransMax* $= 3$, despite the fact that the UEs increase their transmission power at each preamble transmission, which in turn increases the preamble detection probability, P_d. These

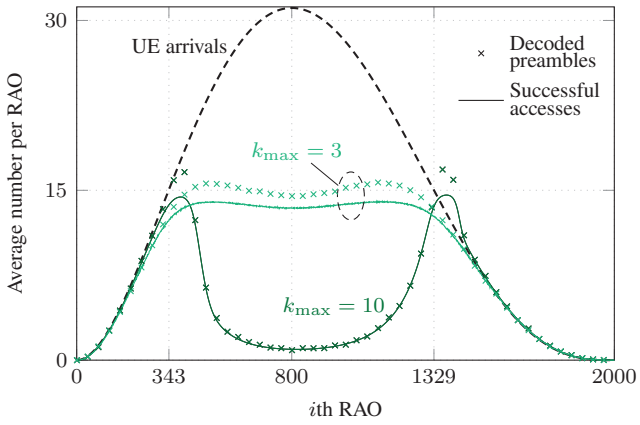


Figure 11. Temporal distribution of M2M UE arrivals, decoded preambles and successful UE accesses, traffic model 2, $N_M = 30000$, uniform backoff, $k_{\max} = \textit{preambleTransMax} \in \{3, 10\}$.

results highlight the importance of reducing congestion in order to enhance performance.

To observe more closely the behavior of LTE-A when *preambleTransMax* $=$ 3, the average number of decoded preambles, and successful accesses per RAO given *preambleTransMax* $\in \{3, 10\}$ are shown in Fig. 11. It can be seen that a higher number of successful accesses per RAO is achieved when *preambleTransMax* $=$ 3, which is due to a lower number of preamble transmissions per RAO. In addition to lowering congestion, which in turn increases the access success probability, reducing the number of preamble transmissions also reduces the energy consumption of UEs in highly congested scenarios. This is highly desirable because the UEs are oftentimes battery supplied.

It is worth noting that by selecting *preambleTransMax* $= 3$ the average number of successful accesses per RAO during congestion is close to the maximum number of uplink grants per RAO that can be sent by the eNB, $N_{UL} = 15$. Hence, a high percentage of the system capacity is being utilized. Nevertheless, the available uplink grants per RAO, N_{UL}, are insufficient for assigning resources to the vast number of UE arrivals. Note that combining the use of an exponential backoff with the reduction of *preambleTransMax* would not be effective, i.e., in the exponential backoff, the upper limit of the backoff time increases with the number of failed preamble transmissions. Thus, the backoff time for the first few preamble transmissions is low.

In Sections V-B and V-C we have shown that either implementing an exponential backoff or reducing the maximum number of preamble transmissions increases the performance of the LTE-A RACH. However, the manipulation of neither of those parameters can prevent the system capacity from being exceeded. Yet another parameter that can be manipulated in an attempt to relieve PRACH congestion is the number of RAOs scheduled per frame. For instance, increasing the number of RAOs per frame would reduce the number of contending UEs per RAO. Nevertheless, this approach has several drawbacks: (i) it implies a reduction of the number of resources available for data transmission and, hence, a contraction of the data transport capacity of the uplink channel; (ii) the total number of RAOs that can be allocated in an LTE-A frame is limited; and (iii) the maximum number of uplink grants that can be sent by the eNB per frame is fixed, so the limitations of the PDCCH remain constant.

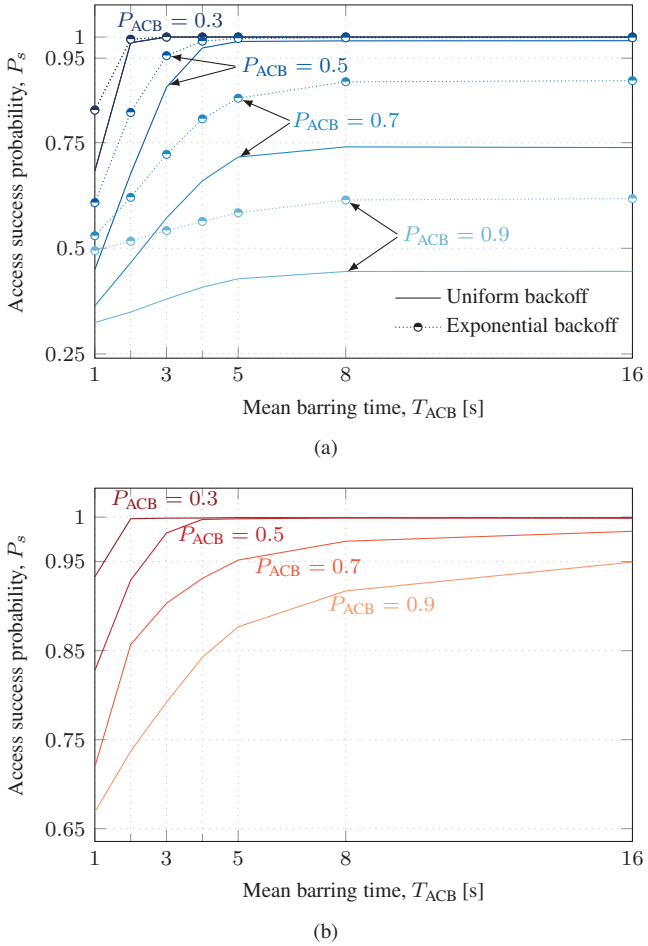


Figure 12. Access success probability of (a) M2M and (b) H2H UEs under the ACB scheme.

