

On the accurate performance evaluation of the LTE-A random access procedure

Israel Leyva-Mayorga, Luis Tello-Oquendo, Vicent Pla, Jorge Martinez-Bauset and Vicente Casares-Giner
ITACA, Universitat Politècnica de València, Spain
email: {isleyma, luiteloq, vpla, jmartinez, vcasares}@upv.es

Abstract—The performance evaluation of the random access (RA) procedure in LTE-A has recently become a major research topic as these networks are expected to play a major role in future 5G networks. Up to now, the key performance indicators (KPIs) of the RA in LTE-A have been obtained either by performing a large number of simulations or by means of analytic models that sacrifice precision in exchange of simplicity. In this paper, we present an analytic model for the performance evaluation of the LTE-A RA procedure. By means of this model, each and every one of the key performance indicators suggested by the 3GPP can be obtained with minimal error when compared to results obtained by simulation. To the best of our knowledge, this is the most accurate analytic model of the LTE-A RA procedure.

Index Terms—Analytic model; LTE-A; performance evaluation; random access (RA).

I. INTRODUCTION

The current LTE-A system has a widely deployed infrastructure, which provides with ubiquitous coverage and global connectivity [1]. As such, LTE-A networks present the best solution for the interconnection of mobile devices (known as user equipments, UEs, in LTE-A) and will serve as a base for the future development of the Internet of things (IoT) [2], [3].

The UEs access the cellular base station (eNB) by means of the random access (RA) procedure; it is performed through the random access channel (RACH) and comprises a four-message handshake: preamble transmission (only allowed in predefined time-frequency resources called random access opportunities, RAOs), random access response (RAR), connection request and contention resolution messages.

The RA procedure of LTE-A was not designed to handle a large number of synchronized access requests. This is a typical behavior in machine-to-machine (M2M) applications, in which the devices communicate autonomously [2], [3], [4]. Consequently, M2M applications may lead to severe congestion in the RACH and, due to the rapid increase in the number of interconnected devices, the frequency and severity of congestion will surely increase in the coming years. In order to develop efficient solutions to congestion in the RACH, the correct performance evaluation of the LTE-A RA procedure is of prime importance.

The performance evaluation of the RA procedure is oftentimes conducted by means of simulations [4] because it is difficult to model analytically. However, simulations may be highly time-consuming and the obtained results are not easily reproducible. One of the first efforts to model the RA procedure was presented in [5], but only the first step:

preamble transmission, is considered. In fact, there are just a few analytic models for the performance evaluation of the complete RA procedure and their accuracy suffers when compared to simulations [6], [7], [8].

The access delay of UEs is the KPI that is most neglected by the existing analytic models. For instance, a general model for the RACH is presented in [8]. Some of the shortcomings of this model, as described by the authors, are: a) the error of the presented model increases at certain traffic intensities and; b) the KPI with the largest relative error is the access delay. Furthermore, only the average access delay is calculated. Clearly, evaluating the access delay by means of its average value is not suitable for time-constrained applications, e.g., health care [9]; instead, the probability mass function (pmf) of delay should be obtained.

To the best of our knowledge, the most detailed analytic model for the performance evaluation of the RA procedure was presented in [6]. While this work was later extended in [7] to incorporate the model of an access control scheme, the basic model of the RA procedure remained. In fact, the model presented in [6] is of similar nature as the one presented in [8]; hence, both models present similar shortcomings. One of the main contributions of [6], [7] is that the probability distribution of access delay can be calculated, but its accuracy is poor. The main reason for this is the use of the expected value of the number of preambles decoded by the eNB instead of its pmf; this issue is described in detail in [10].

In this paper, we present a novel analytic model for the performance evaluation of the RA procedure. We describe the process for calculating the following key performance indicators (selected from the ones suggested by the 3GPP [4]):

- 1) Success probability, defined as the probability to successfully complete the RA procedure within the maximum number of preamble transmissions.
- 2) Probability distribution of the number of preamble transmissions performed by the UEs that successfully complete the RA procedure.
- 3) Probability distribution of the access delay.

The accuracy of our model is evaluated by comparing the results obtained with both, our model and the one presented in [6] with the ones obtained by simulation. Results show that the error obtained by means of our model is minimal and surpasses the accuracy of the model presented in [6]. In addition, results can be obtained within a few tens of seconds.

The rest of the paper is organized as follows. The RA procedure is described in detail and modeled in Section II. Then, the process for obtaining each of the KPIs mentioned above is described in Section III. The accuracy of our model is evaluated in Section IV and conclusions are presented in Section V.

II. MODELING THE RA PROCEDURE

In this section we describe in detail and provide the analytic model of the RA procedure.

The network operates in a slotted channel whose primary time unit is the subframe (of length $t_{sf} = 1$ ms). The time-frequency resources in which preamble transmissions are allowed (random access opportunities, RAOs) occur every t_{rao} subframes [11], [12]; t_{rao} is determined by the parameter *prach-ConfigIndex*, which is broadcast by the eNB through the *System Information Blocks* (SIBs).

