Analysis of PDCCH performance for M2M traffic in LTE

Prajwal Osti, Pasi Lassila, Samuli Aalto, Anna Larmo, Tuomas Tirronen

***Abstract*—As LTE is starting to get widely deployed, the volume of M2M traffic is increasing very rapidly. From the M2M traffic point of view, one of the issues to be addressed is the overload of the random access channel. The limitation in the PDCCH resources may severely constrain the number of devices that an LTE eNB can serve. We develop a Markov model that describes the evolution of the Message 4 queue in the eNB formed by several users performing the random access procedure simultaneously and then study its stability and performance. Our model explicitly takes into account the 4 initial steps in the random access procedure. By utilizing the model, we are able to determine the stability limit of the system, which defines the maximum throughput, as well as the probability of failure of the random access procedure due to different causes. We observe that the sharing of the PDCCH resources between Message 2 and Message 4 with different priorities makes the performance of the whole random access procedure deteriorate very rapidly near the stability limit. However, we can extend the maximum throughput and improve the overall performance by increasing the PDCCH resource size. Furthermore, we estimate the upper limit of the number of devices that can be served by an LTE eNB and determine the minimum PDCCH resource size needed to satisfy a given traffic demand.**

***Index Terms*—LTE, M2M, Markov processes, MTC, PDCCH, Stability**

I. Introduction

Machine-to-machine (M2M) communication or machine type communication (MTC) is the technology that enables several devices to communicate with each other without the need of constant human intervention. Several billions of such devices that use MTC are predicted to exist over the next few years and majority of them are expected to be wireless sensors. This leads to the possibility of developing a wide range of applications over M2M that can potentially generate a huge amount of revenue [1]. Even the existing networks, which are primarily designed for more traditional human-to-human (H2H) traffic, are handling some M2M traffic [2]. However, as the volume of M2M traffic grows more M2M type communication specific provisions should be included in the design of the standards like LTE, which is likely to be the most widely accepted standard for the 4G cellular network. Indeed 3GPP has conducted different studies [3], [4] that attempt to address the issues related to M2M communication in the present systems as well as in the future releases of LTE.

The M2M traffic is different from the traditional voice and data traffic for which most of the existing networks including LTE are optimized. The sensors are typically (but not always) static and need not be optimized for mobile use. A huge number of machines may exist in a cell, which may access the network periodically or in random bursts. They also have a limited power budget which should be used as efficiently as possible. A machine may need a very small portion of the network's resources at a time but as the number of machines grows, their collective resource demand can overwhelm the network very easily, e.g., the radio access network may not be able to handle the momentary surge in random access requests if thousands of devices make random access attempts at the same time.

Indeed, one of the most discussed issues is the problem of random access overload [5], [6] at the edge of the network. Slotted Aloha is the basis of the whole random access procedure which is an inherently unstable protocol [7] and has to be stabilized in some way to make it work [8], [9]. In the H2H case, congestion in the random access channel is rarely a problem because the number of users requesting random access at the same time is almost never so large. However, the deluge of random access requests from the huge number of MTC devices may overwhelm the network's signaling resources. In contention based random access such as Slotted Aloha, when a random access attempt fails, the device may opt for a retransmission possibly after a certain backoff period causing further increase in the traffic and deterioration of performance. Several such Slotted Aloha channels are operating in parallel for random access in LTE and even such multi-channel random access systems are not immune to the inherent instability issues of Slotted Aloha [10]. Recent works (e.g., [11], [12]) have analyzed the performance of backoff algorithms in LTE for stabilizing and optimizing the random access procedure. In fact, the random access will fail in LTE if any one of the four steps of the procedure is unsuccessful, leading to a waste of resources, retrial and ultimate increase in the traffic which may have been prohibitively high to begin with. Moreover, different signaling and data channels are involved in the LTE random access procedure and congestion in any one of them will ultimately affect the performance of the whole procedure.

Manuscript received February 14, 2013; revised .... This work was supported by TEKES as part of the Internet of Things program of DIGILE (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT and digital business).

P. Osti, P. Lassila, S. Aalto are with the Department of Communications and Networking, Aalto University School of Electrical Engineering, Finland. (Email: firstname.lastname@aalto.fi)

A. Larmo, T. Tirronen are with NomadicLab, Ericsson Research, Finland. (Email: firstname.lastname@ericsson.com)

Since the problem is well known there have been various attempts to address this issue. The authors in [5] present a nice overview of the problem when a massive number of devices are to be given access. In the context of assigning separate random access resources to various traffic types, [13] considers the approach of dynamically sharing the random access resource (preambles) between H2H and M2M traffic. In [14] a highly tailored approach to overcome RAN overload is presented by dividing the M2M traffic into several priority classes and providing them different number of random access opportunities. A more dynamic approach for RAN overload control is provided by [15]. It is argued that with the increasing traffic arrival rate the number of subframes that should be allocated for the RACH procedure should also be increased to prevent random access failure. However, we will show here that doing so may not necessarily solve the problem as other bottlenecks come into effect that even reduce the random access throughput.

Fundamentally, the random access procedure is based on parallel Slotted Aloha channels. The Slotted Aloha channel itself has a stability limit that determines the maximum amount of traffic the system can sustain. Almost all the existing works focus in just this first step of the random access procedure (Slotted Aloha), either by employing good backoff algorithms (see e.g., [11], [12]) or by devising efficient ways to share the random access opportunities among different types of traffic in the cell (see [13], [14]). As it turns out in our analysis, the steps following the Slotted Aloha step pose additional limitations on the stability and the performance of PDCCH which is shared between Message 2's and Message 4's with Message 2's receiving the priority, as stated in [16]. These messages are exchanged between the device and eNB in the steps following the initial Slotted Aloha step.

In this paper we analyze the performance of the whole initial random access procedure assuming that the PDCCH has limited resources in the form of CCEs (control channel elements). The initial random access is characterized by 4 steps, as will be discussed in detail later on. A study by 3GPP [17] presents simulation results on the overall success probabilities of the access attempts after all 4 steps. However, our objective is to derive a model amenable to mathematical analysis to obtain fundamental insights. More specifically, we develop a Markov model that describes the evolution of the Message 4 queue and then study its stability and performance (measured by the probability of random access failure). This paper is the first to analyze jointly the impact of all the 4 steps of the initial random access, as far as we know.

In summary, the main contributions of our paper are the following. We provide a tractable model for the analyzing jointly the impact of the steps 1–4 in the LTE random access procedure. From the model we are able to explicitly determine the maximum throughput of the system which gives the upper limit of the arrival rate of the random access requests, i.e., the stability limit. We additionally provide a numerical method to estimate the failure probability of the random access process and its different components at various stages of the random access process. In our extensive numerical examples, we illustrate how the different parameters affect the performance, including the possibility to optimize the maximum throughput for certain parameters. We observe that the failure probability is almost zero when the arrival rate is below the stability limit. However, due to the nature of the priority system, this probability increases very quickly near the stability limit. We additionally provide two examples—determining the maximum number of devices in a cell and dimensioning the PDCCH resource—that highlight the application of our results.

The paper is organized as follows: In the next section we explain the background and role of PDCCH in LTE followed by the steps involved in the random access procedure. This is followed by Section III where we describe our traffic model and identify the points where the random access can fail in our model. Then in Section IV, we present our stochastic model, which describes the evolution of the Message 4 buffer together with an auxiliary variable as a two-dimensional Markov chain. The Markov model is utilized in Section V, where we consider the stability of the Message 4 buffer and derive its maximum throughput. In Section VI, we present the methodology of determining the performance of the system. The methodology is then applied in Section VII, where we give numerical results. Finally, conclusions of the study are presented in Section VIII.

II. LTE BACKGROUND

A. Downlink control information

In LTE, downlink control information is sent over the Physical Downlink Control Channel (PDCCH). The control information includes downlink scheduling assignments, which are used to carry the information needed to receive data on the Physical Downlink Shared Channel (PDSCH), and uplink scheduling grants, which are used to indicate the shared uplink resources (Physical Uplink Shared Channel, PUSCH) the terminal uses to send data to the base station (eNodeB).

The smallest physical resource in LTE time-frequency structure is called a resource element, which consists of one subcarrier during one OFDM symbol. The transmissions are divided into frames of length of $10\,\text{ms}$ which are further divided into subframes of $1\,\text{ms}$. In time domain, a subframe typically consists of 14 OFDM symbols. In every subframe, 1–3 OFDM symbols are reserved for the control region which carries the PDCCHs. Three OFDM symbols is a typical size for the control region and this size is assumed in this work. The total amount of available physical resources during one subframe depends on the number of subcarriers, i.e., the total bandwidth allocated for the LTE carrier. For example, 5 MHz cell bandwidth would correspond to 300 subcarriers.

The total PDCCH resource is measured in control channel elements (CCEs), where each CCE is a set of 36 resource elements. One downlink control message, such as downlink assignment or uplink grant, is carried over PDCCH which uses either one, two, four or eight CCEs. The number required depends on the size of the payload of the message and the coding rate. For the numerical examples, we typically make the assumption that the total resources available for the downlink control information in a subframe is $N = 16$ CCEs, the same number used in 3GPP RAN overload study [17].

Thus, the resources used to schedule data in downlink and used to hand out uplink grants are shared. The capacity

for downlink assignments and uplink grants sent during one subframe depends on the sizes of the PDCCHs used to carry these messages (denoted by N^{Msg2} and N^{Msg4}) and the total number of CCEs (denoted by N). In reality, the number of CCEs one PDCCH (and one downlink control message) uses depends on the channel conditions and is selected by the eNodeB. Using our model, we will later study the effect of different (static) CCE allocation sizes (N^{Msg2} and N^{Msg4}) for different messages as well as the consequence of varying the PDCCH resource size (N).

B. Random access procedure

Below we briefly describe the Contention based Random Access Procedure [18] utilized, e.g., by the M2M traffic.

Step 1: M2M device (UE) initiates the random access procedure by randomly choosing one of the available RACH preambles, and sending the preamble in Message 1 over the Physical Random Access Channel (PRACH). A collision happens if two or more UEs choose the same preamble in the same subframe. However, this collision is realized only in Step 3, i.e., even if two or more UEs use the same preamble for Message 1 and a collision occurs, the base station does not detect this event at this stage.[1] The transmission of a random access preamble is restricted to certain subframes. Let b denote their periodicity, i.e., random access is possible in every bth subframe. In addition, let K denote the total number of available preambles.