Consequently, a congestion control scheme with configurable parameters that can efficiently spread the UE arrivals through time must be implemented to drastically enhance the performance of the LTE-A. Next, we investigate the impact of the ACB congestion control scheme on the network performance.

VI. Performance Analysis of ACB

In this section, we study the impact of the implementation of the access class barring (ACB) scheme on the performance of LTE-A networks with massive M2M traffic. For the sake of simplicity, we assess the performance of LTE-A with an implemented ACB in terms of three KPIs, namely the access success probability, P_s, the access delay, and the average number of preamble transmissions, $\mathbb{E}[k]$, which is closely related to energy consumption. Our main objective is to identify the configuration of ACB parameters that result in an acceptable P_s. Specifically, we aim to identify the combinations of barring rates, P_{ACB}, and barring times, T_{ACB}, that result in $P_s \geq 0.95$ for the M2M UEs.

Fig. 12 shows the P_s of M2M and H2H UEs, given $P_{\text{ACB}} \in \{0.3, 0.5, 0.7, 0.9\}$ and $T_{\text{ACB}} \in \{1, 2, 3, 4, 5, 8, 16\}$ s.

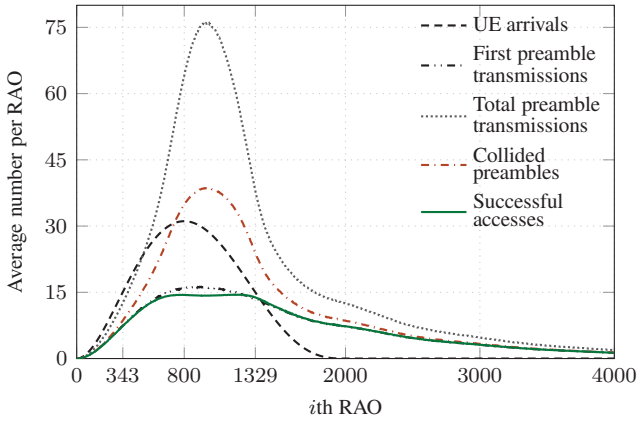


Figure 13. Temporal distribution of M2M UE arrivals, first preamble transmissions, total preamble transmissions, collided preambles and successful accesses, given $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$, uniform backoff.

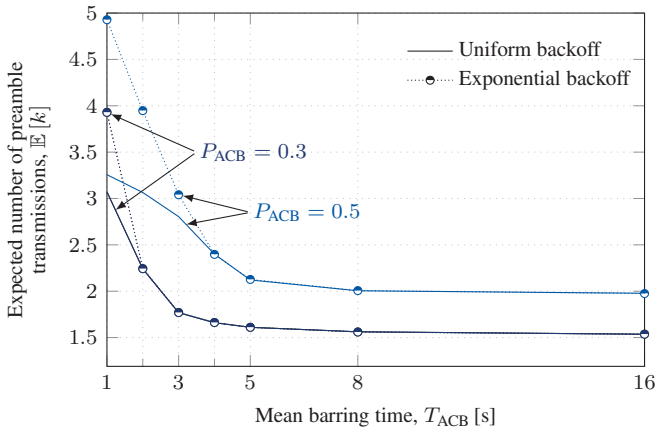


Figure 14. Mean number of preamble transmissions for the successfully accessed M2M UEs under the ACB scheme.

It can be seen that, for every one of the given barring rates P_{ACB}, the access success probability, P_s, increases with the barring time, T_{ACB}. Nevertheless, for each P_{ACB} there exists a maximum value of P_s that is achieved at a certain T_{ACB}. Once this maximum P_s for each P_{ACB} is reached, further increasing T_{ACB} has no observable effect on P_s.

If we compare the P_s of the M2M UEs achieved with the implementation of a uniform backoff with the one achieved with the implementation of an exponential backoff, see Fig. 12a, we observe that, for the latter, shorter barring times are needed to achieve the same P_s. Please note that the H2H UEs always perform a uniform backoff. Therefore, implementing an exponential backoff in the M2M UEs does not lead to a noticeable increase in the P_s of H2H UEs, so these results have been omitted in Fig. 12b.