For the sake of illustration, we use our analytic model to evaluate the performance of the RA in LTE-A under a massive M2M scenario. For this, we follow the recommended RACH configuration and traffic models described in [4]. Specifically, we select the baseline RACH configuration from [4, Table 6.1.1] and the traffic model 2 with $n_{m2m} = 30000$ M2M UEs. This combination leads to severe network congestion and the traffic load (number of UE arrivals per RAO) varies gradually from very low to extremely high and back to very low. Consequently, the accuracy of our model is evaluated in the whole spectrum of traffic loads. Yet another reason for selecting this configuration and traffic model is that most of the studies on the performance evaluation of the RA in LTE-A are performed under these conditions [6], [8], [7]. The selected configuration parameters of the RACH and of the traffic model 2 are shown in Table I on page 6.

Hereafter we denote by i and d , respectively, the number of elapsed RAOs and number of elapsed subframes. That is, the distributions presented in the following and whose domain is time can be given in either RAOs, i , or in subframes, d . The distributions that are given in RAOs are used to model the access of the UEs, whereas the distributions that are given in subframes are used to obtain the distribution of access delay. Recall that the duration of a subframe is 1 ms. In the selected RACH configuration (*prach-ConfigIndex* = 6), the periodicity of RAOs is $t_{rao} = 5$ subframes.

Under the traffic model 2, the UE arrivals follow a Beta(3, 4) distribution over 10 seconds [4]. As a result, the distribution period of the UE arrivals is $t_{dist} = 2000$ RAOs. Let $n_m(i, k)$ be the number of UEs that are about to perform their k th preamble transmission at the i th RAO. The expected number of UEs that are about to perform its first preamble transmission; i.e., the expected number of UEs that begin its RA procedure at the i th RAO is given as

$$\mathbb{E}[n_m(i, 1)] = n_{m2m} \text{Beta}(3, 4) = n_{m2m} \frac{60 i^2 (t_{dist} - i)^3}{t_{dist}^6}; \quad (1)$$

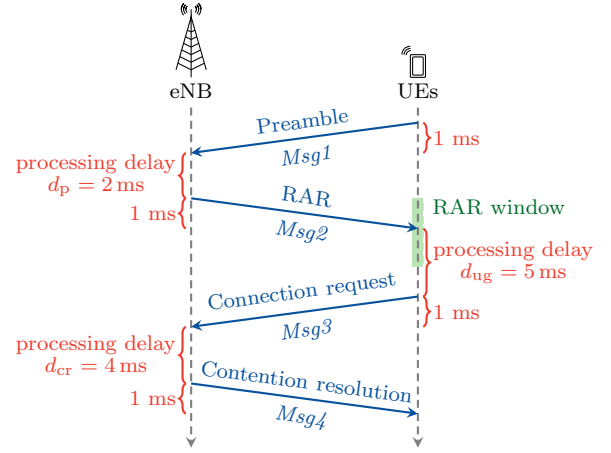


Figure 1. LTE-A contention-based RA procedure.

the expected number of UEs that are about to perform its k th preamble transmission will later be obtained recursively by means of (19).

The RA procedure, as briefly described in Fig. 1, is performed as follows [13], [11], [14], [15].

Preamble (Msg1): At the beginning of the RA procedure each UE randomly selects one out of the n_r available preambles and sends it to the eNB 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. If a preamble is transmitted (with sufficient power) by exactly one UE, it is decoded by the eNB. In this study we assume that if two or more UEs transmit the same preamble at the same RAO, a collision occurs at this point. This goes in line with the 3GPP recommendations for the performance evaluation of the RACH [4] and with most of the literature [6], [7], [8].

To model this step of the RA procedure, we first obtain the pmfs of the preambles transmitted by exactly one (successful transmissions) and by multiple UEs (collisions) for discrete values of the number of UEs that transmit a preamble at a specific RAO. Then we derive these same pmfs for any (continuous) value of the expected number of UEs that transmit a preamble at a specific RAO.

The process of preamble selection and transmission can be modeled as a bins and balls problem, as stated in [6]. For this, let k_{max} be the maximum number of preamble transmissions allowed per UE; this parameter is broadcast by the eNB through the *preambleTransMax* parameter included in the SIB2 [11]. Also let

$$n_m(i) = \sum_{k=1}^{k_{max}} n_m(i, k) \quad (2)$$

be the total number of balls (UEs that select and transmit a preamble at the i th RAO); each ball is randomly placed in one out of the n_r bins (available preambles). Let S and C be the random variables (RVs) that represent the number of bins with exactly one ball and the number of bins with more than one ball respectively; namely, the number of preambles