Step 2: eNodeB replies with Message 2, a.k.a. Random Access Response (RAR), which includes an uplink grant for Step 3. Message 2 is sent over the PDSCH. For this, we need to schedule the user, i.e., send a downlink assignment control message over PDCCH. There may be at most one RAR message in each subframe, but each may have multiple uplink grants (each corresponding to a separate preamble). Let c denote the maximum number of uplink grants per RAR per subframe. Note that in our model, an uplink grant is given for *every used* preamble, whenever not limited by c, as the base station does not detect a collision at this stage and thus makes no distinction between a collided and an uncollided preamble.

Step 3: Next the UE sends Message 3 over PUSCH. The collisions in Step 1 will be realized in Step 3: The two or more UEs that chose the same preamble in Step 1 will all try to utilize the same uplink grant in Step 3 to send their Message 3's. As a result, the Message 3's interfere with each other rendering the signal received at the eNB undecodable and none of the UEs involved will be sent the subsequent Message 4.[2]

Step 4: After receiving Message 3's related to uncolliding preambles generated in Step 1, eNodeB replies with Message 4's using again the PDSCH, which needs to be scheduled on PDCCH. Let N be the size of PDCCH resource (in CCEs), N^{Msg2} and N^{Msg4} be the number of CCEs used to send a Message 2 and a Message 4, respectively. Then a maximum of $M = \frac{N}{N^{\text{Msg4}}}$ Message 4's can be sent in one subframe if Message 2 is not present in that subframe. On the other hand, when a Message 2 is also sent in a subframe then at most $m = \frac{N - N^{\text{Msg2}}}{N^{\text{Msg4}}}$ Message 4's can be sent in that subframe. Although the parameters N, N^{Msg2} and N^{Msg4} are more relevant from the system deployment point of view, our model is greatly simplified when we use the derived parameters M and m.

This messaging scheme is demonstrated in Figure 1. In addition, the (model) parameters introduced in this section are summarized in Table I including a set of typical values for the FDD LTE [17, Table 6.2.2.1.1], where $N = 16$, $N^{\text{Msg2}} = 4$, $N^{\text{Msg4}} = 4$ which leads to $M = \frac{16}{4} = 4$ and $m = \frac{16-4}{4} = 3$.

TABLE I: Model parameters with typical values in [17]

Symbol	Parameter	Typical value [17]
K	Number of preambles	54
b	RACH periodicity	5
c	Max number of UL grants per subframe	3
N	PDCCH resource size in CCEs	16
N^{Msg2}	Number of CCEs used for a Message 2	4
N^{Msg4}	Number of CCEs used for a Message 4	4
M	Max number of Message 4's per subframe (without Message 2)	4
m	Max number of Message 4's per subframe (with Message 2)	3

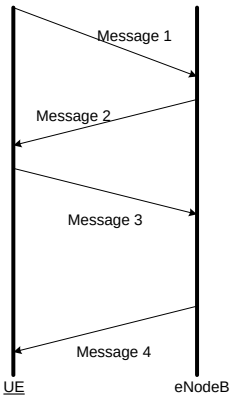


Fig. 1: Message sequence in LTE random access.

III. Performance of the random access procedure

We explore the performance of the random access procedure described in the previous section by modeling and analyzing its success (or failure) probability. To simplify the analysis, we assume that no other traffic except this random access traffic is present in the network. We consider a steady-state traffic scenario where new (i.e., fresh) random access requests arrive according to a Poisson process with a constant rate λ (new requests per subframe). This can be interpreted as a traffic model where the requests are generated independently by a large population of asynchronous M2M devices. Similar scenarios with thousands of devices accessing the system are considered realistic in 3GPP studies, see [4], [17]. If a random access request fails, it needs to be retransmitted again later

[1] Note that this is a more conservative approach (and even realistic from the actual system point of view) than detecting the collision in Step 1 itself.

[2] This is again a conservative assumption as a UE with much stronger signal than others may be selected even in the event of collision due to the *capture effect*.

on. Whenever the station makes an access request (fresh or retransmission), the preamble is selected randomly among the K available ones [19].

Our purpose is to determine the maximum throughput θ^* (successful requests per time unit) of the system. In addition, we study the behavior of the failure probability (and its components described below) as a function of the arrival rate λ of fresh requests.

A random access request may fail due to collision in Step 1 when two or more UEs choose the same preamble. It may also fail due to loss in Step 2 when the number of chosen preambles exceeds the maximum number of UL grants. Here we assume that the excess requests are not buffered but lost. This is reasonable as the UE is expecting a response typically in 5 ms [17] which is of the same order as the interval between transmission opportunities (recall our parameter b). Thus, there is no time for the base station to start buffering the requests from Step 1 to be sent in subsequent time slots. If the Message 2 timer expires, the UE will anyway perform a backoff and eventually makes a retry, see [19]. While Step 3 does not generate any new access failure causes for requests successful in steps 1 and 2, we still have to take into account the final step. Message 2's and Message 4's share the same resources in the PDCCH. But due to a more strict constraint on the return time of Message 2's (5 ms in [17]), we assume that they get the absolute priority over the resource, see also [16]. Only what remains of the resource is then allocated to Message 4's, which have a more lenient time constraint (48 ms in [17]). Thus, in order to avoid additional losses in the final step, there must be a buffer for Message 4's. A failure takes place in Step 4 if the Message 4 corresponding to the original request is delayed in the buffer beyond the threshold that triggers the retransmission timer. Thus, the failure probability consists of the following components:

$$\begin{aligned}\Pr\{\text{failure}\} = \Pr\{&\text{collision in Step 1}\}+\\ &\Pr\{\text{no collision in Step 1, loss in Step 2}\}+\\ &\Pr\{\text{no failure in Steps 1 and 2, delay in Step 4}\}.\end{aligned} \quad (1)$$

It should be noted here that our aim is to analyze the random access procedure itself and therefore we ignore the effect of physical layer impairments on different random access messages for the sake of simplicity.

TABLE II: Summary of the symbols.

Symbol	Parameter
λ	Arrival rate of fresh random access requests
a	Aggregate arrival rate (fresh and retransmitted) of the random access requests
K	Number of preambles
b	RACH periodicity
c	Maximum number of UL grants per subframe in Message 2
M	Maximum number of Message 4's per subframe (without Message 2's)
m	Maximum number of Message 4's per subframe (with Message 2's)
N	PDCCH resource size in CCEs
N^{Msg2}	Number of CCEs used for a Message 2
N^{Msg4}	Number of CCEs used for a Message 4
n	Index of the time slot of the Markov model
A_{nk}	Number of random access requests using preamble k
p_i	$\Pr\{A_{nk} = i\}$ (see (2))
$Y_n^{(1)}$	Number of *successful* Message 1's (see (3))
$\tilde{Y}_n^{(1)}$	Total number of preambles chosen (see (4))
$q_{ij}^{(1)}$	$\Pr\{Y_n^{(1)} = i, \tilde{Y}_n^{(1)} = j\}$ (see (5))
$Y_n^{(2)}$	Total number of *non-colliding* UL grants included in Message 2's
$\tilde{Y}_n^{(2)}$	Total number of UL grants included in Message 2's (see (11))
$q_{ij}^{(2)}$	$\Pr\{Y_n^{(2)} = i, \tilde{Y}_n^{(2)} = j\}$ (see (7))
$q_i^{(2)}$	$\Pr\{Y_n^{(2)} = i\}$ (see (8))
$\tilde{q}_j^{(2)}$	$\Pr\{\tilde{Y}_n^{(2)} = j\}$ (see (9) and (10))
X_n	Queue length of Message 4 buffer at the beginning of the time slot n (see (12))
$Y_n^{(3)}$	Number of *successful* (non-colliding) Message 3's
$Y_n^{(4)}$	Number of *transmitted* Message 4's (see (13))
$\theta^{(1)}(a)$	Throughput of *successful* Message 1's per subframe (see (14))
$\theta^{(2)}(a)$	Throughput of *successful* UL grants per subframe (see (15))
$\tilde{\theta}^{(2)}(a)$	Throughput of *all* UL grants per subframe (see (19))
$\sigma^{(4)}(a)$	Average left-over capacity of Message 4's per subframe (see (20))
θ^*	Maximum throughput of Message 4's (see (23))
a_2^*	Aggregate arrival rate for which the throughput of Message 2's is maximum (see (18))
a_4^*	Aggregate arrival rate for which the throughput of Message 4's is maximum (see (22))

IV. Delay model for Message 4

In this section we develop the Markov chain model used to describe the evolution of Message 4 buffer. The various parameters and the variables of the model are summarized in Table II for quick reference and described more elaborately in the text.

We consider a discrete time model where time slots are indexed by n. The length of one time slot in our model corresponds to the periodicity of RACH opportunities denoted by b, with a typical value of $b = 5$ subframes (cf. Table I), which corresponds to 5 ms in absolute time units. The model does not take into account processing delays but assumes that a message received in a time slot generates a response immediately in the following time slot. Thus, for example, if eNB receives a Message 1 in time slot n, it will reply with a Message 2 in the following time slot $n + 1$.

Consider first the random access channel used in Step 1. For modeling its dynamics, we apply the well-known Slotted Aloha model [7], which is based on the approximative assumption that *all* random access requests together (not only the fresh ones but also the retransmissions) constitute a Poisson process, the rate of which is denoted by a (attempts per subframe). Thus, while the effect of retransmissions on the total traffic is taken into account, the actual retransmission mechanism itself is not explicitly modeled. The Poisson approximation is justified when the fresh requests arrive according to a Poisson process, they do not fail too frequently, and in the event of failure, the retransmissions are sufficiently randomized, i.e., the intervals between successive retransmissions are sufficiently long relative to the time slot duration. As already mentioned,

our scenario of a large population of asynchronous M2M devices allows us to model the arrival process of fresh requests as a Poisson process. Furthermore, as we will see, our 4-step model allows lower traffic rates than the corresponding Slotted Aloha system, and the failure probability remains very small unless the system is operated in close proximity of its stability limit. According to LTE specifications, the Backoff Parameter, which gives an upper limit for the retransmission intervals, can be as high as 960 ms [20]. Moreover, the two retransmission timers in LTE even seem to help (rather than hinder) to make the retransmissions sufficiently random. This logical reasoning for the justification of the Poisson approximation is complemented by a simulation study presented in Appendix B.