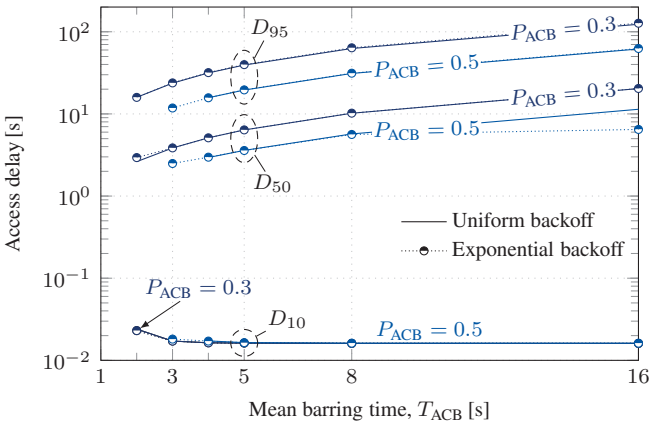
Also note that $P_s \geq 0.95$ for M2M UEs is only achieved when selecting $P_{\text{ACB}} \leq 0.5$. The effect of ACB on the UE arrivals can be closely observed in Fig. 13, where the average number of UE arrivals, preamble transmissions, collided preambles, and successful accesses per RAO given $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$ are shown. This particular combination of barring parameters leads to $P_s = 97.44\,\%$ for the M2M UEs. Such a high P_s is achieved because ACB reduces the UE arrivals per RAO from a maximum of 31.104 to 16.347, which is close to $N_{UL} = 15$ and below $\hat{c}(R) = 20.05$. As a result, we observe a dramatic reduction in the number of collisions and preamble transmissions per RAO when compared with those of Fig. 9.

Next, we proceed to investigate the number of preamble transmissions, k, performed by the UEs that successfully complete the random access procedure. In Fig. 14, we show the mean number of preamble transmissions, $\mathbb{E}[k]$, given $P_{\text{ACB}} \in \{0.3, 0.5\}$ as those barring rates lead to $P_s \geq 0.95$ (except for the lowest values of T_{ACB}, see Fig. 12). It can be seen that both high values of P_{ACB} and low values of T_{ACB} increase $\mathbb{E}[k]$. From Fig. 12 we observed that the implementation of an exponential backoff increases P_s in cases where a $P_s < 0.95$ is achieved by the use of a uniform backoff. On the other hand, from Fig. 14 we observe that, in the mentioned cases, $\mathbb{E}[k]$ also increases. Thus, implementing an exponential backoff scheme may slightly increase P_s at the cost of increasing the energy consumption. In cases where both backoff schemes would lead to $P_s \geq 0.95$, $\mathbb{E}[k]$ is almost identical.

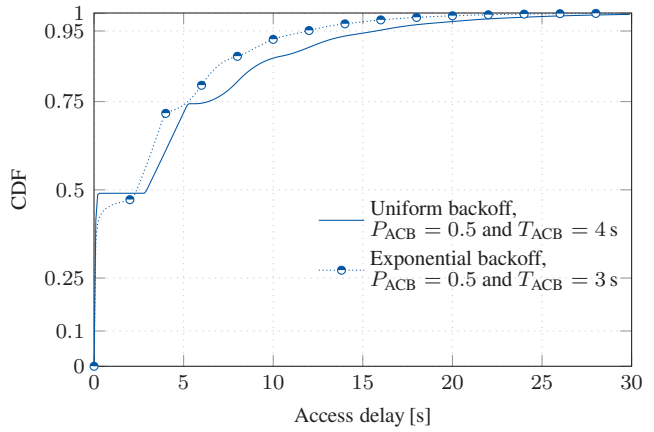
Finally, we studied the access delay when ACB is implemented; Fig. 15 illustrates these results. We calculate the access delay as the time elapsed between the arrival of a UE and the successful completion of its random access procedure, according to the timing values illustrated in Fig. 2 [32, Table 16.2.1-1]. For the sake of simplicity, we evaluate the access delay in terms of percentiles, defined as the maximum delay experienced by the δN UEs with the lowest delay, for the given $\delta \in \{0.1, 0.5, 0.95\}$. It is worth noting that evaluating delay in terms of its maximum achievable value, i.e., the maximum time needed for a UE to successfully complete its random access procedure, is not viable when performing ACB because this value is not upper bounded. In other words, there is no upper limit for the number of ACB checks to be performed by a UE, hence the maximum delay, $\lim_{N_M \to \infty} D_{100} = \infty$. As such, Fig. 15a illustrates the 10th percentile, D_{10}, the 50th percentile, D_{50}, and the 95th percentile, D_{95}, given that $P_s \geq 0.95$.

