Probabilistic Retransmissions for the Random Access Procedure in Cellular IoT Networks

Carlos A. Astudillo, Fernando H. S. Pereira, and Nelson L. S. da Fonseca
Institute of Computing - State University of Campinas, Brazil
castudillo@lrc.ic.unicamp.br, fernandopereira@lrc.ic.unicamp.br, nfonseca@ic.unicamp.br

Abstract—**The collision of multiple MSG$_3$ transmissions due to the selection of the same preamble sequence in the Long Term Evolution (LTE) Random Access procedure is an important problem which can impact on the performance of cellular Internet of Things (IoT) networks. In this paper, we propose a standard-compatible probabilistic retransmission approach to reduce the number of collisions of MSG$_3$ messages in cellular IoT technologies. In our proposal, every Machine-Type Communications (MTC) device with an uplink grant for retransmitting an MSG$_3$ message uses a probability value to decide whether or not to transmit. Two retransmission policies were proposed to reduce the number of simultaneous MSG$_3$ messages received at the base station. To apply these policies, the estimation of the number of MTC devices trying random access in a given Random Access Opportunity is required. A novel method to estimate this value at the device side is proposed based on Random Access Response (RAR) message counting and the Access Class Barring (ACB) barring probability. Results derived via simulations show that the proposed approach decreases the number of collisions of MSG$_3$ messages, reducing the access delay and energy consumption, as well as decreasing the utilization of the Packet Uplink Shared Channel (PUSCH) when compared to conventional LTE Random Access scheme.**

Keywords*—*Internet of Things, LTE-A networks, machine-to-machine communications,, random-access procedure.

I. INTRODUCTION

Machine-Type Communications (MTC) enables communication between geographically distributed devices with reduced or even without human intervention [1]. The support of Massive Machine-Type Communications (mMTC) is one of the main use case categories in the fifth generation (5G) cellular network to attend the expected massive Internet of Things (IoT) market. These networks will use a combination of the evolution of the Long Term Evolution (LTE)[1] technology for cellular IoT networks, such as LTE-Advanced (LTE-A), LTE-MTC (LTE-M) and Narrowband IoT (NB-IoT), as well as new radio technologies. Even though LTE is a key technology for IoT connectivity giving its high capacity, security and ubiquity coverage, the introduction of a huge number of MTC devices transmitting sporadically small packets may overwhelm current cellular networks. Specifically, massive MTC communications can strongly affect the Random Access (RA) procedure operation, jeopardizing the network performance and increasing the energy consumption.

The contention-based LTE RA is a four-way handshake procedure used by the User Equipment (UE) to establish a network connection, request uplink resources, and perform handover. The UE randomly selects a preamble sequence from a set of available preambles and transmit it (MSG$_1$) on the Physical Random Access Channel (PRACH). Upon detecting a preamble sequence, the evolved NodeB (eNB) creates a Random Access Response (RAR) message (MSG$_2$) to be transmitted on the Physical Downlink Shared Channel (PDSCH), including an uplink grant for a transmission on the Packet Uplink Shared Channel (PUSCH). Once the RAR message is received by the UE, it transmits an L2/L3 message (MSG$_3$) on the PUSCH as indicated in the RAR message. In LTE, all PUSCH transmissions, including transmission of MSG$_3$ messages, use synchronous Hybrid Automatic Repeat Request (HARQ) to deal with channel error and interference. Upon successful reception of an MSG$_3$ message, the eNB creates a Contention Resolution (CR) message MSG$_4$ to be transmitted on the PDSCH. The successful reception of the CR message finishes the contention-based RA procedure.

However, if two or more UEs select the same preamble sequence in a certain Random Access Opportunity (RAO), a preamble collision occurs. Different from other technologies such as Radio-Frequency Identification (RFID), an LTE eNB is usually not able to detect such collisions during the preamble detection phase [2]. As a consequence, the eNB allocates an uplink transmission grant using a RAR message (MSG$_2$) for each detected preamble in order to transmit a MSG$_3$ messages, even if a preamble collision has occurred. In this case, the transmissions of MSG$_3$ messages by multiple UE devices collide, which is called the *MSG$_3$ collision problem*. As MSG$_3$ transmissions employ HARQ, used for the regularly scheduled user data transmissions in LTE, the MSG$_3$ transmissions of the involved UE devices may result in further collision events. If no RAR message is received or the maximum number of MSG$_3$ HARQ transmissions is achieved, the UE enters a backoff period and then re-initiates the RA procedure. This is repeated until the UE receives an RAR message or until the maximum number of preamble sequence transmissions is achieved. Indeed, the LTE HARQ protocol was not designed to the situation in which more that one user can transmit on the same resource, quite common in IoT over cellular network scenarios.

Under massive MTC scenarios, RA attempts from several users may occur at the same time, increasing the number of preamble collisions and the impact of the *MSG$_3$ collision problem* on the network performance. Despite the impact of the MSG$_3$ collisions in the performance of IoT cellular networks, it has received little attention in the literature. Most of the proposals are related to the Radio Access Network

---

[1] we use LTE throughout the paper to refer to all technologies based on 3GPP LTE standards (release 8 and beyond).