Also, as discussed earlier, in LTE actually K parallel Aloha channels are used and each time a device makes a retransmission the preamble is selected randomly. This randomization further helps in mixing the fresh random access requests together with the retransmission attempts.

In addition, it is good to notice that the input parameter of the model is not λ, the rate of fresh requests per subframe, but a, the aggregate rate of all requests. In the following section, we will first explain how the throughput θ of successful requests can be determined from the model as a function of a whenever the system is stable. In addition, we give a necessary and sufficient condition for stability. Thereafter we utilize the fact that, in any stable system, the average input rate must be the same as the average output rate, which implies that the arrival rate λ of fresh requests is equal to the throughput θ of successful requests whenever the system is stable. This is how we get the functional relationship between the aggregate request rate a and the arrival rate λ of fresh requests.

Now let A_{nk} denote the total number of random access requests with preamble k (including both the new ones and the retransmissions) in time slot n. Since the aggregate stream of requests (including the fresh ones and the retransmissions) is assumed to follow a Poisson process and the preambles are chosen independently from the uniform distribution, the A_{nk} are IID random variables obeying a Poisson distribution with mean ab/K and point probabilities

$$p_i(a) := \Pr\{A_{nk} = i\} = \frac{(ab/K)^i}{i!} e^{-ab/K}, \quad i \geq 0. \tag{2}$$

This is an immediate consequence of the so called splitting property of the Poisson process, see, e.g., Proposition 6.7 in [21] or Proposition 2.3.2 in [22].

Now define

$$Y_n^{(1)} := \#\{k : A_{nk} = 1, k = 1, \ldots, K\}, \tag{3}$$

$$\tilde{Y}_n^{(1)} := \#\{k : A_{nk} \geq 1, k = 1, \ldots, K\}, \tag{4}$$

where $\tilde{Y}_n^{(1)}$ is referring to the total number of preambles chosen in time slot n, and $Y_n^{(1)}$ is the number of successful (uncolliding) Message 1's. We observe that the joint distribution of the random variables $Y_n^{(1)}$ and $\tilde{Y}_n^{(1)}$ is as follows:

$$q_{ij}^{(1)}(a) := \Pr\{Y_n^{(1)} = i, \tilde{Y}_n^{(1)} = j\} = \binom{K}{i \quad j-i} p_0^{K-j} p_1^i (1 - p_0 - p_1)^{j-i}, \quad 0 \leq i \leq j \leq K, \tag{5}$$

where we have used the multinomial coefficient defined by

$$\binom{i+j+k}{i \quad j} := \frac{(i+j+k)!}{i!\, j!\, k!}$$

and a shorthand notation $p_i = p_i(a)$.

Consider now the dynamics of Step 2. Recall (from Table I) that c denotes the maximum number of UL grants included in a single Message 2. There are at most b Message 2's and, thus, at most bc UL grants per time slot. Message 2's in time slot n are generated by Message 1's of the previous time slot. Let $\tilde{Y}_n^{(2)}$ denote the total number of UL grants included in Message 2 in time slot n, and $Y_n^{(2)}$ the number of successful (uncolliding) UL grants. No losses appear in this step, $\tilde{Y}_n^{(2)} = \tilde{Y}_{n-1}^{(1)}$ and $Y_n^{(2)} = Y_{n-1}^{(1)}$, if the total number of preambles chosen in the previous time slot is sufficiently small, $\tilde{Y}_{n-1}^{(1)} \leq bc$, which is trivially true if $K \leq bc$. But if $\tilde{Y}_{n-1}^{(1)} > bc$, then losses happen so that $\tilde{Y}_n^{(2)} = bc$. We assume that the preambles that are given a UL grant in the latter case are chosen randomly by eNB. Thus, we have (for the non-trivial case $K > bc$)

$$q_{ij}^{(2)}(a) := \Pr\{Y_n^{(2)} = i, \tilde{Y}_n^{(2)} = j\} = \tag{6}$$

$$\begin{cases} q_{ij}^{(1)}(a), & 0 \leq i \leq j < bc, \\ \displaystyle\sum_{k=bc}^{K} \sum_{\ell=i}^{k} q_{\ell k}^{(1)}(a) \frac{\binom{\ell}{i}\binom{k-\ell}{bc-i}}{\binom{k}{bc}}, & 0 \leq i \leq j = bc, \end{cases} \tag{7}$$

with the following marginal distributions

$$q_i^{(2)}(a) := \Pr\{Y_n^{(2)} = i\} = \sum_{j=i}^{bc} q_{ij}^{(2)}(a), \quad 0 \leq i \leq bc, \tag{8}$$

$$\tilde{q}_j^{(2)}(a) := \Pr\{\tilde{Y}_n^{(2)} = j\} = \sum_{i=0}^{j} q_{ij}^{(2)}(a), \quad 0 \leq j \leq bc. \tag{9}$$

By utilizing the definition of $q_{ij}^{(2)}(a)$, we easily find that

$$\tilde{q}_j^{(2)}(a) = \begin{cases} \binom{K}{j} p_0(a)^{K-j} (1 - p_0(a))^j, & 0 \leq j < bc, \\ \displaystyle\sum_{\ell=bc}^{K} \binom{K}{\ell} p_0(a)^{K-\ell} (1 - p_0(a))^\ell, & j = bc. \end{cases} \tag{10}$$

Thus, we have the following representation:

$$\tilde{Y}_n^{(2)} = \min\{B(a), bc\}, \tag{11}$$

where $B(a)$ is a binomially distributed random variable with parameters K and $1 - p_0(a)$.

In Step 3, the collisions (originally due to Step 1) are realized. Those Message 3's in time slot n that are generated by unsuccessful UL grants conveyed in the Message 2's of the previous time slot are colliding in this step. Clearly we have $Y_n^{(3)} = Y_{n-1}^{(2)}$, where $Y_n^{(3)}$ denotes the number of successful Message 3's in time slot n.

Consider finally the dynamics of Step 4. New Message 4's in time slot n are generated by *successful* Message 3's of the previous time slot. Recall (from Table I) that M denotes the maximum number of Message 4's per subframe. Thus, at most bM Message 4's can be carried in a single time slot. Recall

also that the maximum number (per subframe) is reduced from M to m if there is a Message 2 in the corresponding subframe. On the other hand, there may be at most c successful Messages 3's per subframe. Now we have to consider two different cases separately.

1) If $c \leq m$, then all new Message 4's in time slot n are transmitted immediately in the same time slot, implying that the retransmission timer is never triggered due to delay,

$$\Pr\{\text{no failure in Steps 1 and 2, delay in Step 4}\} = 0.$$

2) On the other hand, if $c > m$, then it is not guaranteed that all new Message 4's can be delivered in a time slot. Thus, in this case, a *buffer* is needed for Message 4's in order to avoid additional losses.

Assume now that $c > m$, and let X_n denote the number of buffered Message 4's in the beginning of time slot n. The evolution of X_n is as follows:

$$X_{n+1} = X_n - Y_n^{(4)} + Y_n^{(3)} = X_n - Y_n^{(4)} + Y_{n-1}^{(2)}. \quad (12)$$

Here $Y_n^{(4)}$ denotes the number of transmitted Message 4's in time slot n,

$$Y_n^{(4)} = \min\left\{X_n, bM - \left\lceil \tilde{Y}_n^{(2)}/c \right\rceil (M-m)\right\}. \quad (13)$$

Note that the expression

$$bM - \left\lceil \tilde{Y}_n^{(2)}/c \right\rceil (M-m)$$

on the right hand side refers to the leftover PDCCH service capacity for Message 4's in time slot n.

We observe that $(X_n, Y_n^{(3)})$ is an irreducible and aperiodic two-dimensional Markov chain with state space

$$\mathcal{E} = \{0, 1, \ldots\} \times \{0, 1, \ldots, bc\}.$$

Equation (12) describes the evolution of the first component, while the second one is independent of the previous state of the process,

$$Y_{n+1}^{(3)} \perp\!\!\!\perp (X_n, Y_n^{(3)}),$$

from which the Markov property can be verified. Irreducibility (under assumption $c > m$) and aperiodicity follow easily from the construction.

To summarize, when $c > m$, Message 4's form a queue, and unlike in the previous case we get a nonzero probability of a failure due to queuing delay in Step 4. This probability can be calculated numerically as described later in Section VI.

V. Stability and throughput analysis

In this section we consider the stability of our buffer model. If the system is stable, the buffer for Message 4's does not "explode" and, as already explained in the previous section, the throughput θ of successful requests must be equal to the arrival rate λ of fresh requests. Thus, our purpose is first to find conditions for stability in terms of the total traffic a, and then to determine the throughput of successful requests, $\theta(a)$, as a function of a, as well as the maximum throughput $\theta^* = \max_a \theta(a)$. In order to simplify notation, we assume throughout this section that $K > bc$. The generalization to the case $K \leq bc$ is straightforward.

Recall that we apply the well-known Slotted Aloha model [7] for the random access channel used in Step 1. Thus, the throughput (per subframe) of successful Message 1's as a function of a, the arrival rate of *all* random access requests per subframe, is given by

$$\theta^{(1)}(a) = \mathbf{E}[Y_n^{(1)}]/b = a\, e^{-ab/K}. \quad (14)$$

The maximum throughput in Step 1 is achieved when ab equals the number of available preambles,

$$\max_a \theta^{(1)}(a) = \theta^{(1)}(K/b) = (K/b)\, e^{-1} \approx (K/b) \cdot 0.368,$$

which puts an upper limit for the throughput θ of successful requests, as well as for the arrival rate λ of fresh requests. For greater values of λ, the performance of the random access channel collapses. With the parameter values given in Table I,

$$\max_a \theta^{(1)}(a) = 3.973 \text{ requests/ms}.$$

Let us then consider the throughput in Step 2. Since $K > bc$, the throughput is further reduced by the limited number of UL grants in Message 2. The throughput (per subframe) of successful UL grants as a function of a is clearly

$$\theta^{(2)}(a) = \mathbf{E}[Y_n^{(2)}]/b = \frac{1}{b}\sum_{i=1}^{bc} i q_i^{(2)}(a), \quad (15)$$

where $q_i^{(2)}(a)$ is defined in (8). With the parameter values given in Table I (including $c = 3$), we have, after a numerical optimization of (15) with respect to a, the maximum throughput in Step 2 as follows:

$$\max_a \theta^{(2)}(a) = 2.377 \text{ requests/ms}.$$

Note from (15) that, for all a,

$$\theta^{(2)}(a) < c, \quad (16)$$

implying that

$$\max_a \theta^{(2)}(a) \leq c. \quad (17)$$

On the other hand, since $\theta^{(2)}(a)$ is continuous satisfying $\theta^{(2)}(a) \leq \theta^{(1)}(a)$ for all a, and $\theta^{(1)}(a)$ is bounded with limits $\theta^{(1)}(a) \to 0$ for $a \to 0$ and $a \to \infty$, there is $a_2^* < \infty$ such that

$$\theta^{(2)}(a_2^*) = \max_a \theta^{(2)}(a) \leq \max_a \theta^{(1)}(a) = (K/b)\, e^{-1}. \quad (18)$$