transmitted by one (successful) and by multiple UEs (with collision). The domain of S is $s = 0, 1, \dots, s_{\max}$, where $s_{\max} = \min\{n_r, n_m(i)\}$; the domain of C is $c = 0, 1, \dots, c_{\max}$, where $c_{\max} = \min\{n_r, n_m(i)/2\}$. To solve this problem efficiently, we calculate the joint probability distribution of S and C for a given $n_m(i)$, $p_{S,C}(s, c; n_m(i))$, recursively as

$$\begin{aligned} p_{S,C}(s, c; n_m(i)) &= \left(\frac{n_r - s + 1 - c}{n_r} \right) p_{S,C}(s-1, c; n_m(i)-1) \\ &+ \frac{c}{n_r} p_{S,C}(s, c; n_m(i)-1) \\ &+ \frac{s+1}{n_r} p_{S,C}(s+1, c-1; n_m(i)-1), \\ &\text{for } s = 0, 1, \dots, s_{\max}, \text{ and } c = 0, 1, \dots, c_{\max}, \end{aligned} \quad (3)$$

given the initial condition $p_{S,C}(0, 0; 0) = 1$.

That is, we derive the probability of having s preambles transmitted by exactly one and c by multiple UEs for a given (discrete) $n_m(i)$ from the case in which $n_m(i) - 1$ UEs have already selected its preamble.

The pmf of S for a given $n_m(i)$ is the marginal probability distribution of $p_{S,C}(s, c; n_m(i))$, given as

$$p_S(s; n_m(i)) = \sum_{c=0}^{c_{\max}} p_{S,C}(s, c; n_m(i)). \quad (4)$$

The pmf of S can be calculated once for $n_m(i) \in \{1, 2, \dots, \nu\}$, where $\nu \geq \max\{n_m(i)\}$, $i = 0, 1, 2, \dots$, and stored in a two-dimensional matrix for further use. For the selected scenario, $\nu \approx 350$.

Let $R_S(i)$ be RV that defines the number of preambles transmitted by exactly one and by multiple UEs at the i th RAO respectively, whose (discrete) domain is $r \in \{0, 1, \dots, n_r\}$. We derive the pmf of $R_S(i)$ from the pmf of S by means of the linear interpolation given as

$$\begin{aligned} p_{R_S}(r; i) &= p_S(s; \lceil \mathbb{E}[n_m(i)] \rceil) (\mathbb{E}[n_m(i)] - \lfloor \mathbb{E}[n_m(i)] \rfloor) \\ &+ p_S(s; \lfloor \mathbb{E}[n_m(i)] \rfloor) (1 - \mathbb{E}[n_m(i)] + \lfloor \mathbb{E}[n_m(i)] \rfloor), \end{aligned} \quad (5)$$

where

$$\mathbb{E}[n_m(i)] = \sum_{k=1}^{k_{\max}} \mathbb{E}[n_m(i, k)] \quad (6)$$

is the expected number of UEs that transmit a preamble at each RAO, which is continuous.

The pmf of the RV that defines the number of preambles with collision at the i th RAO, R_C , can be obtained by following an analogous methodology as the one described previously. However, obtaining this KPI is out of the scope of this paper.

Let the event \mathcal{D} be defined as the correct decoding at the eNB of a preamble transmitted by exactly one UE at a given RAO. Due to the use of power ramping, the probability that

the k th preamble transmitted by a UE is correctly decoded by the eNB can be modeled as

$$p_{\mathcal{D};k} = 1 - \frac{1}{e^k}; \quad (7)$$

this power ramping model was presented in [4] and has been adopted in other analytic models such as [6], [7]. By a slight abuse of notation, we denote the average preamble detection probability at the i th RAO by $p_{\mathcal{D};i}$; it is calculated from (7) as

$$p_{\mathcal{D};i} = \frac{1}{\mathbb{E}[n_m(i)]} \sum_{k=1}^{k_{\max}} p_{\mathcal{D};k} \mathbb{E}[n_m(i, k)], \quad (8)$$

Next, let $R_D(i)$ be the RV that defines the number of preamble transmissions that are correctly decoded by the eNB at the i th RAO; its pmf is calculated as

$$\begin{aligned} p_{R_D}(r; i) &= \sum_{\ell=r}^{n_r} \binom{\ell}{\ell-r} (1 - p_{\mathcal{D};i})^{\ell-r} p_{\mathcal{D};i}^r p_{R_S}(\ell; i), \\ &\text{for } r = 0, 1, \dots, n_r. \end{aligned} \quad (9)$$

Hence, the mean number of decoded preambles at the i th RAO is given as

$$\mathbb{E}[R_D(i)] = \sum_{r=0}^{n_r} r p_{R_D}(r; i). \quad (10)$$

RAR (Msg2): The eNB computes an identifier for each of the successfully decoded preambles and schedules the transmission of a RA response (RAR) message (Msg2). It includes, among other data, the uplink grants (reserved uplink resources) for the transmission of Msg3. 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, to receive an uplink grant from the eNB.