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

(RAN) overload problem, which do not tackle directly the *MSG3 collision problem*. For instance, the Third Generation Partnership Project (3GPP) proposed various methods for alleviating this problem in [3], resulting in the inclusion of two contention-avoidance RA techniques in the LTE standard, called Access Class Barring (ACB) and Extended Access Barring (EAB). However, these methods just avoid users to attempt network access at particular times. On the other hand, even though contention-resolution RA schemes effectively address the *MSG_3 collision problem*, they usually require extra signaling and make non-realistic assumptions. Thus, those proposals are difficult to be incorporated into real cellular networks.

In this paper, we propose a probabilistic retransmission approach to ameliorate the *MSG_3 collision problem* in cellular IoT network technologies. Our approach is standard-compatible, which means that it does not require additional signaling to operate in LTE networks. Moreover, it also implies that few modifications at the eNB and UE device are required to implement our proposal in commercial cellular networks. In our proposal, every MTC device requiring a retransmission of a MSG_3 message locally calculates a retransmission probability, which is used to decide whether a given MSG_3 transmission is performed at a given instant. By doing so, our proposal reduces the number of simultaneous MSG_3 messages received at the base station, allowing the successful detection of MSG_3 messages under a preamble collision. Two retransmission policies were proposed based either on the expected number of collided MTC devices per detected preamble or the expected number of collided MTC devices per collided preamble. Moreover, we propose a novel method to estimate the number of MTC devices trying random access in a given Random Access Opportunity at the UE side, which is required to calculate the proposed retransmission policies. This method is based on RAR message counting and the ACB barring probability broadcasted by the eNB. Results derived via simulations show that the proposed approach effectively address the *MSG_3 collision problem*. It decreases the access delay, energy consumption, and the PUSCH resource utilization when compared to the conventional LTE RA scheme.

The rest of the paper is organized as follows. Section II briefly reviews the related work on the MSG_3 collision problem. Section III presents the proposed probabilistic retransmission approach. Section IV shows the performance of the proposed approach via simulations and discusses the simulation results. Finally, we conclude the paper in Section V.

II. RELATED WORK

RA schemes for machine-to-machine communications have recently attracted a lot of attention in the research community because they have an important role in the support of IoT over cellular networks. Two main techniques can be used [4]: the contention-avoidance schemes, which aim at reducing the number of attempts under high loads, impacting the preamble transmission phase of the RA procedure, and the contention-resolution schemes, which aim at resolving the collisions among MTC devices during the RA procedure. In the former, the ACB and EAB are the main approaches [3], both barring some devices to attempt random access during high random access loads. However, in the following, we focus on the review of mechanisms that reduce MSG_3 collisions, including contention-resolution solutions.

Ali *et. al* [4] proposed a contention resolution based on a m-ary contention tree splitting technique. The scheme is based on a non-standard RA technique that allows transmitting the preamble sequence jointly with the UE identity, which is used to identify collided preambles. A binary tree is created for each collided preamble and the RAR message is used to inform the collided UE about the resource to be used in the next RA attempt, sending uplink grants to uncollided preamble only. The process is finalized when all collisions are resolved. Vilgelm *et. al* [5] [6] propose a RA protocol for LTE networks based on binary countdown technique for contention resolution. This protocol introduces micro-slots before MSG_3 transmission for prioritizing MTC devices and resolving contention. This approach also assumes that MTC devices can listen to the transmission of each other. In general, the main problem with the contention-resolution-based RA approaches is that extra signaling between the eNB and the UEs or among UEs is required to introduce the solution into the LTE protocol, making them difficult to be implemented in commercial cellular networks.

Kim *et. al* [2] proposed a RA procedure in which multiple RAR messages per detected preamble are sent in order to reduce the number of MSG_3 collisions due to preamble collision. However, this scheme have some drawbacks. The eNB does not have idea about the number of MTC devices per collided preamble. So, it allocates multiple uplink grants per detected preamble based on the estimated expected value of that variable. However, this may waste a lot of Physical Uplink Control Channel (PUCCH) resources, which decreases the resources availability for actual user data. Magrin *et. al* [7] introduced a method to estimate the number of UEs that chose the same preamble based on machine learning techniques. Even though the authors showed promising results with a synthetically generated dataset, the proposed technique adds additional complexity to the eNB and the collection of a real dataset is difficult with current eNB implementations. Thus, the proposed technique is not easily implementable in existing cellular networks.

The capture effect, which allows the decoding of one of the interfering signals, was also exploited in RA schemes for increasing the MSG_3 detection probability by applying power ramping technique [8] or multiple power levels [9] to the MSG_3 transmissions. However, these approaches increase the MTC device energy consumption [8] as well as the interference that the PUSCH can cause to PUSCH/PRACH of neighboring cells in co-channel deployments.