As expected, a combination of low values of T_{ACB} with high P_{ACB} reduces the access delay. Also, though selecting a long T_{ACB} does not greatly affect P_s, see Fig. 12a, it sharply increases the access delay, as shown in the y-axis of Fig. 15a in logarithmic scale. Hence, a long T_{ACB} should be avoided. In Fig. 15a it can also be seen that, for the cases of interest, the delay experienced by the M2M UEs is almost the same when either a uniform or an exponential backoff scheme is implemented. Also, note that the combination of $T_{\text{ACB}} = 3\,\text{s}$ and $P_{\text{ACB}} = 0.5$ with the exponential backoff leads to $P_s \geq 0.95$. It is in this case that the overall shortest D_{50} and D_{95} are achieved. On the other hand, the shortest delay percentiles for the uniform backoff are achieved by the selection of $T_{\text{ACB}} = 4\,\text{s}$ and $P_{\text{ACB}} = 0.5$. Note that shorter delay percentiles are obtained by selecting $T_{\text{ACB}} = 3\,\text{s}$ and $P_{\text{ACB}} = 0.5$ with the uniform backoff; however, the desired $P_s \geq 0.95$ is not met as can be seen in Fig. 12a. It is worth mentioning that the effect of ACB in the access delay of H2H UEs is almost negligible, as can be seen in Table V for $P_{\text{ACB}} \in \{3, 5\}$.

In Fig. 15b we compare the CDF of access delay between



(a)



(b)

Figure 15. (a) Percentiles of access delay of M2M UEs under the ACB scheme, in logarithmic scale, for the combinations of P_{ACB} and T_{ACB} that result in $P_s \geq 0.95$. (b) Cumulative distribution function of access delay for the combinations that lead to the shortest D_{50} and D_{95}, given $P_s \geq 0.95$.

Table V
ACCESS DELAY OF H2H UES UNDER THE ACB SCHEME

P_{ACB}	T_{ACB} [s]	D_{10} [ms]	D_{50} [ms]	D_{95} [ms]
0.3	2	15.195	20.385	56.823
	3	15.203	20.109	55.171
	4	15.202	18.286	51.187
	5	15.204	18.865	51.951
	8	15.198	16.536	50.785
	16	15.197	15.997	50.730
0.5	2	15.160	20.369	61.278
	3	15.183	20.321	60.685
	4	15.195	20.323	60.235
	5	15.193	19.304	55.264
	8	15.196	19.424	54.080
	16	15.196	17.827	50.915

the selection of a uniform backoff along with $T_{\text{ACB}} = 4$ s, $P_{\text{ACB}} = 0.5$ with that of an exponential backoff along with $T_{\text{ACB}} = 3$ s, $P_{\text{ACB}} = 0.5$. In the former, the initial growth is much more rapid. Nevertheless, in the latter, shorter D_{50} and D_{95} are achieved.

A. Optimal ACB Parameter Configuration

In this section, we evaluate the performance of ACB in terms of delay and energy consumption. Recall that, if a large number of devices try to access the RACH in a short period, the preamble collisions increase significantly, resulting in huge access delays. Besides, in such a congested scenario, the repeated transmission attempts increase the energy consumption of M2M devices, most of which will be energy-constrained. To minimize the adverse effects of congestion mentioned above, the configuration parameters of ACB, P_{ACB}, and T_{ACB}, have to be adjusted adequately. Here, we determine the optimal selection of P_{ACB} and T_{ACB} among those pairs that yields an acceptable P_s for traffic model 2 and $N_M = 30000$. For doing so, we first identify the minimum value of $T_{\text{ACB}} \in \{0.05, 0.1, \ldots\}$ [s] for a given P_{ACB} that leads to an access success probability higher than 0.95, that is,

$$T^*_{\text{ACB}} = \min\{T_{\text{ACB}} \mid P_s(P_{\text{ACB}}, T_{\text{ACB}}) \geq 0.95\}, \qquad (7)$$

then we assess the provided QoS in terms of the expected number of preamble transmissions for the successfully accessed UEs, $\mathbb{E}^*[k]$, and the 95th percentile of access delay, D^*_{95} for the given T^*_{ACB}. The obtained T^*_{ACB}, $\mathbb{E}^*[k]$, and D^*_{95} for each $P_{\text{ACB}} \in \{0.01, 0.02, \ldots, 0.99\}$ are shown in Fig. 16a, Fig. 16b, and Fig. 16c, respectively, with the uniform and exponential backoff. The variability in the curves is caused by the granularity of both P_{ACB} and T_{ACB}.

The results presented in Fig. 16a confirm that, if the exponential backoff is selected, shorter barring times are needed to achieve $P_s \geq 0.95$ when compared to those of the uniform backoff. It can also be seen that there exists a maximum P_{ACB} for each backoff scheme that can be selected in order to achieve $P_s \geq 0.95$: 0.56 for the uniform backoff and 0.64 for the exponential backoff. Hence, the exponential backoff increases the range of P_{ACB} (and also that of T_{ACB}) that can be selected to achieve an acceptable P_s.

If we compare the average number of preamble transmissions, $\mathbb{E}^*[k]$ (see Fig. 16b), with the 95th percentile of access delay, D^*_{95} (see Fig. 16c), we clearly observe the trade-off between these KPIs; i.e., the access delay is high with configurations in which a low number of preamble transmissions are performed and vice versa.