Part of the resources in Step 2 are, however, wasted by colliding UL grants. Let $\tilde{\theta}^{(2)}(a)$ denote the throughput (per subframe) of all UL grants as a function of a,

$$\tilde{\theta}^{(2)}(a) = \mathbf{E}[\tilde{Y}_n^{(2)}]/b = \frac{1}{b}\sum_{j=1}^{bc} j \tilde{q}_j^{(2)}(a). \quad (19)$$

The remaining part of the downlink control channel resources are available for Message 4's. The average leftover service

capacity (per subframe) for Message 4's in time slot n is given by

$$\sigma^{(4)}(a) = \mathbf{E}\left[bM - \left\lceil \tilde{Y}_n^{(2)}/c \right\rceil (M-m)\right]/b = m + (M-m)\left(\tilde{q}_0^{(2)}(a) + \sum_{i=1}^{b-1} \frac{i}{b} \sum_{j=(b-i-1)c+1}^{(b-i)c} \tilde{q}_j^{(2)}(a)\right). \quad (20)$$

where $\tilde{q}_j^{(2)}(a)$ is defined in (9). Note that we clearly have

$$\sigma^{(4)}(a) \geq m. \quad (21)$$

Below we give a necessary and sufficient condition for the stability of the Message 4 buffer, which is the main theoretical result of the paper. For clarity, we have placed the proof in Appendix.

Proposition 1: The buffer for Message 4's is stable if and only if

$$\theta^{(2)}(a) < \sigma^{(4)}(a).$$

If the buffer is stable, the throughput of successful requests, $\theta(a)$, is clearly equal to $\theta^{(2)}(a)$. Define now

$$a_4^* = \sup\{a \leq a_2^* : \theta^{(2)}(a) < \sigma^{(4)}(a)\}. \quad (22)$$

As a direct corollary of Proposition 1, we get the following result.

Corollary 1: The throughput of successful requests is given by

$$\theta(a) = \theta^{(2)}(a)$$

for all $a < a_4^*$, and the maximum throughput by

$$\theta^* = \theta^{(2)}(a_4^*). \quad (23)$$

In addition, we recall from the previous section that the arrival rate λ of fresh requests is equal to the throughput θ of successful requests whenever the system is stable. Thus, for all $a < a_4^*$,

$$\lambda(a) = \theta^{(2)}(a).$$

VI. Performance analysis methodology

In this section, we describe how we use the delay model to determine the performance of the system. The evaluation can only be done numerically and it depends on the following functions $\theta^{(1)}(a)$, see (14) , $\theta^{(2)}(a)$, see (15), $\tilde{\theta}^{(2)}(a)$, see (19) and $\sigma^{(4)}(a)$, see (20). These in turn depend on further definitions given in Section IV. Note that our table of symbols and definitions, Table II, also provides references to the equations characterizing our derived quantities. We start by describing the evaluation of the maximum throughput θ^* and then continue with determining the failure probability and its various components.

A. Determining the maximum throughput θ^*

Consider first the maximum throughput θ^*, which can be thought as the capacity of the system. Functions $\theta^{(2)}(a)$ and $\sigma^{(4)}(a)$ are clearly continuous satisfying

$$\lim_{a\to 0} \theta^{(2)}(a) = 0 < M = \lim_{a\to 0} \sigma^{(4)}(a).$$

While not at all easy to prove, it is, however, intuitively clear that $\theta^{(2)}(a)$ is an increasing function for all $a < a_2^*$ and it intersects with the decreasing function $\sigma^{(4)}(a)$ at most once in the interval $a \in [0, a_4^*]$. If there is no intersection, then there is no stability issue (of the Message 4 buffer) and the maximum throughput is determined by Step 2,

$$\theta^* = \theta^{(2)}(a_2^*) = \max_a \theta^{(2)}(a).$$

The situation is illustrated in Figure 2, where we have plotted functions $\theta^{(1)}(a)$, $\tilde{\theta}^{(2)}(a)$, $\theta^{(2)}(a)$, and $\sigma^{(4)}(a)$ with the (basic) parameter values given in Table I. Note that for these values $c \leq m$ so that there will be no queue at all in the Message 4 buffer. As already mentioned in the previous section, the maximum throughput is

$$\theta^* = 2.377 \text{ requests/ms},$$

which is also visible from Figure 2.

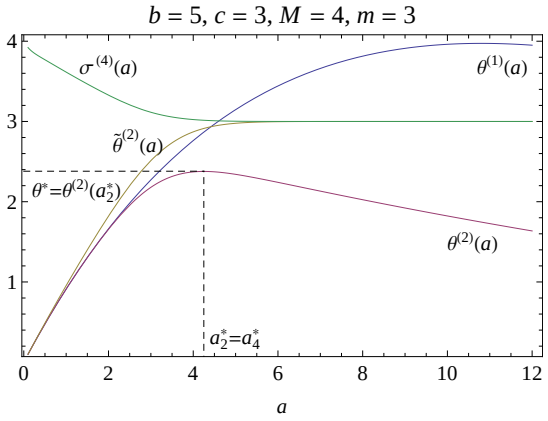


Fig. 2: Illustration of the throughput curves (in different steps) and the maximum throughput θ^* in the case where $\theta^{(2)}(a)$ and $\sigma^{(4)}(a)$ do not intersect. The parameter values are taken from Table I.

On the other hand, if curves $\theta^{(2)}(a)$ and $\sigma^{(4)}(a)$ intersect, then the maximum throughput is equal to the stability limit of the Message 4 buffer,

$$\theta^* = \theta^{(2)}(a_4^*) = \sigma^{(4)}(a_4^*). \quad (24)$$

Figure 3 gives an example of this situation. The parameters used in this case are otherwise the same as in the previous figure but now $c = 6$ (instead of $c = 3$) so that there are more UL grants in Step 2, which makes the stability of the Message 4 buffer an issue. With these parameter values, the maximum throughput is

$$\theta^* = 3.225 \text{ requests/ms},$$

which is calculated by numerically solving (24), where $\theta^{(2)}(a)$ is defined in (15) and $\sigma^{(4)}(a)$ in (20). The numerical result can also be verified by the figure.

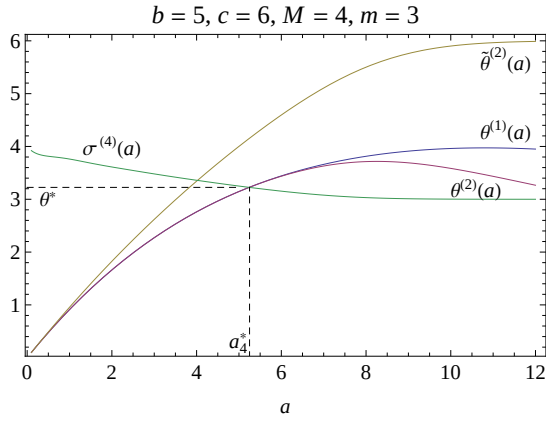


Fig. 3: Illustration of the throughput curves (in different steps) and the maximum throughput θ^* in the case where $\theta^{(2)}(a)$ and $\sigma^{(4)}(a)$ intersect. The parameter values are otherwise the same as in Figure 2 but now $c = 6$ (instead of 3).

B. Determining the failure probability

In addition to the maximum throughput, we are interested in determining the failure probability as a function of the arrival rate λ of fresh requests. Recall from (1) that

$$\begin{aligned}\Pr\{\text{failure}\} = \Pr\{&\text{collision in Step 1}\}+\\ &\Pr\{\text{no collision in Step 1, loss in Step 2}\}+\\ &\Pr\{\text{no failure in Steps 1 and 2, delay in Step 4}\},\end{aligned}$$

The first two probabilities on the right hand side are clearly given by the following equations:

$$\Pr\{\text{collision in Step 1}\} = 1 - \frac{\theta^{(1)}(a)}{a},$$

$$\Pr\{\text{no collision in Step 1, loss in Step 2}\} = \frac{\theta^{(1)}(a)}{a}\left(1 - \frac{\theta^{(2)}(a)}{\theta^{(1)}(a)}\right) = \frac{\theta^{(1)}(a) - \theta^{(2)}(a)}{a}.$$

The third one satisfies

$$\begin{aligned}\Pr\{\text{no failure in Steps 1 and 2, delay in Step 4}\} = &\\ \Pr\{\text{no failure in Steps 1 and 2}\}\times&\\ \Pr\{\text{delay in Step 4} \mid \text{no failure in Steps 1 and 2}\}&\end{aligned}$$

with

$$\Pr\{\text{no failure in Steps 1 and 2}\} = \frac{\theta^{(2)}(a)}{a}.$$

If $c \leq m$, then we know (from Section IV) that

$$\Pr\{\text{delay in Step 4} \mid \text{no failure in Steps 1 and 2}\} = 0.$$

However, if $c > m$, we do not have any explicit expression for the conditional probability

$$\Pr\{\text{delay in Step 4} \mid \text{no failure in Steps 1 and 2}\},$$

but we resort to simulations of the two-dimensional Markov chain $(X_n, Y_n^{(3)})$. In this way, we get an estimate of the conditional probability for any fixed total rate a.

TABLE III: Different scenarios used in the numerical studies. Note that P0 is the basic scenario that is taken from [17] and also mentioned in Table I.

Scenario	b	c	M	m	N^{Msg2}	N^{Msg4}	N
P0	5	3	4	3	4	4	16
P1	5	6	4	3	4	4	16
P2	5	3	4	2	8	4	16
P3	5	6	4	2	8	4	16
P4	5	3	2	1	8	8	16
P5	5	6	2	1	8	8	16
P6	1	3	4	3	4	4	16
P7	1	6	4	3	4	4	16

In the final step, we determine the corresponding arrival rate λ of fresh requests from the equation below,

$$\lambda(a) = \theta^{(2)}(a), \tag{25}$$

which is valid whenever the system is stable as explained in the previous section. By utilizing its inverse function $a(\lambda)$, we are finally able to express the failure probability and its components as a function of the arrival rate λ of fresh requests.