Up to n_{rar} uplink grants can be sent per subframe; each of which is associated to a successfully decoded preamble and the length of the RAR window, w_{rar} , is fixed. Consequently, there is a maximum number of uplink grants that can be sent within the RAR window. Only the UEs that receive an uplink grant can proceed with the transmission of Msg3.

Let $M_U(i)$ be the RV that defines the number of UEs that will receive an uplink grant in response to a preamble transmitted in the i th RAO. Let $n_{\text{ug}} = n_{\text{rar}} w_{\text{rar}}$ be the maximum number of uplink grants that can be sent per RAR window, hence, the domain of $M_U(i)$ is $m \in \{0, 1, \dots, n_{\text{ug}}\}$. The pmf of $M_U(i)$ is given as

$$p_{M_U}(m; i) = \begin{cases} p_{R_D}(m; i), & \text{if } m = 0, 1, \dots, n_{\text{ug}} - 1 \\ \sum_{r=m}^{n_r} p_{R_D}(r; i), & \text{if } m = n_{\text{ug}} \end{cases} \quad (11)$$

and its expected value is

$$\mathbb{E}[M_U(i)] = \sum_{m=0}^{n_{\text{ug}}} m p_{M_U}(m; i). \quad (12)$$

Note that $\mathbb{E}[M_U(i)]$ is indeed the expected number of UEs that successfully complete the first two steps of the RA procedure. Hence the expected number of UEs that successfully complete the first two steps of the RA procedure in its k th preamble transmission can be obtained as

$$\mathbb{E}[M_U(i, k)] = \frac{\mathbb{E}[M_U(i)] \mathbb{E}[n_m(i, k)] p_{\mathcal{D};k}}{\mathbb{E}[n_m(i)] p_{\mathcal{D};i}}. \quad (13)$$

Then, the expected number of failed UE accesses can be easily calculated as

$$\mathbb{E}[M_F(i, k)] = \mathbb{E}[n_m(i, k)] - \mathbb{E}[M_U(i, k)]. \quad (14)$$

Backoff: If multiple UEs transmit the same preamble or if the power used for the preamble transmission is not sufficient, then the preamble transmission fails. If the maximum number of preamble transmissions, k_{\max} (notified by the eNB through the SIB2 [11]), has not been reached, failed UEs ramp up their power and re-transmit a new randomly chosen preamble in a new RAO. For this, the UE waits for a random time, $U(0, b_i)$ ms, and then performs a new preamble transmission at the next RAO; b_i is the backoff indicator defined by the eNB, and its value ranges from 0 to 960 ms. The UEs are only aware of a failed preamble transmission if no uplink grant has been received at the end of the RAR window. As a result, the UEs will not be aware of the failed transmission until

$$d_f = 1 + d_p + w_{\text{rar}} \quad (15)$$

subframes have elapsed; i.e., one subframe is required for preamble transmission, d_p subframes are needed to process the transmitted preambles at the eNB and w_{rar} is the length of the RAR window.

When a UE has transmitted k_{\max} preambles without success, the network is declared unavailable by the UE, a RA problem is indicated to upper layers, and the RA procedure is terminated. The expected number of UEs that terminate its RA procedure at the i th RAO is simply given as

$$\mathbb{E}[M_F(i, k_{\max})] = \mathbb{E}[n_m(i, k_{\max})] - \mathbb{E}[M_U(i, k_{\max})]. \quad (16)$$

Let B be the RV that represents the number of RAOs that a UE has to wait due to backoff. Also, recall that K represents the number of preamble transmissions performed by a UE. If $k = 1$, the UE succeeds in its first preamble transmission and backoff is not performed. Therefore, the conditional pmf of B given $k = 1$ is given as

$$p_{B|K}(i|1) = \delta(i) \equiv \begin{cases} 1, & \text{if } i = 0 \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

It is clear that the conditional pmf of B given $k = 2$ is positive between $i_{B,\min} = \lceil d_f/t_{\text{rao}} \rceil$ and $i_{B,\max} = \lceil (d_f + b_i)/t_{\text{rao}} \rceil$,¹ and is given as

$$p_{B|K}(i|2) = \frac{1}{b_i} \begin{cases} i t_{\text{rao}} - d_f, & \text{if } i = i_{B,\min} \\ t_{\text{rao}}, & \text{if } i_{B,\min} < i < i_{B,\max} \\ d_f + b_i - (i - 1) t_{\text{rao}}, & \text{if } i = i_{B,\max}. \end{cases} \quad (18)$$

¹In an abuse of notation, here and in (18) the backoff indicator, given in subframes, is simply denoted as b_i .