It is worth noting that selecting T^*_{ACB} when $P_{\text{ACB}} \in [0.1, 0.6]$ only causes a slight variation in both $\mathbb{E}^*[k]$ and D^*_{95}, which is highly desirable. In addition, we can observe that the implementation of the exponential backoff increases the number of preamble transmissions but reduces the access delay when compared to the implementation of the uniform backoff.

VII. CONCLUSION

We have performed a thorough study of the massive access of M2M UEs in LTE-A cellular networks. As a baseline, we obtained several key performance indicators (KPIs) to evaluate the performance of LTE-A when M2M arrivals follow either a uniform or a *Beta*$(3, 4)$ distribution as described by traffic models 1 and 2, respectively.

We observed that traffic model 2, which describes the bursty arrivals of a massive number of M2M UEs to an evolved Node B (eNB), leads to severe congestion if the eNB lacks a congestion control scheme. We observed that severe congestion persists regardless of the modification of network parameters such as the maximum number of allowed preamble transmissions, *preambleTransMax*, and the selected backoff

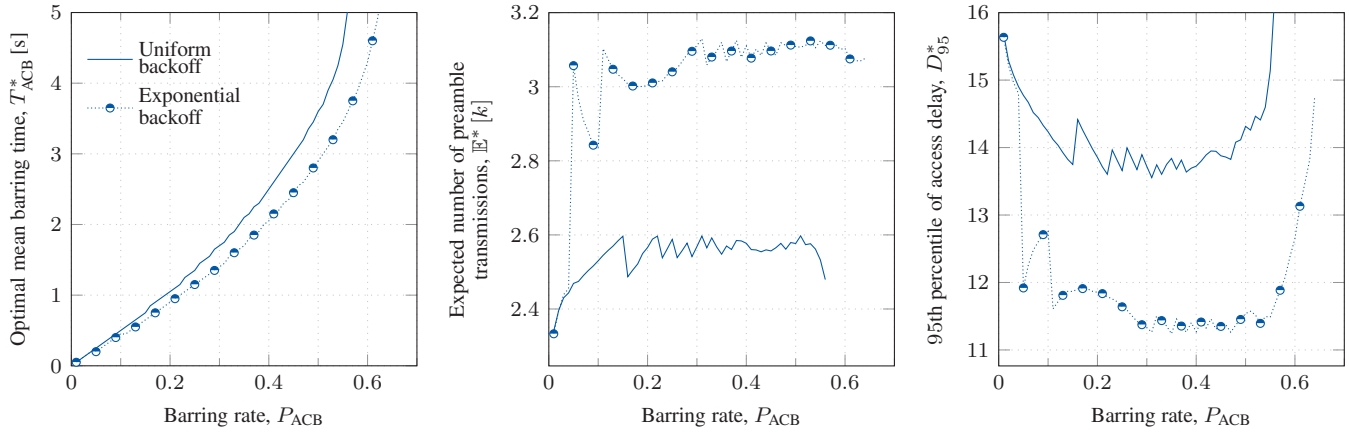


Figure 16. ACB optimal parameter configuration that leads to $P_s \geq 0.95$. (a) T^*_{ACB} defined as (7), (b) $\mathbb{E}^*[k] = \mathbb{E}[k]$ when T^*_{ACB}, and (c) $D^*_{95} = D_{95}$ when T^*_{ACB}, for the given P_{ACB}.

scheme. Furthermore, the severity of congestion increases in cases where collisions occur during the transmissions of *Msg3*.

As such, we have studied the access class barring (ACB) scheme for dealing with PRACH overload and analyzed the impact of its configuration parameters on the network performance. We assume that access success probability, P_s, is the main KPI; hence, we first focus on identifying the combinations of barring rates and barring times for which the system achieves a $P_s \geq 0.95$. Then, we studied other KPIs such as the number of preamble transmissions and the access delay, where we identified a trade-off. Specifically, low barring rates and long barring times increase the access delay but reduce the number of preamble transmissions, hence reducing energy consumption.

It is worth noting that the relevance of energy consumption and access delay highly depends on the traffic characteristics, e.g., the frequency of random access congestion. For instance, if the studied scenario occurs sparingly, these KPIs are not highly relevant, as slight increases in energy consumption will not highly affect the battery life. On the other hand, when this scenario occurs frequently, battery life may be compromised, and highly delayed accesses might cause congestion in subsequent UE accesses.

We also compared the KPIs obtained by implementing a uniform backoff scheme, as described in the LTE-A standard [17], with that of an exponential backoff scheme along with ACB. Results show that an exponential backoff leads to a slightly higher success probability but also increases the mean number of preamble transmissions. Therefore, implementing an exponential backoff may enhance the access success probability at the cost of a higher energy consumption. Moreover, the increase in P_s provided by the exponential backoff allows the selection of lower barring times when compared to a uniform backoff. This fact, in turn, may slightly reduce the access delay.