Ko *et. al* [10] proposes a mechanism that avoids sending multiple MSG_3 messages based on the time-advanced commands received as part of the RAR message. However, this approach can be exploited just in MTC devices with no mobility and when the collided MTC devices are located at different distance from the eNB. Even if the capture effect could be exploited, it allows for the decoding of a maximum of one user transmission per RAR message.

Liang *et. al* [11] proposed the Non-Orthogonal Random Access (NORA) scheme which employs Self-Interference Cancellation (SIC). This scheme introduced a technique to detect preamble collisions and exploits the use of power-domain Non-Orthogonal Multiple Access (NOMA) to decode more than one MSG_3 messages per detected preamble. The main limitations of this proposal includes the reduced chance of detecting a preamble collision in small cells as well as the increase in eNB complexity due to the SIC receiver and the superimposed preamble detection.

In summary, contention-avoidance RA schemes just avoid some MTC devices to send preamble, whereas contention-resolution RA schemes effectively tackle the *MSG₃ collision problem* by avoiding devices to transmit their MSG_3 messages. However, existing techniques to address the *MSG₃ collision problem* requires additional signaling messages and are based on nonstandard compliant procedures that make them difficult to be implemented in real LTE networks.

III. PROBABILISTIC HARQ RETRANSMISSIONS FOR THE RA PROCEDURE IN CELLULAR IOT NETWORKS

In this section, we describe the proposed probabilistic HARQ retransmission protocol, especially designed to be used by MTC devices in the RA procedure of cellular IoT network technologies. Even though we describe a solution for the Frequency Division Duplexing (FDD)-based LTE, LTE-A, and LTE-M technologies in this paper, our proposal can be easily adapted to other cellular IoT technologies such as LTE-based Time Division Duplexing (TDD) and NB-IoT as well.

We tackle the *MSG₃ collision problem* by introducing the concept of *probabilistic HARQ transmissions* into the HARQ mechanism used in the LTE RA procedure. One of the main advantages of our proposal is that no extra signaling is needed to be introduced in the 3GPP standard in order to be implemented. The proposed protocol is based on the standardized RA procedure, RA schemes and HARQ control signaling messages to operate.

A. The proposed probabilistic HARQ protocol

The proposed HARQ protocol aims at increasing the number of MSG_3 messages successfully decoded per RAR message in the RA procedure. A general assumption in Orthogonal Frequency Division Multiple Access (OFDMA)/Single-Carrier Frequency Division Multiple Access (SC-FDMA) systems is that intra-cell interference is avoided by scheduling the orthogonal resources to a single user per Transmission Time Interval (TTI). Thus, existing HARQ protocols for those systems are designed for the event in which uplink transmissions are scheduled by the base station in such a way that the decoding problems are due to channel impairments, noise, and inter-cell interference. However, in the RA procedure, the MSG_3 message transmission from several users can collide when they receive the same uplink grant in the RAR message to transmit their MSG_3 messages (intra-cell interference). This happens because more than one user selected the same preamble in the initial phase of the RA procedure.

Considering that the LTE protocol applies power control to the PUSCH transmissions, capture effect in the MSG_3 reception is difficult to be achieved due to the small differences in the Signal to Interference Ratio (SIR) of the involved interfering signals [9]. Even if the capture effect could be exploited, it allows the decoding of a maximum of one user transmission per RAR message. Thus, the goal of the proposed protocol is to achieve a high number of MSG_3 decoding per RAR message to improve the overall RA procedure performance. By doing this, access delay, energy consumption, and preamble collisions are also decreased.

The proposed HARQ protocol works as follows. Once an MTC device receives the RAR message containing an uplink grant to send the MSG_3 message, the first MSG_3 message is always transmitted using the value of the HARQ probability equals to 1. If this transmission is successfully decoded, the eNB sends back an acknowledgment (ACK) message on the Physical HARQ Indicator Channel (PHICH). The ACK in the first transmission indicates that just one MTC device was transmitting on the allocated resources (neglecting the capture effect) and the HARQ process is finalized as usual in both the UE and the eNB. However, if the first transmission is not decoded, the eNB sends back a negative-acknowledgment (NACK) message. This indicates that more than one MTC device transmitted on the allocated resources or that a decoding error occurred due to channel impairments.