This conditional pmf is of special importance because it allows us to model the backoff process at each RAO by means of the following recursion

$$\mathbb{E}[n_m(i, k)] = \sum_{j=j_{\min}}^{j_{\max}} \mathbb{E}[M_F(i - j, k - 1)] p_{B|K}(j|2), \quad i = 1, 2, \dots, i_{\max}; k = 2, 3, \dots, k_{\max}; \quad (19)$$

where

$$i_{\max} = t_{\text{dist}} + (k_{\max} - 1) i_{B,\max} \quad (20)$$

is the last RAO in which a preamble transmission can occur, $j_{\min} = \min\{i_{B,\min}, i\}$, $j_{\max} = \min\{i_{B,\max}, i\}$ and $\mathbb{E}[n_m(0, k)] = 0$ for $k \geq 2$.

From (18), the conditional pmf of B given K can be calculated recursively as

$$p_{B|K}(i|k) = \sum_{\ell=i_{B,\min}}^{i_{B,\max}} p_{B|K}(\ell|2) p_{B|K}(i - \ell|k - 1), \quad k = 3, 4, \dots, k_{\max}. \quad (21)$$

Let D_{BO} the RV that represents the total number of subframes that a UE has to wait due to backoff during its RA procedure. Clearly, the pmf of D_{BO} conditioned to the number of preamble transmission attempts, k , can be easily calculated as

$$\Pr\{D_{\text{BO}} = i t_{\text{rao}} | K\} = p_{B|K}(i|k). \quad (22)$$

The pmf of the backoff time conditioned to the number of preamble transmissions can be obtained once and used repeatedly.

Connection request (Msg3) and contention resolution (Msg4): After receiving the corresponding uplink grant, the UE adjusts its uplink transmission time according to the received time alignment and transmits a scheduled connection-request message, *Msg3*, to the eNB using the reserved uplink resources. The RA procedure is concluded when the eNB sends the contention resolution message (*Msg4*) to the UEs in response to the connection request message. Hybrid automatic repeat request (HARQ) is used to protect the *Msg3* and *Msg4* transmissions. 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.

Let D_{M3} be the RV that denotes the number of subframes elapsed between the first transmission attempt of *Msg3* by a UE and the successful transmission of a *Msg3* by the same UE, conditioned to the fact that the transmission of *Msg3* will succeed within the maximum number of attempts. The distribution of D_{M3} , $p_{D_{\text{M3}}}(d)$, depends on its round-trip time, d_{m3} , the probability of error during the transmission, $p_{\text{E}_{\text{M3}}}$, and the maximum number of transmission attempts, h_{\max} . To obtain the pmf of D_{M3} , let H be the RV that defines the number of attempts that would be required for the successful transmission of *Msg3*. It is clear that the pmf of D_{M3} given $H = h$ is given as

$$p_{D_{\text{M3}}|H}(d|h) = \delta(d - (h - 1) d_{\text{m3}}), \quad (23)$$

Each *Msg3* transmission has two possible outcomes: successful or not successful, and the number of transmission attempts is limited to h_{\max} . For the sake of simplicity, we consider that the UEs that fail its *Msg3* (or *Msg4*) do not go back to preamble transmission and terminate their RA procedure at this point. This assumption has no impact on the accuracy of our model, since its probability of occurrence,

$$p_{\mathcal{E}_M} = p_{\mathcal{E}_{M3}}^{h_{\max}} + (1 - p_{\mathcal{E}_{M3}}^{h_{\max}}) p_{\mathcal{E}_{M4}}^{h_{\max}}, \quad (24)$$

is very low for typical values of $p_{\mathcal{E}_{M3}}$ and $p_{\mathcal{E}_{M4}}$ (see Table I on page 6).

Therefore, the distribution of D_{M3} alone can be calculated as

$$p_{D_{M3}}(d) = \frac{1 - p_{\mathcal{E}_{M3}}}{1 - p_{\mathcal{E}_{M3}}^{h_{\max}}} \sum_{h=1}^{h_{\max}} p_{\mathcal{E}_{M3}}^{h-1} \delta(d - (h-1)d_{m3}) \quad (25)$$

The distribution of D_{M4} can be obtained in the same manner as D_{M3} just by substituting the round-trip time, d_{m3} , with d_{m4} .

Next, let D_M be the RV that denotes the number of subframes elapsed between the first transmission attempt of *Msg3* and the successful transmission of *Msg4*. The pmf of D_M is given by the sum of D_{M3} and D_{M4} as

$$\begin{aligned} p_{D_M}(d) &= \Pr\{D_{M3} + D_{M4} = d\} \\ &= \sum_{\ell=0}^d p_{D_{M3}}(\ell) p_{D_{M4}}(d - \ell) \end{aligned} \quad (26)$$

Let the RV $M_S(i, k)$ define the number of UEs that successfully transmit their k th preamble at the i th RAO and that will complete the remaining steps of the RA procedure. The expected value of $M_S(i, k)$ is given as

$$\mathbb{E}[M_S(i, k)] = (1 - \Pr\{\mathcal{E}_M\}) \mathbb{E}[M_U(i, k)]. \quad (27)$$