VII. Numerical results

In this section we illustrate the properties of our model through numerical examples. First we observe how the maximum throughput behaves when the parameter c, which is the maximum number of uplink grants in Message 2, is varied. Then we study the queuing behavior of the Message 4 buffer. In addition, we analyze various components of the random access failure probability with varying traffic load. We then provide a method to estimate the maximum number of users that can exist in a cell based on traffic models presented in [17] and finally dimension the PDCCH resource to sustain such traffic.

For the numerical study, we make use of various scenarios formed by different combinations of the parameters b, c, M and m provided in Table III. In every scenario we have a total of $K = 54$ preambles available for the contention based random access, and in every subframe $N = 16$ CCEs are used in the PDCCH for purposes of random access (i.e., sending Message 2's and Message 4's). Scenario P0 is the basic scenario mentioned in the "Typical value" column of Table I. An equivalent parameter set is considered in the RAN overload study [17] by 3GPP. In Scenario P1 we assume $c = 6$ UL grants are given per Message 2 while the rest of the parameters are the same as P1. In Scenarios P2 and P3, $N^{\text{Msg2}} = 8$ CCEs are used for a Message 2 and $N^{\text{Msg4}} = 4$ CCEs for a Message 4, while Scenarios P4 and P5 assume that each Message 2 and Message 4 use $N^{\text{Msg2}} = N^{\text{Msg4}} = 8$ CCEs. The final two scenarios P6 and P7 are *almost* equivalent to Scenarios P0 and P1, respectively, except that they consider the random access opportunity to be available in every subframe (rather than every 5 subframes considered in P0 and P1), i.e., $b = 1$ is assumed in these two final scenarios.

A. Maximum throughput θ^*

From Figures 2 and 3 we observe that the maximum throughput, θ^*, clearly increases with c, the number of UL

grants in Message 2. Maximum throughput is calculated according to Corollary 1 as explained in detail in Section VI-A. Clearly, as c increases, more and more UL grants can be sent in Message 2 of the first subframe (out of b), freeing up the resources to send the Message 4's in the later subframes. Since the number of PDCCH resources is limited, this growth cannot continue indefinitely, and in the limit $c \to \infty$, the maximum throughput approaches a value which is approximately equal to $M - (M - m)/b$, as can be observed from Figure 4.

Intuition behind this approximate limit is as follows. Assume that b is sufficiently small so that the bottleneck of the whole system is due to competition of common PDCCH resources between Message 2's and Message 4's. Now, when $c \to \infty$, all the UL grants (if any) related to a single time slot can be sent in Message 2 of the first subframe (out of b). Thus, there are typically m Message 4's in the first subframe and M Message 4's in the remaining $b-1$ subframes of the time slot in consideration so that the total number of Message 4's per subframe is approximated by

$$\frac{1}{b}(m + (b-1)M) = M - \frac{M-m}{b}.$$

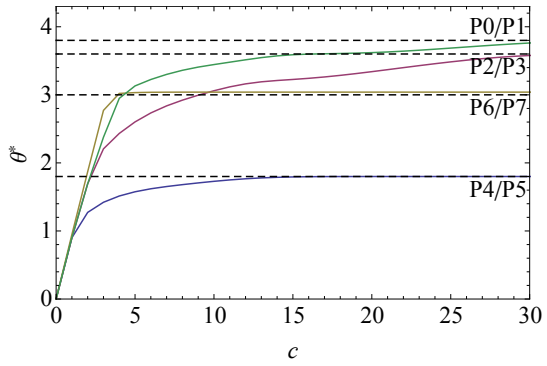


Fig. 4: The maximum throughput as a function of parameter c for different scenarios. We see that the throughput increases with c until it saturates approximately to level $M-(M-m)/b$ (represented by a dashed line). The values of parameters b, M, and m used here are taken from Table III, while parameter c is allowed to vary.

Moreover, from Figure 4, we observe that when a large amount of UL grants are sent in Message 2 (large c), the parameter combination (b, M and m) of P0/P1 performs best in terms of maximum throughput among all the scenarios considered here. Moreover there is also an optimal trade-off between parameters b and c. For a given value of c, increasing b will allow us to send more uplink grants in a time slot (less loss in Step 2) while increasing the collision probability in Step 1. For smaller values of c (typically ≤ 5), different scenarios give optimal performance. For example, when $M = 4$ and $m = 3$ (scenarios P0 and P1), the optimal maximal throughput is a function of the periodicity of the random access opportunities, b, when the number of uplink grants in Message 2, c, is fixed. This can be observed in Figure 5 where we plot the maximal throughput against the periodicity parameter, b for different values of c. In such cases we see that the optimal b is 1 for smaller values of c. This optimal periodicity becomes larger as we increase c.

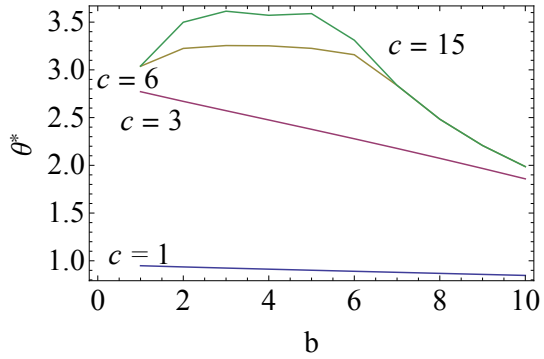


Fig. 5: The maximum throughput as a function of periodicity parameter, b, for different values of uplink grants, c, sent in Message 2. We see that for any given c, there is an optimal value of b for which we get the largest maximal throughput.

B. Behavior of the Message 4 queue

To study the queuing behavior of the Message 4 buffer, the Markov chain (12) was simulated for 10^6 time slots for different values of λ and for Scenarios P1–P7 (except P0 and P6, which do not form a queue) in Table III. Note that in the simulation, the total arrival rate a is the given parameter which then corresponds to a certain rate of new requests $\lambda(a)$ given by (25).

The simulation results are presented in Figure 6, which shows the mean queue length of the Message 4 buffer as a function of λ for different scenarios. As can be seen, it is characteristic for the system that the Message 4 queue remains nicely under control until λ is quite close to the stability limit, θ^*, given by (24). Intuitively, this also means that the likelihood of a user experiencing a timeout event due to buffering delay is very low until the load is close to the stability limit. In the next section, we will observe the contribution of this queuing delay in the timeout event of the random access procedure and compare it with the other causes of failure.

C. Contributions of various components in random access failure

To gain further insight to what are the most likely causes for random access failures, we look at the probabilities given by (1). The three components of the failure probability are depicted, on a logarithmic scale, in Figure 7 as a function of λ for Scenario P1 (with labels "Collision in 1", "Loss in 2", and "Delay in Step 4"). The failure probabilities for collision in Step 1 and loss in Step 2 are obtained numerically as described in Section VI-B, while the conditional probability $\Pr\{\text{delay in Step 4} \mid \text{no failure in Steps 1 and 2}\}$ needed for the third component in (1) has been estimated from simulations of the Markov chain (12) for 10^6 time slots. We can observe that, even with arrival rates close to the stability limit, the collision probability in Step 1 remains relatively low (of

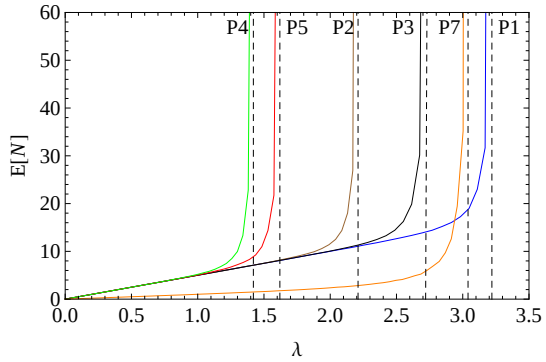


Fig. 6: Mean queue length as a function of λ for various scenarios. The vertical dotted lines represent the stability limit, θ^*, for each scenario. Scenarios P0 and P6 do not produce any queue as $c = m$ in those two cases.

the order 10^{-2}). Also, the probability of loss in Step 2 is considerably lower and makes no difference whatsoever in the system. Thus, the number of preambles K is not a bottleneck and the limitation of c is even less of a bottleneck. Ultimately, as λ grows, the probability of failure becomes dominated by the event that the queuing delay of Message 4 grows too large. However, this happens in a very sharp manner close to the stability limit. For the basic scenario, P0, we get no failures due to delay as discussed in Section IV but we still have two other components as demonstrated in Figure 8.

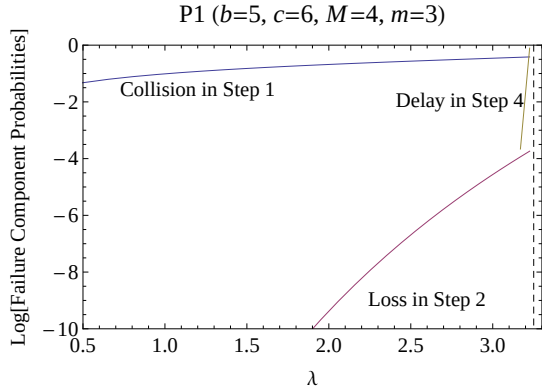


Fig. 7: A breakdown of the failure probability (logarithmic scale) as a function of λ for Scenario P1. The vertical dotted line represents the maximum throughput θ^* for the parameter set used.

On the other hand, in Scenario P6, where a random access opportunity is available in every subframe, a different picture emerges (see Figure 9). Now the probability of loss in Step 2 increases more rapidly than the collision probability in Step 1. Since no queue is formed as $c = m$, there is no loss due to queuing delay. Even at a moderate arrival rate, the limitation of the control channel resource begins to contribute to the failure of the random access procedure more than the collisions in Step 1. This is because very few UL grants can be sent in one Message 2 ($c = 3$) and some users who were successful in Step 1 have to be dropped in Step 2. For Scenario P7, the loss probability in Step 2 remains below the collision probability

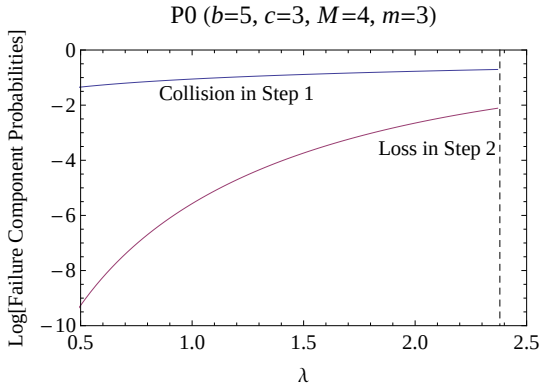


Fig. 8: A breakdown of the failure probability (logarithmic scale) as a function of λ for Scenario P0. The vertical dotted line represents the maximum throughput θ^* for the parameter set used. We see that the probability of failure due to Loss in Step 2 is considerably higher compared to that of Scenario P1.

in Step 1 while approaching it at higher arrival rates. Near the stability limit, the delay in Step 4 again dominates the probability of random access failure (see Figure 10).