As it is not possible to differentiate between the two above-mentioned cases, we propose to perform retransmissions considering the two possibilities. The proposed HARQ probability for *j*-th MSG_3 message transmission of a RA attempt initiated in *i*-th RAO is denoted by $p_{HARQ_{ij}}$ and defined as:

$$p_{HARQ_{ij}} = \underbrace{P_{tx} \times P_{ij}(C|N)}_{\text{collission}} + \underbrace{P_{ij}(S|N)}_{\text{success}}, j > 1 \tag{1}$$

where

$$P_{ij}(S|N) = \frac{P_{ij}(S \cap N)}{P_{ij}(N)} = \frac{{p_e}^{(j-1)} \times P_{s,i}}{{p_e}^{(j-1)} \times P_{s,i} + P_{c,i}} \tag{2}$$

and

$$P_{ij}(C|N) = \frac{P_{ij}(C \cap N)}{P_{ij}(N)} = \frac{P_{c_i}}{{p_e}^{(j-1)} \times P_{s,i} + P_{c,i}} \tag{3}$$

are, respectively, the conditional probabilities that the retransmission comes from an uncollided preamble (S) or a collided preamble (C) transmission given that the MTC device received a NACK message[2] (N) in (*j*-*1*)-th transmission of a RA attempt initiated in *i*-th RAO. The probability of a noncollided preamble $P_{s,i}$ can be calculated as

$$P_{s,i} = \left(1 - \frac{1}{d_i}\right)^{N_i - d_i}, \tag{4}$$

where N_i and d_i are, respectively, the number of users transmitting preambles and the number of successfully detected preamble in *i*-th RAO. The probability of a collided preamble is $P_{c,i} = 1 - P_{s,i}$, and the probability of error of a HARQ transmission p_e is a parameter of the LTE standard, usually defined as 0.1 [3]. Finally, P_{tx} is the retransmission probability given that a collided preamble generated the NACK message received.

If the NACK message was caused by a channel error (with probability $P_{ij}(S|N)$), the optimal retransmission probability

---

[2] The NACK message is received on the PHICH for non-adaptive retransmissions or indicated through a DCI 0 message on the Packet Downlink Control Channel (PDCCH) for adaptive retransmissions.

equals 1, but if the NACK message was caused by simultaneous transmission of various MTC devices on the same PUSCH resources (a collision, with probability $P_{ij}(C|N)$), the retransmission probability is P_{tx}. Note that the optimal value for P_{tx} is the inverse of the number of collided MTC devices with the same preamble. However, the MTC device cannot know this value based only on the HARQ feedback. Moreover, this value is even unknown for the eNB. Thus, in the next subsection, we introduce two policies to determine the P_{tx} value exploring the information available on the device side.

In the proposed HARQ protocol, if the first transmission of a MSG$_3$ message is unsuccessful, the eNB will reserve PUSCH resources for the remaining HARQ transmissions, regardless of the successful detection of a transmission for a given HARQ process. In this way, resources for performing the probabilistic retransmissions are guaranteed and more than one MTC may be detected with the same RAR message. This generates significant radio and energy resource savings since the number of RA attempts is reduced and the network access latency decreased. Note that this maintains almost the same resource utilization as the legacy HARQ process due to the collision of the MSG$_3$ message re/transmission from devices that chose the same preamble. Such overhead will be analyzed numerically in the Section IV.

Upon reception of an ACK message for its HARQ process, the MTC device determines if this ACK message is addressed to it by verifying the status of its last transmission. If it did not perform a transmission in the last scheduled HARQ opportunity (*i.e.,* if the generated random number was greater than p_{HARQ}), the ACK message is intended to another MTC devices that selected the same preamble. Thus, the MTC device continues with the HARQ process as described above. Otherwise, the MTC device finishes the HARQ process unilaterally.

B. Retransmission Policies

We propose two policies for defining the P_{tx} value in (1). The first policy (policy 1) defines that

$$P_{tx_1} = 1/(C_i + 1)\,, \tag{5}$$

where $C_i = N_i/d_i$, which is the expected number of collided MTC devices per detected preamble in the *i*-th RAO in which the MTC device performed the preamble transmission. This policy uses the fact that the MTC devices trying retransmission is a collided MTC device. Thus, $C_i + 1$ is used to obtain the retransmission probability in (5). Every time an ACK is received on the PHICH for the HARQ process of the MTC device in consideration, it indicates that an MSG$_3$ message from other MTC devices in the same collision set was successfully decoded. Thus, if the current C_i value is greater than one, we update the C_i value by subtracting the already decoded MTC device, $C_i = C_i - 1$.

The second policy (policy 2) is based on the expected number of collided MTC devices per collided preamble. Let $k_i = N_i - d_i$ be the number of MTC devices that are in collision. Let R_i denote the number of MTC devices that choose the same preamble given that all the decoded preamble are already selected by exactly one MTC device in the *i*-th RAO. P_{tx_2} is defined as:

$$P_{tx_2} = \sum\nolimits_{c=1}^{N_i - d_i} \frac{1}{c+1} \times P(R = c)\,, \tag{6}$$

where $P(R = r)$ is the probability that r among k MTC devices selected the same preamble and it follows a binomial distribution:

$$P(R = r) = \binom{k}{r} \left(\frac{1}{d_i}\right)^{r} \left(1 - \frac{1}{d_i}\right)^{k-r}\,. \tag{7}$$

To calculate $p_{HARQ_{ij}}$ in (1), the only variable unknown to the MTC device is N_i. Next, we provide a lightweight and standard-compatible method to estimate N_i in MTC devices.