Finally, by adequately selecting the ACB barring rates and barring times, network congestion may be relieved, even for the most congested scenario defined by the 3GPP. As such, ACB was shown to be an efficient scheme for congestion control in the RACH.

APPENDIX
RACH CAPACITY: APPROXIMATIONS AND BOUNDS

Here we derive some approximations and bounds for the system capacity, $c(R)$, defined in [27].

First, we recall that

$$1 - \frac{1}{x} < \log(x) < x - 1, \quad \text{for } x > 0. \tag{8}$$

From (8), it follows immediately that

$$R - 1 < \left[\log\left(\frac{R}{R-1}\right)\right]^{-1} < R. \tag{9}$$

Applying the inequalities in (9) to (3) we obtain

$$\ell_0(R) < c(R) < u(R), \tag{10}$$

where

$$\ell_0(R) \triangleq (R-1)\left(1 - \frac{1}{R}\right)^{R-1} = R\left(1 - \frac{1}{R}\right)^{R}, \tag{11}$$

$$u(R) \triangleq R\left(1 - \frac{1}{R}\right)^{R-2}. \tag{12}$$

From (11) and (12) for $R > 0$, it can be easily seen that

$$\ell_0(R) < \ell_1(R) < u(R), \tag{13}$$

where

$$\ell_1(R) \triangleq R\left(1 - \frac{1}{R}\right)^{R-1} \approx c(R). \tag{14}$$

Now, by observing that $(1 - 1/R)^R$ is increasing and tends to e^{-1}, and $(1 - 1/R)^{R-1}$ is decreasing and tends to e^{-1}, we can see that

$$\ell_0(R) < \ell_2(R) < \ell_1(R), \tag{15}$$

where

$$\ell_2(R) \triangleq \frac{R}{e}. \tag{16}$$

From the above observations, it can also be deduced that if R is sufficiently large, $\ell_0(R) \approx \ell_2(R) \approx \ell_1(R)$. Besides, by numerical evaluation we have verified that $\ell_1(R) < c(R)$.

Table VI
ACCURACY OF THE APPROXIMATIONS AND BOUNDS

		Rel. error (%)			
R	$c(R)$	$\ell_0(R)$	$\ell_2(R)$	$\ell_1(R)$	$u(R)$
10	3.8796	10.1248	5.1755	0.1386	10.9571
20	7.5496	5.0312	2.5427	0.0329	5.2285
30	11.2256	3.3472	1.6855	0.0144	3.4334
40	14.9030	2.5078	1.2605	0.0080	2.5559
50	18.5810	2.0050	1.0067	0.0051	2.0356
60	22.2593	1.6701	0.8380	0.0035	1.6913
70	25.9377	1.4311	0.7177	0.0026	1.4466

Finally, combining the previous derivations we have

$$\ell_0(R) < \ell_2(R) < \ell_1(R) < c(R) < u(R) \tag{17}$$

and the approximations given in (4), i.e., $c(R) \approx \ell_1(R) \approx \ell_2(R)$. As can be seen in Table VI, $\ell_1(R)$ provides an extremely accurate approximation, while $\ell_2(R)$, which is a simpler expression, can be considered as sufficiently accurate for all practical purposes (see also Fig. 6).

REFERENCES

[1] I. S. Association. Internet of things. [Online]. Available: http://standards.ieee.org/innovate/iot/

[2] Ericsson. (2017, Jun.) Ericsson mobility report. [Online]. Available: https://www.ericsson.com/mobility-report

[3] Cisco. (2017, Mar.) Cisco visual networking index (VNI): Global mobile data traffic forecast update, 2016-2021. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[4] 3GPP, *TS 22.368, Service Requirements for Machine-Type Communications*, Mar 2017.

[5] P. K. Verma, R. Verma, A. Prakash, A. Agrawal, K. Naik, R. Tripathi, M. Alsabaan, T. Khalifa, T. Abdelkader, and A. Abogharaf, "Machine-to-Machine (M2M) communications: A survey ," *J. Netw. Comput. Appl.*, vol. 66, pp. 83 – 105, 2016.

[6] Y. Mehmood, C. Görg, M. Muehleisen, and A. Timm-Giel, "Mobile M2M communication architectures, upcoming challenges, applications, and future directions," *EURASIP J. Wirel. Commun. Netw.*, vol. 2015, no. 1, pp. 1–37, 2015.

[7] 3GPP, *TS 23.682, Architecture enhancements to facilitate communications with packet data networks and applications*, Mar 2016.

[8] F. Ghavimi and H.-H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, May 2015.

[9] A. Lo, Y. Law, and M. Jacobsson, "A cellular-centric service architecture for machine-to-machine (M2M) communications," *IEEE Wireless Commun. Mag.*, vol. 20, no. 5, pp. 143–151, 2013.

[10] 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, 2011.

[11] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. C.-C. Hsu, "Overload control for machine-type-communications in LTE-advanced system," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 38–45, 2012.