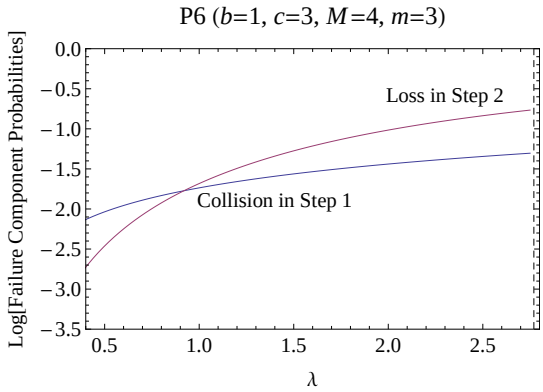


Fig. 9: A breakdown of the failure probability (logarithmic scale) as a function of λ for Scenario P6. The vertical dotted line represents the maximum throughput θ^* for the parameter set used. Note how the limitation of c causes the loss probability in Step 2 to increase beyond the collision probability in Step 1 for higher arrival rates. No queue is formed in this case as $c = m$ so that there is no delay component in the random access failure.

D. Estimation of maximum number of devices in a cell

From our model it is possible to estimate the maximum number of MTC devices that can exist in a cell if the traffic characteristic of the machines is similar to Traffic Model 1 in [17, Table 6.1.1], which assumes that there are a fixed number, D, of devices in a cell each generating a request for random access uniformly over a period of 60 seconds. In our model, this corresponds to arrivals at a rate (or maximum throughput) $\lambda = \theta^* = D/60\,000$ per millisecond, where θ^* is the maximum throughput the respective scenarios support given by (23). Therefore the maximum number of devices can

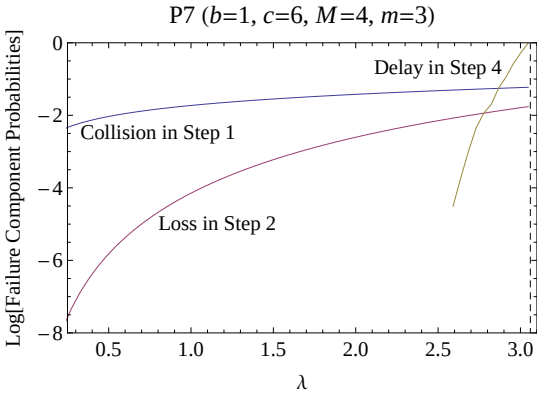


Fig. 10: A breakdown of the failure probability (logarithmic scale) as a function of λ for Scenario P7. The vertical dotted line represents the maximum throughput θ^* for the parameter set used.

TABLE IV: An estimate of the maximum number of devices that can exist in a cell according to our model when the arrival rate of the request follows Traffic Model 1 [17, Table 6.1.1].

Scenario	$D_{\max}$
P0	143 000
P1	194 000
P2	132 000
P3	164 000
P4	85 000
P5	97 000
P6	166 000
P7	182 000

be estimated by the quantity $60\,000{\cdot}\theta^*$. For different scenarios, the maximum number $D_{\max}$ of such devices is tabulated in Table IV.

We can now understand the reason for practically no losses when there are $30\,000$ devices in a cell when Traffic Model 1 [17, Table 6.1.1] is used — according to our model as many as $143\,000$ devices can be served and $30\,000$ is too small a number to produce any kind of discernible failure. On the other hand, when Traffic Model 2 [17, Table 6.1.1] is used, which has an intensity six times that of Traffic Model 1 and is more bursty in comparison, far fewer than $30\,000$ devices may reliably be offered services. Clearly, we will have low probability of random access success with this traffic if $30\,000$ devices are used, as is evident from the packet-level simulation in [17, Table 6.4.1.1.1]. Thus, we can conclude that the model can be used to make predictions about the capacity of a cell as well.

E. Dimensioning the PDCCH resource

In this section we will demonstrate a method to dimension the PDCCH resource to be used in a cell under the two traffic models mentioned in [17], i.e., determine the minimum number of CCEs, $N^{\min}$, that is necessary for the system to operate properly under these two traffic models for scenarios P0–P7. More specifically, we fix the parameters b, c, N^{Msg2} and N^{Msg4} and determine the values of N (and consequently those of M and m as well) for different scenarios that will allow them support the traffic models described in [17].

TABLE V: PDCCH resource size in CCEs needed to support Traffic Model 1 [17, Table 6.1.1].

Scenario	M	m	N^{Msg2}	N^{Msg4}	$N_{\min}$
P0	1	0	4	4	4
P1	1	0	4	4	4
P2	2	0	8	4	8
P3	2	0	8	4	8
P4	1	0	8	8	8
P5	1	0	8	8	8
P6	1	0	4	4	4
P7	1	0	4	4	4

TABLE VI: PDCCH resource size in CCEs needed to support Traffic Model 2 [17, Table 6.1.1].

Scenario	M	m	N^{Msg2}	N^{Msg4}	$N_{\min}$
P0	–	–	–	–	–
P1	4	3	4	4	16
P2	–	–	–	–	–
P3	6	4	8	4	24
P4	–	–	–	–	–
P5	4	3	8	8	32
P6	–	–	–	–	–
P7	4	3	4	4	16

In Traffic Model 1 $30\,000$ devices make random access attempts uniformly over a 60 second period. This means that a maximum throughput of $\theta^* = \frac{30\,000}{60 \text{ sec}} = 0.5$ requests per subframe should be supported. In Table V, we show the minimum number of CCEs necessary to sustain at most $30\,000$ devices. In summary, in scenarios P0 and P1 we need just $N_{\min} = 4$ CCEs in the PDCCH to send Message 2's and Message 4's. In P2, P3, P4, and P5 a minimum of $N_{\min} = 8$ CCEs is sufficient in the PDCCH for sending Message 2's and Message 4's. In the final two scenarios P6 and P7, where we have a random access opportunity in every subframe ($b = 1$), we again need a minimum of $N_{\min} = 4$ CCEs to sustain the arrival rate described in Traffic Model 1. It should be noted that in all our scenarios at least 4 or 8 CCEs are necessary to send each Message 2 or Message 4, which constrains the choice of these minimum number of necessary CCEs to be the multiples of 4 or 8.

The arrival rate is six times higher in Traffic Model 2 compared to the first. Under these conditions, a maximum throughput of $\theta^* = \frac{30\,000}{10 \text{ sec}} = 3.0$ arrivals per subframe should be supported by the system. From (15), we see that with $c = 3$, this throughput is never achieved making it impossible for scenarios P0, P2, P4 and P6 to sustain the arrival rate described in Traffic Model 2 with any number of CCEs. In P1, P3, P5 and P7, where $c = 6$ uplink grants are provided in a Message 2, we require N to be at least 16, 24, 32 and 16, respectively to sustain the traffic. The results for Traffic Model 2 are summarized Table VI.

VIII. Conclusions

The PDCCH of LTE may become a bottleneck when a very large number of devices want access to the network. We have presented a Markov chain model to describe the sharing of the PDCCH resources between Message 2's and Message 4's. Using the model we have calculated the contribution of

various events in the random access procedure's failure. In addition, we have derived a method to determine the maximum throughput of the system, which reveals the upper limit for the arrival rate of fresh random access requests. This method is then used to dimension the PDCCH resource size to support the traffic models studied by 3GPP in [17].

We have observed that near the stability limit, the probability of failure increases very sharply. Indeed, it is easy to see that by admitting more users to the system in Step 2, the capacity (measured by the maximum throughput) of the random access channel can be modestly increased. However, this also increases the probability of the random access failure due to a large queuing delay in the Message 4 buffer. This is because the size of the PDCCH resource is fixed and Message 2's always have priority over Message 4's. Moreover, using our model we are also able to predict the maximum number of devices that can exist in a cell under uniform traffic conditions. Analysis also shows that reducing the number of CCEs in each Message 2 or Message 4 increases the capacity of the random access channel but the behavior of the queue near the stability limit remains more or less the same. So limiting the arrival rates by some kind of admission control is necessary to manage this overload.

Our model can be extended to study the realization of collision in Step 1 of the random access procedure including the impact of the capture effect. These extensions will, no doubt, give better performance bounds but it remains to be seen how much improvement can they really offer. Additionally, the impact of physical layer impairments on the random access procedure can be studied by modifying the model, e.g., by explicitly taking into account the retransmission mechanism and fading effects. As a part of future work, we can also study the impact of sending further messages in the PDCCH after the random access procedure is finished. Moreover, techniques like enhanced PDCCH have been proposed by 3GPP to overcome the overload issues of PDCCH which can also be a subject of further study.

Appendix A
Proof of Proposition 1

Here we give the proof for the necessary and sufficient condition for the stability of the Message 4 buffer.

Proposition 1: The buffer for Message 4's is stable if and only if

$$\theta^{(2)}(a) < \sigma^{(4)}(a).$$

Proof: a) Assume first that $c \leq m$. Then we know from the previous section that there will be no queue at all in the Message 4 buffer (which, of course, is one form of stability). On the other hand, we have in this case

$$\theta^{(2)}(a) < c \leq m \leq \sigma^{(4)}(a)$$

by (16) and (21). So the claim is true whenever $c \leq m$.

b) Assume now that $c > m$, and consider the Markov chain $(X_n, Y_n^{(3)})$ defined on

$$\mathcal{E} = \{0, 1, \ldots\} \times \{0, 1, \ldots, bc\}.$$

The buffer for Message 4's is stable if and only if this irreducible and aperiodic Markov chain is positive recurrent, i.e., there is a unique steady-state distribution

$$\pi_{ij} = \lim_{n\to\infty} \Pr\{X_n = i, Y_n^{(3)} = j\}.$$

Now, depending on a, we have two cases to consider: 1° $\theta^{(2)}(a) < \sigma^{(4)}(a)$ and 2° $\theta^{(2)}(a) \geq \sigma^{(4)}(a)$.