C. Estimation of N_i

Existing methods in the literature for making this estimation are designed to operate at the eNB side [12], [13], [14], [15], [16], [17]. Moreover, most of them involve recursive probability calculation and optimization problems. As the estimation of N_i, denoted by $\hat{N}_i$, needs to be performed by the MTC devices, we propose a simple but still efficient way based on the method proposed by Oh *et. al* [14], by adapting their proposal to the information available at the UE side.

Let M_i be the number of available preambles for the contention-based RA procedure in the *i*-th RAO. Let I_i and d_i be the number of unused preambles and detected preambles as observed by the eNB in the *i*-th RAO, respectively. Even though several proposals assume that an eNB is able to differentiate between a collided preamble and a non-collided preamble, this assumption does not hold in a real eNB implementation [2]. Therefore, $M_i = I_i + d_i$.

The probability of idle preambles in the *i*-th RAO can be computed based on the observed M_i and I_i values as follows [14]:

$$\tilde{p}_{idle,i} = I_i/M_i. \tag{8}$$

This probability can also be calculated as [14]:

$$p_{idle,i} = \left(1 - \frac{1}{M_i}\right)^{N_i} \approx e^{-N_i/M_i}, \tag{9}$$

and $p_{idle,i}$ can be approximated as e^{-N_i/M_i}.

By setting $\tilde{p}_{idle,i}$ equals $p_{idle,i}$, the expected number of concurrent MTC devices in the *i*-th RAO can be estimated as

$$\hat{N}_i = M_i \times \ln\left(M_i/I_i\right). \tag{10}$$

1) Available information in the MTC devices: The estimation method used here are designed to operate in the eNB, which has direct access to I_i and d_i variables at every TTI. However, the UE devices do not have the information about these variables. To deal with this problem, we propose to use the counting of the RAR messages sent by the eNB after preambles detection. As all UE devices that performed preamble transmission in a given RAO monitor the PDCCH for a possible match with its own RA-RNTI within the RAR window, we take advantage of this fact to perform a counting of the number of RAR messages presented in the PDSCH matching its RA-RNTI during the entire RAR window size.

To guarantee that all RAR messages sent are included in the counting, the MTC devices keep performing the RAR counting until the end of the RAR window size, even if it receives a RAR message before the ending of the RAR window. Thus, the resulting counting is equivalent to value of d_i.

2) Estimation under inaccurate information: The eNB generally tries to response all detected preambles by prioritizing RAR messages at the PDCCH scheduler [18] and by allocating more PUSCH resources for sending RAR messages [19]. However, in some cases the eNB may not be able to response to all preambles detected in a RAO because the resources available in a RAR window may not be sufficient to send RAR messages of all detected preambles. For instance, the 3GPP proposed an MTC performance evaluation methodology in [3], which considers that for a 5 MHz cell with RAR window size of 5 ms, the maximum number of RAR messages per RAO equals 15. Thus, when the number of decoded preamble is higher than this value, the MTC devices are not able to make an accurate estimation of N_i from RAR message counting technique proposed above. This inaccurate information is more likely to happen in low bandwidth cells with high Random Access Channel (RACH) loads.

For this reason, we propose a complementary method to obtain the estimated value, based on the access probability p_{ACB} of standardized ACB RA scheme, which was proposed for dealing with the RAN overload problem and signaling storms [3]. Generally, the ACB scheme is activated in a cell independent of the traffic types it supports. The eNB periodically broadcasts the p_{ACB_i} value in the SIB2 [3]. Based on this probability, the ACB scheme limits the number of users trying RA under high signaling and RACH loads. The optimal p^*_{ACB} value which maximizes the RACH throughput was derived in [20] as:

$$p^*_{ACB} = \min\left(1, M/N\right), \tag{11}$$

As both M_i and p_{ACB} are known from the SIB2, the MTC device is able to calculate the estimated value of N from the p_{ACB} received periodically by using (11) when $p_{ACB} < 1$. When $p_{ACB} = 1$, however, N can be estimated from (10). The calculation of $\hat{N}$ should be obtained from (11) when $p_{ACB} < 1$ because the eNB has much hardware resources and more accurate information than the MTC devices. Note that in existing LTE networks, the ACB scheme coexists with other specific solutions for MTC scenarios such as the EAB scheme.

IV. PERFORMANCE EVALUATION

In this section, we assess the performance of the proposed probabilistic retransmission approach by using the LTE Simulator (LTE-Sim) [21]. LTE-Sim is a widely-used event-driven LTE network simulator developed in C++. We used an extended version of the enhanced LTE-SIM module in [22], which implements the RA procedure, PDCCH scheduling, and different RA schemes. Moreover, the RA-Priorized (RAP) PDCCH algorithm proposed in [18] were used. As the RAP algorithm prioritizes MSG$_2$ and MSG$_4$ messages, actual user data transmission is not necessary to properly assess the RA performance of the proposed probabilistic HARQ retransmission approach for the RA procedure [22]. Furthermore, the UE Lauridsen *et al.*'s energy consumption model [23] was implemented in the simulator.