Let D be the RV that defines the number of subframes elapsed since the beginning of the RA until its successful completion. The minimum number of subframes required to successfully complete the RA procedure (minimum access delay) is obtained as

$$d_{\min} = \min\{d \mid \Pr\{D = d\} \geq 0\} = 4 + d_p + d_{ug} + d_{cr}, \quad (28)$$

since 4 subframes are needed for the transmission of *Msg1*, *Msg2*, *Msg3* and *Msg4*; d_p , d_{ug} and d_{cr} are the processing delays of the preamble, uplink grant and connection request messages respectively. Next, let the RV D_{\min} define the minimum access delay; its pmf is given as

$$p_{D_{\min}}(d) = \delta(d - d_{\min}) \quad (29)$$

III. OBTAINING THE KEY PERFORMANCE INDICATORS

In this section we describe in detail the process for obtaining the KPIs for the performance evaluation of the RA in LTE-A. For this, it is necessary to model the RA procedure for each RAO since the beginning of the distribution period, $i = 0$, until the last RAO in in which a preamble transmission can occur, i_{\max} .

The expected number of UEs (out of a total n_{m2m} UEs) that successfully complete the RA procedure (with any number of preamble transmissions) within the whole distribution period is calculated as

$$\mathbb{E}[M_S] = \sum_{i=0}^{i_{\max}} \sum_{k=1}^{k_{\max}} \mathbb{E}[M_S(i, k)]. \quad (30)$$

The probability that a UE successfully completes the RA procedure is the access success probability, given as

$$p_S = \frac{\mathbb{E}[M_S]}{n_{m2m}}. \quad (31)$$

The pmf of the number of preamble transmissions performed by a UE that successfully completes its RA procedure is given as

$$p_{K|S}(k) = \frac{1}{\mathbb{E}[M_S]} \sum_{i=0}^{i_{\max}} \mathbb{E}[M_S(i, k)], \quad \text{for } k = 1, 2, \dots, k_{\max}, \quad (32)$$

hereafter simply denoted as $p_K(k)$. Its expected value can be easily calculated as

$$\mathbb{E}[K] = \sum_{k=1}^{k_{\max}} k p_K(k). \quad (33)$$

Let K_ϕ be the ϕ th percentile of the number of preamble transmissions, i.e., the ϕ percent of the UEs successfully complete the RA procedure with K_ϕ or less preamble transmissions. K_ϕ is calculated by means of a linear interpolation of the CDF of K , $F_K(k)$.

To calculate the pmf of the access delay, recall that the eNB can assign up to n_{rar} uplink grants per subframe; since the RAR window is comprised of w_{rar} subframes, the eNB can assign up to $n_{\text{ug}} = w_{\text{rar}} n_{\text{rar}}$ uplink grants per RAO. Let W be the RV in the domain $d \in \{0, 1, \dots, w_{\text{rar}} - 1\}$ that defines the subframe of the RAR window in which the UEs receive the uplink grant. The pmf of W is calculated as

$$p_W(d) = \frac{1}{\mathbb{E}[M_S]} \sum_{i=0}^{i_{\max}} \max\{0, \min\{n_{\text{rar}}, \mathbb{E}[M_S(i)] - (d n_{\text{rar}})\}\}; \quad \text{for } d = 0, 1, \dots, w_{\text{rar}} - 1. \quad (34)$$

Finally, the pmf of the access delay is given as

$$p_D(d) = \Pr\{D = d\} = \Pr\{D_{\text{BO}} + D_M + D_{\min} + W = d\}; \quad (35)$$

i.e., we calculate the pmf of the access delay as the convolution of the pmfs of the backoff time, D_{BO} , the successful transmission of *Msg3* and *Msg4*, D_M , the minimum access delay, D_{\min} , and the subframe in which the uplink grant is received, W . These pmfs are calculated in (22), (26), (29) and (34) respectively.

From (35), the expected value of the access delay, $\mathbb{E}[D]$ and its CDF, $F_D(d)$, can easily obtained. Let D_ϕ be the ϕ th percentile of the access delay, i.e., the ϕ percent of the UEs successfully complete the RA procedure with a delay that is less than or equal to D_ϕ . D_ϕ is obtained by means of a linear interpolation of $F_D(d)$.

Table I
PARAMETERS FOR THE SELECTED RACH CONFIGURATION AND TRAFFIC MODEL 2.