[12] L. Ferdouse, A. Anpalagan, and S. Misra, "Congestion and overload control techniques in massive M2M systems: a survey," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 3, pp. 1–17, Mar 2015.

[13] 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, Jan 2014.

[14] 3GPP, *TS 36.331, Radio Resource Control (RRC), Protocol specification*, Sep 2017.

[15] 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, 2015.

[16] 3GPP, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Sep 2011.

[17] ——, *TS 36.321, Medium Access Control (MAC) Protocol Specification*, Sep 2017.

[18] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," in *Proc. IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[19] 3GPP, *TS 22.011, V13.1.0, Service Accessibility*, Sep 2017.

[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] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, "Congestion control for bursty M2M traffic in LTE networks," in *Proc. IEEE International Conference on Communications (ICC)*, Jun 2015, pp. 5815–5820.

[22] O. Arouk and A. Ksentini, "General Model for RACH Procedure Performance Analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, Feb 2016.

[23] 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.

[24] G.-Y. Lin, S.-R. Chang, and H.-Y. Wei, "Estimation and Adaptation for Bursty LTE Random Access," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2560–2577, 2016.

[25] J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas, and P. Popovski, "A tractable model of the LTE access reservation procedure for machine-type communications," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[26] Z. Zhang, H. Chao, W. Wang, and X. Li, "Performance Analysis and UE-Side Improvement of Extended Access Barring for Machine Type Communications in LTE," in *Proc. IEEE Vehicular Technology Conference (VTC Spring)*, May 2014, pp. 1–5.

[27] 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, 2014.

[28] 3GPP, *TS 36.211, Physical Channels and Modulation*, Sep 2017.

[29] 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, 2015.

[30] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digit. Commun. Netw.*, vol. 1, no. 1, pp. 1–19, 2015.

[31] 3GPP, *TS 36.213, Physical layer procedures*, Dec 2014.

[32] ——, *TR 36.912, Feasibility study for Further Advancements for E-UTRA*, Mar 2017.

[33] J. E. Wieselthier, A. Ephremides, and L. A. Michaels, "An exact analysis and performance evaluation of framed ALOHA with capture," *IEEE Trans. Commun.*, vol. 37, no. 2, pp. 125–137, Feb 1989.

[34] D. C. Chu, "Polyphase codes with good periodic correlation properties," *IEEE Trans. Inf. Theory*, vol. 18, 1972.

[35] M. M. Mansour, "Optimized architecture for computing Zadoff-Chu sequences with application to LTE," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 2, no. 1, 2009.

[36] F. A. P. De Figueiredo, F. S. Mathilde, F. A. C. M. Cardoso, R. M. Vilela, and J. P. Miranda, "Efficient frequency domain zadoff-chu generator with application to LTE and LTE-A systems," in *Proc. International Telecommunications Symposium, (ITS)*, 2014, pp. 1–5.

[37] C. L. Taylor, D. Nolan, and S. Wainberg, "Priority capabilities in LTE supporting national security and emergency preparedness next generation network priority services," in *Proc. IEEE International Conference on Technologies for Homeland Security (HST)*, Nov 2013, pp. 584–588.

[38] D. Nolan, S. Wainberg, J. R. Wullert, and A. R. Ephrath, "National security and emergency preparedness communications: Next generation priority services," in *Proc. IEEE International Conference on Technologies for Homeland Security (HST)*, Nov 2013, pp. 106–112.

[39] L. Segura, "Access control for M2M devices," Aug. 18 2011, US Patent App. 13/028,093.

[40] H. Thomsen, N. K. Pratas, Č. Stefanović, and P. Popovski, "Code-expanded radio access protocol for machine-to-machine communications," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 355–365, 2013.

[41] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced Radio Access and Data Transmission Procedures Facilitating Industry-Compliant Machine-Type Communications over LTE-Based 5G Networks," *IEEE Wireless Commun. Mag.*, vol. 23, no. 1, pp. 56–63, 2016.

[42] M. Condoluci, G. Araniti, M. Dohler, A. Iera, and A. Molinaro, "Virtual code resource allocation for energy-aware MTC access over 5G systems," *Ad Hoc Netw.*, vol. 43, pp. 3–15, 2016.

[43] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive multiple access based on superposition raptor codes for cellular m2m communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 307–319, Jan 2017.

Luis Tello-Oquendo (S'08) received the B.E. degree in electronic and computer engineering (Hons.) from Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador, in 2010, and the M.Sc. degree in telecommunication technologies, systems, and networks from Universitat Politècnica de València (UPV), Spain, in 2013. He is currently pursuing the Ph.D. degree in telecommunications engineering. In 2011, he was a Lecturer with the Facultad de Ingeniería Electrónica, ESPOCH. From 2016 to 2017 he was a Visiting Research Scholar with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA, USA. He is currently a Graduate Research Assistant with the Broadband Internetworking Research Group, UPV. His research interests include mobile and wireless communication networks, random access protocols, machine-to-machine communications, wireless software-defined networks, LTE-A and beyond cellular systems, Internet-of-Things, machine learning. He is a member of the ACM. He received the Best Academic Record Award from the Escuela Técnica Superior de Ingenieros de Telecomunicación, UPV, in 2013, and the IEEE ComSoc Award for attending the IEEE ComSoc Summer School at The University of New Mexico, Albuquerque, NM, USA, in 2017.