TABLE I. SIMULATION PARAMETERS

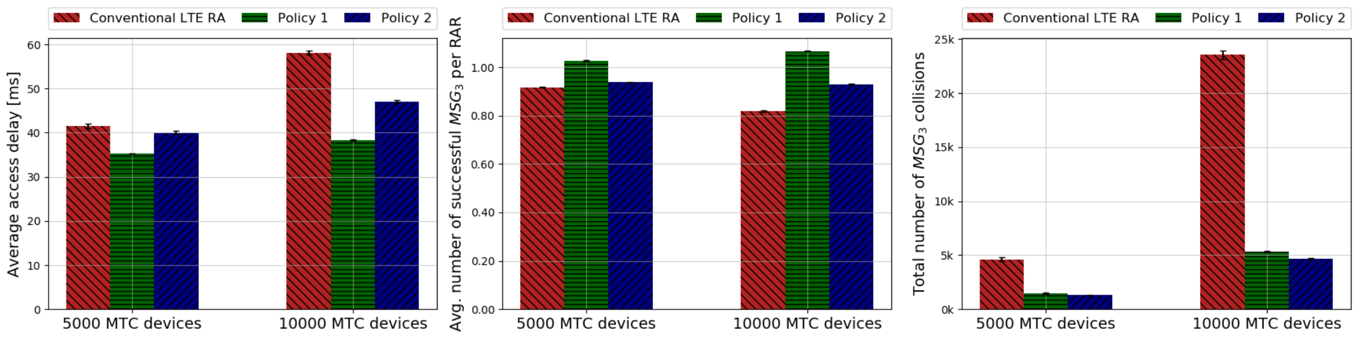
Parameter	Value
System type	Single cell
System bandwidth	5 MHz
Cell radius	0.5 km
PRACH configuration index	6
RA preamble format	0
Contention-based preambles	52
RAR messages per TTI	6
CCEs allocated for the PDCCH	16
CCEs per UE-specific DCI message	4
CCEs per common DCI message	8
Backoff	20 ms
HARQ retransmission probability	10%
maxHARQ-Msg3Tx	5
preambleTransMax	10
RAR Window Size	5 ms
Contention Resolution Timer	48 ms
P_{UE}	23 dBm
Preamble received target power	-118 dBm

We compare the performance of the proposed probabilistic retransmission approach with p_{rx1} and p_{rx2} retransmission policies to that of the LTE RA procedure with the conventional HARQ protocol.

A. Simulation Model

The simulation scenarios comprise a single cell with a 0.5 km radius. An eNB with 5 MHz cell bandwidth in the Frequency Division Duplexing mode is located at the center of the cell with several stationary MTC devices uniformly distributed around it. Scenarios with $5,000$ (low load) and $10,000$ (medium load) MTC devices were executed as proposed by the 3GPP in [3]. Activation of MTC devices follows the $Beta(3, 4)$ distribution within a 10 s interval to simulate an extreme scenario with MTC transmissions highly synchronized as also proposed by the 3GPP in [3]. Once an MTC device is activated, it triggers the RA procedure in order to obtain resources to transmit its uplink data since allocation of PUCCH resources to send scheduling request for several MTC devices is not feasible [24]. A set of 52 preamble sequences are shared among the MTC devices for executing the contention-based RA procedure.

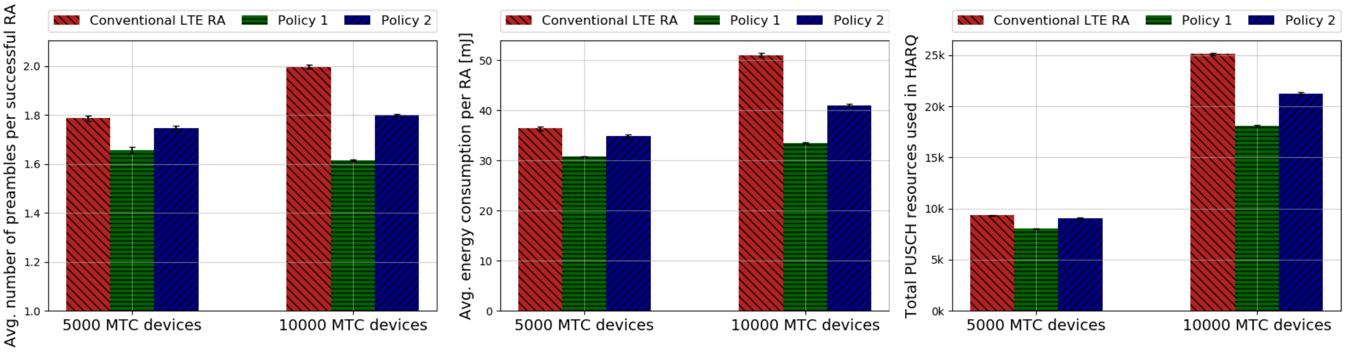
We also assumed that UEs have already received the RA configuration parameters in the beginning of the simulation. A total of 16 Control Channel Elements (CCEs) for 5 MHz bandwidth is available for the PDCCH. Moreover, since all devices are uniformly distributed in the cell, an aggregation level of 4 CCEs per Downlink Control Information (DCI) message is assumed [3]. We also assume six RAR messages are available per TTI. We assume that the HARQ transmission error probability p_e is 0.1 [3] and that any MSG$_3$ message transmission from just one user can be decoded with probability $1 - p_e$. Moreover, we assume that the transmission of MSG$_3$ messages received from two or more users using the same radio resources cannot be decoded, *i.e.*, the capture effect cannot be used to decode one of the users signals. Table I summarizes the main configuration parameters used in the simulations.