Parameter	Symbol	Setting
Number of M2M UEs	n_{m2m}	30000
Distribution period	t_{dist}	2000 RAOs
Distribution of UE arrivals		Beta (3, 4)
PRACH Configuration Index	$prach-ConfigIndex$	6
Periodicity of RAOs	t_{rao}	5 subframes
Subframe length	t_{sf}	1 ms
Available preambles	n_r	54
Maximum number of preamble transmissions	k_{max}	10
RAR window size	w_{rar}	5 subframes
Available uplink grants per sub-frame	n_{rar}	3
Backoff Indicator	b_i	20 ms
Re-transmission probability for $Msg3$ and $Msg4$	$p_{E_{M3}} = p_{E_{M4}}$	0.1
Maximum number of $Msg3$ and $Msg4$ transmissions	h_{max}	5
Preamble processing delay	d_p	2 subframes
Uplink grant processing delay	d_{ug}	5 subframes
Connection request processing delay	d_{cr}	4 subframes
RTT of $Msg3$	d_{m3}	8 subframes
RTT of $Msg4$	d_{m4}	5 subframes

IV. PERFORMANCE EVALUATION

In this Section, the accuracy of our model (with respect to simulations) is evaluated. For this, the model presented by C. H. Wei et. al. [6] is selected as reference and hereafter is simply denoted as the reference model. The parameters for the selected traffic model and RACH configuration are shown in Table I.

The presented simulation results were obtained by means of two independent discrete-event simulators. This allowed us to confirm our results. The first simulator is C-based and the second one is coded in Octave. In each simulation, n_{m2m} UE arrivals are distributed within a period of t_{dist} RAOs, then the contention-based RA procedure is replicated with the parameters listed in Table I. Simulations are run j times until the all the cumulative KPIs obtained up to the j th simulation differ from those obtained up to the $(j-1)$ th simulation by less than 0.01 percent. For all of the KPIs presented in Table II the relative margin of error is less than 0.5 percent at a 95 percent confidence level.

We begin our analysis by comparing the expected number of successful accesses at each RAO, $\mathbb{E}[M_S(i)]$, obtained by simulation, by the reference model and by our proposed model in Fig. 2a; the distribution of UE arrivals, $\mathbb{E}[n_m(i, 1)]$, is also displayed in order to add context. In Fig. 2b we show the absolute error of the calculated $\mathbb{E}[M_S(i)]$ at each RAO. From Fig. 2a and Fig. 2b it can be clearly observed that the results obtained by the two models and by simulation are extremely similar for most of the RAOs. The most notorious exception of this is observed for the reference model in the RAOs where $\mathbb{E}[M_S(i)] \approx 15$, where an absolute error of up to 2 successful accesses per RAO is obtained. The main reason for this is that the number of uplink grants per RAO is calculated from the

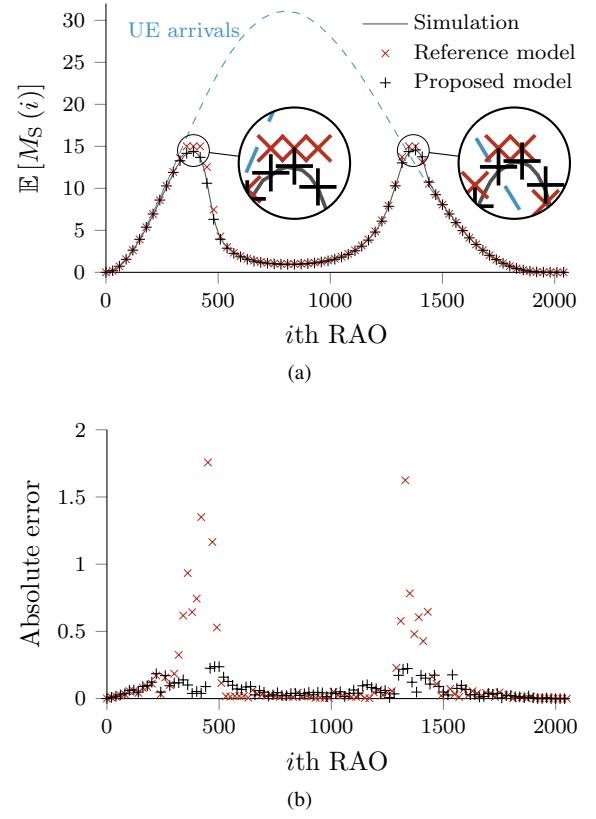


Figure 2. (a) Comparison and (b) absolute error of the expected number of successful accesses at each RAO, $\mathbb{E}[M_S(i)]$, obtained by simulation, by the reference model [6] and by our proposed model.

Table II
KPIs OBTAINED BY SIMULATION AND THE RELATIVE ERROR OBTAINED BY THE REFERENCE MODEL AND BY OUR PROPOSED MODEL.

Key Performance Indicator		Simulation	Relative error (%)	
			Reference	Proposed
Success probability (%)	p_S	31.31	2.76	0.24
Number of preamble transmissions, K	$\mathbb{E}[K]$	3.42	2.29	0.36
	K_{10}	1.00	0.00	0.00
	K_{50}	1.96	2.63	0.73
	K_{95}	8.57	1.01	0.13
Access delay, D (ms)	$\mathbb{E}[D]$	68.32	3.84	1.97
	D_{10}	15.05	26.21	0.36
	D_{50}	46.35	11.73	1.07
	D_{95}	182.42	6.61	0.24

expected number of decoded preambles. As a result of this, the number of successful accesses is overestimated. This issue is described in detail in [8].