1° Assume first that a is such that

$$\theta^{(2)}(a) < \sigma^{(4)}(a), \tag{26}$$

and define $\delta = b(\sigma^{(4)}(a) - \theta^{(2)}(a)) > 0$. For any $(i, j) \in \mathcal{E}$, we have

$$\begin{aligned}
&\mathbf{E}[X_{n+1} + Y_{n+1}^{(3)} | X_n = i, Y_n^{(3)} = j] \\
&\quad = \mathbf{E}[X_n - Y_n^{(4)} + Y_n^{(3)} | X_n = i, Y_n^{(3)} = j] + \mathbf{E}[Y_{n+1}^{(3)}] \\
&\quad \leq \mathbf{E}[X_n + Y_n^{(3)} | X_n = i, Y_n^{(3)} = j] + \mathbf{E}[Y_n^{(2)}] \\
&\quad = i + j + b\theta^{(2)}(a) \\
&\quad = i + j - \delta(1 - \epsilon),
\end{aligned}$$

where $\epsilon = b\sigma^{(4)}(a)/\delta$. In addition, for any $(i, j) \in \mathcal{E}$ such that $i > bM$, we have

$$\begin{aligned}
&\mathbf{E}[X_{n+1} + Y_{n+1}^{(3)} | X_n = i, Y_n^{(3)} = j] \\
&\quad = \mathbf{E}[X_n - Y_n^{(4)} + Y_n^{(3)} | X_n = i, Y_n^{(3)} = j] + \mathbf{E}[Y_{n+1}^{(3)}] \\
&\quad = i - \mathbf{E}\left[bM - \left\lceil \tilde{Y}_n^{(2)}/c \right\rceil (M - m)\right] + j + \mathbf{E}[Y_n^{(2)}] \\
&\quad = i - b\sigma^{(4)}(a) + j + b\theta^{(2)}(a) \\
&\quad = i + j - \delta.
\end{aligned}$$

Thus the non-negative function V defined on $\mathcal{E}$ by

$$V(i, j) = (i + j)/\delta$$

satisfies Foster's criterion:

$$\begin{aligned}
&\mathbf{E}[V(X_{n+1}, Y_{n+1}^{(3)}) - V(i, j) | X_n = i, Y_n^{(3)} = j] \\
&\qquad \leq -1 + \epsilon\, 1_{\mathcal{F}}(x),
\end{aligned}$$

where $\mathcal{F}$ is the finite set

$$\mathcal{F} = \{0, 1, \ldots, bM\} \times \{0, 1, \ldots, bc\},$$

and $1_{\mathcal{F}}(x) = 1$ if $x \in \mathcal{F}$, otherwise 0. It follows from Foster's Theorem (see e.g. [23]) that the Markov chain $(X_n, Y_n^{(3)})$ is positive recurrent under condition (26).

2° Assume now that a is such that

$$\theta^{(2)}(a) \geq \sigma^{(4)}(a). \tag{27}$$

For any $(i, j) \in \mathcal{E}$, we have

$$\begin{aligned}
&\mathbf{E}[|X_{n+1} + Y_{n+1}^{(3)} - i - j| \, | X_n = i, Y_n^{(3)} = j] \\
&\quad = \mathbf{E}[|-Y_n^{(4)} + Y_{n+1}^{(3)}| \, | X_n = i, Y_n^{(3)} = j] \\
&\quad \leq \mathbf{E}[bM + Y_{n+1}^{(3)} | X_n = i, Y_n^{(3)} = j] \\
&\quad = bM + \mathbf{E}[Y_n^{(2)}] = b(M + \theta^{(2)}(a)).
\end{aligned}$$

Thus,

$$\sup_{(i,j)\in\mathcal{E}} \mathbf{E}[|V(X_{n+1}, Y_{n+1}^{(3)}) - V(i, j)| \, | X_n = i, Y_n^{(3)} = j] < \infty,$$

where V is now defined on $\mathcal{E}$ by

$$V(i,j) = i + j.$$

In addition, as shown in 1°, we have, for any $(i,j) \in \mathcal{E}$ such that $i > bM$,

$$\begin{aligned}\mathbf{E}[X_{n+1} + Y_{n+1}^{(3)} - i - j | X_n = i, Y_n^{(3)} = j] \\ = b(\theta^{(2)}(a) - \sigma^{(4)}(a)),\end{aligned}$$

implying, by (27), that

$$\mathbf{E}[V(X_{n+1}, Y_{n+1}^{(3)}) - V(i,j) | X_n = i, Y_n^{(3)} = j] \geq 0$$

for all $x \in \mathcal{E} \setminus \mathcal{F}$, where the finite set $\mathcal{F}$ is defined as in 1°. These conditions together are sufficient (see e.g. [23]) to show that the Markov chain $(X_n, Y_n^{(3)})$ is not positive recurrent under condition (27). ■

APPENDIX B
SIMULATION STUDY ON THE POISSON APPROXIMATION

Recall that due to a large number of machines making independent random access attempts, we assume that the fresh requests arrive according to a Poisson process with an average of λ arrivals per ms. Here we examine, through simulations, the accuracy of our Poisson approximation, which states that the aggregate random access requests (the fresh ones as well as the retransmissions) constitute, approximately, a Poisson process. The simulator mimics the four-step random access procedure in an LTE system and explicitly takes into account a backoff algorithm, which was not *directly* considered in the model presented in Section IV and the subsequent analyses. This mechanism handles the backoffs which may be caused by one of three events — if there is a collision in Step 1 (realized only in Step 3), or if there is a loss in Step 2, or if the delay is beyond the acceptable limit in Step 4 as explained in Section III. More specifically, if a fresh request fails (due to any one of the three events described earlier), our simulator models the backoff mechanism by assigning some probability of retransmission, P_{ReTx}, to that request. This means that such backlogged users will make an attempt for retransmission in the subsequent RACH opportunities with probability P_{ReTx}. After the first failure, a maximum of $N_{\max} - 1$ more retrial attempts can be made to send Message 1 until the attempt is successful. If the request is not successful even after $N_{\max}$ attempts, the request is dropped.

We work at the granularity of time slots of length $b = 5$ ms. A random access request will backoff immediately if it does not receive a Message 2 in the following time slot (loss in Step 2). A backoff due to collision in Step 1 occurs 10 time slots after the corresponding Message 3 is sent (collision in Step 1, realized in Step 3). A request queued up in the Message 4 buffer (and not yet sent) enters the backoff state if the corresponding Message 3 was sent 10 time slots earlier (delay in Step 4). Note that 10 time slots correspond to a time of 50 ms, which is close to the timer value of 48 ms used in [17].

The base station has a wide range of choices for Backoff Parameter values [20, Table 7.2-1]. After a failed transmission attempt, a user will uniformly select a time between 0 and the Backoff Parameter, T_{B}, to remain in the backoff state. This means that if $T_{\mathrm{B}} = 20$ ms (as done in [17, Table 6.2.2.1.1]), a retransmission attempt is made by such a user, on the average, after $\bar{T}_{\mathrm{B}} = 10$ ms (i.e, after two time slots in our simulation model) and a probability of retransmission $P_{\mathrm{ReTx}} = b/\bar{T}_{\mathrm{B}} = 5/10 = 0.5$ can be assigned to all backoff users. Similarly, if the $T_{\mathrm{B}} = 160$ ms (which is also possible according to [20, Table 7.2-1]), a retransmission takes place after of $\bar{T}_{\mathrm{B}} = 80$ ms in average, making $P_{\mathrm{ReTx}} = b/\bar{T}_{\mathrm{B}} = 5/80 = 0.0625$. As the upper bound for the number of trials, we use $N_{\max} = 10$ (as in [17, Table 6.2.2.1.1]) but also $N_{\max} = 5$ to see how varying the maximum number of trials affects the distribution of the aggregate number of random access requests.

Numerical results

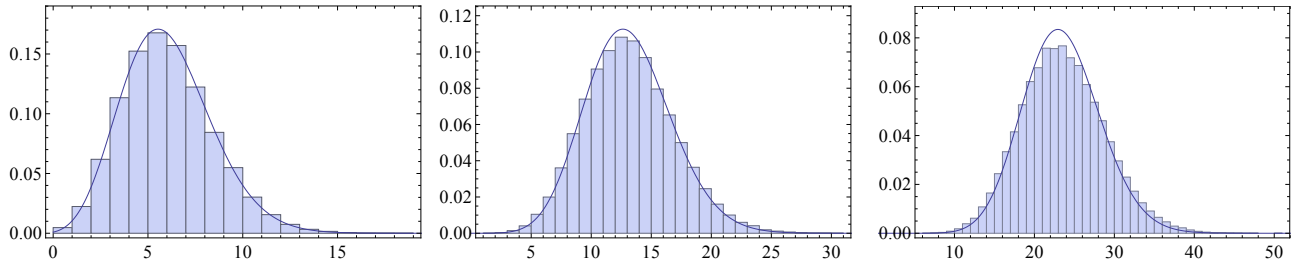
We have run simulations for Scenario P1 described in Table III, where the different parameter values used are $b = 5$ ms, $c = 6$, $M = 4$, and $m = 3$. The additional parameters for the simulation are $N_{\max}$ and P_{ReTx}. Recall that the highest rate of fresh arrivals (λ) that this scenario can support is $\theta^* = 3.225$ requests per ms as observed from Figure 3 and the discussion preceding it. To be as exhaustive as possible, we run the simulation for three traffic conditions — low traffic ($\lambda = 1.0$), medium traffic ($\lambda = 2.0$), and high traffic ($\lambda = 3.0$), where all the arrival rates are expressed per ms. Each simulation run consists of 100 000 time slots.

We first present the results for the case where $P_{\mathrm{ReTx}} = 1/2 = 0.5$ and $N_{\max} = 10$. In Figure 11 (and all the subsequent ones), we have plotted the empirical distribution of the aggregate number of all random access requests arriving in a time slot as the bar chart and overlayed it with the probability mass function of the Poisson distribution with the mean ab predicted from the theoretical model presented in the paper. Recall that a refers to the aggregate arrival rate of all random access requests (per ms). We see that for all the three traffic conditions, the empirical distribution is very close to the corresponding theoretical Poisson distribution. The empirical aggregate request rate $\widehat{ab}$ is also very close to the predicted value ab as shown in Table VII.