(a) Average access delay

(b) Average number of successful MSG_3 transmissions per RAR message

(c) Number of collided MSG_3 transmissions



(d) Average number of transmitted preambles

(e) Energy consumption per successful RA procedure

(f) HARQ protocol overhead on PUSCH

Figure 1. Probabilistic retransmission approach vs conventional LTE RA scheme

B. Simulation Results and Discussion

The figures in this section show mean values derived by using the independent replication method with 10 replication. Confidence intervals of 95 % confidence level are also shown. A simulation runs until all started RA procedures finish either successfully or unsuccessfully. All metrics are shown for the conventional RA scheme and the proposed approach with both retransmission policies.

The proposed approach achieves lower average access delays than does the conventional RA scheme in all settings (Fig. 1(a)). Under low loads, the proposed approach gives access delays between 2 % (policy 2) and 10 % (policy 1) lower than the conventional RA scheme. Under medium loads, the access delay improvement provided by our proposal is even more significant, yielding access delays between 20 % (policy 2) and 30 % (policy 1) lower than those of the conventional scheme. This significant decrease in the access delay occurs because the proposed probabilistic HARQ retransmission strategy considerably reduces the number of collided MSG_3 message transmissions (Fig. 1(c)) and increases the number of successful MSG_3 transmissions per issued RAR message (Fig. 1(b)).

One of the most important results obtained with our proposal is the reduction in the number of collided MSG_3 transmissions (Fig. 1(c)). Under all loads, the proposed approach reduces in more than 75 % the collided MSG_3 transmissions when compared to the conventional RA scheme. This is a direct result of applying the probabilistic retransmission strategy to the MSG_3 transmissions in the RA procedure. Note that, even though the retransmission policy 2 yields a slightly lower number of collided MSG_3 transmissions than does the policy 1, the policy 2 produces higher access delays and lower number of successful MSG_3 transmissions per issued RAR message. This is explained by the fact that the policy 1 updates the expected number of collided MTC devices per detected preamble with every received ACK message, whereas the policy 2 maintains the same value during all retransmissions. This continuous updating adjusts the retransmission probability based on the HARQ feedback for a given process, making the policy 1 less aggressive than policy 2, which is calculated just once each RA attempt. This also evinces that the updating of the retransmission policy impacts positively on the performance of the probabilistic retransmission approach. Moreover, the decreasing in the collided MSG_3 transmissions induced by the proposed approach yields lower average number of transmitted preambles per successful RA procedure when compared to the traditional RA scheme (Fig. 1(d).

Under low loads, the policy 1 gives number of successful MSG_3 transmissions per RAR message 10 % higher than that given by the conventional RA scheme and the policy 2, whereas under medium loads, the proposed policies outperforms the conventional RA scheme, providing number of successful MSG_3 transmissions per RAR message 10 % (policy 2) and 25 % (policy 1) higher than the conventional RA scheme. This is also a result of the probabilistic strategy of our proposal. The adjustment of the retransmission probability performed by the policy 1 also explains this difference between the two policies.

Moreover, the decrease in the average access delay (Fig. 1(a)), collided MSG_3 transmissions (Fig. 1(c)), and number of transmitted preambles (Fig. 1(d)) produced by the proposed approach yields lower device energy consumption for the whole RA procedure when compared to the conventional scheme (Fig. 1(e)). Under low loads, our proposal gives from 5% (policy 2) to 10% (policy 1) energy saving when compared to energy consumed by the conventional RA scheme. Under medium loads, the energy savings increases between 20% (policy 2) and 25% (policy 1). This results show the incapacity of the conventional RA scheme to deal with the *MSG_3 collision problem*, which significantly impacts on the energy consumption of the MTC devices. Moreover, as this is an important key performance indicator for the MTC devices given their limited battery capacity and prohibitive costs incurred with battery replacement in most cases, our approach is relevant for future IoT scenarios over cellular networks. Besides, the above-discussed gains yielded by our proposal are obtained with lower PUSCH resource overhead than that of the conventional RA scheme (Fig. 1(f)). Under low loads, the PUSCH resources utilized by our proposal with retransmission policy 1 is about 5 % lower than that utilized by the conventional scheme, whereas the policy 2 gives almost the same performance as the conventional scheme. However, under medium loads, the proposed scheme achieves between 10% (policy 2) to 25% (policy 1) lower PUSCH resources utilization when compared to those utilized by the conventional scheme. Again, the retransmission policy 1 yields less HARQ overhead because of its update every time an ACK message is received from the eNB for the specific HARQ process, which reduces the number of RA attempts.