Note that, by using the pmf of the expected number of decoded preambles, this error is not present in our model (see (11)), as the absolute error obtained in every RAO is minimal, hence our model is extremely accurate.

To provide with an in-depth look at the accuracy of our model, we show the KPIs obtained by simulation and the relative error between these KPIs and the ones obtained by both of the analytic models in Table II.

From Table II it can be seen that each and every one of the

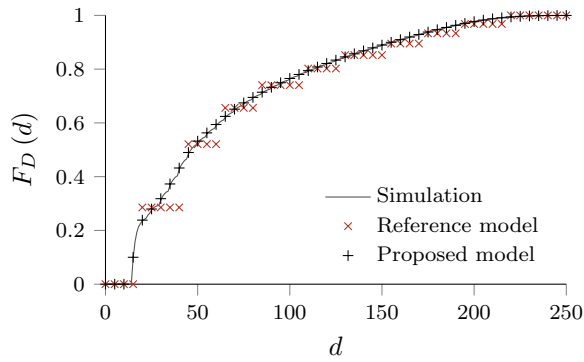


Figure 3. Comparison of the CDF of the access delay, $F_D(d)$, obtained by simulation, by the reference model [6] and by our proposed model.

KPIs obtained by our analytic model are extremely similar to the ones obtained by simulation. In contrast, the reference model leads to an error larger than 2 percent for several KPIs. Specifically, a large error of up to 26 percent is obtained in the percentiles of access delay with the aforementioned model. The reason for such a large error is, once again, the use of expected values instead of the pmf. Specifically, the expected subframe of the RAR window in which the uplink grant is received and the expected delay due to the transmission of *Msg3* and *Msg4* are used, not their pmf. The result of this is the step-like function depicted in Fig. 3.

V. CONCLUSION

In this paper we have presented a novel analytic model for the performance evaluation of the RA procedure in LTE-A. The accuracy of our model has been evaluated with respect to simulation results and then compared with that of the reference model (proposed by C. H. Wei et. al.). Despite the latter was the most accurate model prior to ours, its accuracy drops when the number of successful accesses per RAO approximates the system capacity; i.e., when most of the resources are being utilized. Note that these are the scenarios of highest interest, because the main objective of access control schemes is to reduce congestion while efficiently using the available resources.

Results show that the accuracy of our model surpasses that of the reference model. As such, our model is, to the best of our knowledge, the most accurate analytic model of the RA procedure in LTE-A and its accuracy is not affected by the distribution of the UE arrivals; still, it maintains an acceptable degree of (computational) complexity. For instance,

by implementing our model in Octave, results were obtained within a few tens of seconds.

ACKNOWLEDGMENT

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 I. Leyva-Mayorga was partially funded by grant 383936 CONACYT-Gobierno del Estado de México 2014. 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.

REFERENCES

- [1] 3GPP, "TS 23.682, Architecture enhancements to facilitate communications with packet data networks and applications," Mar 2016.
- [2] 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, 2015.
- [3] P. K. Verma, R. Verma, A. Prakash, A. Agrawal, K. Naik, R. Tripathi, T. Khalifa, M. Alsabaan, T. Abdelkader, and A. Abogharaf, "Machine-to-Machine (M2M) Communications: A Survey," *J. Netw. Comput. Appl.*, vol. 66, pp. 83–105, 2016.
- [4] 3GPP, "Study on RAN Improvements for Machine-type Communications," *TR 37.868*, Jul 2011.
- [5] Ping Zhou, Honglin Hu, Haifeng Wang, and Hsiao-hwa Chen, "An efficient random access scheme for OFDMA systems with implicit message transmission," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2790–2797, 2008.
- [6] 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.
- [7] 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.
- [8] O. Arouk and A. Ksentini, "General Model for RACH Procedure Performance Analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, 2016.
- [9] 3GPP, "Service requirements for Machine-Type Communications," *TS 22.368 V13.2.0*, Dec 2016.
- [10] O. Arouk, A. Ksentini, and T. Taleb, "How accurate is the RACH procedure model in LTE and LTE-A?" in *Proc IEEE International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2016, pp. 61–66.
- [11] 3GPP, "Radio Resource Control (RRC); Protocol specification," *TS 36.331 V13.0.0*, Jan 2016.
- [12] —, "Physical channels and modulation," *TS 36.211 V12.6.0*, Sept 2015.
- [13] —, "Medium Access Control (MAC) protocol specification," *TS 36.321 V13.0.0*, Feb 2016.
- [14] —, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," *TS 36.213 V13.0.0*, May 2016.
- [15] —, "Feasibility study for Further Advancements for E-UTRA," *TR 36.912 V13.0.0*, Jan 2016.