Israel Leyva-Mayorga received the B.Sc. degree in telematics engineering in 2012 and the M.Sc. degree in mobile computing systems with honorable mention in 2014, both from the Instituto Politécnico Nacional (IPN) in Mexico City, Mexico. Since 2015, he has been a Ph.D. student in telecommunications at the Communications Department of the Universitat Politècnica de València, Valencia, Spain, where he was a visiting researcher in 2014. His research interests include wireless sensor networks, communication systems, random access protocols, M2M communications, along with 5G and LTE/LTE-A networks.

Vicent Pla received the Telecommunication Engineering (B.E. & M.E.) and Ph.D. degrees from the Universitat Politècnica de València (UPV), Spain, in 1997 and 2005, respectively, and the B.Sc. in Mathematics from the Universidad Nacional de Educación a Distancia (UNED), Spain, in 2015. In 1999, he joined the Department of Communications at the UPV, where he is currently a Professor. His research interests lie primarily in the area of modeling and performance analysis of communication networks. During the past few years, most of his research activity has focused on traffic and resource management in wireless networks. In these areas he has published numerous papers in refereed journals and conference proceedings, and has been an active participant in several research projects.

Jorge Martinez-Bauset received the Ph.D. degree from the Universitat Politècnica de València (UPV), Valencia, Spain, in 1997. He also received the 1997 Alcatel Spain Best Ph.D. Thesis Award in Access Networks. He is currently a Professor with the UPV. From 1987–1991, he was with QPSX Communications, Perth, Australia, working with the team that designed the first IEEE 802.6 MAN. Since 1991, he has been with the Department of Communications, UPV. His research interests are in the area of performance evaluation and traffic control for multi-service networks.

José-Ramón Vidal is currently an Associate Professor in Telematics at the Higher Technical School of Telecommunication Engineering of the Universitat Politècnica de València (UPV), València, Spain. He obtained the Ph.D. degree in Telecommunication Engineering from the UPV. His current research interest is focused on the area of application of game theory to resource allocation in cognitive radio networks and to economic modeling of telecommunication service provision. In these areas he has published several papers in refereed journals and conference proceedings.

Vicente Casares-Giner (M'75-LM'17) assistant professor (1974), associate professor (1985), and full professor (1991). He obtained the Telecommunication Engineering degree in October 1974 from Escuela Técnica Superior de Ingenieros de Telecomunicación-Universidad Politécnica de Madrid (ETSIT-UPM) and the Ph.D. in Telecommunication Engineering in September 1980 from ETSIT-Universitat Politècnica de Catalunya (ETSITUPC), Barcelona. During the period 1974–1983 he worked on problems related to signal processing, image restoration, and propagation aspects of radio-link systems. In the first half of 1984 he was a visiting scholar at the KTH Royal Institute of Technology in Stockholm, dealing with digital switching and concurrent programming for Stored Program Control (SPC) telephone systems. From September 1, 1994 until August 31, 1995, he was a visiting scholar at WINLAB-Rutgers University-USA, working with random access protocols in wireless networks, wireless resource management, and land mobile trunking system. During the 90's he worked in traffic and mobility models in several European Union (EU) projects. Since September 1996, he is at ETSIT-Universitat Politècnica de València (ETSIT-UPV), Valencia, Spain. During the 00's and 10's he has been involved in several National and EU projects. Professor V. Casares-Giner has authored several papers in international magazines and conferences, such as, IEEE, Electronic Letters, Signal Processing, EURASIP-EUSIPCO, International Teletraffic Conference (ITC), Wireless conferences, IEEE (ICASSP, ICC, ICUPC, ICNC,...). He has served as General co-chair in the ISCC 2005, in the NGI-2006 and as TPC member in several conferences and workshops (Networking 2011, GLOBECOM 2013, ICC 2015, VTC 2016,...). His main interest is in the area of performance evaluation of wireless systems, in particular random access protocols, system capacity and dimensioning, mobility management, cognitive radio and wireless sensor networks.

Luis Guijarro received the M.Eng. and Ph.D. degrees in Telecommunications from the Universitat Politècnica de València (UPV), Spain. He is an Associate Professor in Telecommunications Policy with the UPV. He published the book "The Electronic Communications Policy of the European Union." He researched in traffic management in ATM networks and in e-Government, and his current research is focused on economic modeling of telecommunication service provision. He has contributed in the areas of peer-to-peer interconnection, cognitive radio networks, search engine neutrality, wireless sensor networks, and 5G.