V. CONCLUSION

In this paper, we proposed a probabilistic MSG_3 retransmission approach to ameliorate the *MSG_3 collision problem* in the RA procedure of cellular IoT network technologies. Specifically, we propose the application of a probability value for every MSG_3 retransmission to increase the chance of MSG_3 successful decoding under preamble collisions. Two different retransmission policies were proposed to decrease the MSG_3 collisions; one based on the expected number of collided MTC devices per detected preamble and other based on the expected number of collided MTC devices per collided preamble. To apply the proposed approach at the MTC device side, a novel method to estimate the number of MTC devices trying random access in a given RAO is proposed based on RAR message counting and the probability broadcasted by the eNB in the standardized ACB RA scheme. Simulation results show that the proposed approach effectively reduces the number of MSG_3 collisions. Moreover, our proposal saves energy, reduces the access delay, and yet reduces the PUSCH channel utilization when compared to the conventional RA scheme.

REFERENCES

[1] M.-Y. Cheng *et al.*, “Overload control for machine-type-communications in LTE-advanced system,” *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 38–45, June 2012.

[2] J. S. Kim *et al.*, “Enhancement of LTE RACH through extended random access process,” *Electronics Letters*, vol. 50, no. 19, pp. 1399–1400, September 2014.

[3] 3GPP, “Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-type Communications,” 3rd Generation Partnership Project (3GPP), TR 37.868, September 2011.

[4] M. S. Ali, E. Hossain, and D. I. Kim, “LTE/LTE-A random access for massive machine-type communications in smart cities,” *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, January 2017.

[5] M. Vilgelm, S. R. Linares, and W. Kellerer, “Enhancing cellular M2M random access with binary countdown contention resolution,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–6.

[6] M. Vilgelm and W. Kellerer, “Binary contention resolution for M2M random access prioritization in LTE-A and 5G,” in *2017 IFIP Networking Conference (IFIP Networking) and Workshops*, June 2017, pp. 1–2.

[7] D. Magrin *et al.*, “Enabling LTE RACH collision multiplicity detection via machine learning,” *CoRR*, vol. abs/1805.11482, 2018.

[8] Z. Alavikia and A. Ghasemi, “A multiple power level random access method for M2M communications in LTE-A network,” *Trans. on Emerging Telecommun. Technol.*, vol. 28, no. 6, p. e3137.

[9] J. Kim and J. Lee, “Exploiting the capture effect to enhance RACH performance in cellular-based M2M communications,” *Sensors*, vol. 17, no. 10, 2017.

[10] K. S. Ko *et al.*, “A novel random access for fixed-location machine-to-machine communications in ofdma based systems,” *IEEE Communications Letters*, vol. 16, no. 9, pp. 1428–1431, September 2012.

[11] Y. Liang *et al.*, “Non-orthogonal random access for 5G networks,” *IEEE Trans. on Wireless Commun.*, vol. 16, no. 7, pp. 4817–4831, July 2017.

[12] Y. Pang *et al.*, “Network access for M2M/H2H hybrid systems: a game theoretic approach,” *IEEE Communications Letters*, vol. 18, no. 5, pp. 845–848, 2014.

[13] G. Lin, S. Chang, and H. Wei, “Estimation and adaptation for bursty LTE random access,” *IEEE Trans. on Veh. Technol.*, vol. 65, no. 4, pp. 2560–2577, April 2016.

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

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

[16] G. C. Madueño *et al.*, “Massive M2M access with reliability guarantees in LTE systems,” in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 2997–3002.

[17] S. Vural *et al.*, “Dynamic preamble subset allocation for RAN slicing in 5G networks,” *IEEE Access*, vol. 6, pp. 13 015–13 032, 2018.

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

[19] W. Kim and E. Lee, “LTE network enhancement for vehicular safety communication,” *Mobile Information Systems*, vol. 2017, pp. 8 923 782:1–8 923 782:18, 2017.

[20] S. Duan *et al.*, “D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks,” *IEEE Trans. on Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec 2016.

[21] G. Piro *et al.*, “Simulating LTE cellular systems: An open-source framework,” *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498–513, 2011.

[22] T. P. C. de Andrade *et al.*, “The random access procedure in long term evolution networks for the Internet of Things,” *IEEE Commun. Mag.*, vol. 55, no. 3, Mar. 2017.

[23] M. Lauridsen *et al.*, “An empirical LTE smartphone power model with a view to energy efficiency evolution,” *Intel Technology Journal*, vol. 18, no. 1, pp. 172–193, 2014.

[24] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca, “Impact of M2M traffic on human-type communication users on the LTE uplink channel,” in *7th IEEE Latin-American Conf. Commun. (LATINCOM)*, November 2015, pp. 1–6.