As expected, for low and medium arrival rates the empirical distribution is very close to a Poisson distribution. However, the empirical distribution bears a striking resemblance to a Poisson distribution even at heavy traffic (close to the stability limit θ^*). Similarly, if we reduce the value of $N_{\max}$ to 5, hardly any change is noticed in the empirical distribution. This can be seen in Figure 12. In addition, if we use a longer Backoff Parameter of 160 ms, corresponding to $P_{\mathrm{ReTx}} = 1/16 = 0.0625$, we notice that the distributions are even closer to a Poisson distribution as observed from Figure 13. Moreover, the Backoff Parameter can be as high as 960 ms, in which case $P_{\mathrm{ReTx}} = 1/96 \approx 0.010$, which helps to make the empirical distribution even closer to the Poisson distribution. In summary, we can say that the distribution for the aggregate number of all random access requests may be approximated by a Poisson distribution.

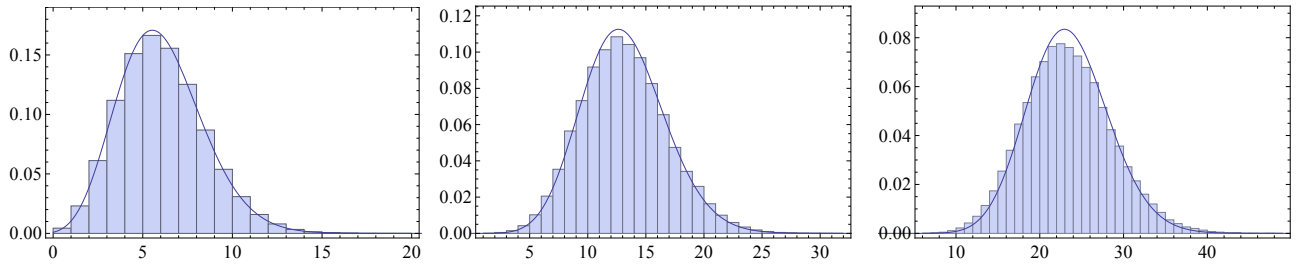
TABLE VII: The predicted value of ab and the corresponding empirical value $\widehat{ab}$ obtained from simulations for different values of parameters λ, P_{ReTx}, and $N_{\max}$.

		$\widehat{ab}$		
λ	ab	$P_{\mathrm{ReTx}}=0.5$, $N_{\max}=10$	$P_{\mathrm{ReTx}}=0.5$, $N_{\max}=5$	$P_{\mathrm{ReTx}}=0.0625$, $N_{\max}=10$
1.0	5.54	5.54	5.65	5.53
2.0	12.64	12.67	12.65	12.63
3.0	22.94	23.07	22.86	23.00



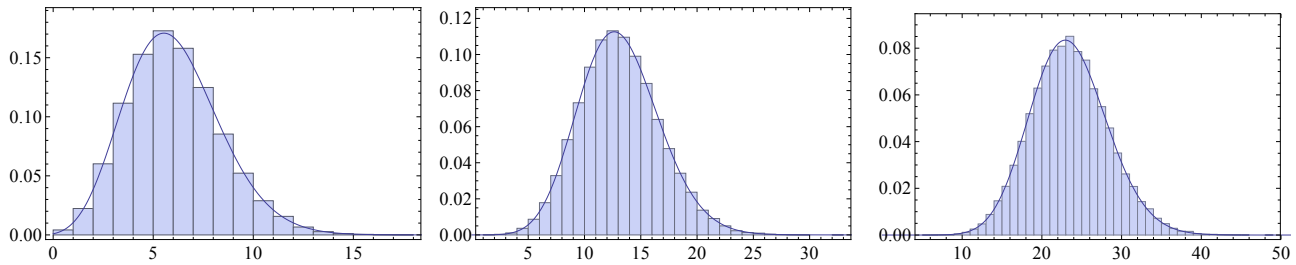
(a) $\lambda = 1.0$, $\widehat{ab} = 5.54$ (b) $\lambda = 2.0$, $\widehat{ab} = 12.67$ (c) $\lambda = 3.0$, $\widehat{ab} = 23.07$

Fig. 11: Empirical distribution of Message 1 when $P_{\mathrm{ReTx}} = 0.5$ and $N_{\max}$= 10



(a) $\lambda = 1.0$, $\widehat{ab} = 5.65$ (b) $\lambda = 2.0$, $\widehat{ab} = 12.65$ (c) $\lambda = 3.0$, $\widehat{ab} = 22.86$

Fig. 12: Empirical distribution of Message 1 when $P_{\mathrm{ReTx}} = 0.5$ and $N_{\max}$= 5



(a) $\lambda = 1.0$, $\widehat{ab} = 5.53$ (b) $\lambda = 2.0$, $\widehat{ab} = 12.63$ (c) $\lambda = 3.0$, $\widehat{ab} = 23.00$

Fig. 13: Empirical distribution of Message 1 when $P_{\mathrm{ReTx}} = 0.0625$ and $N_{\max}$= 10

REFERENCES

[1] J. Conti, "The internet of things," *Communications Engineer*, vol. 4, no. 6, pp. 20–25, Dec.-Jan. 2006.

[2] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wan, "A first look at cellular machine-to-machine traffic—large scale measurement and characterization," in *Proceedings of the 2012 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*. ACM, 2012.

[3] 3GPP, "Backoff enhancements for RAN overload control," 3rd Generation Partnership Project (3GPP), R2 112863, 2011.

[4] ——, "System improvements for Machine-Type Communications (MTC)," 3rd Generation Partnership Project (3GPP), TR 23.888, 2011, V1.6.0 (2011-11).

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

[6] A. Larmo and R. Susitaival, "RAN overload control for Machine Type Communications in LTE," in *GLOBECOM Workshops (GC Wkshps), 2012 IEEE*, Dec. 2012.

[7] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Prentice-Hall, 1992.

[8] M. Rivero-Angeles, D. Lara-Rodriguez, and F. Cruz-Perez, "A new EDGE medium access control mechanism using adaptive traffic load slotted ALOHA," in *Technology Conference. IEEE VTS 54th*, vol. 3, 2001, pp. 1358–1362.

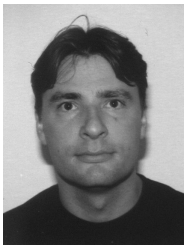
[9] G. Wang, X. Zhong, S. Mei, and J. Wang, "An adaptive medium access control mechanism for cellular based Machine to Machine (M2M) communication," in *Wireless Information Technology and Systems (ICWITS) International Conference on*, September 2010.

[10] I. E. Pountourakis and E. D. Sykas, "Analysis, stability and optimization of Aloha-type protocols for multichannel networks," *Computer Communications*, vol. 15, no. 10, pp. 619–629, 1992.

[11] J.-B. Seo and V. Leung, "Design and analysis of backoff algorithms for

random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 3975–3989, 2011.
[12] ——, "Performance modeling and stability of semi-persistent scheduling with initial random access in LTE," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4446–4456, 2012.
[13] K.-D. Lee, S. Kim, and B. Yi, "Throughput comparison of random access methods for M2M service over LTE networks," in *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, Dec. 2011, pp. 373–377.
[14] J.-P. Cheng, C.-h. Lee, and T.-M. Lin, "Prioritized Random Access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *GLOBECOM Workshops (GC Wkshps)*, Dec. 2011, pp. 368–372.
[15] S. Choi, W. Lee, D. Kim, K.-J. Park, S. Choi, and K.-Y. Han, "Automatic configuration of random access channel parameters in LTE systems," in *Wireless Days (WD)*, Oct. 2011, pp. 1–6.
[16] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A.-C. Hsu, "Overload control for machine-type-communications in lte-advanced system," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 38–45, June 2012.
[17] 3GPP, "Study on RAN improvements for machine-type communications," 3rd Generation Partnership Project (3GPP), TR 37.868, 2011, V1.0.0 (2011-08).
[18] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); Overall Description," 3rd Generation Partnership Project (3GPP), TS 36.300, 2011, V10.6.0 (2011-12).
[19] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.
[20] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), TS 36.321, 2012, V11.0.0 (2012-09).
[21] R. Nelson, *Probability, Stochastic Processes and Queueing Theory — The Mathematics of Computer Performance Modeling*. Springer, 1995.
[22] S. Ross, *Stochastic Processes*, 2nd ed. Wiley, 1996.
[23] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*. Springer, 1993.

Prajwal Osti received his bachelor's degree in electronics and communications engineering from from Tribhuvan University Institute of Engineering, Nepal in 2007 and MSc in communication engineering from Aalto university, Finland in 2011. Currently he is pursuing PhD degree at the Aalto University School of Electrical Engineering. His current research interests include scheduling in wireless networks and Internet of Things communications.



Pasi Lassila is a senior research scientist at the COMNET Department in the Aalto University School of Electrical Engineering. Dr. Lassila received his Ph.D. degree in 2001 and since then has published widely on the mathematical modeling and performance evaluation of networking technologies. His current research interests include flow-level performance of scheduling and resource management methods in cellular networks, capacity limits of multihop wireless networks, mobility modeling and its impact on wireless networks.

Samuli Aalto received his M.Sc. and Ph.D. degrees in Mathematics from the University of Helsinki in 1984 and 1998, respectively. From 1984 to 1997, Dr. Aalto worked as a Research Scientist at VTT Technical Research Center of Finland. Since 1997, he has been with TKK Helsinki University of Technology, which is now part of Aalto University. Currently he acts as Chief Research Scientist leading the Teletraffic and Performance Analysis Group in the Department of Communications and Networking. Dr. Aalto's research interests include queueing theory, teletraffic theory, and performance analysis of modern communications systems and networks.

Anna Larmo is a senior researcher at Ericsson Research in Finland. She received her MSc in communication technology in 2005 from Helsinki University of Technology. She has been with Ericsson since 2004, working with 3G and 4G technologies. Her current research interests include the Internet of Things technologies, radio protocols, performance evaluation, radio resource management, and simulator development. She has also been active in the areas of innovation and patenting.

Tuomas Tirronen received his D.Sc. in Communications Engineering in 2010 from Aalto University. Since 2012 he has been working in Ericsson Research as wireless access networks researcher. His research interests include 4G and 5G, Internet of Things, performance evaluation, radio protocols and resources. He is also active in 3GPP standardization work and innovation and patenting.